Reanalysing education experiments with linked administrative data

An analysis report from the Behavioural Insights Team

November 2022

Alex Sutherland, James Lawrence, Kim Bohling, Tania Locke,
Amber Evans, James Farrington, Anna Bird

Funded by:

ADR UK

UKRI Economic and Social Research Council

THE BEHAVIOURAL INSIGHTS TEAM

Contents

Key terms and abbreviations

Administrative data is information created when people interact with public services, such as schools, the NHS, the courts or the benefits system, and collated by government (ADR UK, no date). It is one of the great untapped resources of 'the information age'. Through processes of governing and being governed, interacting with services and just living life, there are traces of us in numerous administrative datasets. That might be from checking when the bins go out, buying a car, or registering the birth of a child. This by-product of modern bureaucracy has a value of its own, particularly if we can put it to good use.

Randomised controlled trials (RCTs) are a key research design for understanding the causal impacts of policies and interventions. They do this by effectively flipping a coin to decide who receives a new intervention or who receives business as usual (BAU). In doing so, RCTs help to minimise problems relating to self-selection into a treatment or intervention. This in turn, means that we can be more confident that if, on average, an outcome changes in the intervention group following intervention, but does not change in the BAU group, the change was because of (i.e. caused by) the intervention.

Frequent abbreviations	**What they mean / refer to**
ADR UK	Administrative Data Research UK – an ESRC funded programme whose purpose is to put administrative data to use by improving access and depositing of administrative data for use. https://www.adruk.org/our-mission/our-mission/
DfE	Department for Education (UK government department) https://www.gov.uk/government/organisations/department-for-education/about
GDPR	General Data Protection Regulations https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/
LEO	Longitudinal Educational Outcomes dataset - LEO 'connects individuals' education data with their employment, benefits and earnings data to create a de-identified person level administrative dataset'. https://www.gov.uk/guidance/apply-to-access-the-longitudinal-educ

	ation-outcomes-leo-dataset
MoJ	Ministry of Justice (UK government department) with responsibility for courts, prisons, probation and attendance centres https://www.gov.uk/government/organisations/ministry-of-justice/about
NEET	Not in Education, Employment or Training
NPD	National Pupil Database – administrative dataset covering the educational trajectory of all pupils in England. https://find-npd-data.education.gov.uk/
ONS	Office for National Statistics (Independent statistics authority) https://www.ons.gov.uk/aboutus
PNC	Police National Computer – centralised criminal records database covering the UKhttps://www.acro.police.uk/PNC-services
SRS	Secure Research Service – a secure environment for the analysis of administrative data. https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice

This work was produced using statistical data from ONS. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

We debated whether or not to redact so as not to distract from the fact that the project was about data linkage and analysis across several outcomes. However, in consultation with the Education Endowment Foundation, we decided to include intervention names to be more transparent, and because once the EEF archive data are more widely available it will allow others to compare our analyses more easily.

In some of our analyses we have combined treatment conditions in the same trial. This is typically where there were small sample sizes in the trial. We have indicated where this occurred. In some instances we were able to recover *more* outcome data than the original study through the use of administrative data.

1. Executive summary

This project combined *administrative data* with *randomised controlled trials* to understand whether educational interventions led to additional benefits or costs – beyond those initially evaluated. That is what we did, but the real purpose of the project was to demonstrate that such things *can* be done, what the *value* of doing this work can be, and *what needs to happen next*.

What we did

We linked data from the Education Endowment Foundation's trial archive to data held by government departments on other outcomes, to evaluate whether educational interventions showed effects, positive or negative, on outcomes beyond educational attainment.

- Attendance at school (from the National Pupil Database)
- Exclusions (both fixed term and permanent) (from the National Pupil Database)
- Employment status and out-of-work benefits (from the Longitudinal Educational Outcomes database)
- Not in Education, Employment or Training (NEET) status (from the Longitudinal Educational Outcomes dataset)
- Criminal convictions (from the linked NPD and Police National Computer dataset)

We also undertook two other main tasks as part of the project:

- Reviewing and updating theories of change for some of the interventions to assess whether/how there could be pathways to the outcomes we were looking at.
- Asking intervention developers and evaluators for their predictions and rationales for possible impacts on other outcomes, without them knowing the results from reanalysis.

What we found

Across most of the RCTs we looked at there was substantial variation across trials in terms of other outcomes, in itself this is useful given that educational interventions can have spillover impacts in other policy domains.

- In some instances, seemingly negative and positive impacts of interventions *that would otherwise not be known* were observed.
- For attendance, several trials had positive or negative impacts that were statistically significant in isolation, but not once we adjusted for multiple comparisons.
- For exclusions, some interventions significantly raised the risk of exclusion in the years following intervention, whereas others reduced it - even after adjusting for multiple comparisons.
- Project teams consisting of developers and evaluators tended to be overoptimistic about the impact of their interventions on additional outcomes.

As evaluation – and in particular randomised trials – become more central to how policy evaluation is done in the UK, the need for lower friction access to administrative data is paramount. Based on the programme of work leading up to this report (Gibbons et al., 2020; Calcraft et al., 2021) we recommend the following for ONS, ADR UK, What Works Centres and relevant government data owners:

- There is scope for improvements in the administration and interplay between Government departments, the Office for National Statistics Secure Research Service, and external data owners and data requestors. **The systems and processes in place for accessing and working in the SRS and requesting data from administrative data owners need to be streamlined to reduce steps and complexity**.

- **Processes for allowing What Works Centres to routinely track outcomes from government administrative data need to be implemented and standardised to ensure parity and consistency across the What Works Network**. Given this would be a new service, there would be cost implications of setting it up and running it. What Works Centres and the departments that fund them should discuss the demand and costs involved to prevent reinvention of processes and repetition of labour.

- **The use of Privacy Preserving Technology (PPT) – such as synthetic datasets – would dramatically reduce the time taken for data access and analysis across government.** We have demonstrated in other work the value of using low fidelity synthetic data and helped implement it but this should be a core part of how data is made available.

- **The What Works Network should routinely check outcomes from other domains for interventions that have been trialled.** The link between outcomes across What Works Centres – particularly those focused on children and young people – means there is an ever greater need to be able to understand outcomes that cut across policy domains.

- **There are two related methodological issues arising:**
  - **Finding a defensible way of accounting for multiple outcome testing**. This is not a niche point because, whether regarded as 'correct' or not, it will directly relate to decisions about whether an intervention / policy is considered beneficial or harmful. There are many ways in which to approach this issue and it is not settled in the wider research literature. We think the way forward is for What Works Centres to task a specific working group on multiple comparison testing to come to a view on how to undertake this, which then feeds into both What Works Centres and Government departments doing this.
  - **There should be an 'asymmetric appetite' for assessing possible benefits versus harms**. That is, the thresholds for substantive and/or statistical significance (if that is used) should *differ* depending on whether we are assessing the chance of harm or of benefit. Specifically, the threshold for determining chance of harm should be lower than that for assuming a benefit. Put another way, we should be 'more sensitive' to the possibility of harm. Given the variety of outcomes that might be considered this is not something that can be applied in a blanket fashion with one-size-fits-all thresholds, but we believe the *principle* can and should be consistent.
- **There is a need for a routine data request service that allows for asymmetric linkage to administrative data held by departments.** In practice that would mean that trial teams can request basic aggregate statistics for treatment and control groups that allow for a simple assessment of difference/similarity on intermediate or ultimate outcomes while trials are running. This would then bring social policy trials more closely into line with medical trials that routinely monitor outcomes. This isn't currently funded across Government or What Works Centres but would clearly benefit both.

We also have recommendations for the Education Endowment Foundation (EEF) based on the re-analysis work completed:

- **The EEF should default to include both attendance and exclusion alongside attainment as routine outcomes that it will collect data on and monitor**. These could be framed as monitoring during programmes (for example, to look for backfires) or retrospectively collected once trials have completed to understand possible spillovers into other areas of education.

This would not be a large additional burden given that pupil identifiers are collected during EEF trials.

- **The EEF should augment its approach to Theory of Change development to include possible impacts on exclusion and attendance**.
- **The EEF should make its trials archive more easily accessible to researchers**, for example through making a low fidelity synthetic data version of the archive available. This will allow for more work to be completed on long-term follow-ups and spillover effects, as well as evidence synthesis across intervention types and year groups.

2. Introduction

As the What Works centre for educational attainment, the Education Endowment Foundation (EEF) has tested over 190 educational interventions, mainly through randomised controlled trials (RCTs). The primary aim of these interventions was to increase students' attainment, but that might not be the only effect, and effects might not be limited to time spent in compulsory schooling. Some programmes may achieve other desirable student outcomes, such as reducing the chance of criminal conviction or increasing the chance that they go on to full- time education or employment at age 18. Conversely, some interventions that are effective at increasing attainment might have undesirable side effects on other outcomes.[1] Identifying these additional impacts allows decision makers to take a more holistic approach to choosing which interventions should be implemented in a given environment, and mitigation strategies that could limit negative consequences.

The UK now has a clear regulatory framework and infrastructure in place to allow this linkage to take place. The two principal pieces of legislation which underpin this are:

- The Digital Economy Act 2017, which makes provision for "digital government" and data sharing between government departments, and a legal permissive gateway for access to de-identified data for research[2]; and
- The General Data Protection Regulation (GDPR) 2018, which sets out the legal tests and thresholds that need to be met in order for a particular instance of data processing to be carried out.

