School choices: What works at Key Stage 4 in terms of improving GCSE outcomes, two or three years of study?
Evaluation Study Plan

Evaluator (institution): NFER
Principal investigator(s): Simon Rutt, Helen Poet

PROJECT TITLE	What works at Key Stage 4 in terms of improving GCSE outcomes, two or three years of study?
EVALUATOR (INSTITUTION)	NFER
PRINCIPAL INVESTIGATOR(S)	Simon Rutt (impact), Helen Poet (IPE)
STUDY PLAN AUTHOR(S)	Simon Rutt, Helen Poet, Afrah Dirie
STUDY DESIGN	Matched difference-in-differences study
PUPIL AGE RANGE AND KEY STAGE	KS3, KS4
NUMBER OF SCHOOLS	200
NUMBER OF PUPILS	36,000
PRIMARY OUTCOME MEASURE AND SOURCE	1) GCSE English performance (NPD) 2) GCSE Mathematics performance (NPD)
SECONDARY OUTCOME MEASURE AND SOURCE	1) Highest examination category achieved at GCSE and equivalent (i.e. 5 A*-C including English and Maths (and equivalent) from NPD) 2) Number of subjects offered/proportion of pupils taking each subject (NPD)

Study Plan version history

VERSION	DATE	REASON FOR REVISION
1.2 [*latest*]		
1.1		
1.0 [*original*]		*[leave blank for the original version]*

Table of contents

Background and study rationale

The structure of the national curriculum for secondary schools in England (DfE, 2014) is that KS3 runs from Year 7 to Year 9, and KS4 runs from Year 10 to 11. Whilst the national curriculum is compulsory for maintained schools, academy schools (around three-quarters of secondary schools in England) may choose to follow it. There is a broader range of compulsory subjects at KS3 than at KS4, although at KS4 schools must continue to offer pupils access to at least one subject within four 'entitlement areas' (the arts; design and technology; the humanities; and modern foreign languages).

Recent evidence shows that some schools start teaching some or all subjects at KS4 from Year 9 (i.e. up to a year earlier than the curriculum framework states):

- 56 per cent of respondents in NFER's February 2019 Teacher Voice omnibus survey (Lord *et al.*, forthcoming), started teaching the GCSE curriculum for most/all subjects in Year 9 compared to 40 per cent that started in Year 10.

- Similarly the TES (Roberts, 2019) reported that half of ASCL members' schools offered a longer KS4.

- Ofsted's curriculum research (Ofsted and Spielman, 2017) found 10 of 23 schools visited were reducing KS3 to two years and around a quarter of 171 school websites reviewed indicated pupils were selecting GCSE options at the end of Year 8.

- DfE's School Snapshot Survey Winter 2017 (IFF Research, 2018) found that in 35 per cent of schools KS4 began in year 9 for all subjects; with 28 per cent of schools beginning KS4 in year 9 in some subjects for all pupils. In 27 per cent of schools KS4 started in year 10.

The benefits of starting the KS4 curriculum earlier may include improved pupil outcomes at GCSE but there are concerns that this prioritisation of certain parts of the curriculum is driven primarily by school accountability measures (the EBacc, Progress 8 and Attainment 8)

(Brill *et al.*, 2018). In addition, Ofsted has outlined concerns that the focus of the performance measures on academic subjects is particularly restrictive for low attaining pupils which disproportionately affects pupils from low income backgrounds (Spielman and Ofsted, 2017).

Little is known about any similarities and differences between schools adopting the variants. We might expect academies to be more likely to adopt a longer KS4 given their curriculum flexibility. We might also hypothesise the introduction of the new GCSEs (first teaching 2015-17; first testing from 2017-19) has accelerated the adoption of a longer KS4 to provide sufficient teaching time.

There is an evidence gap around whether pupils experiencing a two- or three-year GCSE/KS4 curriculum perform differently at GCSE. Previously NFER conducted an evaluation of the then DCSF's Two Year Key Stage 3 project launched in 2003 (Noden *et al.*, 2007) which aimed to increase the pace of learning in KS3 and open up curricular flexibility, for example, beginning KS4/GCSE curricula earlier through the time saved. The evaluation focussed on performance in Year 7 and 8 (in the then QCA Optional Tests) finding that a shortened KS3 was associated with an increased pace of learning in mathematics.

There is however considerable current policy and practitioner interest and a growing body of research around the breadth of the curriculum which is relevant here. Ofsted's latest inspection framework focuses on both the curriculum as designed and the curriculum as implemented by schools, and intends for schools to offer both breadth and depth to pupils, including a broad Key Stage 3 curriculum where pupils are not expected to 'specialise too early' (Spielman, 2018; and Ofsted and Spielman, 2018).

NFER's forthcoming explorative review on the curriculum in Key Stages 3 and 4 (Lord *et al.*, forthcoming) focuses on the views of a survey panel of secondary school teachers/leaders on the breadth and depth of their KS3/KS4 curriculum offer, whether there are other curriculum areas they would like to provide and if so, any reasons why they currently do not. The authors find most notably that experience of breadth and depth (as reported by surveyed practitioners) seem to be associated with school type and school performance, meaning that students in certain schools may be experiencing a narrow(er) offer (such as starting GCSEs early and a narrower range of subjects) than students in other schools.

Other research argues that curriculum narrowing is having a particular impact on disadvantaged and/or lower attaining pupils. For example, GL Assessment (2019) argues they are 'doubly disadvantaged': deprived of the benefits of a rich and broad curriculum, which their better off peers tend to accumulate outside of school. Furthermore, researchers at Kings College London found that teachers think the nature of the new knowledge-focussed GCSEs is less engaging for these pupils with a greater risk of disaffection (Neumann *et al.*, 2016). On the other hand, the focus on EBacc and Progress 8 subjects has enabled more students – and in particular lower ability and/or disadvantaged students – to access more core curriculum subjects as Gill's (2018) analysis of the impact of the accountability measures on the uptake of qualifications suggests.

At present there is a dearth of evidence about whether a two-or three-year KS4 is better for pupil outcomes and schools, or whether there is no difference. This research will explore the extent and rationale of this practice, and compare attainment outcomes in similar schools offering different lengths of KS4 delivery in some or all subjects, in the context of the Ebacc/Progress 8 policy of encouraging a core curriculum at KS4, and particularly for disadvantaged pupils. The project will also provide contextual information about other factors that schools should consider such as pupil wellbeing, and impacts on timetabling and staffing, through the implementation and process evaluation (IPE).

Intervention

Name:

What works at KS4 in terms of improving outcomes, two or three years of study?

Why:

Some schools have chosen to operate a three year rather than a two year KS4 – this may be related to accountability measures (the research will explore this).

Who and where:

Secondary schools in England and their pupils.

What:

The research project will look at the difference in outcomes in schools that operate a two-year KS4 and those that operate a three-year KS4.

