



Engage in Education
Long term follow-up of exclusions and attainment from a high-risk sample of pupils in the London education and inclusion project

Evaluation Report

July 2019

Independent evaluators: Alex Sutherland, Miriam Broeks, Yulia Shenderovich and Ingrid Obsuth

RAND Europe



EUROPE






The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus (formerly Impetus Trust) and received a founding £125m grant from the Department for Education. Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.

For more information about the EEF or this report please contact:

-  Jonathan Kay
Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP
-  0207 802 1653
-  jonathan.kay@eefoundation.org.uk
-  www.educationendowmentfoundation.org.uk

Contents

Glossary	3
About the evaluator	4
Executive summary	5
Introduction	7
Methods	10
Impact evaluation	17
Conclusion	23
References	26
Appendix A: EEF cost rating	29
Appendix B: Security classification of trial follow-up findings	30
Appendix C: Variables used for randomisation & balance checks	31
Appendix D: Original study questions and summary of original educational and exclusion outcomes	35
Appendix E: Predicting loss to follow-up analysis based on n=738 pupils consented at baseline	41
Appendix F: Ethics, trial registration & data protection	42
Appendix G: Non-educational outcome measures	44
Appendix H: Distribution of outcomes & baseline variable	46
Appendix I: Sensitivity analyses	48
Project team	50
Acknowledgements	50

Glossary

EiE-L	Engage in Education London (the intervention being evaluated)
FIML	Full-Information Maximum Likelihood
GCSE	General Certificate of Secondary Education
ITT	Intention to Treat
LEIP	London Education and Inclusion Project
MDES	Minimum Detectable Effect Size
NPD	National Pupil Database
REML	REstricted Maximum Likelihood

About the evaluator

The project was independently evaluated by a team from RAND Europe with support from Dr. Ingrid Obsuth. The team from RAND included:

Alex Sutherland, Miriam Broeks and Yulia Shenderovich.

The original evaluation was led by Professor Manuel Eisner, Institute of Criminology, University of Cambridge and the protocol and results have been published as journal articles (Obsuth et al., 2014; 2016, 2017).

The lead evaluator for this follow-up study was Alex Sutherland.

Contact details: Dr. Alex Sutherland

Address: RAND Europe, Westbrook Centre, Milton Road, Cambridge CB4 1YG

Tel: 01223 353 329

Email: alex_sutherland@rand.org

Executive summary

The project

The Engage in Education London intervention (EiE-L) aimed to improve the behaviour of pupils in Years 9 and 10 most at risk of exclusion. Pupils met twice a week for 12 weeks for one-to-one and group work. Sessions were led by two external youth workers trained by the delivery team, Catch22. Each of the 12 sessions had a different theme, addressing topics such as effective communication, anger management, and de-escalation. One-to-one support was provided by a keyworker in areas of identified need. In-school sessions were supplemented by home visits and phone calls with parents or care-givers. Teaching staff were also offered training on classroom behaviour management.

Seven hundred and fifty-one pupils in 36 schools were originally nominated by teachers for intervention. These pupils were identified as suitable for the programme based on (i) past problem behaviour, (ii) exclusions, or (iii) unauthorised absences from school. Seventeen schools were randomly allocated into treatment (EiE-L) and 19 to an active control (a one-day careers advice workshop). The primary outcomes were a reduction in the number of exclusions and improvements in academic attainment. The main results of the study have been reported previously (in Obsuth et al., 2016, 2017), and are summarised below. This summary focuses on follow-up data collected two years after the intervention.

The original study was funded by both the EEF and the European Commission (via a grant to the Greater London Authority). This follow-up study was funded by the EEF.

Key Conclusions from the immediate outcome data (reported in Obsuth et al., 2016, 2017)

1. In the short term, there was no evidence that the intervention reduced exclusions. Fixed-term exclusions, as reported by pupils, increased slightly in treatment schools.
2. It was challenging to deliver the programme as intended. Twenty-four sessions were planned with each pupil (12 group, 12 individual) but only around seven of each took place. Although additional support to parents via phone calls and home visits were intended, few took place (n = 164 phone calls and 11 home visits).

Key conclusions from the long term follow up (two years after intervention)

1. There was no evidence that the intervention improved the number of GCSEs achieved (graded A–G), two-years post-intervention follow-up.
2. There was no evidence that the intervention reduced fixed-term school exclusions at two-year post-intervention follow-up. The possible negative finding immediately after the intervention was not evident at the delayed follow-up.
3. It was not possible to draw conclusions about the impact of the intervention on further education uptake, police arrest records, or NEET status because data access and other issues precluded analyses of these outcomes.

Security of findings

Longitudinal follow-up results commissioned by the EEF do not receive a padlock security rating. The original evaluation was not designed to detect the effect on the longitudinal outcomes in this report and was not powered for this purpose. However, the findings need to be considered with caution given the attrition from follow-up. The study uses an experimental design, powered to detect an effect size of 0.35 on educational attainment using assumptions consistent with the original design. Due to 30% attrition from the GCSE outcomes, and 25% from the exclusions outcome, we cannot be certain that the result is attributable to the intervention. The security of the immediate outcome data is discussed further below, and in the published papers (Obsuth et al., 2016, 2017).

Additional findings

After receiving the intervention, students in treatment schools were more likely to self-report that they had been excluded in the previous month (on a fixed-term basis) than students in comparison schools. Moreover, although not statistically significant, official records of exclusion revealed the same direction of effects: more exclusions in the schools that received EiE-L. There were also associations between high levels of compliance and higher self-reported exclusion.

Immediate post-intervention analyses revealed no significant differences between the students in the treatment schools versus control schools on any secondary outcomes.

The original project sought to collect educational attainment data, but a substantial proportion of that data was missing owing to technical difficulties with online testing in schools. As such, the EEF sought to commission a long term follow-up of the original trial participants using administrative data. Long term analyses of attainment and exclusion followed pupils for up to two years post-intervention. This analysis found no significant differences between treatment and control groups in terms of the number of exclusions or the number of GCSEs pupils achieved. However, loss to follow-up means that we cannot be certain this result is wholly attributable to the intervention.

The long term analyses suggest this intervention did not alter pupil exclusion outcomes positively or negatively. As the initial post-intervention analysis suggested potentially harmful effects in terms of self-reported exclusions, this is reassuring. However, owing to the administrative nature of the outcome data collected, we cannot comment on the later impact on unofficial exclusions or other sanctions. There was also no evidence of differences between treatment and control groups in terms of the number of GCSEs achieved.

As one of the first attempts to follow-up a high-risk group participating in an EEF trial, the project also illustrated that standard outcome metrics such as Best 8 were inappropriate as very few of the pupils achieved enough GCSEs to have a Best 8 score calculated. This project illustrates the need for significant amounts of preparatory work to be completed with the relevant data owners when attempting follow-ups of non-educational outcomes.

Cost

The cost per pupil per year (for the 12-week intervention) is estimated to be £881.66. This cost is likely to remain the same if delivery was repeated in subsequent years. This is based on delivering the EiE-L programme to 300 pupils with a total cost of £264,500. External staff managed the group and one-to-one sessions, so delivery time for teaching staff was minimal.

Impact

Table 1: Summary of impact on primary outcomes for long term follow-up

Outcome/ Group	Effect size (95% confidence interval)	Estimated months' progress	EEF security rating	No of pupils	P Value	EEF cost rating
Attainment	-0.08 [‡] (-0.29, 0.12)	n/a	n/a	520	0.43	£ £ £ £ £
Exclusion	0.93 IRR ⁺ (0.59, -1.47)	n/a	n/a	557	0.77	£ £ £ £ £

Notes: [‡]Attainment is measured as the number of GCSEs taken as there were insufficient pupils with eight or more GCSEs to use Progress 8. Please note that there is no clear way to translate into months' progress. The effect size reported for attainment is from a linear model of the number of GCSEs obtained. Re-analysis using a count model produced a non-significant incident rate ratio (IRR) of 0.95 (p = 0.398; 95% CI 0.85–1.07). ⁺ IRR = Incident Rate Ratio. An IRR of 1 means incident counts are equal in the groups being compared.

Introduction

Background evidence

Rationale for evaluation: exclusion and its consequences

In England there are between 100,000 to 150,000 fixed-term exclusions and 10,000 permanent exclusions each year, with an uptick in exclusions since 2015 following several years of decline (see DfE, 2018). Around 25% of exclusions each year are for 'persistent disruptive behaviour' and a further 25% for overtly aggressive behaviour (at least according to how schools classify the reasons for exclusion; see DfE, 2018). This means that around half of all exclusions may arise because of oppositional-defiant or aggressive behaviour that leads to schools imposing this sanction.

Young people who are excluded tend to be 'hard to reach', disruptive, and, in many cases, aggressive towards adults and/or other pupils. They often have communication difficulties, which may compromise their ability to benefit from the curriculum as well as behave in prosocial ways (Clegg et al., 2009). Further, some young people who experience exclusion come from home environments that are far from conducive to educational attainment. Exclusions also do not affect all young people equally. Some over-represented groups include male students, students eligible for free school meals, students with Special Educational Needs (SEN), and young people with an ethnic minority background (for example, Gazeley et al., 2013). These findings suggest that those excluded may not like school in the first place, perhaps partly as a result of finding school difficult due to their (unmet) educational needs. Indeed, previous research has shown that some young people view fixed-period exclusions as akin to school sanctioned holidays (Dupper et al., 2009).

School exclusion has been a fundamental part of school discipline in the U.K. for more than a century, yet as an intervention it is poorly understood, beyond noting the regularity with which certain pupil groups are excluded at higher rates (see DfE, 2012, 2018). What we do know is that exclusion is correlated with many negative outcomes for young people, such as drug use, criminality, imprisonment, and unemployment (see, for example, Home Office, 2018). But there are many unanswered questions about exclusion. For example, we do not know whether exclusion is a marker of a pre-existing behavioural trajectory or if exclusion causes later negative life outcomes (see Challen and Walton, 2004; Gazeley, 2010).

On the one hand, use of exclusion is often justified because of the disruption some pupils can cause the rest of the class (although evidence on that supposed benefit has been challenged as long term exposure to highly punitive contexts may actually be detrimental to non-excluded pupils; see Perry and Morris, 2014). On the other hand, exclusion may end up being more costly to the public purse because both exclusion itself is expensive, as are the long term consequences of it (Sutherland and Eisner, 2014). Given the association between exclusion and the increased risk of poor life outcomes, it seems sensible to intervene to try and prevent and/or reduce the use of exclusion, whilst at the same time contributing to what is a relatively small evidence base (see Valdebenito et al., 2018).

Background: theoretical rationale and existing evidence for the intervention

The Engage in Education London (EiE-L) programme was intended to improve the behaviour of pupils at the highest risk of school exclusion through improving their communication and social skills. As the basis of exclusions are interactions between pupils, their peers, and/or their teachers, improving pupils' communication skills, social skills, and how they handle emotive situations may be one way to prevent or reduce the use of exclusion. Existing meta-analyses (Sandler et al., 2014; Beelman and Losel, 2006; Derzon et al., 2006; Mytton et al., 2006) suggest that social-skills interventions are associated with, for example, reductions in aggression, anti-social behaviour, and violence. However, few interventions have focused on preventing exclusion, and there is little existing evidence on the efficacy of communication- or social skills-based interventions in improving behaviour in school-settings (see Valdebenito et al., 2018, 2019). Whilst EiE-L was based on an existing intervention that had previously been piloted for two years, Engage in Education, there have not been any rigorous evaluations of that approach to date (see Ellis, 2013).

The data reported in this report is the long term follow-up data of the original evaluation study reported elsewhere (Obsuth et al., 2016, 2017). This original study found that immediately post intervention there was no difference between treatment and control pupils in the likelihood of being excluded, nor were there differences on secondary outcomes covered in that evaluation (truancy, arrest, and self-reported behaviour).

Rationale for the long term follow-up

A key motivation for the long term follow-up for LEIP—reported here—in relation to attainment were problems with the collection of attainment during the main study, notably technical problems with school IT systems (see Summary of original educational and exclusion outcomes and Obsuth et al., 2016, 2017). The purpose of the current analyses, therefore, was to explore whether (i) it was possible to access administrative follow-up data for non-educational outcomes, and (ii) what, if any, long term effects were observed two years post intervention using administrative data. Evaluation of potential medium and/or long term effects is particularly important in situations in which null or negative effects were identified immediately following the intervention (for a discussion of the importance of long term follow-up see, for example, McCord, 2003; Sutherland et al., 2017). In this case, it is important to understand whether or not negative effects on school exclusion reported in Obsuth et al. (2016, 2017) continued, or generalised to other outcomes. At the same time, ‘sleeper effects’ of the intervention may also be identified, in which time-lagged positive effects of the intervention may emerge due to the delay required in order for learnt skills to solidify and contribute to positive outcomes (see, for example, Bonell et al., 2018).

Intervention

Brief name: Engage in Education London (EiE-L) is a 12-week-long intervention that targets the communication and broader social-behavioural skills of pupils aged 13–15 deemed by their teachers to be at greatest risk of school exclusion. The aim of EiE-L was to reduce school exclusions and associated problem behaviours.

Why: The basis of exclusions are interactions between pupils, their peers, and/or their teachers, so improving pupils’ communication skills, social skills, and how they handle emotive situations may be one way to prevent or reduce the use of exclusion. Existing research suggests that social-skills interventions are associated with reductions in aggression, anti-social behaviour, and violence.

What: EiE-L consisted of two sessions per week with each pupil: a one-hour group session and a one-hour one-to-one session, over 12 weeks. Core workers from Catch22 were assigned to schools, with each group session being delivered jointly by two core workers and one-to-one sessions delivered independently by one of these core workers. The sessions were structured around specific goals, and aimed to use participative techniques such as pair work, group work, and whole-group discussions. Briefly, the sessions included self-reflection on communication strategies, anger management, de-escalation, ‘code switching’, being assertive, and non-verbal cues/signals (for more detail on the content of sessions see Table 1 in Obsuth et al., 2017). Schools needed to provide a private room for group and one-to-one work, which in some cases proved challenging. Parent(s)/carer(s) were also offered additional support through phone calls and/or home visits.

Who provided: Core workers, with backgrounds in youth work, were hired specifically for the intervention delivery.

How: Face-to-face delivery in both group and one-to-one sessions.

Where: On school premises during term time. Barring English, maths, and science, pupils were excused from lessons in order to participate. Core workers sometimes met pupils at class, but in other cases pupils were escorted or made their own way to intervention sessions.

When and how much: The intervention was delivered in two phases (covering autumn and spring terms in the 2013/2014 academic year) owing to limits on delivery capacity and delays with schools signing up for the trial. In total, approximately 14 sessions (seven group, seven individual) were delivered on average in the 17 schools to around 20 pupils per school, with each group and one-to-one session scheduled to last for one hour.

Tailoring: EiE-L was an adaptation—a shortening—of the existing Engage in Education (EiE) programme being run by Catch22. Whilst there was a core programme of topics to cover in EiE-L, core workers also had flexibility to tailor interventions to individual children’s assessed need (see footnote 1 in Obsuth et al., 2014).¹

¹ “The level of support provided varied according to the needs of each young person involved”: <https://www.catch-22.org.uk/research/london-education-inclusion-project-leip/>

Modifications: There were no modifications made during the course of the intervention that the research team were aware of, but the extent of tailoring was also unclear.