The outcomes that we cover in this report are based on data held by the Department of Education (DfE) and the Ministry of Justice (MoJ). The analysis itself took place in the Secure Research Service (SRS) environment which is operated by the Office for

[1] An example of an educational intervention with *negative* impacts is Achievement for All (Humphrey et al., 2020), where treatment pupils were two months' behind control pupils at the end of the evaluation period. It would be useful, for example, to see whether those negative impacts fade out, as is sometimes observed with positive effects of educational interventions (Bailey et al., 2020; but cf. Bailey and Weiss, 2022a, 2022b on the impact of selective follow-up).

[2] More on the provisions of DEA 2017

National Statistics (ONS). All the researchers who contributed to this analysis were accredited users of the SRS.[3]

Analysing linked, de-identified data in this way can be extremely cost-effective, thanks to the EEF's retention of the original data from the effectiveness trials.[4] If the trials archive were not available for this purpose, then the only way we could evaluate whether there were any effects on these alternative outcome measures would be to run new, large-scale trials. The costs for these are typically in the hundreds of thousands of pounds (to say nothing of the non-financial costs of randomising which interventions are delivered to a large number of educational settings). By using data linkage instead, we remove the need for new trials and make a large cost and convenience saving. As a conservative estimate, for every £1 spent on reanalysis for this project (where we were able to link all trials to a new outcome), we are saving £300 through not having to re-run a trial.[5]

[3] Note that data used in this project was not shared or created under the Digital Economy Act, meaning that it has not resulted in a research ready dataset.
[4]The EEF trial archive is managed by the Fischer Family Trust (FFT). For an example of work linking completed trials for a sub-group of pupils to look at attainment, see What Works Centre for Children's Social Care (no date), see also Verfürden et al. (2021).
[5] See Appendix IV Cost-Efficiency of Trial Reanalysis.

3. Process

Projects included in our reanalysis were selected based on a number of criteria, including a plausible theory of change for at least one of our outcomes of interest. An analysis plan for each of the outcomes and data sets was developed and quality assured as part of our standard internal pre-registration process prior to conducting the corresponding analysis. The process for accessing the data sets varied by data source:

Table 1: dataset, outcome and access route

Dataset	Data / outcomes	Accessed via
Education Endowment Foundation trial archive	Completed randomised trials	EEF and Fischer Family Trust
National pupil database	Exclusions Attendance	Department for Education NPD team
Longitudinal educational outcomes	Benefits claims Employment Post-16 education (FE/HE) Training	Department for Education LEO team
Police National Computer	Criminal justice outcomes - convictions	DfE & Ministry of Justice*
Higher Education data	Application and entry to HE data	Higher Education Statistics Authority

Table note: * The NPD-PNC linkage had been undertaken by Government Departments, facilitating linkage to the EEF archive via NPD. The EEF did not push for this linkage.

We also conducted exploratory research with intervention developers and evaluation teams to gather insights on whether and how they expected the programmes to impact attendance and exclusions.

Definition of Outcomes and analysis specification

We consider four broad categories of outcome. Each has a primary analysis and one or more secondary specifications. We nominate a single primary specification to control the number of primary statistical comparisons being made, given the potentially large number of trials that we are analysing.[6] Furthermore, the estimated p-values from each regression are adjusted using the Benjamini-Hochberg procedure, in order to limit the rate of false discoveries made in the analyses. As an example, since there are 5 EEF trials with available outcome data on convictions, the lowest p-value for the primary outcome must be less than 0.01 to qualify as significant, and so on.

Attendance (34 trials analysed)[7]

The primary analysis uses a quasibinomial[8] model where the outcome is the proportion of available half day sessions which were attended in the year after the intervention started, for each trial. (That is, an increase in this outcome corresponds to an increase in attendance.) For example, if a student has 250 available sessions in a term and attends 225 of them, then this outcome is defined as 225/250 = 0.9.

As a secondary analysis, we use the same outcome but use an Ordinary Least Squares (OLS) regression with heteroskedasticity robust standard errors.

We use attendance taken from the most recent year before the trial started as a control variable (with a logistic transform applied to it for the primary analysis and untransformed for the secondary analysis). We also control for pupil gender, whether the pupil was eligible for free school meals in the last 6 years,[9] and the pupil's SEN status, where these are present in the data.

Lastly, we conduct an omnibus chi-squared test with the null hypothesis that *none of* the interventions had an effect on attendance[10]. If the null hypothesis is rejected, then the standardised beta-hat values can provide a prioritisation for which trials are more likely to have had an effect.

Exclusions (33 trials analysed)

The primary analysis uses an OLS regression with heteroskedasticity robust standard errors on the number of exclusion events experienced by each child

---

[6] This matters because the more comparisons we make, the greater the level of adjustment to statistical significance testing we have to introduce.
[7] Number of trials analysed is for the primary analysis.
[8] We prefer this specification over the OLS because we expect the variance to be higher the further away from 100% attendance a student is. Rather than simply using robust standard errors and making no assumption about how the outcome variance changes, we can encode this knowledge in our choice of regression, hence the quasibinomial choice.
[9] This is the last 6 years as at the time of data collection, which may not be the year the trial started and will vary from trial to trial.
[10] We do this with a chi-squared test: the sum of standardised squared beta-hat values from the primary analysis follows a chi-squared distribution under the null hypothesis, with degrees of freedom equal to the number of beta values.

recorded in the data in all the years between the start of the trial and 2019, for each trial. Exclusion events include fixed-term, lunchtime and permanent exclusions.[11]
As secondary analyses, we ran the following:

- A logistic regression where the outcome is “positive” if a pupil receives any exclusion after the trial starts;
- A quasi-poisson regression where the outcome is the number of exclusions received;
- An OLS regression, where the outcome is log(1 + number of sessions missed) counting only sessions missed due to fixed-term exclusions (with those receiving permanent exclusions dropped from the sample entirely).

As above, we control for the presence of previous exclusions and also run an omnibus test.

Criminal Justice (5 trials analysed)
The primary analysis uses a logistic regression on whether or not a child had any criminal convictions in the 12 months following the actual intervention start date, for each trial. Specifically, we only include offences that occurred in the year after the intervention began and were proven (convicted) at any point. Additional covariates are gender (categorical), FSM6 (Free School Meals within the last 6 years, categorical), and SEN (Special Educational Needs, categorical).

As secondary analyses, we ran the following:

- A robustness check with the same outcome but using a linear regression specification with heteroskedasticity robust standard errors
- A poisson regression, where the outcome is the number of convictions;
- A logistic regression, where the outcome is whether or not a child had any convictions in the 6 months after the intervention started
- A logistic regression, where the outcome is the whether or not a child was convicted of a serious offence. Serious offences constitute those which are deemed violent or sexual, according to the Home Office’s encoding available in the PNC dataset.[12]

[11] In our terminology we refer to all of these as “exclusions”, though the term “suspension” is a common synonym for a fixed-term exclusion and “expulsion” for a permanent exclusion.
[12] PNC data from the last ten years indicates 20% of all recorded offences fall into this category.

Employment: out-of-work benefits; NEET status (3 interventions analysed)

Out-of-work benefits: the primary analysis used a linear regression on the number of months for which the person received out-of-work benefits in the year after the actual intervention start date. As secondary analysis, we ran a logistic regression on whether the person had been in receipt of out-of-work benefits for 3 or more months in the year after the intervention started.

NEET status: here, the primary analysis used a logistic regression on whether the person was NEET (not in employment, education or training) for three months or more, as recorded in the 'Current_Activity_Code' variable in the NCCIS tables.

Selection of interventions for reanalysis

We do not analyse the full set of 190 interventions in EEF's trial archive in this report. We prioritised a list of 35 interventions based on the following criteria:

- **Design:** The intervention was tested using a randomised controlled trial. We excluded any trials that used a waitlist design (e.g., the control group later received the treatment).[13] We also excluded trials with low EEF security ratings (0-1 padlock) because these studies have threats to internal validity or high risk of bias such as attrition or randomisations that have gone wrong.
- **Data sharing:** The data sharing agreements and information shared with study participants enable further analysis and data matching.
- **Data archival:** The data was saved in the EEF archive.
- **Plausible theory of change:** Two experienced education evaluators reviewed the remaining list of interventions and excluded those for which they could not identify any potential causal mechanisms that might link the intervention with changes in attendance or exclusions.[14] This criteria is important, because if we observe an impact for a programme without a plausible link between the intervention and change in outcome, we may take more seriously that the change we see is more likely due to chance than the intervention itself.
- **Trial timing:** The trial needs to be sufficiently far in the past for a reasonable chance of the outcome happening.
    - For attendance and exclusion outcomes, this is not a constraint as these are continuously occurring and measured outcomes throughout

[13] We excluded wait-list trials because, in many situations, that design means that control group participants later receive the intervention. There are exceptions to this, e.g. the cluster-randomised trial of Accelerated Reader RCT designed by Dr. Sutherland (Sutherland et al., 2016) where the control cohort ages out of the trial without intervention meaning they can be traced over time.
[14] We used attendance and exclusions as our primary outcomes of interest, as using all four outcomes would have seriously limited the number of interventions which have plausible links to all of them.

an individual's education career.
  - For criminal convictions, data are more limited for under-18s because there are fewer occurrences of convictions overall, criminal activity in minors is often dealt with through out of court disposals[15] (and children below 10 years old in the UK cannot be charged with a crime at all).
  - Lastly, while education and employment status is well-defined for any young person, a young person cannot be formally employed until age 14 and is required to be in education or work until age 18.[16]
- **We only consider interventions that have some behavioural aspect to them.** Non-behavioural interventions *could* affect any of these outcomes but we considered behavioural trials first. Future work could replicate this process considering other kinds of interventions. For this initial publication, we focus on those interventions more likely to generate an effect since some of the trials were quite limited in sample size, and having to perform a very conservative multiple-comparisons adjustment might hinder our ability to detect real intervention effects.[17]
- **We do not specifically exclude trials with a small sample size, or for which some covariate data is missing**. Where possible we make do without the covariate data (which is generally not required to conclude validity from a RCT, unless we suspect a randomisation failure — its absence adds a little variation to the outcome but not bias). There are some instances of class-randomised trials which are missing the class identifier. In those cases, we are sometimes able to conduct an underpowered analysis by assuming each school has only one treated and one untreated cluster, and sometimes we discard the trial from analysis.