When:
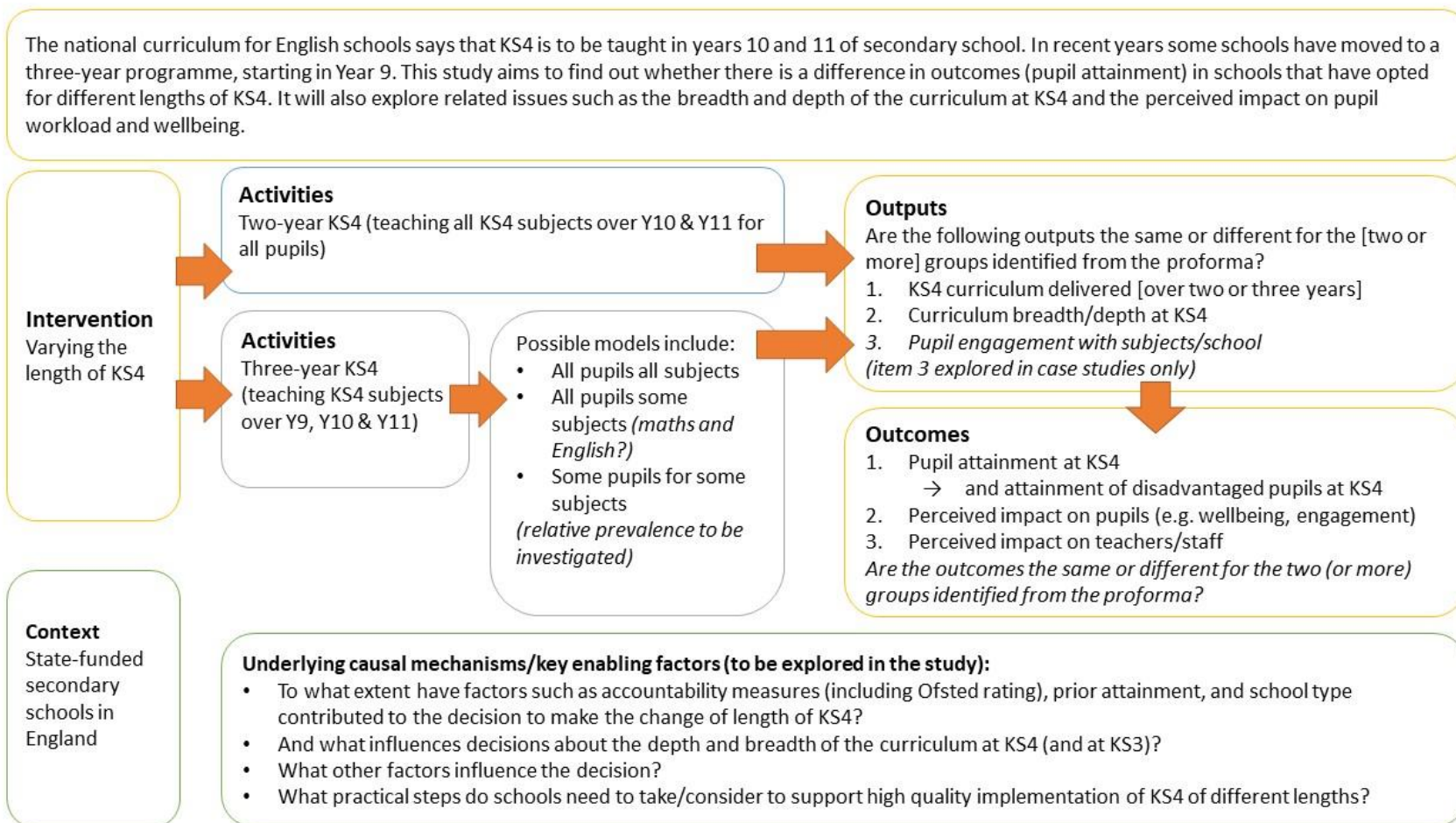
The research will include KS4 results from 2012 to 2019.

Tailoring:

The research will look at how schools have approached their curriculum at KS4 and the variations employed by schools.

A hypothesised logic model is shown on the next page.

Below is a logic model, which outlines the hypothesised activities, outputs and outcomes. It also summarises the underlying causal mechanisms or key enabling factors that could be influencing schools' approaches to the KS4 curriculum. The activities, outputs, outcomes and causal mechanisms will be explored through this project, as outlined in this study plan.

The national curriculum for English schools says that KS4 is to be taught in years 10 and 11 of secondary school. In recent years some schools have moved to a three-year programme, starting in Year 9. This study aims to find out whether there is a difference in outcomes (pupil attainment) in schools that have opted for different lengths of KS4. It will also explore related issues such as the breadth and depth of the curriculum at KS4 and the perceived impact on pupil workload and wellbeing.

Intervention
Varying the length of KS4

Activities
Two-year KS4 (teaching all KS4 subjects over Y10 & Y11 for all pupils)

Activities
Three-year KS4 (teaching KS4 subjects over Y9, Y10 & Y11)

Possible models include:
- All pupils all subjects
- All pupils some subjects *(maths and English?)*
- Some pupils for some subjects
(relative prevalence to be investigated)

Outputs
Are the following outputs the same or different for the [two or more] groups identified from the proforma?
1. KS4 curriculum delivered [over two or three years]
2. Curriculum breadth/depth at KS4
3. *Pupil engagement with subjects/school*
(item 3 explored in case studies only)

Outcomes
1. Pupil attainment at KS4
   → and attainment of disadvantaged pupils at KS4
2. Perceived impact on pupils (e.g. wellbeing, engagement)
3. Perceived impact on teachers/staff
Are the outcomes the same or different for the two (or more) groups identified from the proforma?

Context
State-funded secondary schools in England

Underlying causal mechanisms/key enabling factors (to be explored in the study):
- To what extent have factors such as accountability measures (including Ofsted rating), prior attainment, and school type contributed to the decision to make the change of length of KS4?
- And what influences decisions about the depth and breadth of the curriculum at KS4 (and at KS3)?
- What other factors influence the decision?
- What practical steps do schools need to take/consider to support high quality implementation of KS4 of different lengths?

Impact evaluation

Research questions

1. Do pupils attending secondary schools that teach KS4 over three years perform differently at GCSE to similar pupils in similar schools that teach KS4 over two years? What is the impact on disadvantaged pupils?

2. What is the impact of school-level disadvantage on attainment outcomes and its interaction with length of KS4?

3. Do pupils in schools with two- or three-year KS4 study (on average) different numbers of GCSEs?

4. What evidence is there that the curriculum offered by schools has narrowed, and is there a difference between schools with a two-year KS4 and those with a three-year KS4?

Design overview

Table 1: Design

Design	Matched difference-in-difference study	
Unit of analysis **(school, pupils)**	Pupil	
Number of Units to be included in analysis **(Intervention, Comparison)[1]**	200 schools; 36000 pupils (Business as Usual (2 year KS4):100 schools; 18000 pupils Comparison/intervention (3 year KS4 all/some subjects): 100 schools; 18000 pupils	
Primary outcome(s)	Variable(s)	Dual primary outcomes: GCSE English Performance GCSE Mathematics Performance
	measure (instrument, scale, source)	English GCSE , 0-9 (pre 2017 data to be amended in line with new scoring system), NPD Mathematics GCSE , 0-9 (pre 2017 data to be amended in line with new scoring system), NPD
Secondary outcome(s)	variable(s)	(1) Highest examination category achieved at GCSE and equivalent (i.e. 5 A*-C including English and Maths (and equivalent) from NPD) (2) Curriculum breadth

---

[1] Depending on the method used, the number of units included in the analysis can differ from the pool of potential comparison units. For example, when using matching/weighting the pool of comparisons units could represent all schools in England, but only a certain number of units will be included in the analysis after a suitable match is found. Identifying the precise number of units included might not be possible at the design stage. In these cases Evaluators can speculate on the number of units that are expected depending on the method used.

	measure(s) (instrument, scale, source)	(1) Achieved 5+ A*-C grades including English and Maths, 0-1, NPD (2) Number of subjects offered and/or proportion of pupils studying them per school (NPD)
Baseline for primary outcome	**variable**	Overall KS2 Attainment in English Overall KS2 Attainment in maths
	measure (instrument, scale, source)	KS2 English points score, 0-51, NPD KS2 Maths points score, 0-51, NPD
Baseline for secondary outcome	**variable**	n/a
	measure (instrument, scale, source)	n/a

Data collection

We will use a proforma/survey, sent to all secondary schools in England, to find out the extent to which schools are operating different lengths of KS4. This will enable us to create groups for the main impact analysis of KS4 outcomes.

The proforma will also ask:

- whether the length of KS4 has changed since 2012, and if so when and in what way
  - in schools with a two-year KS4, whether they have had a three-year KS4 at any point since 2012
- if a school has a three year KS4 for at least one subject, to find out which subjects and whether this is the case for all pupils or a sub-set. If a sub-set, then which pupils? (e.g. by high/low ability)
- reasons for the length of key stage.

We will send a school-specific link inviting schools to easily access the proforma (hosted in our survey software Questback). Initially all schools will be contacted via postal and email communications to maximise the response from the first contact. We will monitor response levels carefully via our Survey Administration system allowing us to target reminders to particular school types if required, and ensuring no school will be reminded unnecessarily. As this proforma will be sent to all schools in England, we will also use social media channels (e.g. twitter) and our communication channels (NFERDirect) with schools to promote the survey.

Because of the importance of the quality of proforma responses to the impact design, we will carry out some light touch verification checks. In the proforma we will ask respondents how long they have been at the school. In instances where they joined the school more recently than a change in length of KS4 was made, or if the data for the key question is incomplete, we will look at the school website and prospectus to obtain the information. In a minority of instances we may follow up with a call to the school to check the details of the curriculum.

It is expected that schools will provide a range of dates of when they made the change to deliver KS4 over three years. The expectation is that we will be able to obtain KS4 performance data for at least a couple of years after the first cohort has sat their KS4 examinations, as well as performance data from before the school made that change (also

see population section, below). This will allow a Difference-in-Difference analysis which will be discussed in a later section.