How well (planned/actual): Twenty-four sessions were planned with each pupil (12 group, 12 individual) but only around seven of each took place (14 in total, on average). Although additional support to parents via phone calls and home visits were intended, relatively few took place (164 phone calls and 11 home visits) meaning that just over half of intervention participants' parent(s)/carer(s) were contacted.

Issues that arose during the project

The original LEIP project, as described earlier, focused on short term outcomes such as arrest and exclusion. This is reported in full in Obsuth et al. (2016, 2017) and summarised in Appendix D. However, the following issues arose during the original LEIP project relating to recruitment, implementation, testing, and project timelines.

Recruitment: Initial recruitment of schools by the evaluation team was slow and the target area (Central London) was later expanded to include Greater London to ensure enough participating schools.

Implementation: The number of actual appointments with participants was roughly half that planned. That said, the developer team, Catch22, felt that the intervention was still delivered with sufficient dosage to have an impact.

Testing: The planned attainment tests (Yellis and MIDYIS) suffered from numerous implementation and technological issues in schools, including quite complex login procedures, meaning that far fewer pupils than intended were able to complete tests (418/606, 69%). For the long term follow-up (reported here) there were other challenges faced, primarily about data access and delays. These are described in more detail later in the report.

Timetable slippage: The follow-up study encountered significant delays owing to data access requests and approvals taking considerably longer than anticipated.

Data access: There were significant delays introduced by lengthy negotiations for data access, including review and re-approval of the LEIP ethics approval with the Administrative Data Research Network, which ultimately led to data requests being submitted outside of the ADRN (see **Error! Reference source not found.**). The study was approved by the Institute of Criminology Ethics Board in May 2017.

Evaluation objectives

The original study objectives are given in Appendix D, which also summarises the results from that study and sets out limitations of that study. The follow up study (reported here) sought to use administrative data to track long term outcomes for pupils in the original LEIP study. The long term follow up was driven in part by the loss of attainment data in the LEIP study, owing largely to challenges with data collection during testing that created significant practical barriers for pupils. The long term follow up sought to test the following hypotheses for primary outcomes:

H1: There will be no statistically significant difference in the number of GSCEs achieved (graded A-G) between control and intervention groups.

H2: There will be no statistically significant difference in the attainment 8 GSCE scores between control and intervention groups.

H3: There will be no statistically significant difference in school exclusions between control and intervention groups.

Here, the null hypothesis is that of 'no difference' between treatment and control groups.²

² With a well-executed RCT there should be baseline equivalence between treatment and control groups. These groups (and those sharing their characteristics) constitute the population of interest—hence testing for differences between the two groups is a valid approach, as long as care is taken about the inferences made (see, for example, discussions in Hsu, 2008; Athey and Imbens, 2016).

Methods

Trial design

Trial type and number of arms		Two-arm, cluster-randomised
Unit of randomisation		School
Stratification variable (s) (if applicable)		Free school meal eligibility (FSM); special educational need (SEN); school size; single or mixed gender; baseline pupil behaviour
Primary outcome	Variable	1. How many GCSE's from grades A–G were achieved [ks4_gcse_ag_ptq_ee]. 2. Attainment 8 GCSE score [ks4_att_8]. 3. NPD records of fixed-term and permanent school exclusion. (Note that permanent exclusions were not analysed.)
	Measure (instrument, scale)	
Secondary outcome(s)	Variable(s)	1. Further education uptake 2. Police records 3. Not in Education, Employment or Training
	Measure(s) (instrument, scale)	1. FE uptake from UCAS (applied: yes/no; taken up place: yes/no). 2. Police records: numbers of arrests since the end of intervention. 3. NEET: a yes/no indicator of whether in education, training or employment. <i>Note: no secondary outcomes could be obtained.</i>

The study was a cluster-randomised trial, with schools as the unit of randomisation and outcomes measured at the pupil level. The targeted nature of the intervention, focusing on high-risk pupils, meant that contamination was a real risk for an individually randomised trial within schools, hence a cluster-randomised design was chosen.

To improve recruitment and retention, control schools/pupils were offered an 'active control' condition so that they also benefitted. This was delivered as a one-off careers advice day delivered by external facilitators for all pupils in the relevant year groups. Whilst there was a risk that this one-off event might lead to improved outcomes in terms of education, training, or employment, it was felt that the risk was minimal. However, there is a bigger question about what 'business as usual' looks like with regard to exclusion. This is because school behaviour policy is variable across schools—so much so that the Department for Education commissioned a review into behaviour policy.³

Due to delays with collecting baseline data and constraints on delivery, notably a limit on how many intervention schools could be worked with at once, it was decided to split delivery into two phases—one during the Winter term, and one during the Spring Term.

³ <https://www.gov.uk/government/publications/school-exclusions-review-terms-of-reference>

Participant selection

Description of population including eligibility criteria

In terms of recruitment, both the evaluation and delivery teams were responsible for recruiting schools, but the evaluation team took the lead on this.

Schools were selected based on FSM eligibility (see School Characteristics section in the Appendices). Each school was invited to nominate approximately 20 young people (roughly 10 each from Years 9 and 10) whom they deemed to be at the highest risk for school exclusion and/or becoming NEET, based on (i) past problem behaviour, (ii) exclusions, or (iii) unauthorised absences from school. This resulted in 751 young people overall being nominated by teachers in both treatment and control schools.

The programme targeted young people at greatest risk of school exclusions in highly deprived schools and neighbourhoods in Greater London. Specifically, only schools with a free school meal eligibility rate of 28% or higher were invited to participate in the project.⁴ Eligible pupils were nominated by their schools to take part in the programme. Pupils were not considered 'in' the trial unless they also completed baseline data collection.

EiE-L aimed to improve students' behaviour by developing their communication and broader social skills. More details of recruitment are given in Obsuth et al. (2014; 2017).

Sample size

The project sample size was determined by power calculations (reported in Obsuth et al., 2014). We present the trial power calculations in the section Power calculations: Minimum detectable effect sizes.

Outcome measures

Primary outcome

The original study reported on two educational outcomes: verbal reasoning and maths ability as set out in the study protocol (Obsuth et al., 2014; see Appendix D below). For the follow-up, we use National Pupil Database (NPD) data on educational attainment at Key Stage 4 (GCSE) and school exclusion. The attainment results should be unbiased as they come from high-stakes national testing, where those marking are blind to allocation.

The primary outcome was: a count of how many GCSE's were achieved [ks4_gcse_ag_ptq_ee]. We originally intended to use the Best 8 GCSE score [ks4_att_8] but could not as very few pupils in the study had this measure recorded because they did not achieve eight GCSEs (the Best 8 measure requires at least eight GCSEs to be completed in order for a Best 8 score to be calculated).

For this analysis, we also report on exclusion outcomes (as a primary outcome) following the end of the intervention using the following NPD variables as the basis for measuring exclusions. For each of the following measures, we combined the counts for the two years into a total count, resulting in two exclusion outcomes covering fixed-term exclusions only:

- **the number of fixed term exclusions:** TotalFixedExclusions_ex16 and TotalFixedExclusions_ex15; this is a count of FTEs per pupil in a given academic year, but does not count how long each exclusion period lasted for; and
- **fixed term exclusion sessions (AM/PM):** TotalFixedSessions_ex16 and TotalFixedSessions_ex15; sessions are AM or PM, the time before or after lunch counted as one session, for example, if a pupil

⁴ Please note that the national average of pupils eligible for free school meals in schools was around 15% across England in 2013 (DfE, 2018).

missed one half-day due to FTE that would be one session, irrespective of how many lessons were missed.

Note that there were too few permanent exclusions to conduct an analysis: only n = 15 pupils received were permanently excluded.

NPD data was accessed in MONTH/YEAR format, and informed consent was the only acceptable way to access pupil data at the time.⁵ Where pupils had a missing value for exclusions, we treat those as 'zero': this means we cannot distinguish between pupils who had zero exclusions and those who were missing this information, but our understanding of the NPD data is that, as census data, our interpretation is correct. (The alternative would be that schools would need to submit returns of zero for the majority of pupils.)

In previous analyses we did not correct for multiple outcome testing. Given the preponderance of null results for educational outcomes, this seems a reasonable approach here. However, given that we have two educational outcomes relating to GCSE being tested on the same subjects—number of GCSEs and Best 8—we might be concerned with multiplicity and non-independence of the significance tests (Schulz and Grimes, 2005; Bland, 2004). However, the two measures being used are based on different samples/groups and are not direct derivatives of one another. The 'Best 8' score is only calculated for those pupils who actually achieved eight GCSEs; for those that did not, they do not have a 'Best 8' score—a problem that is likely generalizable to any study of a high-risk group.

Immediate post-intervention outcomes for this study, including official exclusion data within the NPD and both teacher-reported and self-reported exclusions, have previously been reported in Obsuth et al. (2016, 2017).

Changes to planned analysis of secondary outcomes

Analysis and reporting of secondary outcomes was dependent on data being made available. Unfortunately, we were unable to obtain additional outcome data relating to:

- **employment**—we originally planned to include employment status, such year of employment (if applicable), type/sector of employment, and whether or not claiming child benefits from the Department for Work and Pensions (DWP) as our secondary outcomes, however, contrary to what we had understood, this data was not available;
- **university application data**—we were also unable to request data regarding university application status from the Universities and Colleges Admissions Service (UCAS) or via the ADRN; and
- **arrest data**—we submitted a request for arrest data to the Metropolitan Police Service (MPS) in October 2017; we received a reply that asked about the legal basis for the request, and although we replied, were not able to access data in time for this report.

Randomisation

Randomisation was carried out via minimisation (Taves, 2010; Saghei and Saghei, 2011) by one researcher (AS). Minimisation was used in order to achieve balance based on the following variables: free school meal eligibility (FSM) and special educational need (SEN); school size; single or mixed gender; and baseline pupil behaviour measured using self-report data from the young person survey using the Misbehaviour in School Questionnaire (MISQ; see Obsuth et al., 2017). As demonstrated in our published articles (Obsuth et al., 2016, 2017), good baseline balance was achieved across groups at both school and pupil level (please see Appendix C: Variables used for randomisation & balance checks for minimisation allocation details). The one exception to this was pupil sex, which, owing to how single-sex schools were coded for randomisation, meant there were more all-girl schools allocated to control. As a result, pupil sex was added as a pupil-level covariate in all analyses.

⁵ It should be noted that since the NPD request was fulfilled, changes have been made to data protection rules procedures. This does not affect the data collected from the NPD, but does mean a different legal basis would need to be made in order to access data from pupil's who are missing appropriate consent.

Timing of randomisation relative to baseline testing

Teacher screening data was collected for 751 pupils from all 36 schools prior to randomisation; aggregated teacher data was used in the randomisation to ensure balance on pupil problem behaviour (Obsuth et al., 2017). Once pupils had been nominated by teachers, pupils were asked to complete a baseline self-report questionnaire (n = 606 did so). As some schools could not meet the first deadline for pupil baseline data, and because of capacity limits on delivery, randomisation was completed in two phases. Baseline data collection (both teacher and pupil) took place from June to September 2013 and was completed prior to randomisation, meaning that baseline data collection was blinded. Schools were not informed of allocations in their phase until baseline data from all nominated pupils in all schools had been collected.⁶

Number randomised to each arm

Overall, 36 schools and 751 pupils were randomised: 17 schools in the treatment arm (373 pupils) and 19 schools (365 pupils) in the control arm. Of those 751 nominated by teachers, 606 completed the pre-intervention baseline survey: 300 pupils in the treatment arm and 306 pupils in the control arm. For the follow-up study, there were 248 pupils in the treatment arm, and 254 in the control arm.

Follow-up analyses for the main results (exclusion count and attainment count) were first conducted by AS, who was not blind to allocation, but then repeated by YS, who was blind to allocation. The results from these two analyses were consistent in terms of the direction and magnitude of differences (that is, no effect). YS did not replicate the sensitivity analysis.

Statistical analysis

The analyses conducted using the follow up data align with those previously undertaken for the immediate post-test data in terms of model specification and baseline covariates.

Primary outcome analysis

To summarise, our primary intention-to-treat (ITT) analysis model, undertaken on KS4 attainment and exclusion, included pupil sex at pupil level, KS2 maths at pupil level, pre-intervention exclusion, and treatment allocation at school level and phase of intervention delivery, accounting for clustering via multilevel models. Below we set out further detail on our approach.

The analysis of primary outcomes was on an ITT basis for all pupils. This means that once randomised, schools and participants were analysed according to their allocation regardless of whether they received or participated in the intervention or not. The ITT approach is inherently conservative as it captures the averaged effect of offering the intervention for those who do, and those do not, comply.

The unit of analysis is pupil-level outcomes for those in schools with and without EiE-L. We ran our analyses using a multilevel approach, with pupils clustered in schools and schools as random effects. The general equation for the linear multilevel model is:

$$y_{ij} = \alpha + X_{ij}\beta + Z_ib_j + \delta E_i E_j + u_{ij} + u_j \quad i = 1..N, j = 1..M, \quad (1)$$

⁶ 'Phase I schools were randomised and received the intervention in both Year groups in Autumn 2013 and Phase II schools were randomised later and received the intervention in both Year groups in Winter/Spring 2014' (Obsuth et al., 2014: 7). 'Twenty schools with available baseline data were randomised as planned at the end of September (constituting Phase I). The remaining 16 schools (constituting Phase II) were randomised on 15th of November' (*ibid.*).

—where y_{ij} denotes the pupil level outcome; i and j denote pupil and school indexes respectively; and X_{ij} is the $1 \times k$ vector of individual characteristics (here, baseline measures of a specific outcome and pupil sex). $E_i E_j$ is a dummy variable denoting treatment/control group at school level; β and δ are the $k \times 1$ and 1×1 vectors of regression coefficients; and Z_j is a vector of school-level characteristics formed from the variables used for minimisation mentioned previously. u_{ij} is the pupil-level error term; u_j is the school-level error term. In equation (1) α denotes the overall model intercept.

The coefficient δ associated with the $E_i E_j$ dummy constitutes the main result of the trial. ‘All pupils’ represents the main analysis group, so to obtain the main result of the trial, Equation (1) above was estimated for all pupils.

The pupil- and school-level variables included in analyses were:

- **pupil level:**
 - sex—included because of a reported imbalance at pupil level owing to how single-sex schools were allocated, as mentioned Obsuth et al. (2014);
 - a variable for pre-intervention exclusion, as the baseline for the exclusion analyses;
 - a variable for pupil attainment at KS2, as the attainment baseline for both arms of the trial—included as the preferred measure, KS2 points score, was not available for all pupils (note that this was included as a sensitivity analysis); and
- **school-level:**
 - the phase of intervention roll-out, to account for any time-varying effects of the two phases of intervention.

This data was obtained from two sources: the original trial data and an updated NPD data request.

There is some debate about how to analyse the results of trials that used minimisation (Altman, 2008; Taves, 2010; Ivers et al., 2012): this is because there are often several variables used to minimise. However, with small-n trials—as here—the degrees of freedom are limited; in this case, we have 36 schools randomised using five balancing variables. Including all of those variables in the analysis at the school level that might lead to the model being over-determined, as it would be in this case where $n = 36$ with six predictors (including allocation). As such, as with our previously published analyses (for example, Obsuth et al., 2016, 2017), we omitted these variables.

Finally, owing to the small number of schools randomised, we also undertook a sensitivity analysis applying the Kenward-Roger correction for small samples (see McNeish, 2017). All analyses were conducted using Stata v13.0 or higher.