The full list of trials is shown in the table in the Appendix section. This list was agreed by the project team in consultation with the EEF before the analysis took place. It was reduced slightly during analysis in instances where it became clear that the trial data could not be matched to the outcome data in question or was incomplete in such a way as to have severely or totally compromised either the linkage or the identification of which participant was in which cluster.

[15] See page 6 of YJB (2021) for a flowchart showing progression through the youth justice system. Nearly half of arrests of children do not lead to further action. Of the 31,000 arrests that do lead to further action, a police caution is the result in 23%. Of the 24,000 children proceeded against at court, 31% do not lead to a criminal justice outcome.
[16] This is true in England. In the other nations of the UK, with some minor variations, young people are free to leave school at 16 with no further restrictions.
[17] There is a counter-argument that we could have looked at all completed EEF trials - that would have been an efficient approach to looking at overall impacts on attendance and exclusion, but in our view makes the case for linking intervention and outcome harder.

Expert survey

To build upon the theory-led approach to this analysis, we administered a survey to ‘experts’ on each of the interventions. The experts included programme staff who were involved in the development and/or delivery of the intervention, as well as researchers who were involved in the EEF-commissioned evaluations. The purpose of the survey was to gather further hypotheses that might help us understand the results of the data analysis. We also wanted to gather data on whether the experts generally expected their interventions to have an impact on behavioural outcomes that were not measured as part of the evaluations. None of the experts had access to the findings from the archive data analysis prior to taking the survey.

The survey asked respondents to provide responses only for those programmes for which they had expertise. They were asked what sort of impact they expected the intervention to have on our outcomes of interest -- increase, decrease or no impact on [attendance/ fixed-term exclusions/ permanent exclusions]. If they indicated an increase or decrease, they were asked whether they thought the change happened through changes in pupil, teacher, or parent behaviour. They were also given a free text box to expand upon any hypothesised causal mechanisms.

We received a total of 21 responses across 15 interventions (out of 34 in total). Nine responses were from the intervention developers and 10 were from researchers who were involved in the evaluations.

Theory of change mapping

To better understand findings from the analysis, we also engaged in exploratory theory of change (ToC) mapping. This work was guided by two core research questions:

1. Based on the programme model and published findings, is there a hypothesised causal pathway between the intervention and our outcomes of interest (attendance, fixed-term exclusions (suspensions), permanent exclusions)?
2. If yes, what are the possible - plausible - pathways?

The process drew upon all publicly available research studies on the intervention, which were primarily the EEF-funded evaluations. The ToC mapping was conducted by staff who were blind to the results of the archive data analysis, so that they were not influenced by the findings. The staff reviewed the reports for evidence that might indicate an intervention had the means to impact on attendance and/or exclusions. Evidence was primarily found within the implementation and process evaluation (IPE) and generally consisted of findings related to pupil behaviour, motivation, and engagement; parent engagement; changes in teacher practice; and/or how well the implementation partner delivered the programme.

Using the gathered evidence, the analyst generated hypotheses about potential causal pathways and then registered a final hypothesis about whether they expected the intervention to increase, decrease, or have no effect on each of the outcomes of interest. This process was completed for all interventions for which we received a survey response, as well as interventions where we observed the largest effects (positive or negative).

We then reviewed the individual findings across the research activities (archive data analysis, expert survey, ToC mapping) to draw out the synthesised findings presented in the following section.

4. Quantitative analytical findings

Attendance

Following adjustment for 34-fold multiple comparisons, none of the trials had a significant effect on attendance. We note that our conservative analytic approach means that individual evaluations of each trial *might* have found a significant effect where we do not. Indeed, the original evaluation of Trial 67 (Texting Parents) found a small but significant effect on attendance ("pupils who received the intervention had [...] slightly reduced absenteeism (ES = -0.054, p < 0.001).")[18], but this result does not reach significance in our analysis. We found six interventions with larger directionally positive effect sizes on attendance than Trial 67, all of which had significant results on attainment (or related measures) in the original evaluations.[19] That said, many of the interventions with largest *negative* effects on attendance in our analysis also had positive attainment results in the original evaluations.

One trial (Trial 29: Catch-Up Literacy) had a noticeably larger apparent effect size than the others. This trial shows a 0.33+ increase in attendance, which would equate to roughly two and a half extra days of school attendance per school year.[20] We cannot be totally certain whether the Catch-Up Literacy programme increased attendance to this magnitude, and we note that the true effect size is unlikely to be as large as a log-odds ratio increase of +0.33 (unadjusted p-value 0.0025) (see Gelman et al., 2017 for a discussion of 'Type M' errors). This trial was also relatively small (n = 538 in the final analysis sample; individually randomised trial).

The forest plot below shows the estimated effect size (on the log-odds scale) and unadjusted 95% confidence intervals for each intervention. The trials on the far right of the forest plot below merit further exploration for potential promise at boosting attendance, while those on the far left should be reviewed for potential backfires - though we cannot be certain of impact in either case. Below we provide brief overviews of those trials with the greatest apparent effects in both directions (see Case Studies section).

[18] Original EEF report for trial 67. Our results differ to the original trial estimates due to the 34-fold adjustment for multiple comparisons in our analysis, and different regression model specifications.
[19] The six interventions (in order of magnitude of effect size): Catch-Up Literacy, Dialogic Teaching, Flipped Learning, Children's University, Family and Schools Together (FAST), Philosophy for Children.
[20] Given 94.7% average attendance in the control group, the predicted treatment effect in percentage terms is: invlogit(logit(0.947)+0.33) - 94.7 = 1.4pp, which corresponds to an increase of roughly 2.5 more days attended (0.014*190 = 2.66).

The qualitative results from our primary analysis also held with a linear regression specification (secondary analysis), and the omnibus chi-squared test could not reject the null hypothesis that none of the interventions had an effect on attendance

Figure 1: Attendance reanalysis across 34 EEF randomised trials

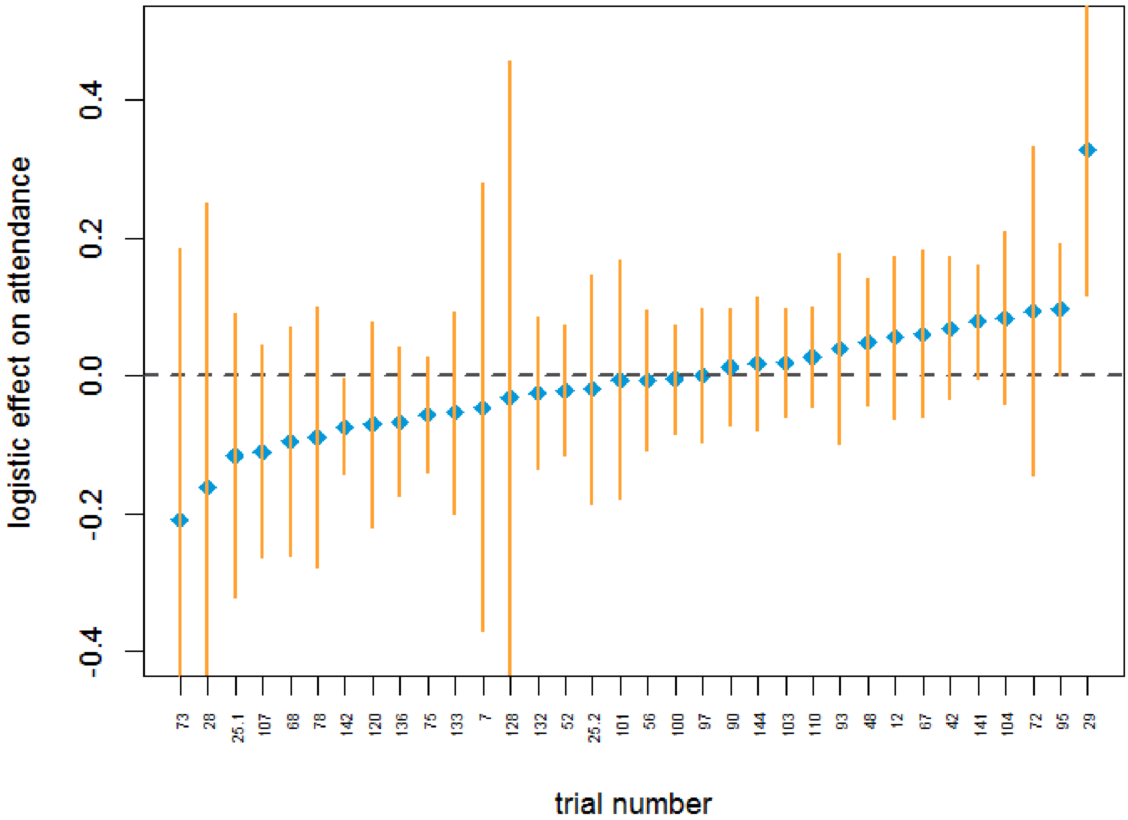


Figure note: Attendance is measured from one year post-intervention.

