Education
Endowment
Foundation

INCLUSIVE (LEARNING TOGETHER)

Evaluation Report

Michael Wigelsworth, Emma Thornton, Patricio Troncoso,
Neil Humphrey, and Louise Black.

May 2023

MANCHESTER
1824
The University of Manchester

Education Endowment Foundation

The Education Endowment Foundation is an independent charity dedicated to breaking the link between family income and education achievement. We support schools, nurseries and colleges to improve teaching and learning for 2 – 19-year-olds through better use of evidence.

We do this by:

- **Summarising evidence.** Reviewing the best available evidence on teaching and learning and presenting in an accessible way.
- **Finding new evidence.** Funding independent evaluations of programmes and approaches that aim to raise the attainment of children and young people from socio-economically disadvantaged backgrounds.
- **Putting evidence to use.** Supporting education practitioners, as well as policymakers and other organisations, to use evidence in ways that improve teaching and learning.

We were set-up in 2011 by the Sutton Trust partnership with Impetus with a founding £125m grant from the Department for Education. In 2022, we were re-endowed with an additional £137m, allowing us to continue our work until at least 2032.

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

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

info@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Department for Education

What Works Network

Contents

About the evaluator

Manchester Institute of Education, University of Manchester

This project was independently evaluated by a team from the Manchester Institute of Education, University of Manchester: Michael Wigelsworth, Emma Thornton, Patricio Troncoso, Neil Humphrey, and Louise Black.

The lead evaluator was Dr Michael Wigelsworth.

Contact details

Dr Michael Wigelsworth
Manchester Institute of Education
The University of Manchester
Oxford Road
Manchester
M13 9PL

Tel 0161 306 1763
Email Michael.wigelsworth@manchester.ac.uk

Acknowledgements

We are grateful for the support provided by the Learning Together Team from University College London who were responsible for conducting the main trial and providing supporting data for this evaluation.

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

Executive summary

The project

INCLUSIVE, also known as Learning Together, is a whole-school programme using a restorative practice approach to reduce bullying and aggression and promote health among secondary school pupils (pupils aged 11 to 16 in Years 7 to 11). The programme includes 2.5 hours of restorative practice training for all school staff, an additional three days of training for five to ten members of staff, and socioemotional skills curriculum materials (five to ten hours per year of lessons for pupils in Years 8 to 10). Schools must also coordinate two action group meetings per term, where a small group of staff and pupils meets to discuss action plans for improving the school climate and practices related to inclusivity informed by the results of needs assessment surveys completed by pupils. In this trial, action groups were supported by an external facilitator for the first two years of programme delivery with schools delivering the programme independently in the third year.

Delivery of the intervention was led by a team from University College London (UCL) and the London School of Hygiene and Tropical Medicine (LSHTM), which conducted an evaluation funded by the National Institute for Health Research (NIHR) examining the impacts of the programme on pupil bullying, aggressive behaviours, mental wellbeing, psychological problems, smoking, alcohol, drug use, contact with police, and health-related quality of life (Bonell et al., 2019). The EEF commissioned the University of Manchester to conduct a supplementary evaluation, reported here, which uses the existing design and data from the trial led by UCL and the LSHTM alongside data from the National Pupil Database (NPD) to examine impacts of the programme on attainment. The evaluation included 6,659 pupils from 40 schools and was set up by UCL and the LSHTM as a cluster-randomised controlled trial. The education outcomes examined in this report include Key Stage 4 (KS4) Attainment 8 scores, maths GCSE results, and English GCSE results for pupils who received the intervention through Years 8 to 10 before sitting their GCSEs in Year 11. While the evaluation led by UCL and the LSHTM included a process evaluation, no qualitative data was collected as part of the EEF evaluation. This efficacy trial began in 2014 and finished in 2018.

Table 1: Key conclusions

Key conclusions
1. Children in INCLUSIVE schools made the equivalent of two months' additional progress in KS4 Attainment 8 scores, on average, compared to children in other schools. This finding has not been assigned an EEF security rating, although it should be noted that there is uncertainty around the results.
2. There is some evidence that INCLUSIVE had a positive impact on pupils' maths and English GCSE results, with pupils in INCLUSIVE schools making an additional month of progress in maths and an additional two months' progress in English, on average.
3. There was no notable difference in the impact of the programme for pupils eligible for free school meals (FSM) compared to other pupils.
4. While the evaluation led by UCL and the LSHTM found that INCLUSIVE reduced bullying in schools—and being bullied has been found to be associated with lower attainment—this report found no evidence that INCLUSIVE had a greater impact on the attainment of pupils who had reported being bullied compared to other pupils.
5. Although the evaluation led by UCL and the LSHTM found that INCLUSIVE had greater effects for boys than girls, there was no evidence in this evaluation that INCLUSIVE impacted children's attainment differently based on gender.