Secondary outcome analyses

It was not possible to conduct secondary outcome analyses as we were unable to obtain the pre-specified variables: further education uptake, police records, and NEET.

Subgroup analyses

We agreed to undertake exploratory subgroup analyses for FSM pupils versus non-FSM pupils, acknowledging that this would be under-powered (see, for example, Newgard and Lewis, 2015). To maximise the sample size for that analysis, we interacted the treatment and FSM indicator variables rather than sub-setting the data. We constructed a variable that indicated if a pupil had ever been ‘ever 6 FSM’ eligible in the years 2012/2013–2016/2017 (inclusive). This led to 82% of pupils in the analysis dataset being classified as ever6fsm eligible.⁷ The result of this analysis is given in

Table , but, as the majority of pupils were FSM-eligible, then there is no meaningful difference to the main result reported in Table 1.

⁷ Taking a lower threshold – such as the first year of data available – made little difference; around 75% of pupils were long-term FSM eligible.

Additional analyses

There were no pre-specified compliance or on-treatment analyses for follow-up and we have not conducted any exploratory analyses along these lines (see the extensive analysis of compliance in Obsuth et al., 2017). Moreover, two-thirds of the treatment group pupils were considered 'compliant' overall (only 93 were not), so we believed there would be little gained from this additional step. As noted, we re-ran analyses using a correction for the small sample size at school level (reported in Appendix I: Sensitivity analyses).

Imbalance at baseline

As previously reported (Obsuth et al., 2017), single-sex schools were randomised without differentiating between all-male and all-female schools, leading to a higher proportion of males in the control arm. As such, we include pupil sex in all of our analysis models to account for this. That aside, we achieved balance on a range of other variables at both school and pupil level (see Obsuth et al., 2017).

Missing data

There was loss to follow-up owing to issues with the consent procedure (discussed below) meaning that the maximum sample size for the long term follow-up was 557 pupils. As we use NPD data for outcome data in the follow-up, missingness on outcomes for pupils we had consent to data linkage for is minimised. We assessed the relationship between loss to follow-up and treatment allocation with a multilevel logistic model. Loss to follow up (yes/no) is being modelled as a function of treatment allocation and other baseline variables. We found that treatment allocation was not associated with loss to follow-up.

Implementation and process evaluation

Results from the implementation and process evaluation are reported in Obsuth et al. (2017), which looked at 'intervention attendance, participants' engagement with the intervention, and treatment fidelity reported by the providers' (p.10).

Timeline

Below in Table 2, we summarise the timeline for the original project and follow-up study (see also Obsuth et al., 2014, [online supplement](#)).

Table 2: Timeline

Date	Activity
May 2013	School recruitment
June—September 2013	Teacher baseline
September 2013	Core worker intervention training (x4 weeks)
September—October 2013	Pupil baseline
September 2013	Randomisation Phase I
October 2013—January 2014	Intervention delivery Phase I
March 2014—May 2014	Outcome data collection Phase I
December 2013	Randomisation Phase II
February 2014—May 2014	Intervention delivery Phase II
June 2014—July 2014	Outcome data collection Phase II
March 2015	Initial discussions with the EEF about follow-up
December 2015—June 2017	Ethics application with ADRN (initial submission to approval)
February 2017	NPD follow-up data requested
October 2017	Request to the NPD for follow-up data approved
October 2017	NPD follow-up data received
March 2015—February 2018	Development work with ADRN regarding data collection

Impact evaluation

Participant flow including losses and exclusions

Figure 1 is the participant flow diagram for the follow-up analysis which shows the loss to follow-up owing to problems with consent based on the numbers originally randomised. In Table we also report on the percent of pupil data available at different stages of the analysis, relative to those randomised and relative to those baselined. Appendix D summarises the results from the original project.

The maximum sample size for follow-up was $n = 557$; for the exclusion data the sample size was 557, for the number of GCSEs the sample size was 520. The loss of cases between the original study and the follow-up is due to issues with obtaining appropriate consent to request and link pupils' official records to survey data. Upon auditing consent forms for follow-up, we discovered that some consent forms had been unsatisfactorily or incorrectly completed and, as such, could not be used for data requests. Specifically, some forms were incomplete but had been signed by pupils; in other cases, signatures had been written on sheets other than the signature sheet and the relevant boxes not ticked to permit data requests. Overall, for follow-up there were 557 pupils; 280 of these pupils were in the treatment group and 277 in the control group. Within this sample, approximately 80% of pupils were 'ever6fsm' eligible at some point between 2012 and 2017 (using, for example, the EVERFSM_6_SPR12 variable from the National Pupil Database). For the attainment data, a combination of pupils missing attainment data in the NPD and having incomplete data on other variables in the analysis meant a further loss of data, meaning the maximum sample size for that analysis was 520 pupils.

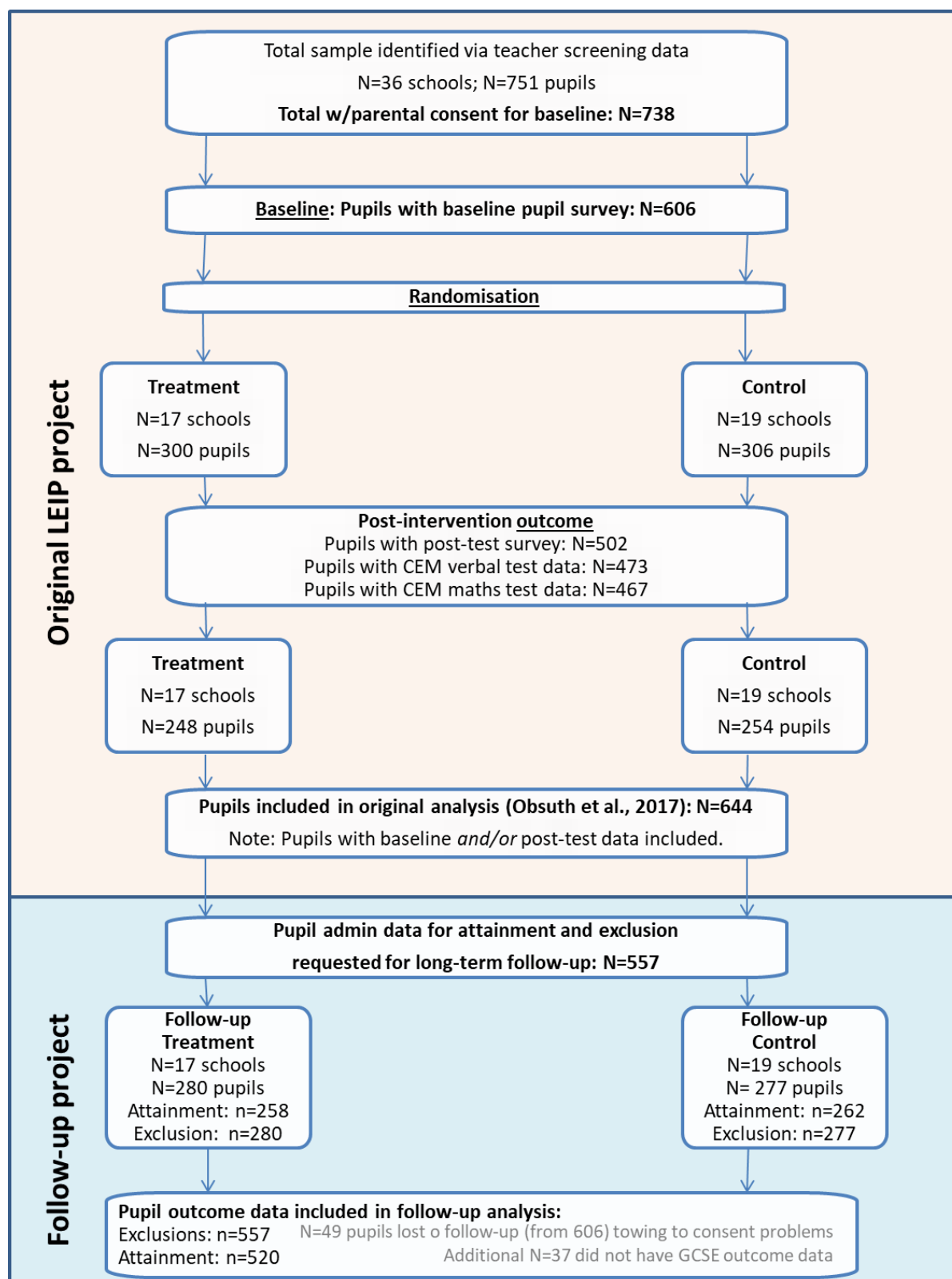
Note that for the original analysis in Obsuth et al. (2017), data from a total of 644 pupils was used. This included any pupil with pre- *and/or* post-intervention data available.

Attrition

Table 3: Loss to follow-up as percentage of pupils randomised

	<i>Treatment</i>	%	<i>Control</i>	%	<i>Total</i>
Identified pre-randomisation →	751				
Baselined & randomised with consent	300	80.4%	306	83.8%	606
Follow-up data requested (i.e. appropriate consent obtained/available)	280	75.1%	277	75.9%	557
Follow-up data received	280	75.1%	277	75.9%	557

Figure 1: Loss to follow-up for long term outcomes data



Power calculations: Minimum detectable effect sizes

Prospective sample size/minimum detectable effect size (MDES) calculations were given in the study protocol (Obsuth et al., 2014: 10). Initial MDES calculations were undertaken using Optimal Design (Raudenbush et al., 2011) but we recalculated these for Table 4 using PowerUp! (Dong and Maynard, 2013). Funding and delivery capacity limits meant that a maximum of 40 schools (20 treatment) could be included in the trial. With 40 schools (J); 20 pupils (n) within each school; a desired power of 0.80; an alpha of 0.05; an assumed ICC of 10%; assuming no correlation between baseline

and post-intervention data; and no level two measures. Using these parameters the (conservative) estimate of MDES was $d = 0.346$. Given these parameters, subgroup analyses for FSM pupils would be under-powered. Perhaps more importantly, the majority of the sample (82%) were FSM-eligible, which suggests that a 'subgroup analysis' would not add much to the existing analyses. At the point of randomisation we had recruited 36 schools. At the point of baselining pupils with consent (which was post-randomisation but prior to schools knowing allocation), there were $n = 606$ pupils in 36 schools (MDES $d = 0.376$). At follow-up, there were 557 pupil records available from the 36 schools. Keeping assumptions the same, the MDES was $d = 0.380$. Table 4 sets out the power calculations for each of these stages.^{8,9}

Table 4: Minimum detectable effect size at different stages

		Protocol	Randomisation	Analysis
MDES		0.346	0.376	0.380
Pre-test/post-test correlations	Level 1 (pupil)*	0.00	0.00	0.00
	Level 2 (class)	n/a	n/a	n/a
	Level 3 (school)	n/a	n/a	n/a
Intraclass correlations (ICCs)	Level 2 (class)	n/a	n/a	n/a
	Level 3 (school)	10%	10%	10%
Alpha		0.05	0.05	0.05
Power		0.8	0.8	0.8
One-sided or two-sided?		Two	Two	Two
Average cluster size ⁺		20	16	16
Number of schools	Intervention	20	17	17
	Control	20	19	19
	Total:	40	36	36

⁸ Post-hoc power analyses for moderator effects reported in Obsuth et al. (2017) suggested the study could detect differences as small as a standardised regression coefficient of $\beta 0.20$. Similarly, the ICC assumption for the power calculations was more conservative than necessary. The published outcome analyses reported ICCs of 5% and close to zero, which is supported by the results from other U.K. behavioural trials in education that also report ICCs close to zero (for example, Bonell et al., 2015; 2018). However, with a binary outcome such as exclusion or a skewed outcome such as number of GCSEs passed, power would be lower owing to the distribution of the outcome variable. Using the protocol sample sizes and basing the detectable difference on the observed prevalence of exclusion (9%) as the proportion of exclusions in one group, the detectable differences for 40 schools, 20 pupils, $\alpha 0.05$; $\beta .80$; ICC 10% would be: 12pp (increasing outcome) and 7pp (decreasing outcome).

⁹ Stata command: `clustersampsi, detectabledifference binomial p1(.09) m(20) k(20) rho(0.10) size_cv(0.0) base_correl(.0)`.

Number of pupils	Intervention	400	300	280
	Control	400	306	277
	Total:	800	606	557

Table note: * pre-post correlation assumed to be zero for all calculations. + Average cluster sizes rounded.

Pupil and school characteristics

We report on the school and neighbourhood characteristics from the original study in School Characteristics and Neighbourhood Characteristics sections of the appendices. The original descriptors for young person characteristics are reported in Obsuth et al. (2017), but below, in Table 5, we provide the descriptive statistics for the characteristics of the follow-up sample of 557 participants. In spite of the loss to follow-up, balance between groups held, with the exception of pupil sex, which, as originally reported in Obsuth et al. (2014), was imbalanced because of how single-sex schools were handled in the randomisation.

Table 5: Descriptive statistics for treatment and control group participants in follow-up sample (n = 557)

Panel A: Sex	Control	Treatment	Total
Male	207	182	389
Female	70	98	168
Total	277	280	557
Panel B: Ethnicity	Control	Treatment	Total
White British	57	76	133
Other White	7	12	19
Black African, Black Caribbean	110	92	202
Asian (e.g. Chinese, Vietnamese, Korean)	7	4	11
South Asian (e.g. Indian, Pakistani)	28	28	56
Latin American	4	3	7
Dual heritage	34	27	61
Missing	13	23	36
Other	17	15	32
Total	277	280	557
Panel C: KS2 Maths Points	Control	Treatment	Total
0	32	52	84
15	19	21	40
21	57	47	104
27	123	118	241
33	38	33	71
Missing	8	9	17
Total	277	280	557
Panel D: Pre-intervention exclusion?	Control	Treatment	Total
No	187	176	363
Yes	90	104	194
Total	277	280	557
Panel E: Ever6FSM	Control	Treatment	Total
No	50	48	98
Yes	227	232	459
Total	277	280	557

Outcomes and analysis

There were two main outcomes assessed in this study: (i) the number of fixed-term exclusions (as a count variable) and (ii) the number of GCSEs of grades A–G achieved by the pupil (as a count variable). Table 6 and Table 7 below give the sample sizes, variances, and standard deviations by treatment group so that the effect size calculations for GCSE given can be reproduced. The intra-class correlation (ICC) for GCSE was 3.8%.¹⁰

The primary analysis shows that there was no discernible difference between treatment groups in terms of the number of fixed-term exclusions or the number of GCSEs achieved. Note that for exclusions, we report the effect size as an incident rate ratio (IRR) as the outcome variable is a count and is skewed. IRRs are centred on one, with ‘one’ indicating ‘no difference’, and range from zero to infinity. IRRs compare the incident rate in one group to another, giving an indication of how frequently an event occurs in one group versus another, so an IRR of 2.0 would mean that the rate in group 2 was twice that of group 1.

For attainment, we report results from a linear mixed model for the number of GCSEs between grades A–G, converting this into an effect size following Hedges (2007) and the EEF evaluator guidance (EEF, 2018). As with exclusions, there was no difference between groups. Noting that whilst the distribution of the outcome is far less skewed than exclusions (see graphs in Appendix H), the number of GCSEs is still a count measure, we therefore also re-ran this analysis using a model appropriate for count data, as well as additional sensitivity analyses, all of which show consistent null results: there was no effect on outcomes and this was robust to the analysis approach taken (see Appendix I: Sensitivity analyses).