Exclusion

The exclusion results follow a similar pattern to the attendance results in that while none of the results are statistically significant after correcting for multiple comparisons, there is one trial (Trial 37) with a noticeably larger effect size which still has a relatively small unadjusted p-value. In this case, the intervention is the Nuffield Early Language Intervention and the direction of the effect is to *increase* the average number of exclusions by 0.27 per pupil, with an unadjusted p-value of 0.011. The raw figures reinforce the magnitude of difference – there was one exclusion event in the control group (n=129 pupils), and seventy-four exclusion events in the treatment group (n=265 pupils).[21] As with Trial 29, we cannot be certain whether the programme increased exclusions or not on a statistical basis, due to adjustment for 34-fold multiple comparisons and generally low base rates / sparse data. This trial was also relatively small (n=394). We provide further exploration of this result below (see case studies section). Other results are also notable: Trial 100 appears to reduce exclusions by 0.20 per pupil, and this effect was statistically significant before the multiple comparisons adjustment.[22]

Figure 2: Exclusion reanalysis across 33 EEF randomised trials. Estimated effects by OLS (see figure note).

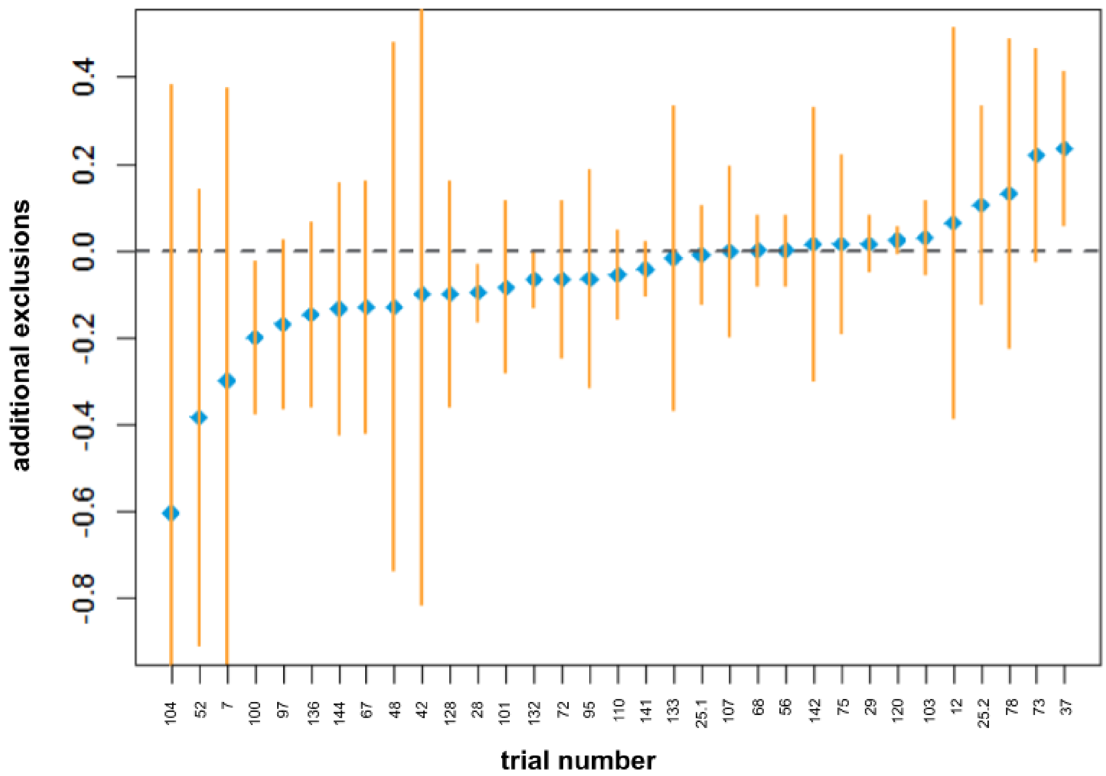


Figure note (text is intentionally italicised and underlined): *For each trial, there are different follow-up periods over which the total number of exclusions are measured.* The total number of exclusions measures all lunch-time, fixed term, and permanent exclusions in all the years between the start of each trial and 2019.

[21] The seventy-four exclusions in the treatment group are accounted for by 12 pupils in seven schools from when the intervention began in 2012, up until 2019, with one pupil recording twenty-one temporary exclusions over the period.
[22] There is also a strong observed reduction noted for Trial 104 [Children's University] - a reduction in exclusions of 0.60 episodes - but given the wide confidence interval in this point estimate, we view this as a finding in need of further verification.

Criminal Justice

Results from the primary analysis show that none of the interventions from the five trials in scope demonstrated any significant impacts on criminal convictions in the twelve months after the intervention actually started. The prevalence of criminal convictions is generally very low for pupils in the EEF archives, reflecting the low prevalence overall in the school-age population (5% of all pupils have a caution or conviction, MoJ-DfE, 2022). Specifically, in our analysis sample, the average proportion of pupils who are convicted within 12 months of intervention start dates is well below 1%. This sparse outcome data means intervention effects are quite noisily estimated, reflected in the wide confidence bands in the first forest plot in Panel A. However, the point estimates for interventions in trials 12 and 56 do reflect potentially large effect sizes (log-odds of 0.50, equivalent to those in the treatment group being roughly 1.6 times less likely to be convicted in the year after the intervention start date). Of course, the lack of statistical significance and sparsity of this outcome data means we cannot attribute these effects to each of the interventions with confidence. The null findings are robust to logistic and linear regression specifications (secondary analysis) and hold when looking at convictions over a shorter follow-up period of 6 months (exploratory analysis).

As additional analyses, we separately tested intervention effects on the *total number,* and *severity,* of criminal convictions within one year after the intervention starting. Again, there were no statistically significant effects.[23] As expected, the results for both of these outcomes follow similar patterns to the primary outcome: whether there were *any* convictions in the twelve months after the intervention actually started. Panel A displays the results for the three key outcomes of interest in this reanalysis. The forest plots show the estimated effect sizes (on the log-odds scale) and unadjusted 95% confidence intervals for each intervention

Serious offences (with a sexual or violent classification) constitute a minority (one in five) of all offences. Therefore, incidences of serious crimes are even more sparse when looking at the year after each of the interventions started in the EEF archive data, which hinders statistical precision when it comes to estimating intervention effects. Indeed, trial 56 had to be omitted from this analysis due to insufficient counts of serious offences. Elsewhere, raw point estimates indicate a large positive intervention effect in trial 12, in that serious offending is less likely, but a potential backfire in trial 103.

---

[23] Following Sutherland (2013) we looked at both reconviction yes/no, frequency of reconviction and severity of reoffending.

Figure 3: Convictions reanalysis across 5 EEF randomised trials. Panel A (left) reconviction yes/no in one year; Panel B (middle) serious crime conviction yes/no in one year; Panel C (right) number of reconvictions in one year