Population

All state-funded secondary schools in England are in scope for this project because we want to find out what is happening in schools across the country, in the current academic year. Previous studies, cited above, were based on samples or panels. We will report descriptive analysis based on the responses from the proforma to contribute to the evidence base on school autonomy and the curriculum.

At this stage we anticipate that there will be up to three main groups[2]:

1. Schools which start KS4 in year 9 for all pupils and in all subjects
2. Schools which start KS4 in year 10 for all pupils and in all subjects
3. Schools which start KS4 in year 9 for all pupils in some subjects.

We need a minimum of 200 matched schools for the impact analysis. As the data collection is not linked to an incentive and incentives are not being offered, we expect the response rate to be around 10 per cent. Therefore we need to invite all secondary schools to complete the proforma.

Once we receive the information from the proforma, we will run frequencies of the data to check whether the groupings we expect are present. Depending on the size of the response[3] we will determine whether a two or three-arm comparison is possible, but at this preliminary stage we are of the view that a two-arm is more likely to be possible. This is because, based on the evidence outlined in the background section, we estimate that 50-60 per cent of the proforma returns will be from schools offering some form of a three-year KS4 (groups 1 and 3 above combined). We also require responses from enough schools that made the change to a three-year KS4 early enough to be able to conduct the event-study analysis outlined below. Once the proforma responses have been collected, cleaned and analysed we will be able to determine whether we are able to include the three groups listed above, or whether we need to treat all schools that start KS4 in year 9 as a single group (i.e. whether they start early for all or some subjects).

While all schools will be included in the *descriptive* analysis, we will only select schools for the *impact* analysis that will have data that covers the period we wish to analyse. We propose to use data between 2012 and 2019. Depending on when the schools made the change to a three year program we will have data prior to this event and data for post event, as shown in Table 2.

We realise that any decision to move from a two to a three year KS4 will take a different amount of time in different schools. However for the purposes of this analysis, we are assuming that the final decision and move to implementation would take place in the academic year before the first Year 9 cohort start studying KS4.

---

[2] There are likely to be other groups, such as schools which start KS4 in Year 9 but only for some pupils studying some subjects. We will capture this scenario using the proforma. However impact analysis of the KS4 outcomes of this group would be dependent on collecting pupil lists from schools, which is beyond the current scope of this project. Should this scenario be a more significant group than anticipated we will discuss options with EEF.
[3] Power calculations are based on achieving 100 schools in an individual arm. On receipt of the proformas we will investigate whether there are enough schools to provide sufficient power for a three armed evaluation.

Table 2: Eligibility for event analysis (date of change to a three year KS4)[4]

decision made (Academic Year)	Year 9s start studying KS4	first year of exams for 3 year KS4	
2008-09	2009-10	2011-12	*[not enough pre-event data available]*
2009-10	2010-11	2012-13	
2010-11	2011-12	2013-14	
2011-12	2012-13	2014-15	
2012-13	2013-14	2015-16	eligible for event impact analysis
2013-14	2014-15	2016-17	
2014-15	2015-16	2017-18	
2015-16	2016-17	2018-19	*[not enough post-event data available]*
2016-17	2017-18	2019-20	data not available yet
2017-18	2018-19	2020-21	
2018-19	2019-20	2021-22	
2019-20	2020-21	2022-23	

One example shown above is that a school might inform us that they made the change in the academic year 2013/14. The first relevant Year 9 cohort, taking a three-year KS4 within these schools, would have started their KS4 studies in 2014/15 and would have sat their GCSEs in the summer of 2017. A school who made the change a year later would see their first GCSE cohort in the summer of 2018.

It is only following the return of the proformas from schools that we will be able to identify the cohorts available for analysis – these two examples above would be included, but a school making the change more recently, in 2017/18 for example, would be excluded from the impact analysis as their KS4 data would not yet be available. Furthermore schools, for example, that only opened in the academic year 2016/17 and therefore have no data prior to their opening[5] will be excluded from the impact analysis. Ideally we would be able to identify a significant number of schools who all made the decision to change in the same academic year and we have a similar number of schools who have continued with a two year program throughout this period.

We will select a comparison group of schools for each of the event years that are *similar to the intervention schools at the time the decision was made* (also see the section on 'identifying the groups for comparison', below). So for group 1 in Table 3 below, we would select a comparison group of schools with a 2 year KS4 that are similar to group 1 schools in 2012/13.

---

[4] Due to the timeline of this project and the length of time taken to receive bespoke data from the NPD, we opted to use these particular years as we have data for these years available in the SRS for use.
[5] This is an issue that will primarily affect academy schools. New academy schools who open in a specific year but are linked to a single previous school will remain within the analysable cohort. New academy schools that either have multiple previous schools or have no historical information will be excluded from the analysis.

Table 3: Available cohort data for three-year KS4 schools

Group	2012/13	2013/14	2014/15	2015/16	2016/17	2017/18	2018/19
1	School decide to change	First cohort starting 3 year GCSE		First cohort take 3 year GCSE	Second cohort take GCSE	Third cohort take GCSE	Fourth cohort take GCSE
2		School decide to change	First cohort starting 3 year GCSE		First cohort take 3 year GCSE	Second cohort take GCSE	Third cohort take GCSE
3			School decide to change	First cohort starting 3 year GCSE		First cohort take 3 year GCSE	Second cohort take GCSE

We will use available administrative data to compare the groups by key characteristics (school type (in particular if an academy or not), percentage eligible for free school meals (FSM), region, school size, KS2 intake performance data and KS4 outcomes).

At this stage, we have opted to be conservative for the reasons above, and assume enough responses for a two-arm comparison, i.e. a 'treatment' group of schools operating a three-year KS4 (all subjects/some subjects), and a 'business-as-usual' group operating a two-year KS4. The remainder of this study plan is written for a two-arm comparison. The exploratory/review stage is important to be able to understand whether the main analysis outlined below is possible in practice. After this initial exploratory work, we will review our planned analysis (below) in light of the available data and discuss this further with EEF[6].

In terms of pupils, we will only select pupils who have been at school for the full two or three years of KS4. Pupils who have joined after the start of KS4 (September of year 9 for three year KS4 schools or September of year 10 for two year KS4 schools) may dilute the treatment effect as they could potentially move from a two year KS4 school to a three year KS4 school in Year 10. We will identify whether a pupil has completed a full two or three years of KS4 by using the pupil date of entry variable which is available in the NPD.

Sample size calculations

Table 2: Sample size calculations

Sample size calculations as highlighted in the table below are based on a number of important assumptions. The most important of these are the intra class correlation and the correlation between the baseline measure of attainment and the outcome. The ICC for this design has assumed an ICC of 0.2[7] indicating the amount of outcome variance attributable to the school. The other consideration is the correlation between KS2 and GCSE and whilst no data is currently available for the new GCSE gradings previous EEF guidance suggested

---

[6] For example, if we exceed the expected response rate, we will investigate whether it would be possible to introduce a third-arm to the analysis, e.g. to split the three-year KS4 group into 'all subjects' and 'some subjects'.

[7] This is slightly more conservative than a figure of 0.19 suggested by EEF for English GCSE outcomes and analysis of the 2014 GCSE data

a correlation of 0.65 for English GCSE outcomes. We have been slightly conservative to ensure a well powered design.

To fully characterise the effect of the length of KS4 programme on attainment, it is unlikely to use a single outcome so we have opted for two primary outcomes, analysed separately. If these two outcomes were analysed independently at the significance level of 0.05, the probability of finding at least one false positive increases to 0.098 (familywise error rate). To control for this, we have adjusted the p-value using Beonferroni correction (Vickerstaff *et al*., 2019).

		Study Plan
		OVERALL
Minimum Detectable Effect Size (MDES)		0.15
Pre-test/ post-test correlations	level 1 (pupil)	0.65
	level 2 (class)	
	level 3 (school)	
Intracluster correlations (ICCs)	level 2 (class)	
	level 3 (school)	0.2
Alpha		0.025
Power		0.8
One-sided or two-sided?		2
Average cluster size		180
Number of schools	Business as Usual (2Y KS4)	100
	Comparison 1 (3Y KS4 all subjects)	100
	Total	200
Number of pupils	Business as Usual (2Y KS4)	18000
	Comparison 1 (3Y KS4 all subjects)	18000
	Total	36000

Outcome measures and other data

Baseline measures

No baseline measures.
Primary outcome

We already have access to de-identified[8] national pupil database (NPD) data, including KS4 pupil and exam level data going back to 2012[9].