EEF security rating

These findings have not been assigned an EEF security rating. The main INCLUSIVE trial, which focused on health outcomes, was not set up by the EEF nor designed based on EEF guidelines and did not have attainment as the primary outcome. As a result, there are challenges in applying the EEF's security rating criteria to the attainment outcomes presented in this report and with comparing the results of this trial to those of other EEF trials.

While a security rating will not be applied to the findings, it should be noted that there are limitations that should be considered when interpreting the results. The trial was a well-designed, two-armed, randomised controlled trial but was powered to a lower level than most EEF trials because this evaluation was supplementary to an existing trial that had been powered for evaluation of health outcomes rather than attainment. 23% of pupils who started the trial were not included in the final analyses of attainment outcomes because of issues with pupil data matching. Due to data limitations, it was not possible to evaluate whether pupil attainment outcomes varied based on whether the school had delivered the intervention as intended. The trial included some features of both efficacy and effectiveness trials but is considered by the EEF to have been closest to an efficacy trial design.

Additional findings

Pupils in INCLUSIVE schools made, on average, two additional months' progress compared to those in the control group equivalent. This is our best estimate of impact. As with any study, there is some uncertainty around the result: the possible impacts of this programme also include small positive effects of one month of additional progress and positive effects of up to three months of additional progress.

These findings provide tentative evidence that INCLUSIVE may be effective in improving pupil attainment, in addition to improving pupil health and behavioural outcomes as reported previously by Bonell et al. (2019). The intervention is primarily designed to reduce bullying and aggression but may also support wider outcomes, including pupil learning. The positive impact of INCLUSIVE on pupil attainment observed in this evaluation is comparable to impacts that have been observed for other programmes with a focus on psycho-social outcomes (Corcoran et al., 2018).

There remain some uncertainties about how the intervention leads to impacts on pupil attainment. There is wider evidence that being bullied is linked to lower educational attainment (Brown and Taylor, 2008; Glew et al., 2005; Risser, 2013; Woods and Wolke, 2004). However, this study did not find evidence that the attainment of pupils with prior experience of being bullied improved more than that of other pupils, suggesting this might not be the mechanism through which the programme impacts on attainment. Other potential explanations for the impact on attainment include that INCLUSIVE improved pupil mental health and wellbeing or pupil commitment to school, with pupils feeling more included and invested in their school due to involvement in whole-school decision-making via action groups. It is possible that these changes may have contributed to improvements in pupil engagement with learning. Alternatively, the intervention may have improved teachers' classroom management skills, enabling more learning to take place at school.

This was an innovative evaluation, using data collected for a randomised controlled trial focused on health outcomes in combination with attainment data to enable evaluation of the educational outcomes of the intervention. Consequently, some aspects of the evaluation differed from the EEF's usual approach. For instance, although the UCL-led trial included the collection of interview, survey, and observation data, the EEF-funded study did not include an implementation and process evaluation. This study also deviates from the EEF's typical approach to assessing impact on the attainment of pupils eligible for free school meals. The EEF's usual analysis approach was not deemed appropriate because of the nature of this evaluation, which built on the existing INCLUSIVE trial design. As a result, the EEF is not presenting a separate 'months' progress' figure for pupils eligible for FSM. However, the evaluation results suggest there was no difference in the impact of the programme for pupils eligible for Free School Meals (FSM) compared to other pupils.

Cost

The average cost of INCLUSIVE for one intervention school was around £50,244, or £58 per pupil per year when averaged over three years. This is an estimate of the additional costs incurred by schools in the intervention group above those of control schools incurred as part of their usual practice dealing with bullying in school. It includes costs of staff time spent dealing with bullying as well as those associated with programme activities. These cost estimates were calculated based on NIHR protocol rather than EEF's cost analysis guidelines, so caution should be taken when comparing costs across EEF evaluations.

Impact

Table 2: Summary of impact on primary outcome(s)

Outcome/ group	Effect size (95% confidence interval)	Estimated months' progress	EEF security rating	No. of pupils	p value	EEF cost rating
GCSE (KS4 Attainment 8 scores)	0.14 (0.05; 0.23)	2	n/a	5,128	0.004	£ £ £ £ £

Introduction

Background

Bullying in schools is a public health concern given the associated prevalence and long-term consequences for wellbeing and achievement often associated with these behaviours (WHO, 2014). Definitions vary; however, bullying is commonly seen as unwanted, and sometimes repeated, acts of verbal, physical, or psychological aggression inflicted on an individual with the aim of inducing harm, fear, intimidation or distress, and usually involves an uneven distribution of power (Olweus, 2013).