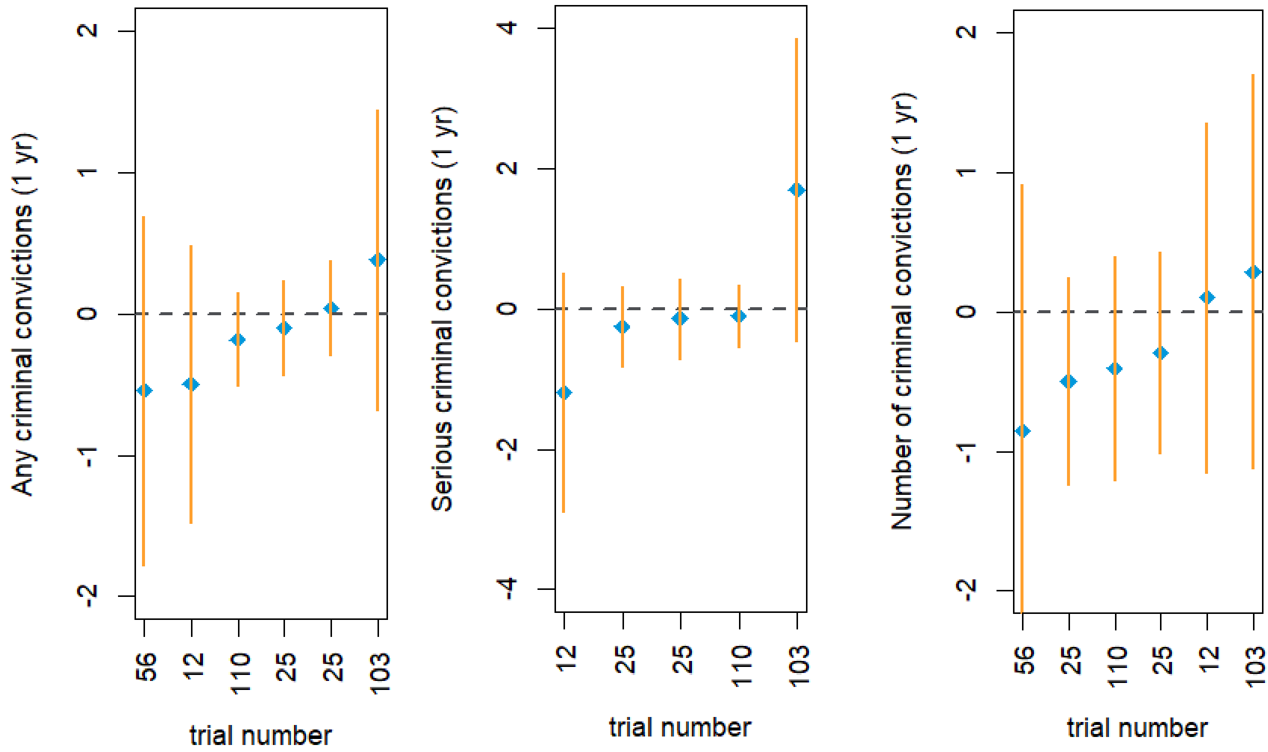


Figure note: Estimated effect sizes are reported on log-odds scales

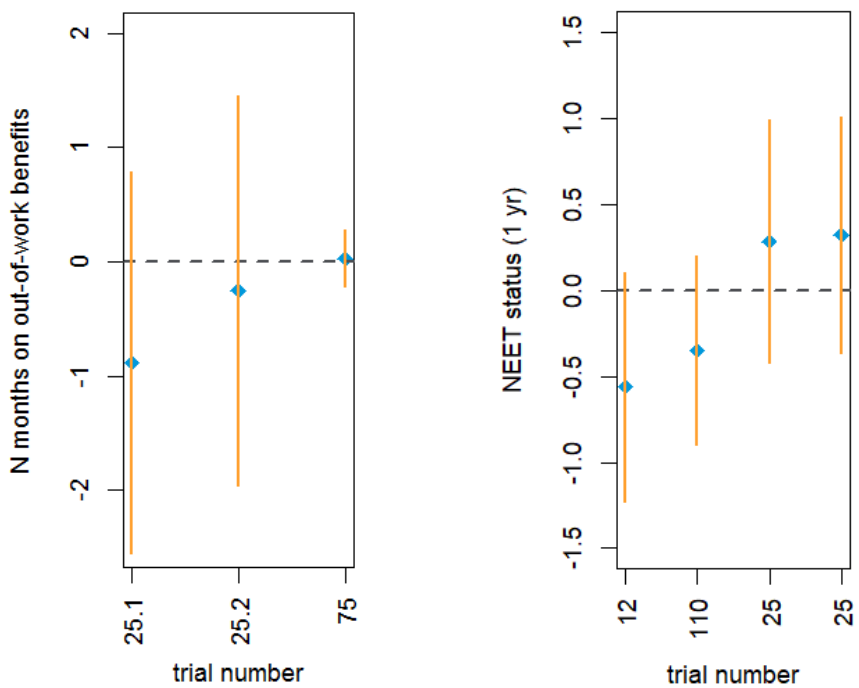
Employment: out-of-work benefits

The analysis on out-of-work benefits included only two trials: trial 25 Increasing Pupil Motivation, which had 2 intervention arms; and trial 75 Paired Reading. Both trials had statistically insignificant effects on the number of months a person was on out-of-work benefits from when they finished Year 13 to the end date in the LEO dataset (5 April 2019), though the interventions from trial 25 directionally reduced months spent on such benefits. A secondary robustness check (logistic regression on whether a person was out-of-work benefits for 3 months or more) also gave null results. Figure 4, Panel A displays the results from the primary analyses on out-of-work benefits and NEET status.

NEET status

Beyond the analysis on out-of-work benefits afforded by the LEO data, there was sufficient matched NPD data on NEET status for three trials. The ONS estimates around 10% of young people aged 16-24 are NEET in the UK. Based on the young people involved in the trials available to us, we estimate a similar but slightly lower figure, likely at least partly because the young people in our trials are towards the lower end of the age bracket borne from a lack of corroborating data to evaluate those whose statuses are missing. Looking at Panel B, and similar to the out-of-work benefits results, we do not observe significant intervention effects on the binary outcome: whether or not the young person was in NEET for three months or more. There is some indication that interventions in trials 12 and 110 had positive effects on reducing NEET, but since these effects are statistically insignificant we cannot be confident that this is a causal relationship and could instead be the result of chance. Both of the interventions in trial 25 are associated with higher rates of NEET, but the small absolute effect size and wide confidence intervals mean we should *not* interpret this result as a convincing backfire.

Figure 4: Panel A (left) Out of work benefits reanalysis on 2 EEF randomised trials (x3 effects in total); Panel B (right) Not in Education, Employment or Training (NEET) reanalysis on 3 EEF trials



5. Case studies & expert survey results

In this section we provide a deeper look at six of the reanalysed EEF interventions that merit further exploration based on our results. We focus on:

- Interventions with the largest effects on the outcomes evaluated.
- Interventions where there was a particularly strong expectation of impact (based on mechanisms and expert hypotheses).

For each of these mini case studies, we summarise the intervention, original trial results, hypothesised mechanisms for impact (based on the expert survey, where applicable), and re-analysis results. Taken together, this information can be used to better understand the overall impacts of the interventions and whether, how, and why particular interventions may be more strongly linked to behavioural outcomes beyond attainment.

Trial 29 - Catch-Up Literacy: Largest positive effect on attendance

This trial had by far the largest effect size on attendance (a log-odds ratio of +0.33, equivalent to roughly two and a half extra days of attendance per school year - a large effect compared to most attendance interventions),[24] but it was not significant after adjusting for multiple comparisons. The intervention is a one-to-one programme to address gaps in pupils' literacy, which also encourages parents to support children's learning. The original evaluation reported a significant positive impact on attitudes to school, self-assessed reading, confidence, and enjoyment in writing - although there was no measured impact on reading scores. The improvement in attitudes to school and confidence could explain the potential impact on attendance. Programme designers in our expert survey also hypothesised a positive impact: *"Pupils are better able to access the curriculum, leading to less frustration and more willingness to attend school."*

However, after adjusting for multiple comparisons, we cannot be confident that this observed effect has not simply arisen by chance. This may be due to the sample size not being large enough to detect an effect of the size estimated. Moreover, we

[24] Rogers, T., & Feller, A. (2018). Reducing student absences at scale by targeting parents' misbeliefs. *Nature Human Behaviour*, *2*(5), 335-342.

did not find any effect from this trial on exclusions (slight directional decrease, but not significant). We note that the larger re-trial of Catch-Up Literacy (trial 133) appears left of centre on our forest plot with no observed impact on attendance. The original evaluation did not find significant results either - potentially due to programme changes required to scale up the programme.

Original trial details	
Year	2012-2015
Randomisation	School-level
Sample size	• 141 schools. • N = 538 pupils. Treatment = 286; control = 271.

Trial 95 - Dialogic teaching: Second largest effect on attendance

The intervention aims to improve KS2 children's learning by improving the quality of classroom communication. It provides teaching resources and teacher mentoring to help teachers spend more class time on meaningful dialogue that encourages pupils to reason, discuss, speculate, argue and explain, rather than simply give the expected answers. This trial had the second largest directionally positive effect on attendance, though it was not significant (a log-odds ratio of +0.1, which would equate to roughly 1 extra day of school attendance per school year). The original evaluation found a positive impact on attainment in English, science and maths (1 to 2 months+). The consistent results across subjects suggest that the approach may have improved children's overall thinking and learning skills. Teachers also reported positive effects' on pupil confidence and engagement. The reported increase in confidence and engagement could have contributed to a potential increase in attendance, though we cannot be sure if this trial had any impact on attendance given the lack of statistical significance. We did not observe any impact from this trial on exclusions (middle of forest plot, very slight directional decrease, but not significant).

Original trial details	
Year	2014-2017
Randomisation	School-level
Sample size	• 69 schools. Treatment = 31; control = 38. • N: 3912 pupils. Treatment = 1832; control = 2080.

Trial 142 - Thinking, Doing, Talking Science: Potential negative impact on attendance

This intervention is a science programme designed to make lessons more practical, creative and challenging. This trial had a weak negative effect on attendance (equivalent to roughly 0.5 day less of school attendance per school year) that was not significant after adjusting for multiple comparisons.[25] This was a small trial with relatively few clusters (41 schools, randomised at school level). Therefore the evidence on whether this small negative result is directly due to the intervention, or chance, is not conclusive but given the small sample size the probable interpretation is that this is due to small sample variation.

The original trial had positive results on science attainment (3 months' additional progress) with indications of particular benefits for pupils on free school meals, and reports of a positive impact on attitudes to science. EEF also tested a scaled-up version of the intervention, which did not have a significant impact.
There is no obvious logic to explain a potential backfire effect on attendance. The trial had no significant effect on exclusions, though the directional trend was to increase exclusions (the fourth largest increase detected). Again, this result is likely to be down to chance given the lack of statistical significance.

Original trial details	
Year	2013
Randomisation	School-level
Sample size	• 41 schools. Treatment = 21; control = 20. • Total N = 1,264. Treatment = 655; control = 609.

Trial 37 - Nuffield Early Learning Intervention: Largest increase in exclusions

This intervention involves weekly tuition for small groups of children identified as having relatively low language skills and is designed to improve listening, narrative and vocabulary skills for KS1 students. The original trial found positive impacts on

[25] Interestingly this result would be significant at 5% if it was being looked at in isolation. As noted, we took a very conservative approach to significance testing in this report but how to approach this when conducting long-term follow-ups is an area of methodological development.

children's language skills (2 months' extra progress for the 20-week version and 3 months' extra progress for the 30-week version).