---

[8] De-identified means that direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.

[9] As long as we are granted permission to do this. If not, or if we require data prior to 2012, then this would be obtained through the normal NPD application process.

Over the last few years there have been changes to how GCSE performance is measured which makes the longitudinal aspect of the design a little more challenging. We have decided to use English and maths grades[10] as primary outcomes (analysed separately) as there is a higher level of consistency over the years in how these have been measured, although some adjustment will be made to account for a new points scoring system introduced in 2017.

Therefore GCSE grades in English and Mathematics[11] will be used as co-primary outcomes, calculated and analysed separately. Due to changes in the scoring system of GCSE grades between the years included in the analysis, we will be creating variables that compute the highest grade achieved on the same scale which will be described in more detail in the final study plan. The following variables will be used from the NPD:

1) English: KS4_APENG, KS4_APENG_PTQ, KS4_APENG_PTQ_EE

2) Mathematics: KS4_HGMATH, KS4_ HGMATH _PTQ, KS4_ HGMATH _PTQ_EE

Analysis will identify the average effect across all years following the time of when the decision was made. Interaction terms will be included to determine if there are differential time effects.

Secondary outcomes

A secondary outcome will be created that identifies the breadth of the curriculum offered by a school at KS4. It will be analysis of this variable, over multiple years, which will determine the extent to which the curriculum has narrowed, if at all. Much of the discussion about curriculum breadth relates to KS3, and specifically about narrowing that occurs in year 9 in schools that start KS4 early. At the moment this is outside of the scope of this project as information about subjects offered at KS3 is not being collected[12]. However one hypothesis is that schools running a three-year KS4 may enable pupils to study more GCSEs, over the longer period of time, and this is something we will be able to explore.

Data currently available will allow analysis that identifies the range of subjects offered at KS4 and the relative take up of these qualifications within each school. There is a degree of data exploration required prior to determining the most useful outcome variable and this process will be discussed with EEF during the exploration period. This variable could possibly be a count of the number of subjects offered, or the number of subjects offered within particular subject groupings. An alternative would be to identify the proportion of students taking each subject. Whilst a school may offer a similar number of subjects to that offered in previous years, if students are guided[13] towards particular subject groups then we may see a smaller proportion[14] of students taking subjects that fall outside the 'buckets'. Exploratory analysis will identify whether a combination of subject count and the proportion of students taking a subject is the best way to identify the breadth of curriculum offer. The primary aim is to be able to identify whether a school's decision to move to a three-year programme of KS4

---

[10] Thought had been given to using Attainment 8 as the outcome measure but given changes to what is included within this measure it may be correlated with group membership and so it was decided not to use this as an outcome. Furthermore, recent changes in GCSEs (including, but not limited to changes to the scoring system and changes to the equivalencies of some qualifications) meant Attainment8 would not provide a single consistent measure over the time period this study requires.

[11] Adjustments will be made to account for changes in the GCSE scoring system.

[12] Although it will be explored with the case study schools as part of the IPE, below.

[13] It might be the case that whilst schools offer many subjects only students that are expected to achieve the higher grades are encouraged to take subjects that might not contribute to their Attainment 8 score.

[14] This statistic may need to be weighted to account for variable cohort sizes.

teaching is associated with a change in the breadth of subjects offered for study and whether this change is similar in the groups of interest or if there has been divergence.

To investigate the impact of the length of a KS4 programme on pupils' GCSE attainment across a broader range of subjects, a further secondary attainment outcome will be analysed looking at the highest examination category a pupil has received. As mentioned earlier, there is no single measure for this available from the NPD due to changes in the GCSE scoring system. As a result of this, we will be creating our own measure by combining KS4_EXAMCAT, KS4_EXAMCAT_PTQ, and KS4_EXAMCAT_PTQ_EE from the NPD. We will then create a binary variable indicating whether a pupil achieved 5A*-C including English and maths (and equivalent) or not.

Selection Mechanism

As outlined in the background section above, there is little evidence at present about which schools are more likely to adopt a three rather than two year KS4. One of the aims of this project will be to provide descriptive statistics about the characteristics of schools adopting the differing approaches and their stated reasons for adopting one or the other, for both the most recent academic year and the years prior. This will be collected using the proforma. Analysis of the proforma data (specifically, the questions about the length of KS4 curriculum) will inform our approach to the matching process.

MATCHING/WEIGHTING

Identifying the groups for comparison

Based on our expected response rate, we anticipate that we will have enough schools for a two arm comparison, which would be split based on each school's KS4 practice for our co-primary outcomes of English and maths. In this scenario, the groups would be:

1. Schools with a two-year KS4 for maths and English
2. Schools with a three-year KS4 for maths and English.

In this scenario, schools that have only introduced a three-year KS4 for other subjects (but have a two-year maths and English KS4) would be in group 1 (above) as we are assuming a low response rate and that we will need to include them in that arm[15].

On receipt of the proforma data it will be matched to NFER's register of schools to link in the school characteristics of the responding schools. We will run descriptive analysis of the data to check which programmes/models of KS4 are in our dataset. We will also check that we have sufficient numbers of schools that made the change in the years eligible for the event impact analysis (see Table 2), in both arms.

This study plan will be updated once the responses to the proforma have been received. The final decision about the number and make-up of the arms will be based on being able to achieve a suitable number of similar schools that have either changed their program to a three year delivery or have remained within the more traditional two year model. If proforma information allows for a third arm (i.e. to split the three year KS4 arm into two groups, namely three year KS4 some subjects and three year KS4 all subjects) then this will be discussed with EEF and the advisory group.

---

[15] However it should be noted, that we will review all of this, depending on the results to the proforma, and the ideal scenario would be that we have 'pure' groups in which schools have all subjects in a two-year KS4, and all subjects three-year KS4. (Furthermore, if responses to the proforma allow, we would introduce a third arm made up of schools with a three-year KS4 for some subjects, including maths and English)

Matching process if selection is needed

If we receive a large number of responses from schools then the matching process will proceed as follows.

Schools from the three-year KS4 arm will be matched with schools from the two-year KS4 arm that were similar (in terms of the observable baseline characteristics listed below) at the time the decision was assumed to be made (i.e. the year prior to the first Year 9 cohort starting KS4 study; Table 2). At the time of writing we are assuming that we will need to include multiple years (i.e. schools that changed to a three-year KS4 in 2013-14, 2014-15, and 2015-16) and consequently we will include a flag in the dataset to indicate when the change to a three year took place.

We will use freely available school-level data to identify relevant characteristics. We will consider using the following observable characteristics[16] of the schools that have introduced the change to a three-year programme until we obtain a matched dataset:

- school size (as a continuous variable)
- school FSM
- school type (maintained/ single academy trust/multi academy trust/free school)
- performance data (KS4 results)

The reason for a matching process is to ensure that the difference in difference analyses is run using a list of two year KS4 schools that are as comparable as possible to three year KS4 schools. In addition to the variables identified above it may also be possible to include information from the proforma as variables within the matching process. For example if a large proportion of schools state that a reason for their decision was based on teacher workload issues then this could be used to create comparable schools, which would be important as we would expect teacher workload to be related to our main outcome variable, i.e. GCSE performance.

The outcomes from these analyses will be discussed with EEF prior to updating the study plan. After agreement, measures will be introduced in to the matching method described above.

According to Little (2014), there are three main decisions affecting a matched dataset: the choice of measuring distance; the choice of matching strategy; and choice of algorithm to perform matching.

There are many different ways of measuring distance (D_{tc}) between the observable characteristics of study groups, the most common are:

1) Exact[17]:
    - $D_{tc} = 0$ if $\mathbf{X}_t = \mathbf{X}_c$
    - $D_{tc} = \infty$ if $\mathbf{X}_t \neq \mathbf{X}_c$
2) Mahalanobis[18]:
    - $D_{tc} = \sqrt{(\mathbf{X}_t - \mathbf{X}_c)' \, \mathbf{S_X}^{-1}(\mathbf{X}_t - \mathbf{X}_c)}$

---

[16] Initial investigations of available data will look to identify the most suitable measures to be used in the matching process. The purpose will be to identify those factors that it is reasonable to assume are associated with the outcome of interest and predict propensity to be treated.