There is substantial evidence to show that the consequences of being bullied at school or being a perpetrator of bullying results in mental and physical harm, extending far beyond the immediate incidents and into adulthood, including increased likelihood of mental health difficulties including anxiety and depression, self-harm, and suicidal ideation (Arseneault, 2017). Exposure to bullying is also firmly associated with poorer educational attainment (Brown and Taylor, 2008; Glew et al., 2005; Risser, 2013; Woods and Wolke, 2004) such as. For example, correlations have been reported between being a bullying victim and poorer GCSE attainment (Department for Education, 2018b). Further, those who experienced frequent bullying while at school had lower qualification levels in midlife compared to their peers (Takizawa et al., 2014). Bullies themselves are more likely to have lower qualifications in adulthood (Wolke and Lereya, 2015) and those who engage in aggressive behaviour at school are also likely to have poorer educational attainment (Risser, 2013; Vuoksimaa et al., 2021).

In England, bullying is prevalent in schools. Although reported rates of bullying vary between surveys (in part due to inconsistencies in definitions of bullying across questionnaires), the prevalence of those who were frequently bullied in England in the 2015 PISA survey, 14.2%, was higher than the average for countries in the Organization for Economic Cooperation and Development, which was 8.9% (OECD, 2017).[1] Data from the Health Behaviour in School-aged Children (HBSC) project indicates that 32% of young people experienced some form of bullying within two months of data collection in 2014 (Brooks et al., 2015). Using data from the Office for National Statistics (the ONS) annual crime survey, the DfE has estimated the prevalence of bullying in England to be consistent from 2013 to 2018, remaining stable at 17% despite dropping to 15% in 2015/2016 (DfE, 2018a). With respect to frequency, there are reports indicating that one in ten students report being bullied every day (Kelly et al., 2010). The prevalence of young people reportedly bullying others is also high. Figures from the HBSC project indicate that 18% of young people had bullied another young person (Brooks et al., 2015).

Given the prevalence of bullying, and the negative long term impacts it can have—on both victims and perpetrators—prevention is a high priority in English education policy and all schools are legally obliged to have pro-active strategies in place (DfE, 2017). These policies should be communicated transparently with parents, pupils, and staff to ensure that when instances of bullying do occur, they are addressed quickly and effectively (DfE, 2017). There is growing evidence for effective strategies in addressing bullying that schools can adopt and adapt to their own specific circumstances. These include taking a whole-school approach (Langford et al., 2014) by which the focus of intervention is change at an institutional level, for instance adopting polices that promote empathic climate and culture. Changing individual perceptions and attitudes is thereby a consequence of systematic intervention rather than as a direct result of a specific curriculum or set of instructions. However, this does not preclude a more direct approach. There are also several evidence summaries noting the value of social and emotional learning (SEL) in addressing attitudes and behaviours related to bullying (for example, emotional regulation; Wigelsworth et al., 2021). Restorative practice techniques have been seen to address bullying, anti-social, and aggressive behaviour (Lloyd et al., 2006; Skinns et al., 2009; YJB, 2004), although there have not yet been any randomised trials examining the impact of this form of intervention in schools (Bonell et al., 2018).

Concerning the wider context, whereas government policy is starting to recognise some of these elements, there remains an evidence-to-policy gap: current advice and guidance does not reflect emergent evidence for what might be successful in preventing and intervening in bullying. For instance, while Department for Education guidance (2017) does note the importance of an 'ethos of good behaviour', it also states that schools 'should apply disciplinary measures to pupils who bully in order to show clearly that their behaviour is wrong' (p.13) without reference to further empathetic or

[1] In this context, 'frequently bullied' indicates those pupils among the top 10% of students with the highest values in the index of exposure to bullying across all 79 countries within the PISA dataset (OECD, 2019).

restorative type practices. This is particularly true in the secondary sector where a more rationalist approach to schooling can mean that SEL interventions are a ‘harder sell’ and often meet greater attitudinal and logistic challenges to implementation than in primary education (Lendrum et al., 2013). There is, therefore, a need to further build the evidence base for effective approaches in order to more fully inform policy and practice.

INCLUSIVE is a whole-school approach that incorporates restorative practices and SEL. The Steer Report recommended that schools should adopt restorative practices when tackling bullying and aggression (Steer, 2009), where efforts are made to repair damage rather than assigning blame and inflicting punishments on perpetrators (Wright, 1999). Such practices can involve ‘circle time’, designed to promote positive relationships, and ‘conferencing’ whereby the participants, including teachers and sometimes parents or external professionals, are involved in resolving issues (Morrison, 2005). SEL involves explicitly teaching social and self-regulation skills to children and young people. Such skills are drawn on by children, for example, to calm themselves down when they feel angry. In teaching these skills to children and young people, it is hoped that they will be able to recognise and manage emotions, develop and maintain positive relationships with others, and possess important life skills for handling challenging situations (Merrell and Gueldner, 2010). Further, interventions that focus on changing the school environment and systemic processes within the school (whole-school interventions) have promising effects on bullying and victimisation compared with curriculum-based interventions (Langford et al., 2014; Smith et al., 2004; Vreeman and Carroll, 2007), possibly because they frame bullying as a universal problem in schools rather than a problem with any individual student, thus avoiding any stigma associated with bullying or victimisation (Smith et al., 2004; Vreeman and Carroll, 2007).