In this reanalysis, the trial appears to have significantly *increased* the average number of exclusions from 2012 to 2019 by 0.27 per pupil, with an unadjusted p-value of 0.011 (p-value > 0.05 after adjustment for multiple hypothesis testing). Although this result suggests a backfire effect, it is hard to be certain that the programme actually caused more exclusions compared to the control group. Moreover, because this trial was small (total n = 394), the analysis is somewhat underpowered and any estimated effects may suffer from Type S (Sign) and or Type M (Magnitude) Errors (see Gelman et al., 2017). Nevertheless, the measured effect warrants closer scrutiny, especially given the current scale up of this intervention and previous effectiveness trial (Sutherland et al., 2019), as well as the 'at risk' nature of the target group: the material costs of implementing a backfiring intervention in this context would be detrimental.

Original trial details	
Year	2012-2016
Randomisation	Pupil-level
Sample size	• 68 schools • 20-week Programme: Total N = 236; Treatment N = 121; control N = 115 • 30-week Programme: Total N = 229; Treatment N = 114; control N = 115.

Trial 104 - Children's University: Expected impact on attendance (due to pupils' interest in extracurricular activities at school), but no statistically significant impact observed. Potential reduction in exclusions

This trial tested the impact of social learning opportunities, after-school clubs, and cultural visits on educational attainment provided to KS2 pupils in Year 5 and Year 6. The intervention ran for 2 years in Year 5 cohorts, and 1 year in Year 6. The original trial had a range of positive outcomes, including: reading and maths scores (maths only for pupils eligible for free school meals); "teamwork" and "social responsibility" (small impacts); and a range of non-cognitive outcomes, such as future aspirations, communication, empathy, and resilience. This led programme experts to hypothesise an impact on attendance:

> *"Anecdotally, we are told that in some circumstances children want to attend school so they can then attend the club or activity. Children's University is also designed to encourage a love of all learning in children, so it is hoped an enthusiasm for learning outside the classroom is reflected in an enthusiasm for learning in it too."*

We observed a potential reduction in exclusions, which was not significant after adjusting for multiple comparisons, and we observed a directional improvement in attendance (fourth largest increase), but this result was also not statistically significant. We do note, however, that there would likely be positive spillover effects from greater interest in extracurricular activities among treated pupils. That is, control group pupils would likely become more interested in activities if other pupils attend.

Original trial details	
Year	2014-2016
Randomisation	School-level
Sample size	• 68 schools. Treatment = 36; control = 32. • Total pupils N = 1,224. Treatment = 650; control = 574.

Expert prediction results

The idea of eliciting responses from experts was to help to understand how well experts were calibrated to the direction / magnitude of possible impacts on outcomes other than attainment; and also to understand whether it is possible to 'read across' impacts and make predictions based on existing evidence. Previous research illustrates that in *some cases* experts are no better than chance and predicting effectiveness (Tetlock, 2005) but *some experts* are able to make reliably accurate predictions about outcomes (Tetlock and Gardner, 2015).

In total we received responses covering 15/34 trials from experts involved in either the delivery or evaluation of a given intervention. Table 2 below sets out the comparison of expert responses to actual outcomes observed for attendance and exclusions. The cells highlighted in green are where the predictions from experts and the results align. We focused on the *direction of results* rather than the magnitude or statistical significance, meaning that we are making a judgement about whether predictions were accurate. Overall, experts were more accurate at predicting the results for attendance (60% accuracy) rather than exclusions (53% accuracy). At face value it may be that attendance is less *effort* to link to interventions because it has a more positive frame (Steiger and Kuhberger, 2018) - i.e. interventionists tend to be optimistic about positive outcomes (e.g. Chalmers and Matthews, 2006; Sharot, 2011).

We might wonder what benefits come from such predictions. One is that it is costly to repeat experiments, as we have noted, and in some cases where sample sizes are particularly small and trials 'uninformative' (Lortie-Forgues and Inglis, 2019), being able to extrapolate may prove fruitful. Another reason is that through further testing of this approach it may be possible to identify individuals who can make reliably accurate predictions about early-stage interventions, which may help to streamline decision-making about what to progress to further testing (or not). Finally, this exercise of forecasting effectiveness is also, usually implicitly, being undertaken all the time when it comes to the implementation of new ideas, particularly when those ideas are imported from other contexts or countries (Cartwright and Hardie, 2012). Practising prediction can build skill in predicting effectiveness, as well as build humility; we will often be wrong because we are overconfident. Reflecting on how similar projects have gone in the past is already encouraged to reduce optimism bias in infrastructure projects, but would also be valuable with social programmes which are implemented much more frequently (Hallsworth et al., 2018).

Table 2: Expert elicitation results (x15 trials)

Intervention	Attendance		Exclusions	
	Hypothesis	Actual	Hypothesis	Actual
68	No impact	No impact	No impact	No impact
103	No impact	No impact	Decrease	No impact
29	Increase	Potential Increase	Decrease	No impact
78	No impact	No impact	No impact	No impact
104	Increase	No impact	No impact	Potential decrease
110	No impact	No impact	No impact	No impact
141	No impact	No impact	Small decrease	No impact
25	Increase	No impact	Decrease	No impact
97	No impact	No impact	No impact	No impact

128	No impact	No impact	No impact	No impact
100	No impact	No impact	No impact	No impact
93	Increase	No impact	No impact	No impact
101	Increase	No impact	Decrease	No impact
67	Increase	No impact	Decrease	No impact
142	Increase	No impact	No impact	No impact
Accuracy	**60%**		**53%**	

6. Discussion

The purpose of this project was to link completed randomised controlled trials to administrative data to demonstrate that this can be done, what value it can bring to data owners, and its utility for policy/policy evaluation. We have been successful as a demonstration project in being able to link data and conduct further analysis on completed trials. We have also helped to test and develop processes and systems for accessing and using administrative data, but there are still many frictions to doing this efficiently for government and external teams (Gibbons et al., 2020).

The value to policy-making

We cannot build knowledge of policy effectiveness without data, and policy cannot wait years for outcomes and analysis. Combining administrative data with randomised controlled trials brings both short- and long-term benefits. In the short-term, it may allow trial teams to understand if and how trial participants progress with the systems they interact with (e.g., do they remain in education and how consistent is their attendance if so?). This means that rather than waiting to look at educational attainment several years post-intervention, we can look at whether *attendance* is affected in the short-term, including negatively, by the new intervention. In the long-term, we can better understand whether policies designed for one specific outcome generate additional benefits, involve trade-offs between outcomes, or have no impact on other outcomes of interest. For example:

- We increase reading ability and this also leads to additional gains in terms of attendance as well as reductions in behaviour problems.
- We can intervene to improve attainment and this has no impact (negative or positive) on other outcomes.
- We know that if we implement a specific intervention with attainment gains, we will negatively affect other outcomes and therefore have to make a judgement about whether to implement or not.

This cross-policy thinking fits with, for example, the shared outcomes fund and the work of the evaluation task force,[26] meaning that this project has substantial relevance to impact evaluations conducted within government and by independent teams. Working across departments ahead of policy roll-out, knowing the outcomes will be measured in different policy areas will add different dimensions to policy design and thinking. That effects may vary by outcome requires more thought about the mechanisms through which those changes might occur (Cartwright and Hardie, 2012). That in turn relates to the growing recognition that average treatment effects may tell us less than we think they do, and that we should specifically look for how policy effects vary across groups as another way to refine and hone policies (Vivalt, 2015; Wager and Athey, 2015).

Being able to track outcomes in near real-time also opens up the possibility of government departments using empirically-based ‘stopping rules’ when piloting new policies. Stopping rules are criteria that determine that if an outcome or measure differs by a pre-specified amount at a certain time, the trial is stopped (see Phillips, 2015; Pocock, 1983). This would mean that an intervention with negative consequences could be much more quickly detected and halted.

Making data sharing easier and faster will make policy better and cheaper

There is a pressing need *now* to rigorously evaluate the effectiveness of policy in the UK (NAO, 2021). Part of the barrier to evaluation is the time it takes for outcomes to be realised or measured. Being able to ‘fail fast’ and adjust programmes in real time and understand the impacts on intermediate outcomes are uncommon but valuable for policy making. Being able to do so allows for much faster policy evolution, as well as a wider range of policy options to be tested in parallel, through, for example, adaptive trials (Hopkins et al., 2020).

[26] https://www.gov.uk/government/publications/shared-outcomes-fund-round-two; https://www.gov.uk/government/organisations/evaluation-task-force

Balancing the bureaucratic burden of data access with legal requirements and the rights of citizens will require regular attention. In particular the *processes* governing access are at risk of being gummed up with 'sludge' (Sunstein, 2021). Data owners need to regularly reassess the burdens they place on requesters and seek to harmonise processes and aggressively minimise burdens (see e.g. Appendix V).

Future iterations and versions of data sharing can reduce some of the barriers we faced through the use of privacy preserving technologies such as synthetic data (Calcraft et al., 2021). There may also be other approaches that do not rely on bi-lateral individual data sharing in and out of secure settings, for example, what we term 'asymmetric linkage'. This is where trial participant information in Department A is sent to Department B to be linked to additional outcomes but instead of linked data being returned, Department B send means and standard deviations for the treatment and control groups i.e. enough information to assess whether the groups differ.

Working in this way to link data quickly to see and share interim results does not have to be limited to inter-governmental projects. What Works Centres could also be using a similar approach to track intermediate outcomes more quickly. The volume of trials completed and underway in What Works Centres, if reanalysed, could quickly build a substantial body of knowledge about what is or is not effective across multiple outcomes that benefit their target populations now and in the future.

