KEEP Teaching
Statistical Analysis Plan

Evaluator: UCL Institute of Education
Principal investigator(s): David Wilkinson

Education Endowment Foundation

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, Sam Sims
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school (teacher) level
TRIAL TYPE	Efficacy
TEACHER CAREER STAGE AND SCHOOL TYPE	Early career teachers in secondary schools
NUMBER OF SCHOOLS	207
NUMBER OF TEACHERS	207
PRIMARY OUTCOME MEASURE AND SOURCE	Teacher Job satisfaction – (Bespoke Teacher survey)
SECONDARY OUTCOME MEASURE AND SOURCE	Teacher retention in the profession; Teacher retention in the schools (School Workforce Census - SWC)

Table of contents

Table of contents 2
SAP version history 3
Introduction 4
Design overview 5
Sample size calculations overview 6
Analysis 6
RQ1 6
RQ2 and RQ3 7
RQ4 8
Imbalance at baseline 8
Missing data 9
Compliance 9
Effect size calculation 9
References 10

This analysis plan was written post-randomisation and prior to analysis of the final outcome data and deals only with the statistical analysis for the main trial and the longitudinal analysis. This document has been written based on information contained in the study Evaluation Protocol (amended) (uploaded 29 June 2022) published on the EEF website, in which full details of the background and design of the trial are presented.

SAP version history

Any changes made to the protocol which impact on the SAP, and any changes made to the SAP after its initial publication, will be specified here. There are no such changes to note to date.

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	16/02/2023	Creation of original document

Introduction

KEEP Teaching is a is a programme developed by the Institute of Physics (IOP) that aims to improve the job satisfaction of Physics Newly Qualified Teachers (NQTs) by increasing the proportion of time they spend teaching Physics. Improvements in job satisfaction as a result of spending more time teaching Physics are expected to work through reduced workload and improved pedagogical content knowledge. Improved job satisfaction may also result in better retention of Physics NQTs. In addition, there is some evidence that pupils being taught by teachers who accumulated more experience teaching a specific subject are more effective (Cook and Mansfield, 2016) and that for primary teachers improved mathematical knowledge is significantly related to student attainment gains (Hill, Rowan & Ball, 2005). Unfortunately, due to the cancellation of GCSEs in 2020 and 2021, we do not have the data to assess pupil attainment as an outcome. KEEP Teaching seeks to increase the proportion of time spent teaching Physics through better aligned timetabling. Tailored guidance will be provided by the IOP to schools through e-mail exchange, phone calls and some face-to-face liaison.

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[1] 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.

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

Design overview

Table 1: Study design overview

Trial design, including number of arms		Two-armed cluster randomised controlled trial
Unit of randomisation		School / NQT pairing
Stratification variables (if applicable)		None
Primary outcome	variable	NQT job satisfaction
	measure (instrument, scale, source)	The data will be collected from a short online survey in the summer term of the NQT year
Secondary outcome(s)	variable(s)	NQT retention in profession NQT retention in school
	measure(s) (instrument, scale, source)	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.

Participation in the trial requires a school to hire a physics NQT. 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. Recruitment was across England and covered three cohort of NQTs: those who started their NQT year in September 2019, 2020 and 2021. The trial is a two-armed random assignment with only one NQT per school so assignment is at the school/NQT level. KEEP Teaching is compared to a business as usual control group.

Randomisation was conducted using simple randomisation with no stratification using a coin toss for each pairing recruited sequentially. Recruitment was expected to be 200 schools, with 207 schools actually recruited, with equal allocation to treatment and control groups.

We wanted 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. This approach is sometimes critiqued because it can introduce biases when the recruiter knows what the next allocation will be. In this study, to avoid this problem, the evaluation team conducted the randomisation and the recruiter did not know what the next allocation would be. For more details, see the evaluation protocol.