INCLUSIVE was subject to a three-year cluster randomised trial, which involved a two-year facilitated intervention and a further year of observation of schools continuing the intervention without outside facilitation. The trial was conducted by University College London (UCL) and the London School of Hygiene and Tropical Medicine (the LSHTM) with funding by the NIHR; it examined self-reported experiences of bullying victimisation and perpetration of bullying (referred to collectively as ‘health outcomes’). Results of this primary trial show INCLUSIVE to be effective in reducing bullying victimisation, though there were no effects in reducing aggression perpetration (Bonell, Allen, Opondo, et al., 2019; Bonell et al., 2018).

This report draws on data from the trial, combining primary data collected by UCL and the LSHTM (health outcomes) with academic data drawn from the National Pupil Database to determine the impact of INCLUSIVE on the academic attainment of pupils in participating schools.

Intervention

INCLUSIVE is a whole-school, multi-component intervention incorporating restorative practice and a socioemotional skills curriculum, thus combining curricular and environmental components of interventions, which are usually classified separately in reviews in the field (Blank et al., 2010). The intervention also included action groups comprising staff and students that examined results of a needs assessment survey which informed choice of locally decided actions and coordination of intervention delivery. This was supported by an external facilitator. From the perspective of Humphrey’s (2013) SEL taxonomy, it may be described as a hybrid programme in terms of its prescriptiveness, offering both ‘manualised’ content of core components and flexible, needs-led delivery of non-core components. Interventions such as INCLUSIVE can be seen as part of a growing attempt to make schools central to efforts to improve the mental health and wellbeing of students in the United Kingdom.

In order to provide a comprehensive and transparent description of INCLUSIVE, we utilise an adapted version of the Template for Intervention Description and Replication (TIDieR; Hoffmann et al., 2014), drawing upon other comprehensive descriptions of the intervention, namely protocol and publications arising from the main trial examining bullying and aggression outcomes (Bonell et al., 2014; Bonell, Allen, Opondo, et al., 2019; Bonell et al., 2018).

Name of intervention

INCLUSIVE (also referred to as ‘Learning Together’).

Rationale

Bullying, violence, and aggression among children and adolescents is a public health concern (Armitage, 2021; Bellis et al., 2012). Experiences of bullying (both perpetration and being a victim: Glew et al., 2005; Risser, 2013; Woods and Wolke, 2004) and perpetrations of aggressive behaviour (Risser, 2013; Vuoksimaa et al., 2021) are related to an increased risk of poor educational attainment. Schools are well placed to target bullying and aggression, given their wide reach and central role for children (Greenberg, 2010). Further, systematic reviews have shown whole-school interventions to be effective in reducing these behaviours, compared to interventions that focus on the curriculum (Langford et al., 2014; Smith et al., 2004; Vreeman and Carroll, 2007). A logic model for INCLUSIVE can be seen in Figure 1.

Recipients

INCLUSIVE has been designed for secondary schools. School staff receive training and additional resources (see below) while a curriculum is delivered to pupils during Years 8 to 10.

Materials

School staff receive five to ten hours of teaching and learning per year on restorative practices, relationships, and social and emotional skills based on the Gatehouse Project curriculum from external education facilitators and were given written summaries of the material covered in these training sessions. All schools receive a manual containing guidelines for action groups (see below for details of these), and an external facilitator for these action groups, for the first two years of the intervention. Action groups continued in the third year but with internal facilitation. Schools were sent a report on student needs each year following the cohort from Year 7. There were findings from an annual survey of students aged 11 to 12 about their attitudes to, and experiences of, school and their experiences of bullying, aggression, and other risk behaviours. Schools are also given lesson plans and slides to aid teachers' delivery of the social and emotional skills curriculum (five to ten hours per year of lessons on social and emotional skills for students in Years 8 to 10 (aged 12 to 15).

Schools had to provide five hours or two units, minimum, per year. Units were as follows with lesson numbers indicated (each lesson was one hour).

Year 8

- Classroom connections x 6
- Belonging x 2
- What if…? X 5
- Ups and downs x 4
- What's going on here? X 4
- Expectations x 5

Year 9

- Positive climate x 4
- Skills x 4
- Goal setting x 5
- Universal code I x 4
- Universal code II x 4

Year 10

Schools could draw on any previous, unused units or deliver any of the following ten-minute 'work-out' sessions:

- Anxiety
- Positive thinking
- Life's ups and downs
- Helpful thinking
- Feeling down
- Managing expectations
- Things adults say
- Success and failure
- Self-control
- Moral code
- Avoiding conflict
- Resolving conflict
- Striving for excellence
- Peer pressure
- External pressure
- Internal pressure
- Verbal put downs
- Dealing with pressure

Procedure

All staff in schools undertake two and a half hours of training in restorative practises during the first year of the intervention. This training was provided by trainers accredited by the U.K.'s Restorative Justice Council. Five to ten staff members at each school also received in-depth training in restorative practice and implemented this in the form of restorative meetings and conferences. This additional in-depth training was from certified providers and takes place over three days.