Appendices

Appendix I: Table A1. EEF trials analysed & outcomes assessed

Trial Number	Trial Name	Year	Sample Size	Exclusions	Attendance	NEET	Justice
7	Changing Mindsets - Pupil workshops	2013	Number schools = 10 Number pupils = 174	Y	Y	-	-
12	LIT Programme	2012	Number schools = 34 Number pupils = 4,413	Y	Y	-	Y
25	Increasing Pupil Motivation	2012	Number schools = 63 Number pupils = 10,250	Y	Y	Y	Y
28	REACH	2013	Number schools = 21 Number pupils = 202	Y	Y	-	-
29	Catch-up Literacy	2013	Number schools = 15 Number pupils = 557	Y	Y	-	-
37	Nuffield Early Learning Intervention	2012	Number schools = 68 Number pupils = 394	Y	N	-	-
42	Philosophy for Children	2012	Number schools = 48 Number pupils = 1,529	Y	Y	-	-
48	Shared Maths	2012	Number schools = 79 Number pupils = 2,786	Y	Y	-	-

52	Chess in Primary Schools	2013	Number schools = 100 Number pupils = 3,865	Y	Y	-	-
56	Let's think Secondary Science	2013	Number schools = 47 Number pupils = 5,882	Y	Y	-	Y
67	Texting Parents	2014	Number schools = 29 Number pupils = 15,697	Y	Y	Y	-
68	Abracadabra (ABRA)	2013	Number schools = 48 Number pupils = 1,591	Y	Y	-	-
72	Flipped Learning	2013	Number schools = 24 Number pupils = 1,129	Y	Y	-	-
73	Graduate Coaching Programme	2013	Number schools = 4 Number pupils = 291	Y	Y	-	-
75	Peer Tutoring in Secondary Schools	2013	Number schools = 20 (60 classes) Number pupils = 1,306	Y	Y	Y	-
78	Changing Mindsets - Inset	2013	Number schools = 24 Number pupils = 885	Y	Y	-	-
90	Parenting Academy	2014	Number schools = 16 Number pupils = 1,895	-	Y	-	-
93	ReflectED Metacognition	2014	Number schools = 30 (70 classes) Number pupils = 1,570	-	Y	-	-
95	Dialogic	2015	Number	Y	Y	-	-

	Teaching		schools = 69 Number pupils = 3,912				
97	Learner Response System	2013	Number schools = 94 Number pupils = 2,829	Y	Y	-	-
100	Research Learning Communities	2014	Number schools = 116 Number pupils = 4,966	Y	Y	-	-
101	Switch On Reading (re-grant)	2015	Number schools = 183 Number pupils = 902	Y	Y	-	-
103	Best Practice in Grouping Students Intervention A: Best Practice in Setting	2015	Number schools = 115 Number pupils = 3,322	Y	Y	-	Y
104	Children's University	2014	Number schools = 68 Number pupils = 1,224	Y	Y	-	-
110	Embedding Formative Assessment	2015	Number schools = 140 Number pupils = 25,393	Y	Y	-	Y
120	Family skills	2016	Number schools = 102 Number pupils = 1,985	Y	Y	-	-
128	Maths Counts	2015	Number schools = 35 Number pupils = 291	Y	Y	-	-
132	IPEEL (re-grant) 1-year programme	2015	Number schools = 83 Number pupils = 2,465	Y	Y	-	-
133	Catch-up	2016	Number	Y	Y	-	-

	Literacy (re-grant)		schools = 141 Number pupils = 1,006				
136	IPEEL (re-grant) 2-year programme	2015	Number schools = 78 Number pupils = 2,196	Y	Y	-	-
141	Families and Schools Together (FAST)	2015	Number schools = 115 Number pupils = 4,221	Y	Y	-	-
142	Thinking, Doing, Talking Science	2012	Number schools = 41 Number pupils = 1,264	Y	Y	-	-
144	Changing Mindsets (re-grant)	2016	Number schools = 101 Number pupils = 4,437	Y	Y	-	-

Appendix II: Table A2. EEF trial reanalysis: key outcome summary statistics by intervention group

Trial No	Trial Name	Year	Trial group sizes (N)		Exclusions (total number of exclusions)		Exclusions (N with any exclusions)		Attendance rate (% sessions attended)		PNC (N with any convictions)		PNC (total number of convictions)		NPD (NEET status)		LEO (months on out-of-work benefits)	
			C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T
7	Changing Mindsets - Pupil workshops	2013	99	96	133	83	18	19	mean: 95.5 sd: 4.7	mean: 95.4 sd: 4.9								
12	LIT Programme	2012	2073	3403	1060	1897	332	621	mean: 95.5 sd: 5.0	mean: 95.7 sd: 4.6	8	8	11	20	162	134		
25	Increasing Pupil Motivation	2012	5553	T1:2689 T2:2407	330	T1:184 T2:238	235	T1: 133 T2: 154	mean: 93.6 sd: 10.6	T1 mean: 92.5 T1 sd: 12.3 T2 mean: 92.6 T2 sd: 11.5	109	T1: 48 T2: 49	623	T1: 183 T2: 201	450	T1: 189 T2: 198	mean: 5.7 sd: 12	T1 mean: 4.8 T1 sd: 10.8 T2 mean: 5.2 T2 sd: 11.8
28	REACH	2013	64	133	76	142	16	38	mean: 94.9 sd: 6.2	mean: 94.4 sd: 7.5								

Trial No	Trial Name	Year	Trial group sizes (N)		Exclusions (total number of exclusions)		Exclusions (N with any exclusions)		Attendance rate (% sessions attended)		PNC (N with any convictions)		PNC (total number of convictions)		NPD (NEET status)		LEO (months on out-of-work benefits)	
			C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T
29	Catch-up Literacy	2013	282	302	249	284	63	63	mean: 94.8 sd: 6.8	mean: 96.2 sd: 4.1								
37†	Nuffield Early Learning Interventio n	2012	129	265	1	74	1	12										
42	Philosophy for Children	2012	1708	1472	1481	1144	373	321	mean: 95.9 sd: 4.4	mean: 96.0 sd: 4.3								
48	Shared Maths	2012	3177	3289	2359	2082	511	461	mean: 95.7 sd: 4.6	mean: 95.9 sd: 4.3								
52	Chess in Primary Schools	2013	1952	2051	2381	1664	464	475	mean: 96.0 sd: 4.8	mean: 96.0 sd: 5.2								
56	Let's think Secondary Science	2013	4036	3980	2487	2479	617	587	mean: 95.2 sd: 5.9	mean: 95.2 sd: 5.9	7	4	19	8				
67	Texting Parents	2014	8103	7557	3180	1968	908	676	mean: 94.6 sd: 7.6	mean: 94.9 sd: 8.0							mean: 0.9 sd: 3.3	mean: 1.1 sd: 3.9

Trial No	Trial Name	Year	Trial group sizes (N)		Exclusions (total number of exclusions)		Exclusions (N with any exclusions)		Attendance rate (% sessions attended)		PNC (N with any convictions)		PNC (total number of convictions)		NPD (NEET status)		LEO (months on out-of-work benefits)	
			C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T
68	Abracadabra (ABRA)	2013	1157	1081	135	133	36	38	mean: 96.0 sd: 4.0	mean: 95.7 sd: 4.4								
72	Flipped Learning	2013	636	578	324	244	82	71	mean: 96.0 sd: 4.1	mean: 96.5 sd: 4.3								
73	Graduate Coaching Programme	2013	187	186	208	219	62	52	mean: 96.5 sd: 4.6	mean: 95.7 sd: 5.8								
75	Peer Tutoring in Secondary Schools	2013	1335	1244	362	372	165	142	mean: 95.5 sd: 5.1	mean: 95.2 sd: 5.3							mean: 0.8 sd: 2.3	mean: 0.9 sd: 2.3
78†	Changing Mindsets - Inset	2013	610	425	411	311	94	76	mean: 96.1 sd: 5.7	mean: 95.6 sd: 4.4								
90	Parenting Academy	2014	1341	1241					mean: 95.9 sd: 3.8	mean: 95.9 sd: 4.0								
93	ReflectED Metacognition	2014	859	994					mean: 96.2 sd: 4.2	mean: 96.4 sd: 4.1								

Trial No	Trial Name	Year	Trial group sizes (N)		Exclusions (total number of exclusions)		Exclusions (N with any exclusions)		Attendance rate (% sessions attended)		PNC (N with any convictions)		PNC (total number of convictions)		NPD (NEET status)		LEO (months on out-of-work benefits)	
			C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T
95*	Dialogic Teaching	2015	2583	2489	974	830	273	260	mean: 95.9 sd: 4.6	mean: 96.3 sd: 4.1								
97†	Learner Response System	2013	3255	2812	3134	2299	719	576	mean: 95.7 sd: 4.7	mean: 95.7 sd: 4.8								
100	Research Learning Communities	2014	2777	2617	1494	929	355	317	mean: 96.5 sd: 3.7	mean: 96.5 sd: 3.6								
101†	Switch On Reading (re-grant)	2015	2711	6185	392	505	110	170	mean: 95.2 sd: 4.8	mean: 95.2 sd: 4.4								
103*	Best Practice in Grouping Students Intervention A: Best Practice in Setting	2015	12906	11793	5920	5733	1644	1594	mean: 95.2 sd: 6.9	mean: 95.3 sd: 6.9	6	8	14	17				
104*	Children's University	2014	1671	2163	2284	1242	281	234	mean: 96.2 sd: 4.3	mean: 96.8 sd: 3.8								
110	Embedding Formative Assessment	2015	13277	12600	2974	2220	1205	1055	mean: 94.3 sd: 8.3	mean: 94.5 sd:8.4	82	65	349	221	470	423		