[17] $\mathbf{X}_t$ is a vector of observable characteristic values for the treatment group and $\mathbf{X}_c$ for the control group

[18] $\mathbf{S}^{-1}$ is the covariance matrix of the observations

3) Propensity score[19]:

- $D_{tc}(\mathbf{X}_t, \mathbf{X}_c) = |\pi_t - \pi_c|$

The exact method is the most straightforward way but it is not ideal in our case as we have some continuous observable characteristics and it is unlikely that the value for these covariates is exactly the same for both study groups. An extension of exact matching is coarsened exact matching (CEM), which allows continuous or ordinal data to be segmented into strata. However, if the strata are too complex, this makes it more likely it will result in failed matches as CEM requires an exact match ($D_{tc} = 0$).

The Mahalanobis method is not ideal in our case as we have observable characteristics that include several dichotomous variables (e.g. region) and the Mahalanobis method may not be the most suitable method for such variables (Little, 2014). Using propensity scores overcomes this through collapsing the vector of observable characteristics into a scalar propensity score.

A propensity score is the probability of participating in a given intervention, given a set of observable baseline characteristics. In our case, the relevant propensity is for a school to undertake a three-year programme of GCSE study. We have chosen to estimate the propensity scores using a logistic regression model[20]. The outcome of interest in the estimation of propensity scores is the binary indicator of whether a school is part of the main group of interest, i.e. have a three-year programme of GCSE study.

The last decision to affect a matched dataset is the type of matching algorithm used. Our chosen matching strategy is a 1:1 matching strategy without replacement using the nearest neighbour matching algorithm and with the caliper set to 0.2. This assigns a set of nearest propensity scores (neighbours) to a treatment school. Since each treatment school is matched based on a minimum distance between its propensity score and the score of its nearest neighbours, the overall heterogeneity of the matched dataset is reduced.

Matching imbalance will be investigated looking at the standardised mean differences between three year and two year KS4 schools[21], before and after matching, as this is the most used technique to explore the balance of covariate distributions between treatment groups. We will also identify the extent to which matched groups are similar through data visualisations. SMD and plots will be produced using the MatchIt (Ho *et al.*, 2013) and cobalt (Greifer, 2020) package in R (R Core Team, 2017).

We will be computing propensity scores as well as creating a matched dataset using the MatchIt (Ho *et al.*, 2013) package in R (R Core Team, 2017). Once a matched sample has been formed, the diff-in-diffs treatment effect can be estimated by comparing the outcomes between treatment schools and comparison schools through the use of regression models using the lme4 package in R (Bates *et al.*, 2015).

Matching process with minimum number of schools

The above description of the matching process assumes a large pool of schools where selection is necessary. If a smaller proportion of schools return the proforma then we may

---

[19] π_t is the probability of belonging in the treatment group , given the observable characteristics and π_c is the probability of belonging in the control group

[20] A propensity score $e(X_i)$ can be estimated from logistic regression of the treatement condition on the covariate x $\log\left(\frac{e(X_i)}{1-e(X_i)}\right) = \boldsymbol{\beta} X_i$ (Pan and Bai, 2015)

[21] The SMD is given by the difference in mean outcomes between groups divided by the standard deviation of outcome.

only have enough schools to populate the evaluation. In this eventuality we would use the same matching process described above to identify the propensity for a school to be in the group that has introduced a three-year program. These propensities would be used within the analysis as weights (computed as the inverse of the propensity score). It is also worth noting that schools would be matched using data for the years prior to the year of change (also see the section on population above). Similarity between the treatment and comparison schools will be based on their characteristics at the point when the decision to switch to a three year programme is made (assumed to be in the year prior to implementation, see Tables 2 and 3), although we acknowledge that such decisions are, in fact, made over the course of many months or years.

DIFFERENCE-IN-DIFFERENCES (DD)

Primary outcome analyses

Primary analyses would use GCSE grades in English and maths as the main outcomes and will be conducted at pupil level. Outcomes for English and maths will not be combined into a single effect. Pupil level linear regression models with school level random or fixed effect[22] will be run and will control for a variety of pupil and school level factors[23] whilst including a treatment dummy variable aiming to capture the effect of being in a school with a three-year GCSE programme compared to a two-year programme. Analysis would look to identify whether the long-term trend of school performance has changed since the introduction of a three-year GCSE programme through the use of multilevel models analysed using the lme4 packages (Bates *et al.*, 2015) in R.

The regression model for the primary ITT analysis is given by

$$Y_{it} = \beta_0 + \beta_1 \text{treatment}_i + \boldsymbol{\beta_2 postchangeyears} + \boldsymbol{\beta_3}\text{treatment}_i * postchange\text{years}_t + \boldsymbol{X'_{it}\beta} + \epsilon_{it}$$

where Y_{it} is the GCSE grade of a pupil in treatment i at year t, treatment_i is a dummy variable set to 1 if the pupil is in the treatment group[24] at any year t, postchangeyears_t is a binary variable with the years prior to the first cohort sitting their GCSE's acting as the reference group. This is further explained below.

The vectors of coefficients $\boldsymbol{\beta_2}$ and $\boldsymbol{\beta_3}$ represent the average change in GCSE attainment for the years after making the introduction to a three year program compared to the years before that change, and the average change in GCSE attainment for the interaction between the treatment group and the indicator for years after the change. The coefficient for $\boldsymbol{\beta_3}$ will identify the average effect post policy change of a three year programme at KS4 over schools that operate a two year programme of KS4.

Other covariates for this model ($\boldsymbol{X'_{it}}$) include: *gender* (male as the reference group), *region* (North as the reference group), *ethnicity* (white British as the reference group), *SEN level* (1 as the reference group), and *pupil FSM* as binary variables, academic *year* as a categorical variable, *KS2 prior attainment*, *school size* and *school FSM* as continuous variables. We have included such covariates as it is likely that there will still be some variance left after

---

[22] The robust Hausman test will be run to determine if the coefficient estimates from the multilevel model are statistically significantly different from the fixed effects estimates (assumed to be unbiased estimates).

[23] There may be residual differences in baseline characteristics between two and three year KS4 schools after propensity score matching, these will be added in the model to reduce bias that arises due to residual differences (Austin, 2011)

[24] Treatment group being schools undertaking a three year programme of KS4 study.

selecting the comparison group of schools by matching eligible schools to the observable characteristics of the treatment schools.

The error associated with the ith pupil at year t is given by ϵ_{it} and this model will be run in R using the lme4 (Bates *et al.*, 2015) package with a full syntax trail.

If the outcome of running this analysis suggests that there is a statistically significant difference in GCSE grades achieved before and after the decision to change the length of KS4 programme was made, we will consider running an additional analysis that identifies whether there is a continuous improvement in GCSE grades after the year that the decision was made for all years after the change.

Further analyses

Extension of the Primary outcome analyses

If results from the primary analysis indicate that the length of a KS4 programme has an impact on pupils' GCSE attainment, then we would investigate at which year after the change took place had the most impact. A *year* variable would be created to identify the year in which a cohort actually sat their GCSE examinations. For example, a school that started a three-year programme in the academic year 2014/15 would see this first cohort sit their examinations in the summer of 2017. In this case the GCSE's of 2016 would act as the baseline year and years following this would be recorded as +1, +2 etc. Pupils within comparison group schools would receive the same flag. The years preceding 2016 would be recorded as -1, -2 etc. The positive years would be combined into a single dichotomous variable to indicate the GCSE years since the change (*postchangeyears)*, as indicated in the regression equation above.

The year variable would be added to the model to determine any differential effects between the years after the change took place. For analysis purposes a number of dichotomous variables would be created for each year. Interaction terms would be introduced between these dichotomous variables and the treatment variable; the coefficients on these variables are those that identify the estimated effect of interest. This follows a similar analytical design to Eyles (2017) and their analysis of academies using the date of conversion. This design would allow us to see the trajectory prior to the introduction of the three-year programme and the trajectory after the change. With comparisons to the trajectory of schools maintaining a two-year programme, we will be able to identify the differential effect of the new programme. The above analytical design can be considered a Difference-in-Differences design. Prior to undertaking the primary analysis, outcomes will be graphically displayed to identify the common trends pre-treatment. These plots will also be introduced post analysis to identifying longitudinal trends.

The main covariates that will be included in the model have been discussed above and the covariates will be discussed and included in the final study plan prior to any analysis taking place.

As mentioned in the matching section if the match does not result in a comparable group of schools propensity weights will be applied during the analysis to ensure analysis has accounted for observable differences. This will be discussed in a revision of the study plan once the final matching exercise has taken place.