Table 6: Primary analysis

Raw means				Effect size			
Intervention group				Control group			
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n in model (intervention; control)	Hedges g (95% CI)	p-value
Exclusions: Fixed-Term session count	280 (20)	0.74 ^a (0.58, 0.90)	277 (29)	0.87 (0.66, 1.08)	557 (280, 277)	0.93 ^b (0.59, 1.47)	0.766
Attainment: GCSEs A–G count	267 (33)	5.28 (4.90, 5.66)	270 (36)	5.59 (5.24, 5.93)	520 (258, 262)	-0.08 ^c (-0.29, 0.12)	0.427

Table notes: ^a Mean and 95% CI for exclusion based on Poisson distribution as the outcome is a count variable. ^b Incident rate ratio (IRR) is the relative frequency of exclusions in the two groups, comparing group two to group one. The IRR is centred on one. Values above one mean the rate is higher in the second group, whereas values below one mean the rate is higher in the first group. ^c Effect size is calculated by dividing the model coefficient (0.255) by the pooled standard deviation for the outcome (3.01) (actual calculation: 0.2550221/3.009364). Substituting the weighted variances into the denominator as per the EEF guidance (EEF, 2018:4, fn 11) produces a very similar effect size (0.085 vs 0.084).

¹⁰ For a summary of the previous reported outcomes from the trial, see

Table 7: Effect size estimation parameters

Intervention group					Control group			
Outcome	Unadjusted differences in means	Adjusted differences in means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance	Population variance (if applicable)
GCSE count	-0.304	-0.255	267 (33)	9.85	270 (36)	8.26	9.05	n/a

Free school meal pupil subgroup analysis for attainment outcome

We found no significant differences for FSM pupils. Below, in Table 4, we report exploratory analyses for FSM pupils. To conduct this analysis we interacted FSM status with treatment allocation, so the interaction is a cross-level one (which underlines the limited statistical power given that $n = 36$ at school level). The interaction term is not-significant, as expected given that there are no significant main effects (Jaccard and Turrisi, 2003), but leaving that aside, the resulting point estimate is very small ($b = -0.15$).

Table 8: Exploratory FSM x treatment allocation analysis for number of GCSEs ($n = 520$)

Outcome: number of GCSEs grades A*-G achieved		95% CI					
		Point estimate	se	z	p	Lower	Upper
Treatment: 0=Control, 1=Treatment		-0.12	0.63	-0.19	0.85	-1.36	1.12
KS2 Maths points (zero is ref. cat.)	15	-1.47	0.57	-2.56	0.01	-2.59	-0.34
	21	-0.13	0.44	-0.31	0.76	-1.00	0.73
	27	0.65	0.38	1.71	0.09	-0.09	1.40
	33	1.48	0.47	3.12	0.00	0.55	2.41
Ever6fsm: 0=No, 1=Yes		-0.57	0.47	-1.23	0.22	-1.49	0.34
Treat X FSM interaction		-0.15	0.67	-0.23	0.82	-1.46	1.15
Delivery phase: 1 / 2		0.34	0.32	1.07	0.29	-0.28	0.96
Female: 0=male, 1=female		0.24	0.29	0.83	0.41	-0.33	0.82
Intercept		5.15	0.72	7.20	0.00	3.75	6.55

Cost

Catch22 stated that the cost of EiE-L was £881.66 per pupil when bidding for the EEF funding for the EiE-L intervention. This is based on delivering the EiE-L programme to 300 pupils with a total cost of £264,500. Actual delivery costs may have been higher if, for example, staff time for escorting to or from sessions was included, but we do not have information on this.

Conclusion

Key conclusions from the immediate outcome data (reported in Obsuth et al., 2016, 2017)

1. In the short term, there was no evidence that the intervention reduced exclusions. Fixed-term exclusions, as reported by pupils, increased slightly in treatment schools.
2. It was challenging to deliver the programme as intended. Twenty-four sessions were planned with each pupil (12 group, 12 individual) but only around seven of each took place. Although additional support to parents via phone calls and home visits were intended, few took place (164 phone calls and 11 home visits).

Key conclusions from the long -term follow up

3. There was no evidence that the intervention improved the number of GCSEs achieved (graded A–G), two-years post-intervention follow-up.
4. There was no evidence that the intervention reduced fixed-term school exclusions at two-year post-intervention follow-up. The possible negative finding immediately after the intervention was not evident at the delayed follow-up.
5. It was not possible to draw conclusions about the impact of the intervention on further education uptake, police arrest records, or NEET status because data access and other issues precluded analyses of these outcomes.

Interpretation

Based on the previous study, we specified hypotheses based on a ‘null’—that we would not find differences. As such, all three hypotheses were supported: in terms of the main analysis, the results presented were that we did not find evidence of impact on either exclusion or attainment. These results were robust to different model specifications and were replicated by an analyst blinded to treatment allocation.

H1: There will be no statistically significant difference in the number of GSCEs achieved (graded A–G) between control and intervention groups.

H2: There will be no statistically significant difference in the attainment 8 GSCE scores between control and intervention groups.

H3: There will be no statistically significant difference in school exclusions between control and intervention groups.

For exclusion, our conclusion can be tempered by the fact that we did not collect evidence from pupils or teachers as with the published papers. However, the result from Obsuth et al. (2016) relied on self-report data from pupils who were aware of their participation in the intervention. This does not mean that this self-awareness means they were inflating estimates of exclusion: it could mean that those pupils were simply more aware of what exclusion actually was, or were counting situations such as ‘internal exclusion’ as ‘exclusion’. Those types of sanctions do not feature in official records (yet) and so there may have been occasions where pupils were subjected to punishment and counted those towards their self-report. Similarly, given the overlap between other ways to exclude pupils—such as off-rolling or managed moves (Thompson, 2018)—we must also acknowledge that official records of exclusion underestimate the true level of sanction against pupils. In the context of this trial, that under-estimate should be balanced across treatment and control groups, but what is a challenge for education researchers interested in this area is how best to collect such information in a systematic way that balances risk of bias, cost, and scale.

Reflecting on the original results where treatment pupils self-reported a higher rate of sanction than controls, the intervention also involved removing pupils from regular lessons (barring English, maths, and science). As such, one can also imagine that this would increase opportunities for friction and disruption within the school that could have resulted in further sanctions. The process evaluation conducted at the time was not able to unpack this level of detail but where interventions are removing pupils from lessons it seems sensible to try and capture information pertaining to this from subject teachers and intervention staff post-hoc. (Attempting to ask about this during an intervention raises the risk that, for example, prompting a focus on lesson disruption actually generates sanction-responses from that teacher.)

Overall the null effect on exclusion for the long term follow-up fits with the original published results from the study and the few other studies from the U.K. in this field (for example, Bonell et al., 2018) and overseas (such as Augustine et al., 2018).¹¹ However, the systematic review by Valdebenito et al. (2018, 2019) found that interventions, primarily from the U.S., *could* lead to reductions in exclusions (caveats reported in that review about outcome measurement and risk of bias notwithstanding). The lack of evidence of effective interventions to reduce exclusion, or rather, the inability of researchers to keep up with what schools are doing, means there is a risk that this knowledge gap will continue or widen and schools will continue using exclusion, an intervention that has little basis in evidence, or finding their own equally poorly evidenced alternatives.

Previously published statistics on excludees makes clear that they are much more likely to underachieve at KS4 (DfE, 2019). Even a single fixed-period exclusion was associated with a 30 percentage point difference in the proportion of pupils achieving five A*–C GCSEs in 2015/2016, when compared with pupils who had never been excluded (DfE, 2019: 26). As such, focusing on KS4 was appropriate. The results from this study show that in the long term there was no difference between treatment and control groups in terms of the number of GCSEs achieved. But one challenge faced was the fact that not all pupils in this study did not have a 'Best 8' measure for KS4. This is because this high-risk group did not all achieve enough GCSEs for a Best 8 score to be calculated. As we discuss in the report, this required an adaptation to use the number of GCSEs achieved as our measure of attainment. That there was this gap in the data raises questions about how best to measure educational attainment in studies focusing on pupils at high risk of underachievement. In this situation, something akin to 'Best 8' could have been calculated but the difficulty is that the 'cut point' for the 'best X' would be arbitrary. The most inclusive approach would have been to take the single highest GCSE attained and convert that to a score, but that throws away information. We opted for a simple count of GCSEs achieved on the basis that this is a suitably flexible measure for the study participants that captures variation in 'attainment' in the broadest way possible using the data at hand.

Yet we should also acknowledge that GCSE scores or the number of GCSEs are some distance from the programme theory, which emphasised communication skills and behaviour. More fundamentally, this study raises questions about what interventions might be used to improve the educational outcomes of pupils at high risk of exclusion. Here, an approach intended to improve communication skills and emotional regulation did not lead to differences in the number of GCSEs for pupils. The lack of effect in both the short and long term, caveats notwithstanding, suggest that more attention may be needed with regard to programme theories for interventions for young people at risk of exclusion or underachievement. That seems particularly relevant given the recent publication of the Timpson Review into exclusion (Timpson, 2019). There is a good deal of discussion in Timpson regarding the role of school culture and expectations about behaviour, but the review of existing evidence (Graham et al., 2019) mixes weak and strong research designs, and there is little detail on *how* to implement approaches that might be applicable and successful, even if the evidence were stronger. Both of these points speak to the paucity of rigorous research on effective approaches to reduce or prevent exclusion, but also on the gaps in knowledge about implementation of approaches.

We also need to consider the potential impact of implementation on the results (as observed by a peer reviewer of this report). As noted, approximately half of planned sessions actually took place with the high risk group that was the focus of the intervention. One might reasonably ask whether such a level of implementation might be expected to lead to change—particularly when coupled with a focus on exclusion or attainment. However, this sceptical view of success is countered by the view that the training for the intervention team lasted four weeks, a substantial investment; as such, it seems reasonable that the intervention could have been effective, even with imperfect implementation. That it did not again speaks to the need for a greater focus on intervention development, but also on how to prevent exclusion from being necessary in the first place.

Finally, this is one of the few long term follow-ups of an EEF trial, and is also one of the few EEF trials focusing on high-risk secondary pupils. There are few long term follow-ups of educational interventions more widely, so even with the limitations we discuss above and below, the study adds to the evidence base.

¹¹ In the Bonell et al. (2018) there appeared to be baseline differences in the number of exclusions between treatment and control groups (148 in control had been excluded, 100 in treatment; around 1.5pp difference between groups), which was also observed at outcome (310 in control, 255 in control; again around 1.5pp).

Limitations

There are several limitations to the follow-up study, some of which stem from issues that only became apparent during the follow-up.

Substitution of educational baseline measure: The preferred baseline measure (KS2 attainment) was not available for all pupils. Rather than not have a baseline, we substituted maths points level. Although that may not fully capture all the relevant differences between groups, this seemed a reasonable approach given circumstances and alternatives.

Substitution of educational outcome measure: We originally intended to use the 'Best 8' measure—a score based on the highest eight GCSEs a pupil has achieved. However, very few pupils in the sample actually achieved eight GCSEs, meaning that their 'Best 8' score could not be calculated. As such, to ensure that some measure of attainment was included we used the number of GCSEs achieved as the measure of attainment.

Selection into the study: We have to acknowledge that, as with any school volunteering for an RCT, there was selection into the original trial sample. This does not necessarily limit the generalisability of the findings (other issues notwithstanding) but it means we should be clear about the target population. The original trial targeted high deprivation areas and, as we set out in the Neighbourhood Characteristics section of Appendix C, the schools recruited for the study came from neighbourhoods with above-average levels of deprivation and crime. The pupils in those schools had rates of FSM eligibility substantially higher than the national average at the time.

Attrition: There was missingness owing to changes to data protection and problems with how consent was collected by field staff, but this is balanced across study arms. Whilst there was nothing to be done to recover the data lost because of consent issues, we attempted to assess the impact on results and undertook analyses to assess result sensitivity. Leaving out baseline measures for the attainment analysis increased the sample size to 557 and did not affect the results. Similarly, extensive sensitivity analyses based on bootstrap re-sampling and full information maximum likelihood model (FIML) methods did not produce different results.

Future research and publications

We propose that in situations where pupils are likely to fall below the 'Best 8', 'Progress 8', or other alternative national benchmarking measures of attainment, should be used, such as the number of GCSEs from grades A–G achieved. But a research question arises from this: what is the 'best' measure of pupil attainment for low-achieving pupils? Should GCSEs remain the standard benchmark or should social and emotional skills—those skills that will facilitate employment post-education—be given more prominence?

Similarly, the use of stand-alone testing with pupils who are deemed at high risk of exclusion, or being the focus of targeted interventions, is difficult for both schools and researchers. For schools, this means managing that pupil's behaviour well enough so that they can sit an additional 'test' and the subsequent disruption that may bring. One solution may be to collect pre- and post-intervention data from the whole class so as to reduce the stigma attached to individual testing. Whole-class testing may also create an opportunity to understand if there had been any treatment spillovers (negative or positive) on other pupils. Whilst disruptive to the timetable, this may also make the process more manageable from the school's perspective.

Given the paucity of evidence about non-educational outcomes following education-focused trials, it seems sensible to revisit this process once the processes for data access under ONS are clearer. In Appendix G, we set out both our efforts at collecting some of the longitudinal outcomes and our current understanding of how other researchers might access these.

Finally, given the wide variety of behaviour that could lead to exclusion and the many ways schools could consider responding to such behaviour, interventions to reduce exclusion could, and would, be varied in nature, and there is certainly a need for more rigorous evaluation in this field. As recently reported by Valdebenito et al. (2019), there are few well-evidenced interventions focusing on reducing exclusion, so further intervention studies to try and prevent or reduce exclusion, directly or indirectly, would be appropriate and are much needed.