Schools develop action groups, which consist of a minimum of six students and six staff members (including at least one member of the senior leadership team, teaching, pastoral, and support staff). Schools were encouraged to select a diversity of students, including those with a history of misbehaviour or who struggled academically. Two action groups should take place in each term with a total of six meetings in each school year. Action groups are intended to derive an action plan to organise the delivery of INCLUSIVE. Elements should include:

- reviewing and amending related existing school policies—those pertaining to discipline, behaviour management, and staff-student communications—so that they incorporate restorative practises;
- implementing restorative practises throughout the school to prevent, and respond to, bullying and aggression;
- additional tailored actions to address local priorities; and
- delivering the social and emotional skills curriculum for Years 8 to 10.

Action groups review the report on student needs to inform decisions during action group meetings. Action groups ensure that implementation is appropriate for students at the local level and enable local tailoring of the intervention. In the first two years of the intervention, these action groups were externally facilitated; in the final year there was no external facilitator.

Schools delivered classroom-based social and emotional skills education in 'stand-alone' lessons, for example, personal, social, and health education (PSHE) lessons, or integrated it into tutor time or various subject lessons (for example, English) to students in the trial cohort as they moved through Years 8 to 10 (aged 12 to 15 years). They received five to ten hours of teaching and learning across the school year on restorative practices, relationships, and social and emotional skills based on the Gatehouse Project curriculum.

Schools selected units for each year from 'establishing respectful relationships in the classroom and the wider school', 'managing emotions', 'understanding and building trusting relationships', 'exploring others' needs and avoiding conflict', and 'maintaining and repairing relationships'.

Restorative practices are also delivered by school staff throughout the course of the intervention. Primary restorative practices include the use of respectful language when challenging or supporting behaviour and 'circle time' to build relationships. Secondary restorative practices involve staff carrying out restorative 'conferences' to address more serious behavioural issues.

Implementers and mode of delivery

External trainers, who are members of the U.K.'s Restorative Justice Council, provided training to staff members to ensure teachers understood the necessary skills to engage in restorative practice. Action group meetings comprised at least six students and six staff and were led by a member of the senior leadership team in the first two years of the intervention and by external facilitators and in the final year. The social and emotional skills curriculum was delivered by school staff, guided by externally provided lesson plans.

Tailoring

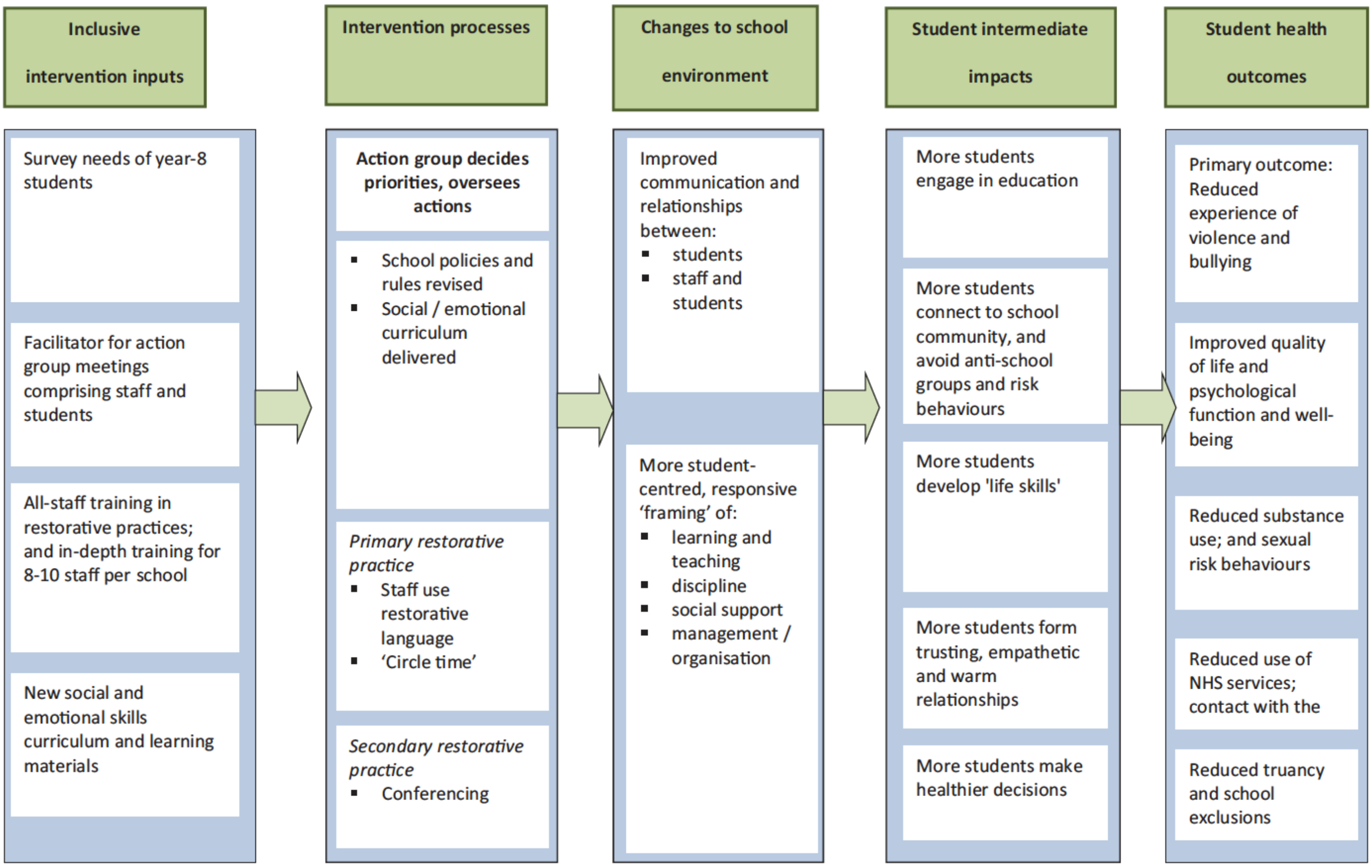
INCLUSIVE enables local tailoring, as informed by the action group meeting. Action groups ensure that implementation of the intervention is suitable at the local level in each school, for example, regarding the revisions to policies and rules, deciding which units of the social and emotional skills curriculum to deliver in each year, and implementing decisions from action group meetings to improve relationships and student participation—for example, student peer mentors or providing restorative practice training to staff who had not attended as part of the intervention.