Analysis will also be restricted to pupils who have been in each school for the correct length of KS4 teaching. For example a pupil that arrives in a school at the start of Year 11 will be

excluded from the analysis. This will ensure analysis is based only on pupils who have received the full two or three years of KS4 within the same school.

Secondary outcome analyses

Analysis of the variable that identifies the curriculum offer, as discussed above, will be a school level analysis rather than at pupil level and will, therefore, not need to account for the clustering of pupils within schools. These would include a school-level indicator for the proportion of pupils eligible for free school meals, school type, geographical location, school size and a number of other factors associated with curriculum provision. An important factor to include here would be the whether the school is an academy and if it belongs to a MAT, as the increased autonomy of these schools may be a determining factor in a changing curriculum offer. The same term used for the primary outcome would be introduced into the model to identify the year of change. The introduction of these terms will identify whether there has been a change in the school offer and whether this change is associated with the change from a two-year GCSE programme to a three-year programme of study.

Secondary analysis of an additional GCSE outcome, 5+ A*-C grades including English and maths (and equivalent), will be used to investigate whether the length of a KS4 programme has an effect on the examination category a pupil receives. Models with this outcome would be run using the same matched sample and covariates as identified in the primary analysis. We will be running multilevel logistic regression models using the lme4 package in R (Bates *et al.*, 2015).

Subgroup analyses

Interaction terms between the treatment variable and the variable for eligibility for free school meals, and between the treatment variable and the variable for school deprivation will be introduced into the model to identify any differential FSM and school deprivation effects between pupils studying KS4 for three or two years. Following EEF's guidance protocols, analysis will be run separately on a dataset containing only pupils eligible for free school meals[25].

Missing data

Prior to commencing any analysis we will look at the amount of missing data, but as all the data is coming from statutory testing we do not envisage this being problematic. If there is less than 5% missing data, we will consider carrying out a complete cases analysis as this is unlikely to be biased.

However, if there is more than 5% missing data, following EEF guidelines, the extent of missingness and the pattern of missingness are important when analysing missing data. The number of complete cases (a complete case would be a case with complete data on all variables of interest) will be reported. The extent of missingness will be also quantified and reported. Based on these numbers we will discuss the adverse effects of the missing data on sample size and thereby on statistical power and other implications.

We will identify likely missingness mechanisms; Missing completely at random (MCAR), Missing not at random (MNAR), and Missing at Random (MAR). To test these assumptions we will conduct diagnostics to establish any measurable predictors of missingness from the data. Initially, we will look for any imbalances between the groups (attrition and non-attrition)

---

[25] FSM pupils are EEF's key target group and, although the findings will be less secure for this group when the trial is not powered to detect an effect on FSM, it is important that it is reported (with appropriate caveats).

through cross-tabulations. If our groups are not equivalent (i.e. statistically significantly difference on any measure, carrying out a 'complete case analysis' (using only cases with complete data) may be biased as the study groups may not be representative of the original sample. This analysis also shows us whether cases with particular characteristics are more likely to have dropped out (biased attrition).

We will also run logistic regression models with missingness on each variable of interest (1=missing, 0=otherwise) as the dependent variable and observable characteristics as independent variables. Tests will be conducted at both the student and school levels, as there are relevant school-level variables available.

We will carry out Little's MCAR test which will be performed in R using the package BaylorEdPsych (Beaujean, 2012) as we can examine multiple variables simultaneously. The null hypothesis for Little's MCAR test is that the data is MCAR.

Given the need to compare estimates of the effect of the length of KS4 programme under different missing data assumptions, in order to generate sensible estimates (although less precise) under the MCAR assumption, our first step will be to perform a complete case analysis for each model. If, for each model, only the endpoint outcome is missing, we can then add any covariates predictive of non-response to this model and thereby produce valid estimates under the MAR assumption.

If we have MAR outcome data and covariates at endpoint, the best strategy is multiple imputation which will produce estimates under an MAR assumption. Multiple imputation uses complete cases to make multiple estimates of each missing value which are then used to make a single best estimate. The variables used are generally those to be included in the primary models and those which are associated with missingness. We will use the mice (Buuren and Groothuis-Oudshoorn, 2011) package in R.

Sensitivity analysis will then be used to test the sensitivity of these results to the possibility that the data are missing not at random (MNAR). We will use sensitivity analysis to assess the robustness of the analysis we have conducted under the MAR missing data assumption via multiple imputation. We will consider imputing 'extreme' values to the missing cases for the relevant variables (such as all 0s or all 1s for a binary variable) and seeing how, if at all, this changes the conclusions based on MCAR and MAR estimates.

As part of the process evaluation, we will explore the extent to which pupils who were on the school roll at the start of the KS4 programme did not take their GCSEs at the same school. The analysis described earlier will only look at pupils who have been at school for the full two or three years of KS4. Analysis will identify whether a particular programme results in an increased likelihood of not completing their GCSEs in the school they started the KS4 programme. This would be achieved by undertaking a logistic model on all those pupils on roll in Year 9 or Year 10 at the start of KS4 and whether or not they took their GCSEs in that school.

Robustness checks

To ensure that we obtain a sample of schools with a 2 year KS4 programme that are as comparable as possible to schools with a 3 year KS4 programme, we will consider running additional matching models with alternative matching methods. Potential alternative matching methods could be changing the caliper setting or changing the matching strategy so that more than one school with a 2 year programme matches to one school with a 3 year programme.

By comparing the distribution of continuous baseline covariates between the two groups of schools after matching obtained from different models, through data visualisations, we would be able to determine the most efficient matching model. The most successful matching model will be the one that has the distribution of continuous baseline covariates of the two year group as similar as possible to the distribution of continuous baseline covariates of the three year group (Austin, 2011).

Robust Hausman tests[26] will be run to determine if the results from the multilevel model are biased as there are unobserved school characteristics related to both the decision to offer 3 year GCSEs and the outcome variables. If the test does show this to be the case then the primary analysis model including fixed effects will take precedence. Other possible sources of robustness checks will be investigated at the time of analysis, for example the Mundlak (1978) correction.

An additional covariate might come out of the secondary analysis looking at the narrowing curriculum. If a significant effect is found in this analysis then we would look to include a covariate in an additional secondary analysis of GCSE outcomes to see if this is associated with variation in GCSE outcomes.

A range of models will be run with each building on the previous version. The first would just contain treatment group variables, followed by pupil level covariates and a final model introducing school level covariates. The inclusion of additional covariates will be presented as sensitivity analyses to assess the robustness of findings of the primary analysis.

Effect size calculation

Using the statistical analysis guidance provided by EEF, the formula for calculating effect size is given by:

$$\mathbf{ES} = \frac{(\mathbf{\overline{Y}_T} - \mathbf{\overline{Y}_C})\mathbf{adjusted}}{\sqrt{\sigma^2}}$$

The numerator for the effect size calculation will be the coefficient of the treatment group from the regression model. All effect sizes will be calculated using the square root of the population variance of the two groups, as the denominator which was calculated using the multilevel model without any covariates.

Confidence intervals for each effect size will be derived by multiplying the standard error of the interaction coefficient by 1.96. These will be converted to effect size confidence intervals using the same formula as the effect size itself.

Implementation and process evaluation (IPE)

Research questions

- IPE_RQ1: What are the softer impacts of having a two- or three-year programme at KS4 on pupils? (in particular, but not exclusively, related to wellbeing and ability to manage workload)
- IPE_RQ2: What are the softer impacts of having a two- or three-year programme at KS4 on teachers and their lesson planning?

---

[26] We will be running a robust Hausman test as the traditional Hausman test cannot be used in the presence of heteroscedasticity which is likely to occur (Hoechle, 2007).

- IPE_RQ3: What are the reasons for operating different lengths of KS3/4? (e.g. pupil outcomes/accountability measures/resourcing factors) What were the factors that influenced the decision to change the length of KS3/4?
- IPE_RQ4: What form do different lengths of KS3/4 take? What variation occurs within groups?
- IPE_RQ5: What strategies and practices are used to support high quality implementation of different lengths of KS3/4?
- IPE_RQ6: Which factors affect the breadth of subjects that schools offer for study at KS4 and at KS3?

This project is not an evaluation of an intervention but instead research into whole school decisions about the organisation of their curriculum, and consequently several of the dimensions affecting implementation of an intervention (i.e. the TIDieR) are not directly relevant (e.g. reach, fidelity, dosage, quality, responsiveness). The factors affecting implementation (Humphrey *et al.*, 2017) are more helpful here, which have been used as the framework for some of the research questions outlined above.