Sample size calculations overview

Table 2: Sample size estimations

		Protocol OVERALL	Randomisation OVERALL
Minimum Detectable Effect Size (MDES)		0.26	0.26
Mean of Outcome Measure		3.67	3.67
Standard Deviation of Outcome Measure		0.72	0.72
Alpha		0.05	0.05
Power		0.8	0.8
One-sided or two-sided?		2	2
Number of schools	intervention	100	103
	control	100	104
	total	200	207

Based on cohort 1 data (available at the time of writing the protocol) the mean score of job satisfaction was 3.67 with a standard deviation of 0.72, which we take as our estimate for the full sample.

All estimates are based on standard EEF assumptions of 80% power and 5% significance level. We additionally assume that 30 per cent of the variation in the outcome is explained by covariates in the model, based on analysis of teacher job satisfaction in Sims and Jerrim (2020). The minimal detectable effect size is estimated to be 0.26 with the small increase in realised sample above the expectation at protocol stage (from 200 to 207 pairings) makes no discernible difference. No subgroup analysis is to be conducted.

Analysis

Analysis will follow the EEF's (2022) most recent guidance[2]. All analyses will be conducted in Stata v17.

Primary outcome

Research question 1

The estimated impact will be estimated on an intention-to-treat (ITT) basis, using all schools in the treatment and control group to which they were randomised irrespective of whether or not they actually received the intervention. We will estimate outcomes using a linear regression model including a dummy variable indicating trial arm allocation.

The primary outcome will be NQT job satisfaction. This measure will come from an online survey implemented towards the end of participant's NQT year. The measure is based on Thompson and Phua (2012), who systematically developed and validated this instrument

[2] Please see the Statistical Analysis Guidance.

based on analysis of 901 quantitative and 28 qualitative studies of job satisfaction. Respondents are asked how far do they agree with these statements about their 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

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

The equation to be estimated is:

$$Y_i = \alpha + \beta_1 Treat_i + \beta_2 \gamma_i + \beta_3 s_i + \varepsilon_i$$

where i is the NQT, Y_i is the mean job satisfaction score for NQT i, $Treat_i$ is our treatment indicator (a dummy variable where 1 represents being allocated to receive the intervention and 0 represents allocation to the control group), γ_i is a set of two dummy variables indicating in which year the NQT started teaching, s_i represents a vector of other control variables (teacher gender, school type -whether the school is an academy, Ofsted rating, region of school, urban/rural location, number of pupils, percentage FSM pupils, percentage EAL pupils, percentage SEN pupils, and average school level performance at end Key Stage 4) and ε_i is an error term.

Estimated impact in terms of pupil's outcomes will be converted into a Hedges' g effect size (Hedges, 1981) with 95% confidence intervals (CI). For details, see Effect size calculation section.

Secondary outcome

Research question 2 and 3

The secondary outcomes are both measures of retention, so require a different type of analysis. We will use a survival analysis approach; and will estimate Cox Proportional Regression models following the same specification above. This approach 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 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.

We will use data from the SWC to identify whether NQTs are employed in state funded education in England each year after their NQT year for up to two years. We will also report results from a Weibull proportional hazard model as a robustness check.

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 SWC data.

Given that this analysis relies on data observed up to two years after the completion of the intervention and available one year after that, it will be published in an addendum report three years after the main evaluation report.

Research question 4

Additional regression analysis will further examine the relationship between the 'matchedness' of timetables and teacher job satisfaction and teacher efficacy. More specifically, timetables collected from each school/NQT pairing will give us three measures of timetable 'matchedness':

- Specialism – the proportion of classes which are within the teacher's specialism. Note that the specialism may be physics, or physics with maths, depending upon training route and is self-reported. In addition, where a teacher teaches combined science, the rota of topics is examined to estimate the proportion of lessons which are physics across the year.
- Repeats – 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.
- Groups – 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.

This analysis will not exploit the random allocation of NQTs into treatment and control group. It will include an additional 30 NQTs who were recruited into the trial after the deadline for allocation' hence were not assigned to a treatment or control group. The same data from timetables and surveys were collected for these NQTs allowing us to estimate models with a slightly bigger sample size.