References

- Athey, S. and Imbens, G. W. (2016) 'The Econometrics of Randomized Experiments', chapter 5 in the *Abdul Latif Jameel Poverty Action Lab Handbook of Field Experiments* (eds Banerjee and Dufló), Cambridge, MA: J-PAL. <https://www.povertyactionlab.org/handbook-field-experiments>
- Altman, D. G. (2008) 'Allocation by minimisation in a cluster randomised trial', presentation to the Third Annual Conference on Randomised Controlled Trials in the Social Sciences: Methods and Synthesis, September/October 2008. <http://www-users.york.ac.uk/~mb55/talks/alloct4a.htm>
- Augustine, C., Engberg, J., Grimm, G. E., Lee, E., Wang, E. L., Christianson, K. and Joseph, A. A. (2018) 'Can Restorative Practices Improve School Climate and Curb Suspensions? An Evaluation of the Impact of Restorative Practices in a Mid-Sized Urban School District', Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RR2840.html
- Beelmann, A. and Losel, F. (2006) 'Child social skills training in developmental crime prevention: Effects on antisocial behaviour and social competence', *Psicothema*, 18, pp. 603–610.
- Bland, M. (2004) 'Multiple significance tests and the Bonferroni correction', available from: <https://www-users.york.ac.uk/~mb55/intro/bonf.htm>
- Bonell, C., Fletcher, A., Fitzgerald-Yau, N., Hale, D., Allen, E., Elbourne, D. et al. (2015) 'Initiating change locally in bullying and aggression through the school environment (INCLUSIVE): a pilot randomised controlled trial', *Health Technol Assess*, 19 (53).
- Bonell, C., Allen, A., Warren, E., McGowan, J., Bevilacqua, L., Jamal, F., Legood, R., Wiggins, R., Opondo, C., Mathiot, A., Sturgess, J., Fletcher, A., Sadique, Z., Elbourne, D., Christie, D., Bond, L., Scott, S. and Viner, R. (2018) 'Effects of the Learning Together intervention on bullying and aggression in English secondary schools (INCLUSIVE): a cluster randomised controlled trial', *Lancet*, 392, pp. 2452–64.
- Challen, M. and Walton, T. (2004) 'Juveniles in Custody: A unique insight into the perceptions of young people held in prison service custody in England and Wales', London: Her Majesty's Inspectorate of Prisons.
- Clegg, J., Stackhouse, J., Finch, K., Murphy, C. and Nicholls, S. (2009) 'Language abilities of secondary age pupils at risk of school exclusion: A preliminary report', *Child Language Teaching and Therapy*, 25 (1), pp. 123–140.
- DfE (2012) 'A profile of pupil exclusions in England', Research Report DFE-RR190, London: Department for Education. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/183498/DFE-RR190.pdf
- DfE (2018) 'Permanent and fixed-period exclusions in England: 2016 to 2017', London: Department for Education. <https://www.gov.uk/government/statistics/permanent-and-fixed-period-exclusions-in-england-2016-to-2017>
- DfE (2019) 'Timpson review of school exclusion: technical note', London: Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/799910/Technical_note.pdf
- Derzon, J., Jimerson, S. R. and Furlong, J. J. (2006) 'How effective are school-based violence prevention programs in preventing and reducing violence and other antisocial behaviors? A metaanalysis', in S. R. Jimerson and J. J. Furlong (eds), *The handbook of school violence and school safety: From Research to Practice*, New York: Routledge (pp. 429–441).
- Dupper, D. R., Theriot, M. T. and Craun, S. W. (2009) 'Reducing Out-of-School Suspensions: Practice Guidelines for School Social Workers', *Children and Schools*, 31, pp. 6–14.
- Ellis, T. (2013) 'Engage in Education Final Evaluation', <https://cdn.catch-22.org.uk/wp-content/uploads/2013/11/Engage-in-Education-Final-evaluation-Executive-summary-June-2013.pdf>

- Gazeley, L. (2010) 'The role of school exclusion processes in the reproduction of social and educational disadvantage', *British Journal of Educational Studies*, 58 (3), pp. 293–309.
- Home Office (2018) 'Serious Violence Strategy', London: Home Office.
<https://www.gov.uk/government/publications/serious-violence-strategy>
- Hsu, L. M. (2008) 'Random Assignment Procedures', in Nezu, A. M. and Nezu, C. M. (eds), *Evidence Based Outcome Research: A Practical Guide to Conducting Randomized Controlled Trials for Psychosocial Research*, Oxford: Oxford University Press.
- Ivers, N. M., Halperin, I. J., Barnsley, J., Grimshaw, J. M., Shah, B. R., Tu, K., Upshur, R. and Zwarenstein, M. (2012) 'Allocation techniques for balance at baseline in cluster randomized trials: a methodological review', *Trials*, 13:120. <http://www.trialsjournal.com/content/13/1/120>
- Jaccard, J. and Turrisi, R. (2003) *Interaction effects in multiple regression*, London: Sage.
- Gazeley, L., Marrable, T., Brown, C. and Boddy, J. (2013) 'Reducing inequalities in school exclusion: Learning from good practice', in 'A report to the Office of the Children's commissioner for the centre for innovation and research in childhood and youth'.
- Graham, B., White, C., Edwards, A., Potter, S. and Street, C. (2019) 'School exclusion: a literature review on the continued disproportionate exclusion of certain children', London: Department for Education.
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/800028/Timpson_review_of_school_exclusion_literature_review.pdf
- Hedges, L. V. (2007) 'Effect Sizes in Cluster-Randomized Designs', *Journal of Educational and Behavioral Statistics*, 32 (4), pp. 341–370.
- King, G. and Zeng, L. (2001) 'Logistic regression in rare events data', *Political Analysis*, 9, pp. 137–63.
- Long, J. S. and Freese, J. (2006) *Regression Models for Categorical Dependent Variables Using Stata* (2nd edn), College Station, TX: Stata Press.
- McCord, J. (2003) 'Cures that harm: Unanticipated outcomes of crime prevention programs', *The Annals of the American Academy of Political and Social Science*, 587 (1), pp. 16–30.
- McNeish, D. (2017) 'Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction', *Multivariate Behavioral Research*, 52 (5), pp. 661–670, DOI: 10.1080/00273171.2017.1344538
- Mytton, J., DiGiuseppi, C., Gough, D., Taylor, R. and Logan, S. (2006) 'School-based secondary prevention programmes for preventing violence', *Cochrane Database Syst Rev*. Art. No.: CD004606.
- Newgard, C. D. and Lewis, R. J. (2015) 'Data handling and statistics essentials', in *Emergency Medical Services: Clinical Practice and Systems Oversight* (2nd edn), pp. 439–448.
- Obsuth, I., Sutherland, A., Nordby, L., Scott, S., Valdebenito, S., Hall, R. and Eisner, M. (2014) 'London Education and Inclusion Project (LEIP): a cluster-randomised controlled trial protocol of an intervention to reduce antisocial behaviour and improve educational/occupational attainment for pupils at risk of school exclusion', *BMC Psychology*, 2: 24.
- Obsuth, I., Cope, A., Sutherland, A., Nordby, L., Murray, A., Eisner, M. (2016) 'London Education and Inclusion Project (LEIP): Exploring negative and null effects of a cluster-randomised school-intervention to reduce school exclusion – findings from protocol-based subgroup analyses', *PLoS One*, 11 (4): e0152423.
doi:10.1371/journal.pone.0152423
- Obsuth, I., Sutherland, A., Cope, A., Nordby, L., Murray, A. and Eisner, M. (2017) 'London Education and Inclusion Project (LEIP): Results from a cluster-randomised controlled trial of an intervention to reduce school exclusion and antisocial behaviour', *Journal of Youth and Adolescence*, 46 (3), pp. 538–557.

- Perry, B. L. and Morris, E. W. (2014) 'Suspending progress collateral consequences of exclusionary punishment in public schools', *American Sociological Review*, 79 (6), pp. 1067–1087.
- Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X. and Martinez, A. (2011) *Optimal design software for multi-level and longitudinal research* (software, Vn. 3.01). <http://www.wtgrantfoundation.org>
- Saghaei, M. and Saghaei, S. (2011) 'Implementation of an open-source customizable minimization program for allocation of patients to parallel groups in clinical trials', *Journal of Biomedical Science and Engineering*, 04 (11), pp. 734–739.
- Sandler, I., Wolchik, S. A., Cruden, G., Mahrer, N. E., Ahn, S., Brincks, A. and Brown, C. H. (2014) 'Overview of meta-analyses of the prevention of mental health, substance use and conduct problems', *Annual Review of Clinical Psychology*, 10, pp. 243–273.
- Schulz, K. F. and Grimes, D. A. (2005) 'Multiplicity in randomized trials I: Endpoints and treatments', *The Lancet*, 365 (9470), pp. 1591–1595.
- Sutherland, A. (2014) 'A Methodology for Reconviction using Police National Computer (PNC) data', SSRN.
- Taves, D. R. (2010) 'The use of minimization in clinical trials', *Contemporary clinical trials*, 31 (2), pp. 180–184.
- Thompson, D. (2018) 'The links between exclusions, alternative provision and off-rolling', London: Education Datalab. <https://ffteducationdatalab.org.uk/2018/11/the-links-between-exclusions-alternative-provision-and-off-rolling/>
- Timpson, J. (2019) *Timpson Review of School Exclusion*, London: HMSO. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/799979/Timpson_review_of_school_exclusion.pdf
- Sutherland, A. and Eisner, M. (2014) 'The Treatment Effect of School Exclusion on Unemployment', SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2380956
- Sutherland, A., Ariel, B., Farrar, W. and De Anda, R. (2017) 'Post-Experimental Follow-Ups - Fadeout versus Persistence Effects: The Rialto Police Body-worn Camera Experiment Four Years On', *Journal of Quantitative Criminology*, 53, pp. 110–116.
- Valdebenito, S., Eisner, M., Farrington, D., Ttofi, M., Sutherland, A. (2018) 'School-based interventions for reducing disciplinary school exclusion: a systematic review', *Campbell Systematic Reviews*. <http://campbellcollaboration.org/library/reducing-school-exclusion-school-based-interventions.html>
- Valdebenito, S., Eisner, M., Farrington, D., Ttofi, M., Sutherland, A. (2019) 'School-based interventions for reducing disciplinary school exclusion: a systematic review and meta-analysis', *Journal of Experimental Criminology* (online first).

Appendix A: EEF cost rating

Cost ratings are based on the approximate cost per pupil per year of implementing the intervention over three years. More information about the EEF's approach to cost evaluation can be found [here](#). Cost ratings are awarded as follows:

Cost rating	Description
£ £ £ £ £	<i>Very low:</i> less than £80 per pupil per year.
£ £ £ £ £	<i>Low:</i> up to about £200 per pupil per year.
£ £ £ £ £	<i>Moderate:</i> up to about £700 per pupil per year.
£ £ £ £ £	<i>High:</i> up to £1,200 per pupil per year.
£ £ £ £ £	<i>Very high:</i> over £1,200 per pupil per year.

Appendix B: Security classification of trial follow-up findings

Longitudinal follow-up results commissioned by EEF do not receive a padlock security rating. The original evaluation was not designed to detect the effect on the longitudinal outcomes in this report and was not powered for this purpose. However, this trial has low-to-moderate security. The study uses an experimental design, powered to detect an effect size of 0.35 on educational attainment using assumptions consistent with the original design. Due to 30% attrition from the GCSE outcomes, and 25% from the exclusions outcome, we cannot be certain that the result is attributable to the intervention. The security of the immediate outcome data is discussed further in the body of the report, and in the published papers (Obsuth et al., 2016, 2017).

Appendix C: Variables used for randomisation & balance checks

Table 2 sets out the school-level variables derived from administrative data used for randomization. Each variable was entered as binary or categorical in the randomization itself, and we show summary statistics from the original variable for each grouping.

Table 2: School-level variables used for randomization (minimization)

Panel	Split	Frequency
Panel A. FSM group	Mean % FSM eligible	
0 < 37%	32	17
1 > =37%	45	19
Total	38.9	36
Panel B. SEN group	Mean % SEN	
0 < 12.05%	8.0	18
1 > =12.05%	19.3	18
Total	13.7	36
Panel C. School sex	% of schools	
Mixed sex	72	26
Single sex	28	10
Total	100	36
Panel D. Year 9/10 Cohort sizes	Mean n pupils	
Small <250	747	8
Medium 250-400	900	13
Large >400	1,246	15
Total	1,010	36

Table 3 shows the balance of characteristics used for randomization. This includes the survey-derived variable (ASB – anti-social behaviour), which was a composite score created by a principal components analysis (PCA) of the baseline survey data.

Table 3: Balance following allocation

Allocation	BALANCE FACTORS											
	FSM		SCHOOL SEX		ASB (PCA)		SEN		School size TQ			Total
	fsm< 37%	fsm>= 37%	Mixed sex	Single sex	<mean ASB	>=mean ASB	SEN< 12.05%	SEN>= 12.05%	Large	Medium	Small	
Intensive	8	9	12	5	10	7	9	8	8	6	3	17
Light	9	10	14	5	9	10	9	10	7	7	5	19
Total	17	19	26	10	19	17	18	18	15	13	8	36

Young person characteristics

Table 4 documents the demographic characteristics of the pupils in the original study sample (adapted from Obsuth et al., 2017). The baseline differences in participant sex were driven by randomisation, so pupil sex was included as a covariate in all analyses.

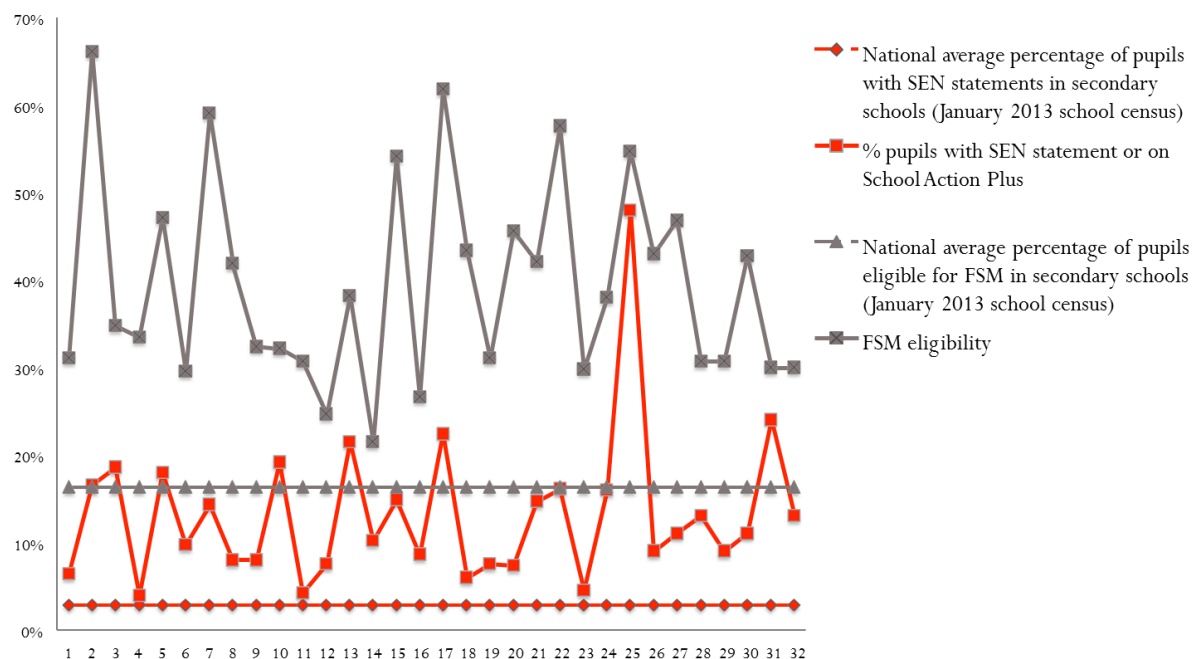
Table 4: Demographic information for young people in final sample

Allocation	Treatment; n (%) n = 300	Control; n (%) n = 306	Total; n (%) n= 606
Sex			
Male	196 (65.3%)	234 (76.5%)	430 (71%)
Female	104 (34.7%)	72 (23.5%)	176 (29%)
Race			
British European (i.e. White)	90 (30%)	61 (19.9%)	151 (24.9%)
Other European (i.e. White Non-British)	17 (5.7%)	14 (4.6%)	31 (5.1%)
Black African, Black Caribbean (i.e. Black)	108 (36%)	136 (44.4%)	244 (40.3%)
Asian (i.e. Chinese, Vietnamese, Korean etc.)	6 (2.0%)	8 (2.6%)	14 (2.3%)
South Asian (i.e. Indian, Pakistani, etc.)	31 (10.3%)	34 (11.1%)	65 (10.7%)
Latin American (i.e. Hispanic)	4 (1.3%)	5 (1.6%)	9 (1.5%)
Mixed race	29 (9.7%)	39 (12.7%)	68 (11.2%)
Missing	15 (5%)	9 (2.9%)	24 (4%)
Students' living situation			
I live with...			
...my biological mother and father	139 (46.3%)	124 (40.5%)	263 (43.4%)
...only one biological parent	138 (46%)	161 (52.6%)	299 (49.3%)
...non-parental care	16 (5.3%)	16 (5.2%)	32 (5.3%)
Missing	3 (1%)	3 (1%)	6 (1%)
Other	4 (1.3%)	2 (0.7%)	6 (1%)
Were you born the UK?			
Yes	247 (82.3%)	246 (80.4%)	493 (81.4%)
No	49 (16.3%)	5 (1.6%)	54 (8.9%)
Missing	4 (1.3%)	9 (2.9%)	13 (2.1%)

School Characteristics

Consistent with the requirements of the EEF, all schools had a rate of FSM equal to or higher than 28%. Of the 36 schools, 12 were academies, 17 were state funded community schools and seven were mixed funding schools, incorporating Foundation schools and Voluntary Aided Comprehensives. All 36 schools were situated in urban areas and all 36 had higher than the national average numbers of pupils with SEN statements (see **Error! Reference source not found.**) and FSM eligibility.