Fidelity

Intervention fidelity is considered viable if schools complete the follow actions:

- at least five members of staff complete in-depth training;
- six action group meetings are held per year;
- policies and rules are reviewed;
- locally decided actions are implemented;
- group members assess that action groups had a good, or very good, range of members;
- members assess that action groups are led well or very well;
- schools deliver at least five hours, or two units, of the social and emotional skills curriculum each year; and
- at least 85% of staff report that if there was trouble at school, staff respond by talking to those involved to help them get on better.

Figure 1: Logic model for the INCLUSIVE intervention (reproduced from Bonell et al., 2018)


Inclusive intervention inputs
Survey needs of year-8 students
Facilitator for action group meetings comprising staff and students
All-staff training in restorative practices; and in-depth training for 8-10 staff per school
New social and emotional skills curriculum and learning materials
Intervention processes
Action group decides priorities, oversees actions
▪ School policies and rules revised
▪ Social / emotional curriculum delivered
Primary restorative practice
▪ Staff use restorative language
▪ 'Circle time'
Secondary restorative practice
▪ Conferencing
Changes to school environment
Improved communication and relationships between:
▪ students
▪ staff and students
More student-centred, responsive 'framing' of:
▪ learning and teaching
▪ discipline
▪ social support
▪ management / organisation
Student intermediate impacts
More students engage in education
More students connect to school community, and avoid anti-school groups and risk behaviours
More students develop 'life skills'
More students form trusting, empathetic and warm relationships
More students make healthier decisions
Student health outcomes
Primary outcome: Reduced experience of violence and bullying
Improved quality of life and psychological function and well-being
Reduced substance use; and sexual risk behaviours
Reduced use of NHS services; contact with the
Reduced truancy and school exclusions

Evaluation objectives

The aim of this supplementary evaluation was to assess the impact of the INCLUSIVE intervention on the distal outcome of academic attainment in participating schools. Specifically:

- Does INCLUSIVE produce effects on attainment that are comparable with those of existing SEL programmes—that is, an effect size (ES) of 0.46 or larger (following Sklad et al., 2012)?
- Does INCLUSIVE produce positive effects on attainment that are 'meaningful—that is, an ES of 0.4 or larger (following Hattie, 2009)?

In respect of the primary aims of the current study, it is important to note that improvements in academic attainment are not included in the original logic model for the intervention as specified by Bonell and colleagues (2018) and shown in Figure 1. This had implications for assessing adequate power for this supplementary evaluation as the data drawn from the trial was powered to detect more proximal health outcomes, allowing for a much larger cluster effect and overall ES. As a result, the current evaluation picked established empirical benchmarks (Sklad et al, 2005; Hattie, 2009) in order to consider effect size rather than rely on statistical significance as an indicator of impact.