The equation to be estimated is essentially the same as described for the primary outcome except that we will not include the treatment indicator variable:

$$Y_i = \alpha + \beta_1 \gamma_i + \beta_2 s_j + \beta_3 \text{specialism} + \beta_4 \text{repeats} + \beta_5 \text{groups} + \varepsilon_{ij}$$

where i is the NQT, Y_i is the mean job satisfaction score for NQT i, γ_i is a set of two dummy variables indicating in which year the NQT started teaching, s_i represents a vector of other control variables (teacher gender, school type -whether the school is an academy, Ofsted rating, region of school, urban/rural location, number of pupils, percentage FSM pupils, percentage EAL pupils, percentage SEN pupils, and average school level performance at end Key Stage 4), specialism, repeats and groups are the 'macthedness' indicators described above and ε_i is an error term.

Imbalance at baseline

School and teacher level characteristics will be summarised descriptively by randomised group, both as randomised and as analysed (to check for balance and attrition). This will include teacher gender and specialism and school characteristics: Ofsted rating, school type (academy status), urban/rural location, number of pupils, percentage FSM pupils, percentage EAL pupils, percentage SEN pupils, and average school level performance at end Key Stage 4).

We will present characteristics on the basis of:

- Participating schools / NQTs as at the point of randomisation
- Participating schools / NQTs pupils in the final analysis sample

Reporting will follow the standard EEF template, with means and standard deviations reported for continuous variables and counts and percentages in each category given for categorical variables.

We will assess balance by calculating absolute standardised differences presented as Hedges' g effect sizes (Imbens & Rubin, 2015) between the treatment and control groups. Differences of greater than 10% will be considered as indicative of imbalance. If imbalance is observed, we will run an additional sensitivity analysis incorporating any variables on which imbalance is present as additional covariates into the primary outcome model.

Missing data

We will report the number of complete cases (i.e., those without missing data). The amount of missingness and its distribution will be explored and summarised by treatment arm in the report. In the event of greater than 5% missing data, we will conduct further investigation into the mechanisms of missingness. We will investigate the extent to which school and NQT characteristics are correlated with missingness, using a logistic regression, where the dependent variable is a binary indicator for missingness. If this shows significant associations with any of the characteristics, we would conduct an additional analysis including those covariates in the primary analysis model to assess the robustness of the main results.

Compliance

A Complier Average Causal Effect (CACE) analysis is not planned as part of the evaluation. We know that all schools received the timetabling guidance (the intervention), but there will be differences in the degree to which greater 'matchedness' of timetables is possible and in the extent to which timetables have been modified as a result of the guidance. We will assess this through the Implementation and Process Evaluation (IPE) where we will examine the average difference in 'matchedness' (dosage) between the treatment and control groups.

Effect size calculation

Effect sizes will be calculated using Hedges' g, following the standard approach for EEF trials as set out in the EEF analysis guidance. This will therefore be calculated as:

$$\mathrm{ES} = \frac{(\overline{\mathrm{Y}}_{\mathrm{T}} - \overline{\mathrm{Y}}_{\mathrm{C}})_{\mathrm{adjusted}}}{sd_{pooled}}$$

Where $(\overline{\mathrm{Y}}_{\mathrm{T}} - \overline{\mathrm{Y}}_{\mathrm{C}})_{\mathrm{adjusted}}$ is the regression adjusted difference in means between the treatment and control groups as recovered from the regression model, and sd_{pooled} is the pooled unconditional variance of the treatment and control groups. All relevant parameters will be provided in the report so that readers are able to compute alternative definitions of effect sizes.

A 95% CI for the effect size will be calculated by calculated by inputting the lower and upper confidence limits for the coefficient on the treatment variable from the regression model into the effect size formula.

References

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.

Cook, J. and Mansfield, R. (2016), Task-specific experience and task-specific talent: Decomposing the productivity of high school teachers, *Journal of Public Economics*, 140, issue C, p. 51-72.

Education Endowment Foundation (2022) Statistical analysis guidance for EEF evaluations, EEF October 2022

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. Journal of Educational and Behavioral Statistics , 1981, vol. 6, issue 2, 107-128

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.

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

Thompson, E. and Phua, F. T. T. (2012) A Brief Index of Affective Job Satisfaction. Group and Organization Management 37(3): 275-307, June 2012. DOI: 10.1177/1059601111434201