Figure 2: FSM eligibility and SEN statements at LEIP schools



Note: Information available for 32 of 36 LEIP schools with publicly available Ofsted records.

Neighbourhood Characteristics

The schools in the study were typically recruited from deprived catchment areas. Table 8 shows some of these characteristics. Catchment areas are comprised of a number of Wards. Data for the six largest wards in each school's catchment area were compiled to generate an average score. As Table 8 **Error! Reference source not found.** shows, these neighbourhoods feature higher than average levels of socially rented houses, families without work, foreign born and ethnic minority individuals, childhood obesity and crime rates. The level of socially rented households is nearly double the national average and drug, violence against the person, robbery, and theft and handling rates are over double the national average. These statistics show that the young people in the study live in neighbourhoods characterised by high crime levels and deprivation.

Table 5: Average neighbourhood characteristics

Measure (Source of data)	LEIP School Average* (Range)	London Average	UK Average
% with no qualifications (2011 census)	19.01 (9.95; 30.85)	17.6	22.7
Crime rate (Metropolitan Police Service)	105.03 (75.92; 154.65)	92	66
Violence against the person rate (Metropolitan Police Service)*	21.78 (17.08; 29.75)	18	10.6
Robbery rate (Metropolitan Police Service)	4.93 (2.49; 7.91)	4.2	1.2
Theft and handling rate (Metropolitan Police Service)	43.86 (24.16; 77.65)	38.9	24.6
Drug rates (Metropolitan Police Service)	8.04 (3.23; 14.94)	6.2	3.7
% Black, Asian and Minority Ethnic (2011 UK Census)	46.44 (22.57; 84.07)	40.2	14

% Not born in UK (2011 UK Census)	44.22 (26.22; 59.1)	36.7	13.4
% Children age 10-11 who are obese	24.10 (18.68; 29.66) (Department of Health Cross-Government Obesity Unit NCMP Dataset)	22.1 (2011 UK Census)	19 (2011 UK Census)
% Employment age 16-74 (2011 UK Census)	62.18 (54.39; 69.71)	65.4	64.5
% Households social rented (2011 UK Census)	34.34 (17.82; 49.67)	24.1	17.6
% dependent children (0-18) in out-of-work households (HM Revenue & Customs)	27.9 (15.68; 36.28)	21.7	18.1
% of lone parents not in employment-2011 (2011 UK Census)	51.68 (45.02; 65.74)	47.8	40.6

Notes:

* Range of LEIP school catchment area ward averages shown in brackets.

+ Data reported based on the Metropolitan Police Service are rates per 100,000.

Appendix D: Original study questions and summary of original educational and exclusion outcomes

Original study research questions

The original study sought to answer the following research questions.¹² Does the intervention affect the:

1. Behaviour of participants in terms of officially recorded truancy, temporary and/or permanent exclusions?
2. Self- or teacher-reported disruptive behaviour of participants?
3. Educational attainment of participants in terms of GCSE or other formal tests (e.g., SATs)?
4. Communication skills of participants in terms of their expressive language, understanding, language processing, and/or social communication skills?
5. Self-reported and officially recorded delinquent and/or criminal behaviour of participants?
6. Likelihood of being Not in Education Employment or Training (NEET) once the children complete compulsory schooling?

The London Education and Inclusion Project (LEIP) was an evaluation of a 12-week-long intervention, Engage in Education-London (EiE-L), delivered by the charity Catch22, in collaboration with ICAN. The intervention was aimed at young people in secondary schools who were most at risk of school exclusion. It targeted their communication and broader social-behavioural skills with the aim to reduce school exclusions and problem behaviours.

The aim of the trial was to evaluate the short and long term effectiveness of the intervention on school exclusions and academic outcomes; as well as the short term effects on behavioural and interpersonal outcomes. The evaluation was a cluster-randomised controlled trial involving pupils in 36 schools in London. Schools were randomly allocated into the intervention or control group. Pupils in the 19 control group schools received a one-off workshop focusing on employability skills delivered by trained corporate volunteers.

Summary of original educational and exclusion outcomes

The original LEIP study collected measures of exclusion and aptitude in Maths and English. For the former, this consisted of pupil and teacher self-report, and official records of exclusion, before and after the intervention. For the latter, this consisted of in-school testing completed online using the Centre for Evaluation and Monitoring (CEM) Yellis and MidYIS tests at baseline and follow-up. We present summaries of the results from these outcomes below. For the exclusion data, we report these as presented in previously published papers (Obsuth et al., 2016, 2017). Data were analysed through intent-to-treat analyses and Full-Information Maximum Likelihood (FIML) was utilised to address missing data; resulting in a sample of 615 for CEM analyses, 644 for self-reports, 685 for teacher-reports and 710 for official records.

Exclusion outcomes showed mixed results and were largely null in the original study

The results for **exclusion** reported in Obsuth et al. (2017) suggested a small but statistically significant negative effect on the primary outcome of exclusion and null effects for the secondary outcomes that measured educational, behavioural and socio-emotional outcomes. Specifically, at post-intervention, students in treatment schools were slightly more likely to self-report that they had been excluded in the previous month than students in control schools. Although not statistically significant, the teacher reported exclusion data as well as the official records of exclusion revealed the same direction of effects: more exclusions in the treatment schools post-intervention. With respect to the additional outcomes, the analyses revealed no significant differences between the students in the treatment schools versus control schools on any of the 15 secondary outcomes including communication skills, prosocial behaviors, student-teacher relationships, antisocial behavior, delinquency, official arrest records, and academic attainment in Maths and English (see Obsuth et al., 2016, 2017).

¹² https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEF_Project_Protocol_EngageInEducation.pdf

Additional analyses reported in Obsuth, et al. (2016) were carried out to examine a set of pre-specified characteristics as potential moderators of the effects of EiE-L on the primary outcome—school exclusion. These analyses revealed few significant moderator effects. Specifically, young people who attended up to 75% of the total intervention sessions were more likely to report exclusions than controls, in contrast to non-attenders and High Attenders (>75% of sessions attended), who were no different to controls. Low engagement in group and one-to-one sessions was associated with young people self-reporting more exclusions than controls. As with our findings related to attendance, young people who attended sessions and were highly engaged were no different to controls. Furthermore, only one of the four assessed baseline characteristics – problem behaviour, communication, school bond, teacher-student relationship – revealed a significant moderating effect. Specifically, and perhaps counter-intuitively, young people in the treatment group with *more serious baseline behaviour problems* were reported by their teachers to receive *fewer* exclusions than comparable pupils in the control group. On the other hand, young people with *less serious baseline behaviour problems* who received the intervention were reported to receive *more* exclusions than their comparable counterparts in the control schools. A similar pattern of findings was found in the self-report data (see Obsuth et al., 2016 for further details).

Finally, analyses of the demographic characteristics – sex and age/year group – revealed a significant moderating effect relating to the year group of the young people but not in relation to the young people's sex. Teachers of Year 9 pupils in treatment schools were more likely to report exclusions than teachers in control schools; whereas teachers of Year 10 pupils in treatment schools were less likely to report exclusion than in control schools. The interaction was not significant for the young person reported exclusions. While the positive Year 10 results may seem encouraging, it is tempered by the fact that we did not observe statistically significant reductions in exclusions according to self-reports. Furthermore, results for this moderator derived from official records showed the opposite direction of effect, meaning that we can be less confident about this finding. We must also acknowledge that there could be differences according to source because of gaps in how exclusions are recorded (e.g. not all exclusions are officially recorded).

Academic aptitude results were null in the original study for both vocabulary and maths

The results for **attainment** were as follows: there were no significant differences between pupils in treatment and control schools for either verbal ability ($b = 1.386$, se 1.430, $p = 0.333$; SMD 0.079) or maths ($b = 2.425$, se 1.518, $p = 0.110$; SMD 0.138) (see Table 7 in Obsuth et al., 2017). However, there were limitations to the data collection that led to both baseline imbalance on CEM data, and attrition at follow-up (discussed below).

Background information on CEM verbal and maths tests

Pupils completed a computer-administered aptitude test measuring academic potential developed by the Centre for Evaluation and Monitoring at Durham University.¹³ This measure provides a standardised, adaptive, curriculum-free assessment of the young people's maths and verbal abilities. It was selected based on its adaptive format meaning it is free from floor effects. At baseline the measure was administered by the fieldworkers and scored by CEM staff who provided standardised scores for aptitude in Maths and English.

The vocabulary test measures pupils' knowledge and understanding of words. Pupils are presented with a target word and need to choose which one of four option words has the nearest meaning to the target word. The maths test covers a range of competencies in maths with relatively short closed questions. Both subtests have been shown to be highly predictive of later academic achievement (<http://www.cem.org/secondary>). Results are scored in the same way as intelligence tests, i.e. with a mean of 100 and a standard deviation of 15 in the reference population sample. The tests are designed to adapt to the skills level of each pupil. In other words, the difficulty of each new question presented reflects the estimated skills level of the pupil based on their responses to previous questions. At the end of each test a standard error and convergence statistic is computed for the test score. The standard error depends on the number of questions asked and how closely the difficulty of those questions were matched to the pupil's ability – the greater the number of questions asked, and the closer the match between the questions' difficulties and pupil ability, the smaller the error. The convergence is the difference in the pupil's ability estimates between their penultimate question and their final question. As such it provides an indicator of whether the estimates of a pupil's ability had stabilised by the test's end.

¹³ <http://www.cem.org>

Reliability

Because each individual responds to a specific sequence of tasks no conventional reliability scores can be computed. However, CEM provided an overall reliability score for the measure via email communication on 22 October, 2014. According to this, the Rasch Person Reliabilities (Model Reliabilities) for the computer-based subtests in 2012/13 were provided as: Vocabulary= 0.892; Mathematics= 0.941.

Non-consistent responses

For quality control purposes the process developed by CEM includes an algorithm that excludes results where certain criteria have not been met. For example the number of questions answered, convergence and time taken are all used to determine whether a test result can be considered valid. In line with CEM recommendations we accepted all tests where CEM had computed scores as valid estimates. Across both waves of data collection no overall scores were computed for 18 vocabulary tests and 29 maths tests.

Test Duration

CEM provided the research team with measures of the duration of the vocabulary and maths subcomponents of the test. The vocabulary test took subjects an average of 5.0 minutes (sd = 2.3 min) and had a range of between 1 and 12.6 minutes. The maths test required an average of 10.2 minutes (sd = 5.9 min) and a range of between 1.1 and 29 minutes. Especially for the maths test we found a strong correlation between the duration of the test and the test score of $r = .63$, suggesting that young people who spent more time on trying to solve the problems scored higher. A similar but weaker association was found for the vocabulary subtest ($r = .30$).

Correlations between test scores

Both pre and post assessment scores in vocabulary and maths skills were correlated ($r_{pre} = .536$; $r_{post} = .533$) and there were relatively strong test-retest stabilities of $r = .583$ and $r = .552$ for verbal and maths skills respectively. These stability estimates are lower than those reported by CEM for the overall test results. The lower stability may be due to the utilization of only two subtests in this study and the fact that in the target population young people with learning difficulties and behavior problems were over-represented, meaning that test scores may be less reliably measured than in normative populations.

Challenges with test delivery

There were several issues with test delivery, which are relevant for considering the validity of the measures in this study.

At the baseline data collection, the CEM tests were completed on the CEM server and the Durham team uploaded the participant details when these had been made available to the evaluation team by the school before baseline data collection. By pre-loading this information, it meant the participants could select their own name from a drop down list, easing administration. The test results were automatically saved as the participant completed the test (incomplete results were also saved).

At the post-intervention data collection, the same procedure was followed in the first 12 Phase I schools, which completed data collection in March. However, the evaluation team was informed by CEM at the beginning of March (i.e. during data collection) that their online assessment would be shut down as of 30th April for maintenance and preparation for the new academic year. This unforeseen change resulted in several additional steps being required in the remaining seven Phase I schools. As this alternative procedure was time-consuming and potentially more error-prone at multiple steps, CEM agreed that we would be able to go back to the online administration from the server for the Phase II schools. As the standardisation was slightly altered during the month of May, when the system was updated, a formula was applied to the data collected after May 2014 to make it comparable with the information collected at the baseline.

These technical problems created difficulties in the schools when administering the test. This affected the young people's willingness to complete the tests. For instance, the alternative procedure required that 1.) computer settings allow installation of external material, 2.) installation of the CEM tests from school specific USB keys on each of the computers (up to 10), 3.) downloading completed CEM tests back onto the USB keys, 4.) ensuring that the data was downloaded correctly and has been deleted from the computer, 5.) sending all collected CEM tests to the CEM office for processing. All of these required time and IT support, which led to multiple sittings, rescheduling and disruptions. Not surprisingly some young people lost motivation and often did not attend a rescheduled session – meaning that fewer than 500 pupils overall completed post-intervention CEM testing.

Baseline imbalance on CEM tests appeared to be due to higher missingness in the control group

The original research team undertook exploratory analyses of the CEM data to understand the imbalance found at baseline and missingness on outcome testing.

Table 6 presents baseline and post-intervention means and sample sizes for treatment and control groups for the educational aptitude data. It is notable that there were imbalances at baseline on both the verbal and maths scores at, but owing to the asymmetrical missingness (missing in control) we cannot be sure whether this represents a true baseline difference or is an artefact of missing data.

Table 6: Baseline and post-intervention CEM mean scores by allocation.

	Baseline CEM		Post-intervention CEM	
	Intervention Group	Control Group	Intervention Group	Control Group
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Maths	83.39 (16.44) N=300***	88.10 (14.64) N=258***	82.51 (16.47) N=238	83.07 (18.36) N=229
Verbal	90.31** (16.55) N=304	93.82** (13.59) N=264	90.10 (17.06) N=239	91.04 (17.25) N=234

Note: * $p < .05$; ** $p < .01$; *** $p \leq .001$. Acknowledging CONSORT advises against significance testing at baseline.

Overall assessment of data quality

We believe that the difficulties with the test administration described above may have had some limited negative impact on the quality of the collected data. We also note that the subtests used in this study are relatively short and may therefore be somewhat less reliable than more extensive tests. On the other hand, the test instrument provided by CEM is a widely used non-curricular test for academic abilities that has inbuilt mechanisms for discovering inconsistent responding. Overall, issues with baseline administration that suggested imbalance and attrition from the testing led to the conclusion that it would not be appropriate to base the assessment of effectiveness on the CEM data. As such, the decision was taken to defer outcome analysis until both year groups had completed GCSEs.