Research methods

In addition to understanding if there is an impact on pupil outcomes, it is important to understand whether there are any other contextual factors that schools thinking about changing the length of KS4 should also consider, in particular softer impacts on pupils and teachers. As this project is, at this stage, exploratory, the IPE is relatively light and it is designed to aid understanding of other factors that may be driving decisions around length of KS4 (and the extent to which factors relating to the length of KS3 are influencing this decision).

The first source of information will be the proforma: in addition to informing the impact analysis, we will include an open question asking the reasons why the school chose to have a two/three year KS4. We will code the responses to this question using a coding frame developed from the data. This information will provide additional context and insight, and where possible will be used to inform case study selection (below).

Qualitative case studies will be conducted to understand how different lengths of KS4 are implemented in schools. Their purpose will be to understand why schools have selected that length of KS4 and the perceived impact on timetabling, staffing and pupil choice of subjects. The proforma will conclude with a question asking the respondent if they will be willing to participate in a case study, and to provide contact details. When arranging the case study visits, we will be flexible and responsive to each school's needs and timetable.

Based on the evidence above, we anticipate that there will be two to three main groups in the sample: schools with a two-year KS4 for all pupils; schools with a three-year KS4 for all pupils; and schools with a three-year KS4 for some subjects. We will review these groupings once we have received the data from schools via the proforma. If the impact analysis uses two groups we will still aim to include some case studies of schools operating other KS4 models in order to aid understanding of the breadth of models in schools. (For example, if possible we would also aim to include schools that have considered a three-year KS4 but decided against it, or those currently operating a two-year KS4, but who have offered a three-year KS4 in the past/are considering doing so. This is dependent on receiving proforma responses from schools that meet those criteria.) We will use purposive sampling in order to cover a range of school types (including, for example, rates of FSM) and KS4 outcomes. The sampling will also be informed by the responses to the proforma question about why they operate their model of KS4. Where appropriate we will use the matching

variables to identify case study schools. The sample will be selected from the schools responding to the proforma in spring 2020, therefore we will review and update the case study sampling strategy in summer 2020.

We will conduct twelve face to face case studies, consisting of interviews with the following people:

- Headteacher
- Two form tutors/pastoral staff in KS4
- One form tutor/pastoral staff member in KS3.

The interview with the head teacher/senior leader (selected in consultation with the school about who would be most appropriate to speak to) will explore:

- the context of the school and the reasons behind why and how they have chosen to operate KS4 (including internal and external factors such as recent curriculum changes, accountability measures and the influence of other bodies such as MATs, Ofsted).
- *who* made the decision to change and *how* the process of making the decision and implementing the change took place in practice.
- the perceived relative benefits and challenges of a two/three year KS3/4 for the schools, teachers and pupils.
- the breadth and depth of the curriculum offered at KS3/4 (in terms of numbers and types of subjects offered and studied)
- breadth/depth of the curriculum within subjects (for example does additional time in a three year KS4 on the GCSE curriculum mean more rehearsal for exams, or covering wider range of topics/in greater depth).

We will request from all case study schools that, where possible, the interview with the senior leader is with a member of staff that worked in the school at the time a change was made to the length of KS4, and who was involved in the decision making process.

For each case study we will also interview three form tutors/pastoral staff of students in KS4/KS3. These interviews will explore the softer outcomes for pupils studying in different lengths of KS4, such as the wellbeing of their pupils and, in particular, how pupils in their classes are managing with the demands of KS4 work. The interviews will also explore the perceived impacts upon the teaching body such as timetabling, curriculum and lesson planning (including the issue of curriculum breadth/depth – see above) and views on relative benefits and drawbacks of changing the length of KS4.

Throughout the case study interviews we will explore the perceived impact upon disadvantaged pupils compared to their peers (particularly in the form tutor interviews).

As part of the research question on curriculum narrowing/curriculum breadth we will ask schools for information about the subjects offered in KS3 and KS4 and the amount of time dedicated to different subjects across the timetable. We will also use the case studies to explore whether schools have made changes to their extra-curricular offer as a result of their KS4 (for example whether some non-EBacc subjects like music or art are available as clubs in response to changes in the offer at KS4).

The project team will design the topic guides in the spring of 2020, drawing on available evidence in the field. We will review them in the summer, allowing us to include probes or additional questions to explore or test any emerging findings from the impact analysis with schools and the coding of the open question in the proforma (above). We will review the topic guides after the first 1-2 visits, and update them if required (for example to clarify

wording) – the project leader will conduct at least half of these initial visits. All of the interviews will be conducted by an experienced member of the research team, supervised by the project leader.

Analysis

Table 3 outlines how the IPE will be analysed and the research question(s) covered.

The qualitative data will be analysed initially using a top-level coding frame(s), developed from the semi-structured interview schedule. The interview schedules have not yet been developed but are intended to cover the topics outlined in the section above. After the first round of coding, sub-codes will be created and assigned to the text as the data is analysed and themes emerge (for example relating to reasons for choosing a particular length of KS4). We will analyse the data across roles, as well as the 12 case study units, and analyse the dataset by key factors including length of KS4 and (sample permitting) school type.

We will report the findings thematically, but where possible we will also include anonymised vignettes to illustrate examples of practice.

Table 3: IPE methods overview

Research methods	Data collection methods	Participants/ data sources (type, number)	Data analysis methods	Research questions addressed
School proforma	Online survey	Sent to all state-funded secondary schools	Frequencies of closed questions (see impact section), inductive coding of open question	IPE_RQ3
Case study visits in schools	Semi-structured interviews	12 case studies, 4 interviews per CS (1xHT and 3x pastoral staff per school)	Mixed inductive/ deductive coding; thematic analysis	IPE RQs 1-6

Ethics

Ethical agreement for participation within the project (completing the proforma and participating in case studies) will be provided by the headteacher of the school.

All data gathered during the trial will be held in accordance with the data protection framework established by the Data Protection Act 2018 and the General Data Protection Regulation (EU) 2016/679, and will be treated in the strictest confidence by the NFER and EEF. Our legal basis for gathering and using this data is legitimate interests, through our work as a research organisation. The project will only use de-identified pupil data from the NPD; we will not be collecting any personal data about pupils. If a school would like to take part in a case study, we will collect personal (contact) data from teachers/senior leaders in order to arrange the visit. Role and length of time working at the school will be collected as part of the light-touch validation process of the survey (see the section on Data Collection, above).

Data protection

The legal basis for processing personal data is covered by:

GDPR Article 6 (1) (f) which states that 'processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party except where such interest are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of the personal data'.

We have carried out a legitimate interest assessment, which demonstrates that the evaluation fulfils one of NFER's core business purposes (undertaking research, evaluation and information activities). It has broader societal benefits and will contribute to improving the lives of learners by providing evidence about the impact of the length of KS4 on pupil attainment, and the perceived impacts on pupils' wellbeing. It is therefore in our legitimate interest to process and analyse personal data for the evaluation. We have carefully considered all of the personal data being collected and it is all necessary to achieve the aims of the research (see Ethics, above). We have also balanced any potential impact on the data subjects' rights and find that our activities will not do the data subject any unwarranted harm. Privacy information will be provided to all data subjects in an accessible and transparent manner.

We will not use pupil or teacher names or the name of any school in any report arising from the research.

NFER will access the NPD data for analysis through the SRS secure online system. The SRS system does not allow users to remove or copy data from its servers.

At the end of EEF evaluations data is archived to allow for further secondary analysis. Only school-level data will be added to the EEF archive; none of the personal data collected will be added to the EEF archive.

We will not share personal data collected through interviews or in the proforma/survey with other organisations.

Further information will be available in the project privacy notice: https://www.nfer.ac.uk/media/3916/eesc_privacy_notice.pdf.