In addition, subgroup effects were examined for (1) children eligible for free school meals, (2) pupils identifying as being bullied at baseline, and (3) gender differences. These subgroups were included following the brief by the EEF to consider differential impact in respect of FSM eligibility and the identification of significant findings from the UCL/LSHTM trial—notably reductions in bullying behaviour and gender differences in impact (Bonell et al., 2018; Bonell, Allen, Warren, et al., 2019) as specified in the statistical analysis plan.

An implementation and process evaluation (IPE) conducted by the implementers that examined different aspects of implementation (trial context and fidelity) took place throughout the main trial period. Full details can be found in Bonell et al. (2018).

Trial context was assessed in intervention and control schools, including discipline systems, staff training, social and emotional skills curricula, and student participation in decision-making. This drew on interviews with intervention facilitators and trainers, members of action groups in intervention schools, staff on school senior leadership teams (SLTs), and other staff in the intervention and control arms; there were also focus group discussions with students and staff in schools selected as case studies.

Trial arm fidelity was assessed using:

- follow-up surveys with staff and students;
- structured researcher observation of action group meetings and staff training;
- surveys of adults leading curriculum implementation and implementing restorative practice;
- interviews with adults delivering the curriculum;
- structured diaries kept by facilitators of action group meetings and by trainers of all-staff training; and
- administrative documents such as minutes and attendance sheets.

Fidelity was scored out of eight points (see Fidelity section above).

In brief, the process evaluation involved interviews with an SLT member and two other members of staff (first two intervention years) and an SLT member (third year) in control schools. These interviews focused on practises and services pertaining to bullying, discipline, and social and emotional skills education in control schools. Data pertaining to implementation was to form part of the independent evaluation in respect to fidelity (as above and reported here), however, difficulties in data transfer between institutions meant this data was not available for analysis (see 'changes to statistical analysis plan' below).

The independent evaluation protocol and SAP can be found on the EEF website.

Ethics and trial registration

The INCLUSIVE trial and collection of data therein was approved by the UCL Ethics Committee (ref 5248/001). This approval, alongside the fact that the University of Manchester (UoM) was not conducting any primary research, served as ethical approval for both institutions. Written, informed consent was obtained at school level (headteacher) for random allocation and intervention and at the individual pupil, staff, and intervention-facilitator level for data collection. Information sheets and consent forms for student surveys were identical in intervention and control schools and did not refer to the intervention.

As part of pupil-level consent, written consent was obtained from each student for linkage of data to the NPD. Note that consent for participation in the survey was obtained in each of the three waves; consent for linkage to NPD was obtained in wave 2 (Year 9).

The trial registration number for the health-related outcomes is ISRCTN10751359 (UCL/LSHTM trial).

Data protection

In respect of data protection, the General Data Protection Regulation (GDPR) came into effect post data collection (2016). In order to fulfil our legal requirements as far as feasibly possible, we sent a privacy notice detailing data rights to all participating schools requesting dissemination to the relevant cohort of pupils. Limitations on the data with respect to presenting participants their rights under GDPR after data collection means that no subsequent data archiving was possible for this data.

Data was processed under Section 537A of the Education Act (1996) which permits the sharing of NPD data with third parties for the purpose of 'promoting the education or well-being of children in England are conducting research or analysis, producing statistics or providing information, advice or guidance'. In respect of the GDPR, the data was processed under the legal basis of 'public task' as the project was intended to improve increasing current knowledge about how social and emotional skills relate to academic outcomes and whether this relationship varies as a function of child characteristic (for example, gender) with a view to improving child outcomes. In order to do so, processing of special categories of data was necessary for archiving, scientific, historical research or statistical purposes (Article 9 (2) (j)).

Project teams

University of Manchester evaluation team

Michael Wigelsworth: principal investigator;

Emma Thornton: research associate and main analyst;

Patricio Troncoso: specialist in trial design and analysis, author of the SAP;

Neil Humphrey: co-investigator; and

Louise Black: research associate.

Developer team

University College London

Jennifer McGowan, Leonardo Bevilacqua, Farah Jamal, Meg Wiggins, Anne Mathiot, Grace West, Deborah Christie, and Russell M. Viner.

The London School of Hygiene and Tropical Medicine

Chris Bonell, Elizabeth Allen, Emily Warren, Zia Sadique, Rosa Legood, Charles Opondo, Joanna Sturgess, Sara Paparini, Tara Tancred, and Diana Elbourne.

In association with:

Adam Fletcher, Cardiff University;

Stephen Scott, King's College London;

Lyndal Bond, Victoria University; and

Miranda Perry (freelance).

Methods

Trial design

The current report details the independent evaluation of secondary data analysis following the original INCLUSIVE trial data to assess the efficacy of the intervention for educational outcomes.