Figure 3 (overleaf) shows the distribution of the CEM maths (Panel A) and verbal ability (Panel B) scores at baseline, comparing treatment and control groups. For maths (Panel A), it is clear that the control distribution does not mirror the treatment distribution. There is a substantial gap in the control distribution that results in two peaks for that group. This is reflected in the means in

Table 6 above. For verbal ability, Panel B, it seems clear that the treatment and control peaks are not aligned, with the treatment group having a lower peak (again reflected in the means in

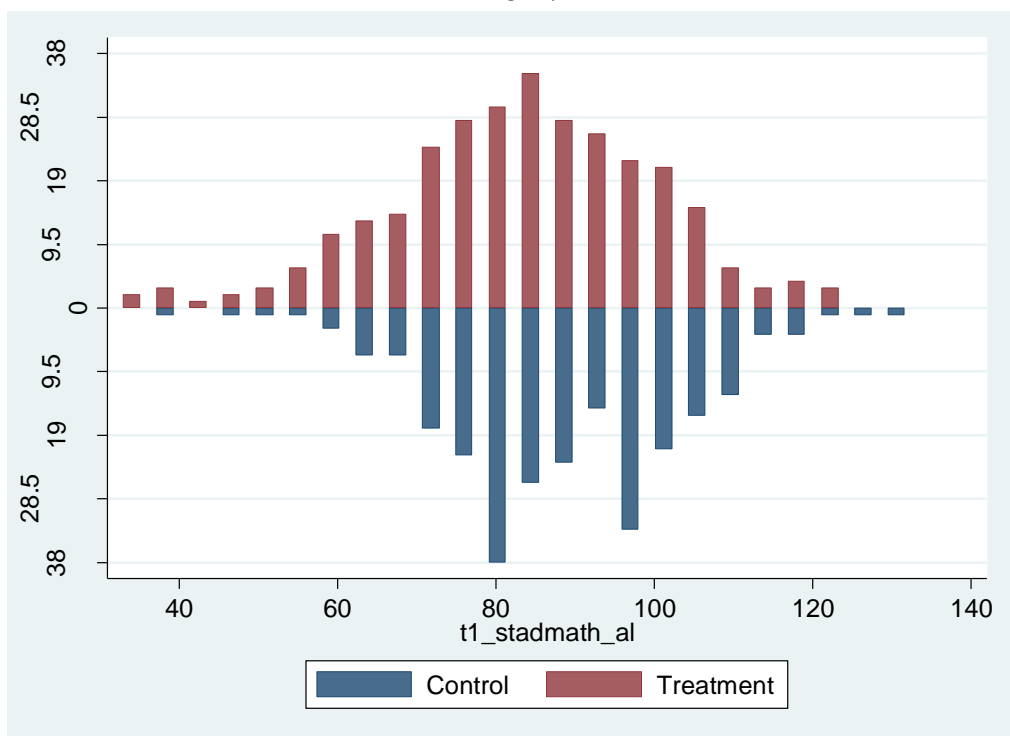
Table 6 above).

Overall assessment of data quality

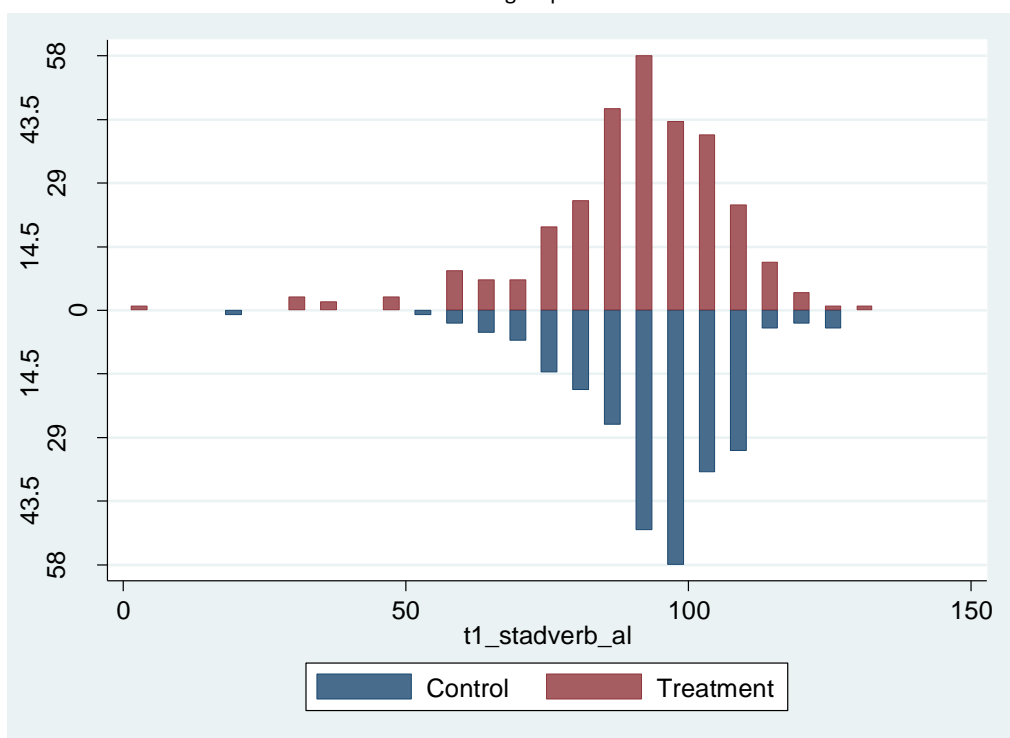
We believe that the difficulties with the test administration described above may have had some limited negative impact on the quality of the collected data. We also note that the subtests used in this study are relatively short and may therefore be somewhat less reliable than more extensive tests. On the other hand, the test instrument provided by CEM is a widely used non-curricular test for academic abilities that has inbuilt mechanisms for discovering inconsistent responding. Overall, issues with baseline administration that suggested imbalance and attrition from the testing led to the conclusion that it would not be appropriate to base the assessment of effectiveness on the CEM data. As such, the decision was taken to defer outcome analysis until both year groups had completed GCSEs.

Figure 3: Baseline distribution of CEM maths (panel A) and verbal (panel B) scores

Panel A: Distribution of CEM maths scores between groups



Panel B: Distribution of CEM verbal scores between groups



Appendix E: Predicting loss to follow-up analysis based on n=738 pupils consented at baseline

We assessed whether loss to follow-up was predicted by the variables used to randomize, allocation itself, or pupil sex (or missing pupil sex). This is important to understand, because if there is a relationship between these variables and loss to follow-up it tells us that we may need to include specific measures in subsequent analyses.

Our initial analysis included all the variables listed in the previous paragraph. Our analysis was based on the initial screened sample of pupils with available data – the maximum sample size at that point was n=751 – but we had complete data for 738 pupils (98%), 241 of whom were not included in the follow-up analysis. However, very wide confidence intervals for some estimates raised concerns about collinearity. Given the approach taken to randomization (minimization) the most likely candidates for collinearity were the variables used to randomize being used alongside allocation itself. Re-running the analysis without those variables (Table 7) shows that there was no relationship between allocation or pupil sex and loss to follow-up. There was a very weak relationship between pupil sex being missing pre-allocation and loss to follow-up but the small number of cases (n=70) could be problematic for estimate as logit models do not perform well at the tails of the distribution or with rare events (see e.g. King and Zeng, 2001). Similarly, even without the minimization variables included the bounds for the school-level ‘allocation’ variable are very wide, but given the small sample-size at school-level this likely reflects the uncertainty around this estimate for the specific outcome. Further analyses of pre-intervention baseline characteristics is possible, but this information would only be available for pupils who completed baseline surveys (n=606). For example, running such a model with the baseline behaviour problem measure (not tabled) shows that was not related to loss to follow-up. The other counter-argument to repeatedly testing for loss-to-follow up is that this is effectively data-mining and the more testing done, the more likely one is to find a difference by chance.

Table 7: Multilevel logistic model predicting loss to follow-up using pre-randomisation data (n=36 schools, n=738 pupils)

Outcome:	95% CI					
No follow-up	Odds Ratio	Std. Err.	Z	P>z	Lower bound	Upper bound
School level						
Treatment: 0=Control, 1=Treatment	0.68	1.32	-0.20	0.84	0.02	30.52
Pupil level						
Girl	0.92	0.40	-0.20	0.84	0.39	2.18
Sex missing	0.06	0.09	-1.97	0.05	0.00	0.99

Note: Odds ratios are centred around 1, with 1 meaning “no difference between groups”, values above/below 1 mean that the outcome is more/less likely in one group or another (depending on which group is the reference category). ‘Sex missing’ n=70.

Appendix F: Ethics, trial registration & data protection

Trial registry number: ISRCTN23244695

Ethics/code of conduct

The original LEIP project and the consent procedure were approved by the Institute of Criminology Ethics Review Committee on 20 May 2013 (approval letter available upon request). All schools involved in the study signed data sharing agreements with the University (example data sharing agreement available upon request).

The follow-up study was approved by the Institute of Criminology Ethics Review Committee (University of Cambridge) on the 7th of May 2017.

Teacher consent

Teachers were asked to complete an informed consent form when filling out the online and paper versions of the study questionnaire.

Parental consent

Both opt-out parental consent and pupil assent were sought, which introduced the possibility that parent(s) would agree, but pupil participants would refuse. Letters to parents were sent out by schools, signed by headteachers. Parents were given one week to advise the research team or school by contacting the school (either by post or phone) and indicating that they wished to opt their child out of the study. Of the approximately 800 letters sent to parents, 15 parents/guardians opted their child out of the project.

Young person assent

Young people were presented with an Information Sheet and Assent Form prior to completing any questionnaires. Fieldworkers read out the study information portion of the assent form to the group and make sure that each young person understood what was being asked of them, allowing time for questions to be asked. As we intended to follow up participants beyond the life of the study and link their data to government records, they were asked to tick separate boxes to consent to linking self-reported data to official records. We reiterated that we would not share any of this information with the school, their parents, the police or anyone else, that their responses to the questionnaires are confidential, and that all of their information will be anonymised. Once this was completed and the young people had the opportunity (or been prompted) to ask questions, they were asked to confirm their willingness to participate by signing the forms. When parental consent and young person assent were opposed then parental consent was considered.

School identification and recruitment

The study protocol (Obsuth et al., 2014) describes both the school and pupil recruitment strategies in detail, but in brief, Inner London schools with a free school meal rate equal to or greater than 28% were approached to take part. Special schools, pupil referral units and schools already involved in similar initiatives were excluded from the trial. Schools were approached via letter, detailing the study and invitation to participate in the study, following up via email and telephone. Interested schools were invited to send back to us an Expression of Interest (EOI) document, which was followed up via telephone. As initial recruitment was slow, the team began a second phase of school recruitment in a small number of Outer London boroughs. Schools were eligible on the basis of (1) the school having a FSM prevalence $\geq 28\%$ (2) the number of schools in a given borough; and (3) physical proximity to schools already in the study. Interested schools were invited to an Information Event held in central London, during which the study was further explained to them. At the end of recruitment 29 of the 36 schools included in the study were present at the initial Information Event.

Pupil identification and recruitment

Year 9 and 10 pupils at high risk for fixed-term exclusion ('suspension') from school during the 2013/14 academic year in select schools in London were the target population. The intention was to focus on pupils in the top 5% in the 9/10 Year group in terms of problem behaviour. Schools were asked to nominate between 16-24 young people (8-12 per year) at the greatest risk for exclusion, with the number of pupils nominated dependent on school size. The guidelines asked schools to select the young people who are at high risk for school exclusion and/or becoming NEET based on a) having had previous school exclusions, b) unauthorized absences, and c) having engaged in behaviours that lead to other disciplinary measures previously being used.

Data protection

The original study was conducted under the Data Protection Act (1998), so consent was sought for data-requests from participants. The follow-up study was also begun prior to the new GDPR regulations being introduced, as such, some initial data requests were on the basis of consent and governed by the DPA (1998). Eventual data requests for the National Pupil Database fell after GDPR implementation. This meant that requests for additional data were sometimes refused on the basis that the 2013 consent / project information was not compliant with GDPR. Follow-up requests post-GDPR for arrest data, for example, were not responded to.

Appendix G: Non-educational outcome measures - suggestions for EEF teams conducting follow-up studies

A key objective for the project was to test out the feasibility of collecting non-educational follow-up data, and ideally collect that data for use in the project. To that end, and several months prior to starting the project, the team began discussing data access with the Administrative Data Research Network (ADRN) and submitting an application for the data set out in Table 8 below, covering UCAS applications, arrest/criminal history, and (originally) truancy and exclusion data. As a new entity explicitly aimed at researchers trying to access administrative data, the ADRN held promise as a 'one stop shop' – including data of interest to this project. As such, there was a good rationale for pursuing data via the ADRN. However after correspondence, including submission and re-approval of the project ethics by the ADRN, it became clear that the data being sought could either not be accessed via the ADRN, or could be requested directly from the respective data holders. Following that, the team submitted applications to the data holders set out in Table 8 but were unable to obtain the additional data. We provide a summary of issues encountered with data access in the table, but provide some more detail here.

UCAS data: We had wanted to understand whether there were any impacts on applications to university and/or to understand more about what information UCAS may hold. Our focus was on whether LEIP participants and controls differentially applied to university, and if they did, whether there were differences in offers or acceptance. (With hindsight, looking at *university* rather than FE applications was an optimistic request in the timeframe given the eventual GCSE results of the pupils.) Through discussions with UCAS, we learnt that a limited amount of university application data could be requested via the 'Strobe service'. This was initially understood by the research team as being akin to the Ministry of Justice 'datalab', where data could be submitted and either analysis outputs or data returned to the results team. However, the UCAS 'Strobe' service was only able to offer aggregated summary data for each of the treatment and control group, but could not undertake any analysis. Not being able to return person-level data meant that we would not have been able to include baseline data and/or variables used for randomisation, or indeed the treatment allocation. As such, it was clear that this service would not be suitable for this project. If the Strobe service were able to provide group-level means, standard deviations and other outcome-appropriate descriptive statistics then some form of comparison would be possible. However, without the inclusion of stratifying variables or accounting for the data structure, for example, differences between groups may be exaggerated. Specifically, through the underestimation of standard errors and subsequent confidence intervals there is an increased risk of falsely rejecting the null hypothesis of no difference.

Arrest/conviction data: there have been several different channels through which to request criminal justice data. The 'backbone' of justice data is the Police National Computer (PNC) (Sutherland, 2014). PNC is a central database that contains information on aspects of offence type, date of offence, arrest, charge and – if available – criminal justice outcome(s). The original project analysed data on arrest, and we requested this follow-up data initially from the Metropolitan Police Service as we had done previously. There was a substantial delay in receiving a response from the MPS about the request, and the eventual response was that the request could not be processed using consent as the legal basis. Following that, we also made enquiries and eventually submitted a request to the Home Office (via the Ministry of Justice) in autumn of 2018 but the response was that as this was not a government sponsored project the request would not be a priority.

Those interested in non-educational follow-ups would need to invest a substantial amount of time discussing data access with both the organisation holding source data, and any organisations purporting to facilitate access to that data. The ONS secure data service is one such avenue, but there is a lack of clarity around timelines, and the breath of data access is also unclear at this time.