Personnel

Name	Institute	Roles and responsibilities
Simon Rutt (SR)	NFER	Project Director, responsible for leading the NFER team and project delivery.
Helen Poet (HP)	NFER	Project manager, responsible for overseeing the day to day running of the project
Afrah Dirie (AD)	NFER	Statistician, responsible for statistical analysis
Kathryn Hurd (KH)	NFER	Test and Schools administration lead, responsible for overseeing design, dispatch and processing of the proforma

Risks

Risk	Likelihood/ impact	Mitigation
Insufficient schools complete the proforma	Medium/High	We have anticipated that the response rate is likely to be lower than for trials, and adjusted the sample size accordingly. Careful monitoring of response rates and targeted communications and reminders to schools will be used.
There are insufficient schools in one or more of the expected curriculum groups which affects the power of the analysis	Medium/High	We will look at the number of schools in each group and calculate if the analysis is possible across two, rather than three groups (the three groups being a. KS4 starts in Y9 all pupils, all subjects, b. KS4 starts in Y10 all pupils, all subjects, c. KS4 starts in Y9 all pupils, some subjects)
Researcher loss (due to sickness, absence or staff turnover)	Medium/Medium	NFER has a large research department with numerous researchers experienced in evaluation who could be redeployed.
Case study visits no longer possible due to Covid-19/social distancing measures	Medium/Medium	If case study visits are no longer possible, we will use telephone or video interviews instead.

Timeline

The original timeline was delayed by 3 months due to Covid-19 after the launch of the survey. A 'relaunch' is planned in the summer term 2020.

Date	Activity	Staff responsible/ leading
Nov-Dec	Draft study plan, Advisory Board meeting	SR/HP
w/c 20th Jan 2020	Project agreed	SR/HP
w/c 27th Jan	Project start-up, privacy notice. Finalise proforma and confirm sample	SR/HP
Feb - April	Analysis of NPD data to determine performance matching variables	SR/AD
3rd Feb	Agreed proforma entered onto online survey system, sample set up	KH/HP
w/c 24th Feb	Proforma (survey) sent to schools (after Feb half term)	HP/KH
March	Reminders to schools (email, social media)	NFER, EEF
	Survey paused due to Covid-19	

Date	Activity	Staff responsible/ leading
10th June	Survey relaunched (+1 reminder)	NFER, EEF
3rd July	Survey period ends	KH
July	Data cleaning	KH
Mid-July - Aug	Initial descriptive analysis, PSM, match proforma data to NPD	SR/HP/AD
Aug - mid-Oct	Review study plan for impact analysis in light of available data from proforma. Advisory Group meeting (if required) Agree any changes with EEF.	SR/HP
– Mid-Oct - Jan	Impact analysis	SR/HP/AD
Nov – Jan 2021	Case studies (IPE)	HP
Feb	IPE analysis	HP
Mar- May	Report writing	SR/HP/AD
7th May	Draft report sent to EEF	SR/HP
Early June	Advisory Group meeting	NFER/EEF
30 July 2021	Publication	NFER/EEF

References

Andrews, J., Perera, N. Eyles, A., Heller Sahlgren, G. Machine, S., Sandi, M. and Silva, O. (2017). *The Impact of Academies on Educational Outcomes* [online]. Available: https://epi.org.uk/wp-content/uploads/2017/07/EPI_-Impact_of_Academies_Consolidated_Report-.pdf [20 May, 2020].

Austin, P.C. (2011). 'An introduction to propensity score methods for reducing the effects of confounding in observational studies', *Multivariate Behavioral Research*, **46**,3, 399-424 [online]. DOI 10.1080/00273171.2011.568786.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015). 'Fitting linear mixed-effects models using IME4', *Journal of Statistical Software*, **67**,1, 1-48. [online]. DOI 10.18637/jss.v067.i01.

Beaujean, A. A. (2012). *BaylorEdPsych: R Package for Baylor University Educational Psychology Quantitative Courses. R Package Version 0.5* [online]. Available: https://CRAN.R-project.org/package=BaylorEdPsych [20 May, 2020].

Brill, F., Grayson, H., Kuhn, L. and O'Donnell, S. (2018). *What Impact Does Accountability Have on Curriculum Standards and Engagement in Education. A Literature Review* [online]. Available: https://www.nfer.ac.uk/media/3032/nfer_accountability_literature_review_2018.pdf [20 May, 2020].

Department for Education (2014). *The National Curriculum in England: Key Stages 3 and 4 Framework Document* [online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/381754/SECONDARY_national_curriculum.pdf [20 May, 2020].

Gill, T. (2018). *The Impact of the Introduction of Progress 8 on the Uptake of Qualifications in English Schools – an Update for 2016/17*. [online]. Available: https://www.cambridgeassessment.org.uk/Images/519667-the-impact-of-the-introduction-of-progress-8-on-the-uptake-of-qualifications-in-english-schools-an-update-for-2016-17.pdf [20 May, 2020].

GL Assessment (2019). *Dangerous Diet: How Exam Rations Endanger a Broad and Balanced Curriculum* [online]. Available: https://www.gl-assessment.co.uk/news-hub/reports/dangerous-diet-how-exam-rations-endanger-a-broad-and-balanced-curriculum/ [20 May, 2020].Greifer, N. (2020). *Cobalt: Covariate Balance Tables and Plots. R Package Version 4.1.0* [online]. Available: https://CRAN.R-project.org/package=cobalt [20 May, 2020].

Ho, D., Imai, K., King, G. and Stuart, E. A. (2013). 'MatchIt: nonparametric preprocessing for parametric causal inference', *Journal of Statistical Software*, **42**, 8, 1-28 [online]. DOI 10.18637/jss.v042.i08.

Hoechle, D. (2007). 'Robust standard errors for panel regressions with cross-sectional dependence', *The Stata Journal,* **7**, 3, [online]. DOI 10.1177/1536867X0700700301.

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R. and Kerr, K. (2016). *Implementation and Process Evaluation (IPE) for Interventions in Education Settings: an Introductory Handbook* [online]. Available: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_Handbook.pdf [20 May, 2020].

IFF Research (2018). *The School Snapshot Survey: Winter 2017* [online]. Available:https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/727807/The_school_snapshot_survey_winter_2017.pdf [20 May, 2020].

Little, T. D. (2014). *The Oxford Handbook of Quantitative Methods in Psychology*. Oxford: Oxford University Press.

Lord, P., O'Donnell, S., Williams, M. and Hillary, J. (forthcoming). *The Breadth and Depth of the Curriculum: Tensions and Intensions*. Slough: NFER.

Mundlak, Y. (1978). 'On the pooling of time series and cross section data', *Econometrica*, **46**, 1, 69-85. [online]. DOI 10.2307/1913646.

Neumann, E., Tower, E., Gewirtz, S. and Maguire, M. (2016). *A Curriculum for All? The Effects of Recent Key Stage 4 Curriculum, Assessment and Accountability Reforms on English Secondary Education* [online]. Available: https://www.researchgate.net/publication/311455824_A_Curriculum_for_All_The_effects_of_recent_key_stage_4_curriculum_assessment_and_accountability_reforms_on_English_secondary_education [20 May, 2020].

Noden, P., Rutt, S., Schagen, S. and West, A. (2007). *Evaluation of the Two Year Key Stage 3 Project* (RR836*)* [online]. Available: https://webarchive.nationalarchives.gov.uk/20130323000849/https://www.education.gov.uk/publications/eOrderingDownload/RR836.pdf [20 May, 2020].

Office for Standards in Education, Children's Services and Skills (2018). *An Investigation into how to Assess the Quality of Education through Curriculum Intent, Implementation and Impact. Phase 3 Findings of Curriculum Research*. [online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/766252/How_to_assess_intent_and_implementation_of_curriculum_191218.pdf [20 May, 2020].

Office for Standards in Education, Children's Services and Skills and Spielman, A. (2017). *'HMCI's Commentary: Recent Primary and Secondary Curriculum Research'*, [online]. Available: https://www.gov.uk/government/speeches/hmcis-commentary-october-2017 [20 May, 2020].

Pan, W. and Bai, H. (2015). *Propensity Score Analysis, Fundamentals and Development.* New York, NY: Guildford Press.

The R Foundation (2019). *R: A Language and Environment for Statistical Computing* [online]. Available: https://www.R-project.org/ [20 May, 2020].

Roberts, J. (2019). 'Exclusive: Heads fear Ofsted will penalise three-year GCSEs*', TES,* 9 January [online]. Available: https://www.tes.com/news/exclusive-heads-fear-ofsted-will-penalise-three-year-gcses [20 May, 2020].

Spielman, A. (2018). *Letter to Public Accounts Committee* [online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/752721/HMCI_PAC_letter_311018.pdf [20 May, 2020].

Vickerstaff, V., Omar, R. and Ambler, G. (2019). 'Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes', *BMC Medical Research Methodology,* **19,** 129, [online]. DOI 10.1186/s12874-019-0754-4.