Trial No	Trial Name	Year	Trial group sizes (N)		Exclusions (total number of exclusions)		Exclusions (N with any exclusions)		Attendance rate (% sessions attended)		PNC (N with any convictions)		PNC (total number of convictions)		NPD (NEET status)		LEO (months on out-of-work benefits)	
			C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T
120	Family skills	2016	1048	942	2	26	2	6	mean: 95.4 sd: 5.8	mean: 95.1 sd: 5.8								
128	Maths Counts	2015	148	149	32	18	9	8	mean: 95.3 sd: 6.3	mean: 95.0 sd: 5.2								
132	IPEEL (re-grant) 1-year programme	2015	1335	1347	722	664	162	174	mean: 96.2 sd: 5.1	mean: 96.1 sd: 4.7								
133	Catch-up Literacy (re-grant)	2016	580	634	100	108	36	46	mean: 95.6 sd: 6.0	mean: 95.1 sd: 5.0								
136	IPEEL (re-grant) 2-year programme	2015	1363	1453	499	317	119	113	mean: 96.5 sd: 3.8	mean: 96.2 sd: 3.8								
141	Families and Schools Together (FAST)	2015	3723	3525	433	279	104	95	mean: 95.5 sd: 4.4	mean: 95.8 sd: 4.1								
142	Thinking, Doing, Talking Science	2012	4402	4341	712	828	193	231	mean: 96.7 sd: 3.7	mean: 96.3 sd: 4.3								

Trial No	Trial Name	Year	Trial group sizes (N)		Exclusions (total number of exclusions)		Exclusions (N with any exclusions)		Attendance rate (% sessions attended)		PNC (N with any convictions)		PNC (total number of convictions)		NPD (NEET status)		LEO (months on out-of-work benefits)	
			C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T
144	Changing Mindsets (re-grant)	2016	2455	2420	1114	745	239	206	mean: 96.3 sd: 4.3	mean: 96.4 sd: 4.3								

** Arm sizes differ from original EEF trials after data linkages.*

- ***95 Dialogic Teaching**: Our sample in the control arm is larger than the analysis sample and the number of pupils randomised in the trial into control. We do not know why there are more pupils but this is the number of control pupils in the archive.*
- ***103 Best Practice in Grouping**: Our samples are larger than the analysis sample in the original trial. For the original, many pupils randomised were dropped because follow-up outcome measures could not be collected. With the administrative data linkage we were able to collect data on more pupils.*
- ***104 Children's University** has an imbalance: "The initial survey included all pupils in the Years 4 and 5 cohorts – a total of 3,840 (2,166 in 36 treatment schools and 1,674 in 32 control schools). Because 68, rather than 80, schools were recruited the developers wanted slightly more schools to work with in the first stage, and this unequal division of the 68 was agreed with the evaluators and funders."*

† Interventions from these trials have been grouped together for the purposes of the reanalysis

- ***37 NELI**: We grouped the 20 week and 30 week versions for this project meaning the treatment sample is larger overall.*
- ***Changing Mindsets**: We grouped the teacher training and the pupil workshop interventions for this project.*
- ***97 Learner Response System**: we grouped the samples for the 1- and 2-year programmes.*
- ***101 Switch on reading**: we grouped 'Switch-on reading', and 'switch-on reading & writing' interventions, explaining (N=3,000 in each).*

Appendix III: Theory of Change process description:

Research Questions:

- Based on the programme model and published findings, is there a hypothesised causal pathway between the intervention and our outcomes of interest (attendance, fixed term exclusions, permanent exclusions)?
- If yes, what are the possible causal pathways?

Process to complete the template below:

1. Start by filling out the background info section.
2. Populate your resource list (use hyperlinks). The most common are:
    a. EEF project page -- Final report (most useful), study protocol (probably less useful). Note: May be both an efficacy trial report and an effectiveness trial report.
    b. Studies cited in the final report
    c. Programme webpage
3. Begin looking for evidence. What to look for:
    a. Logic model or theory of change (also copy a screenshot of it below the correct heading)
    b. IPE findings related to:
        - pupil behaviour
        - pupil motivation
        - pupil engagement
        - parent engagement
        - changes in teacher/TA practice (particularly around classroom or behaviour management)
        - skills of deliverer
    c. Testimonials on webpage (these will obviously be positively biased but may indicate other hypothesised mechanisms)
4. Populate the ‘Evidence’ table.
    a. Feel free to copy/paste findings. Just be sure to include a citation with page number.
    b. Indicate if it is a + or - finding and which actor it corresponds to (i.e., who is changing their behaviour?)
    c. Highlight key terms in the finding.
5. Generate hypotheses.

   a. Review your evidence table and start to generate all possible hypotheses about ways the intervention might increase or decrease our outcomes of interest.
6. Add any notes on further research that might be useful to explore causal mechanisms.
7. Outputs saved [here]

Template

Programme name:

Author:

Background info:

- Population:
- Intervention:
    - Deliverer: (Teacher, tutor, TA, external provider)
    - Mode of delivery: (full classroom, 1:1, small group)
    - Dosage: (length of sessions, frequency, etc)
- Outcomes (from trials): [copy/paste the main findings]

Hypotheses based on evidence review:

Attendance:

- May **[increase/decrease] attendance through [pupils/teachers/parents]**

Exclusions

- May [**decrease/increase] exclusions through [pupils/teachers/parents]**

Suggested further research

- Explain here

Resource list:

- Efficacy trial**: (**EEF page**)** ← hyperlink
- Effectiveness trial **(**EEF page**)**

Evidence:

Evidence:	Direction:	Actor:

Programme logic model:

(copy/paste with citation)

Project & evaluation staff:

(copy/paste from report -- in case we want to do targeted follow-up)

The deliverable for this stage is a tailored project initiation document with:

- a definition of the problem;
- KPIs and outcome measures;
- an updated project plan;
- a map of stakeholders and data sources; and
- an initial assessment of risks and opportunities (which will be continuously updated throughout the duration of the project as part of our project management).

Note: Explain why this deliverable format is suitable/appropriate

Quantitative datasets	Policy documents, existing studies and academic literature	Interviews	Observations

Appendix IV: Cost efficiency of trial reanalysis calculations

We base our cost efficiency estimate on how much EEF randomised trials cost, covering both delivery (intervention) and evaluation. The total project grants for each project are available on the EEF website,[27] but we base our calculations on cost of delivery being in the region of £500,000 and evaluation costing £250,000, so £750,000 in total. For some projects that would be an overestimate, and for some, an underestimate. We think it would particularly underestimate projects that required significant input from EEF staff, because their time would not be reflected in the grant amount. Similarly, projects requiring bespoke data collection from pupils or schools would also be much more expensive, as such exercises can cost upwards of £100,000 depending on how many schools and pupils there are.

The other factor that isn't included in costs is the timeline for delivering such a project. Typically the process for procuring and recruiting delivery and evaluation teams begins almost a year before delivery begins, with delivery typically starting in the first term of the school year (Sept-Dec). Interventions may then 'run their course' over a school year or longer, and data collection would take place at the end of the school year following intervention. If administrative data is used it might then take six months for that to be available, with analysis and write up taking another six months depending on the project complexity. All in all, the process is roughly three years from start to finish for projects that deliver interventions in a relatively short period of time.

In table A3 below we set out estimates of how much money has been saved by not having to procure and re-run RCTs of the same interventions to look at these outcomes. The total number of re-analysed trials varies because in some cases pupils were not old enough to be included in the reanalysis for a given outcome. Two examples are (i) pupils were not old enough to appear in criminal convictions data because they were younger than 10 years old at the point PNC data were available, and (ii) had not yet finished compulsory education so would not feature in data looking at employment.

It is important to note that the cost savings will only increase as more trials reach 'maturity' for inclusion (i.e. pupils are old enough to be included in all reanalyses). This means that across the four outcomes included, if all trials were reanalysed (n=32), **the maximum cost saving would be more than £500 per £1 spent on reanalysis.**

[27] https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects

Table A3: Cost efficiency calculations for reanalysis

Outcome	Number of re-analysed trials[+]	Total cost of re-running trials	Cost saving from reanalysis per outcome*
Attendance	34	34 x £750,000 = £25,500,000	£25,500,000 / £183,800 = £138.74
Exclusions	33	£24,750,000	£134.66
NEET	3	£2,250,000	£12.24
Convictions	5	£3,750,000	£20.40
Total	n/a	£54,000,000	£293.80

Table notes: [+] This is the number of trials analysed for primary analysis for each outcome. * Total cost of re-running trials X number of trials analysed, divided by the total grant funding awarded to BIT by ESRC (a fixed cost).

Appendix V: Example of administrative process improvement

We set out below the steps in the process for extending project timelines within the SRS (left) and (right) suggested approach.

	Current process steps	Proposed process steps
1	ONS SRS emails to remind that deadline is close	ONS SRS emails team to remind that deadline is close & includes option to request extension and sets out required information, cc'ing data owner
2	Reply to ONS to say extension desired	Requesting team replies to all with requested information
3	ONS replies to say extension needs to go to data owner (DfE)	Data owner considers request
4	Email to data owner to request extension	Once a decision is made, data owner replies to the requesting team & ONS with confirmation of request & existing paperwork is automatically updated & sent out to the requester team. No signature is needed because this is an extension of an existing project from an accredited organisation with accredited researchers working on the project.
5	Data owner sends form to request reasons for extension	
6	Paperwork returned for extension request	
7	Request considered by data owner	
8	Decision made - paperwork sent to the requester team for completion and signature.	