Table 8: Overview of data collection efforts

Data source	Outcome(s) of interest	Process	Lessons learned	How to access the data now
Administrative Data Research Network (ADRN)	<ul style="list-style-type: none"> School truancy and exclusion data Application rates to university, entry into higher education Arrest records 	One organisation to access data from different data holders. Applications then proceed using the same process as data holder.	The scope of the ADRN has changed significantly, focusing more on survey, registration and census data.	See the ADRP website for details of which datasets are available: https://esrc.ukri.org/research/our-research/administrative-data-research-partnership/
Universities and Colleges Admissions Service (UCAS)	<ul style="list-style-type: none"> Data on application to university. Data on success rate. (Received offer) Data on offer uptake. (Accepted offer) 	UCAS data needs to be accessed via the STROBE service.	Strobe service is not appropriate because they are unable to conduct the necessary analysis on behalf of requestor. E.g. aggregate means by group are returned but no analysis is undertaken.	https://www.ucas.com/about-us/policies/terms-and-conditions/strobe-service-terms-supply The main issue remains that STROBE does not allow statistical model output to be returned. Requesting e.g. means, standard deviations (etc) by treatment group would allow comparison to be undertaken but would not account for clustering.
Metropolitan Police Service (MPS)	<ul style="list-style-type: none"> Arrest records: date of offence and arrest, offence type. Criminal justice outcome associated with an offence. 	Data first requested via MPS. Follow-up request made via Ministry of Justice & Home Office PNC service desk.	Routes to accessing arrest and PNC are not clear. There are significant pressures on access to arrest/PNC data and non-governmental research is not a high priority.	Police National Computer (PNC) data can be requested via the Police Database Information Access Panel (PIAP), run by the Home Office. Note that PNC data is being migrated onto National Law Enforcement Data Services but the timeline is not confirmed.
Department for Work and Pensions	<ul style="list-style-type: none"> Not in Education, Employment or Training 	This was not available through the ADRN at the time of application.	n/a	It may be possible to access via the Longitudinal Educational Outcomes (LEO) dataset. https://www.gov.uk/government/statistics/graduate-outcomes-2015-to-2016

Appendix H: Distribution of outcomes & baseline variable

In this section we briefly present figures showing distributions for the outcomes and baseline measure used in our analyses. These help to demonstrate to the reader why we chose to model outcomes in a particular way, and also why we approached our sensitivity analyses as we did.

The outcomes used in the analyses presented in the main body of the report were count measures: counts of exclusions, counts of missed sessions owing to exclusion, and counts of GCSEs. As such, it is important to examine the distribution of these outcomes to help to inform the modelling approach. For example, it is important to understand whether a count variable has an excessive number of zeros or is over-dispersed, both of which have implications for the analysis approach eventually used (see Long and Freese, 2006).

In Figure 4 and Figure 5 below, we set out the distribution of these two exclusion outcomes. Both of these figures illustrate that the outcomes are very skewed, as might be expected with count variables. There are also a substantial number of zeros in the data, which suggests that a modelling approach that deals with that may be more appropriate than a standard Poisson model.¹⁴

For the number of GCSEs achieved Figure 6, this is much less skewed, which suggests that a standard count (or even linear) model may be appropriate for analysis.

For the substituted baseline measure – KS2 maths Figure 7 – the figure shows that there were very few values in the data (only five). As such, this variable was treated as categorical in analysis, meaning that we included dummy variables for n-1 categories.

¹⁴ See <https://stats.idre.ucla.edu/stata/dae/zero-inflated-poisson-regression/>

Figure 4: Distribution of fixed-term exclusion count outcome (n=557)

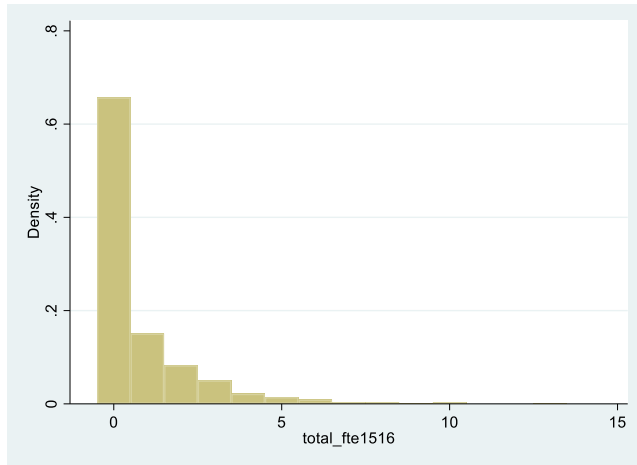


Figure 5: Distribution of FTE missed sessions outcome (n=557)

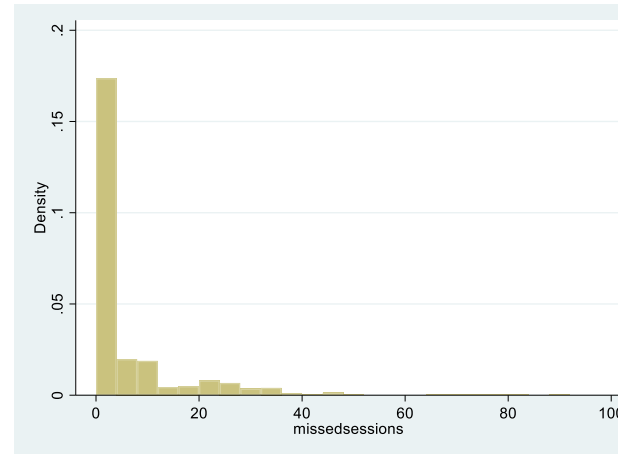


Figure 6: Distribution of attainment outcome – GCSE A-G count (n=537)

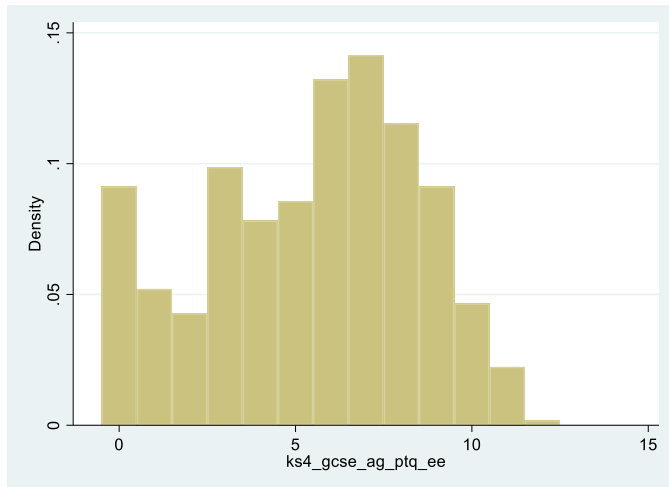
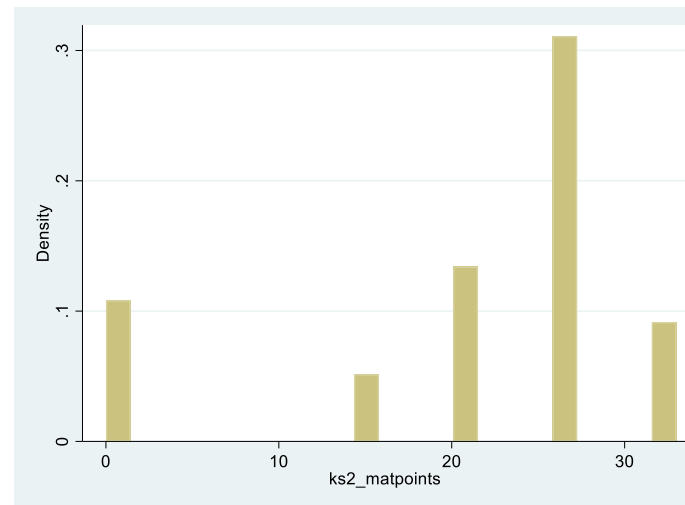


Figure 7: Distribution of baseline attainment measure maths points score (n=540)



Appendix I: Sensitivity analyses

This section sets out the results from sensitivity analyses conducted for the two main outcomes: exclusion and the number of GCSEs sat. The reason for doing this is to assess whether our results change if we use a different approach to our analysis. By ‘change’ this could be the direction of effect, or the magnitude of effect. To summarise the results from these additional analyses: there was no detectable effect for either outcome, and these results did not vary considerably by the type of model used. Below we provide additional details on the analysis approaches used.

For exclusion, we assessed this by operationalising the outcome as (i) the total count of FT exclusions across the two years of data and (ii) as the total number of sessions (AM or PM) missed in the same time period. For the number of GCSEs sat, we only operationalised this in one way (a count of GCSEs sat), but used several model specifications to assess robustness to specification. The results should be regarded as exploratory in nature.

Exclusion

We tested out several model specifications for exclusion. Table 9 below presents results from these analyses, giving the point estimate and confidence intervals for the treatment allocation variable from each model (other variables omitted). Model 1 restates the main results from Table 4 (a multilevel Poisson model). The remaining models mirror the same non-significant result using different specifications. Model 2 is the result from 500 bootstrap replications of the main model result. Models 3 and 4 are results from single-level zero-inflated Poisson models (with cluster-robust standard errors). (Zero-inflated because the graph of the number of fixed-term exclusions suggests an excess of zeros.) Table 10 is the result of the same model but substituting the number of missed AM/PM sessions owing to exclusion. As with the main result and sensitivity analyses in Table 9, the result for treatment allocation was not different from one (IRR 1.09; se 0.21; p. 0.65; 95% CI 0.75-1.60).

Table 9: Sensitivity analyses for number of fixed-term exclusions outcome (n=557)

Model specification	Model number / notes	IRR	se	z	p	95% CI	
						Lower	Upper
Poisson (Multilevel)	1. Result reported in Error! Reference source not found.	0.93	0.22	-0.30	0.77	0.59	1.47
	2. Bootstrapped Poisson estimate (n=500 replications)	0.93	0.17	-0.38	0.71	0.65	1.33
Zero-inflated Poisson (Single level)	3. Zero Inflated Poisson	0.81	0.12	-1.46	0.15	0.60	1.08
	4. Bootstrapped ZIP (n=500 replications)	0.81	0.12	-1.46	0.15	0.60	1.08

Table note: All models control for: baseline exclusion, phase of roll-out, pupil sex. Single-level ZIP models have SE adjusted for clustering by school. Note that the single-run and bootstrapped ZIP models returned the same results in terms of the point estimates and standard errors for the treatment indicator, but there were some small differences for the standard errors for other predictors (not tabled).

Table 10: Results for number of missed sessions owing to exclusion (n=557)

Zero-inflated Poisson model					95% CI	
Outcome: number of missed sessions	IRR	se	z	p	Lower	Upper
Treatment: 0=Control, 1=Treatment	0.91	0.17	-0.47	0.64	0.63	1.33
Baseline exclusion?: 0=No, 1=Yes	0.83	0.10	-1.48	0.14	0.65	1.06
Delivery phase: 1 / 2	1.22	0.22	1.11	0.27	0.86	1.74
Female: 0=male, 1=female	0.94	0.18	-0.31	0.76	0.65	1.37
Intercept	13.52	4.42	7.97	0.00	7.13	25.64

Attainment

We also re-examined the results from the attainment analyses presented earlier. Table 11 sets out the results from these analyses, starting by restating the results presented earlier in **Error! Reference source not found.** All models presented control for delivery phase, KS2 maths points (entered as a categorical variable with zero as the reference category owing to the nature of its distribution – see Figure 7) and pupil sex. The consistent theme across different model specifications is that treatment allocation was unrelated to the number of GCSEs sat. The first four results are from multilevel linear models, with result 1 as per **Error! Reference source not found.** above. In Model 3, we apply the Kenward-Roger correction for small samples (see McNeish, 2017) but as expected this affects the standard error rather than the point estimate. Models 5 and 6 use a count specification (Poisson) as the outcome is a true count variable, with again both sets of results indicating a null result. Finally, model 6 is a full information maximum likelihood model (FIML), which uses all available information to estimate results. The point estimate and other statistics are consistent with the other null results reported from linear models reported (M1-M4).

Table 11: Sensitivity analyses for number of GCSEs grades A*-G achieved

Outcome: number of GCSEs grades A*-G achieved								95% CI	
#	Type	Comment	n	Point estimate	se	z	p	Lower	Upper
1	Linear	Multilevel (Error! Reference source not found.)	520	-0.26	0.32	-0.79	0.43	-0.88	0.37
2	Linear	Multilevel (n=500 bootstrapped replications)	520	-0.26	0.34	-0.78	0.44	-0.95	0.43
3	Linear	Multilevel w/small sample correction	520	-0.26	0.34	-0.78	0.44	-0.95	0.43
4	Linear	Multilevel w/small sample correction (n=500 bootstrapped replications)	520	-0.26	0.25	-1.05	0.29	-0.75	0.23
5	Count	Multilevel Poisson model ⁺	520	0.95	0.06	-0.85	0.40	0.85	1.07
6	Count	Multilevel Poisson model + Bootstrap ⁺	520	0.95	0.05	-1.05	0.29	0.87	1.04
7	Linear	Full information maximum likelihood	557	-0.26	0.32	-0.80	0.43	-0.88	0.37

Table note: ⁺Reported result is an incident rate ratio (IRR), centred on 1 and ranging from zero to infinity.

Project team

Team member	Affiliation	Roles/responsibilities
Dr. Alex Sutherland	RAND Europe	Senior researcher for original LEIP evaluation. Lead of the long term follow-up study.
Dr. Ingrid Obsuth	University of Cambridge	Project manager / lead researcher for original LEIP evaluation.
Miriam Broeks	RAND Europe	Project manager for follow-up study
Dr. Yulia Shenderovich	RAND Europe	Review analyst for quantitative analysis

The original LEIP project was co-funded by the EEF and European Social Experimentation Grant (EC reference VS/2012/0345) with funding for the programme provided to the Greater London Authority. The original evaluation was led by Professor Manuel Eisner, Institute of Criminology, University of Cambridge and the protocol and results have been published as journal articles (Obsuth et al., 2014; 2016, 2017).

Acknowledgements

We would like to express our gratitude to Catch22, the intervention partner on the original LEIP project. Thanks also to the European Commission, which provided the funding to the Greater London Authority for this specific evaluation project, in collaboration with Professor Manuel Eisner, University of Cambridge. We are grateful to Professors Eisner and Gelsthorpe at the Institute of Criminology for allowing access to the LEIP dataset for the follow-up study.

We would also like to acknowledge and thank the Education Endowment Foundation who not only provided financial support for the implementation of the Catch22 intervention, but also building on their many years of experience sponsoring evaluation research, provided guidance and help, particularly with hard to reach schools.

Next, we would like to thank the leadership team, humanities teacher Lauren Bailly and the young people from The Manor School in Cambridge, who provided us with invaluable feedback on our questionnaires during our pilot project. Further, we would like to acknowledge and thank all the LEIP fieldworkers who demonstrated diligence, sensitivity and professionalism in working with the young people and schools in this project.

Finally, our deepest gratitude goes to the 36 LEIP schools, teachers and young people who participated in this project and shared their time and experiences with us. Without them this project would not have been possible.

You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

To view this licence, visit <https://nationalarchives.gov.uk/doc/open-government-licence/version/3> or email: psi@nationalarchives.gsi.gov.uk

Where we have identified any third-party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at <https://educationendowmentfoundation.org.uk>



The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP

<https://educationendowmentfoundation.org.uk>

 [@EducEndowFoundn](https://twitter.com/EducEndowFoundn)

 Facebook.com/EducEndowFoundn