The original INCLUSIVE trial was a three-year, stratified cluster RCT with two arms and schools as the unit of randomisation. A total of 40 schools in the southeast of England were randomised to either the intervention or control arms. All students in the school who were at the end of Year 7 (aged 11 to 12) at baseline (wave 1) were included. Two follow-ups took place, the first at 24 months (when pupils were at the end of Year 9—wave 2) and the second at 36 months (when pupils were at the end of Year 10—wave 3). Schools in the intervention group received the INCLUSIVE programme over three years between 2014 and 2017 while schools in the control group continued with their usual practice for the duration of the trial, receiving no extra input. Control schools were provided with £500 for any costs associated with taking part in the trial (for example, administrative costs). The contract signed by headteachers in the control group committed the school to not taking part in any similar whole-school interventions during the trial period. Further particulars for the trial design can be seen in Table 3.

Table 3: Trial design

Trial design, including number of arms		Cluster randomized controlled trial, two arms, over 3 years
Unit of randomisation		School
Stratification variable (s) (if applicable)		Single sex vs. mixed sex school (dichotomous categorical) School level deprivation as measured by percentage of students eligible for EVERFSM (low/moderate 0 to 23%; high >23%, with 23% being the median for England) School contextual value-added attainment (CVA) in GCSE exams (above and below median for England of 1,000)
Primary educational outcome	Variable	Pupil attainment (Attainment 8 score (KS4_ATT8))
	Measure (instrument, scale, source)	Attainment 8 raw score for July 2018 examinations, as provided by the National Pupil Database. Range: 0-92
Secondary educational outcome(s)	Variable(s)	Mathematics GCSE, English GCSE
	Measure(s) (instrument, scale, source)	MATHS GCSE results for July 2018 examinations (as provided by the NPD (KS4_APMAT); Range: 0-9 ENGLISH GCSE results for July 2018 examinations (as provided by the NPD (KS4_APENG); Range: 0-9
Baseline for primary outcome	Variable	KS2 Attainment (Reading and Mathematics)
	Measure (instrument, scale, source)	KS2_KS2READAPS (from 12/13 school year) Range: 0-39 KS2_KS2MATPS (from 12/13 school year) Range: 0-39
Baseline for secondary outcome(s)	Variable	KS2 Attainment in Reading and Mathematics (average point score)
	Measure (instrument, scale, source)	KS2_KS2READAPS (from 12/13 school year) Range: 0-39 KS2_KS2MATPS (from 12/13 school year) Range: 0-39

Changes to the statistical analysis plan

Due to concerns around the integrity and completeness of the secondary outcome data (specifically, teacher-estimated Key Stage 4 attainment), 'stop/go criteria' were introduced. These concerns were as a result of a change in the mandatory return of teacher estimates of KS4 attainment during the course of the trial. Prior to the change, these were used as an interim measure of impact (as they would have been collected ahead of national testing). Instead, changes in government policy made return of this data voluntary. Stop/go criteria stipulated that if returns of teacher-estimated KS4 scores did not meet the minimal criteria (maximum 25% of missing pupil-level data), then no inferential analyses involving this variable would take place (see SAP). Returns did not exceed 20% (more than 80% of pupil data was missing) meaning that this data was omitted from the analysis (as per the revised SAP).

The SAP stated that all analyses would be conducted in MLwiN. However, this decision was made prior to the management of the NPD being taken over by the Office for National Statistics (ONS), which implemented a secure environment for data analysis, limiting access to certain software. This, combined with staff changes with respect to the main analyst who was familiar with the R environment available within the ONS secure environment, meant that all analyses were conducted in Rstudio (RStudio Team, 2022).

The SAP refers to a series of planned analyses investigating the moderating influence of fidelity of implementation on intervention outcomes. A statistical model utilising Complier Average Causal Effect estimation (CACE) was intended utilising implementation data from the original trial conducted by UCL and the LSHTM. In order to fulfil requirements for secure data transmission and to produce a dataset eligible for analysis (that is, combing primary data from UCL with NPD attainment data), a data-sharing protocol was established between the ONS, UoM, and UCL. Data was provided by UCL to the NPD, which would conduct data matching (utilising variables that the NPD would need, but not UoM, for example, child DOB) and return an appropriate dataset to UoM. As unique variables that would have enabled more robust data matching—such as Unique Pupil Reference—were not part of UCL's original trial data, fuzzy matching was employed by the NPD instead. This took the form of matching UCL and NPD records through matched variables, specifically, school name, pupil name, and pupil data of birth.

Unfortunately, the resultant dataset provided to UoM was not adequate. The trial design, though intended to be blinded to condition, was provided unblinded. Further, errors in the stages of data sharing meant that implementation data was not accurate—with information missing for INCLUSIVE schools and present for control schools. This meant that UoM did not have reliable implementation data. Given the data sanitation between institutions, there was no facility to correct or amend this error and, as a result, CACE analysis could not be completed. The logic model (tested using Structural Equation Modelling; SEM) was conducted on complete cases for those in the intervention condition. We were unable to conduct the SEM on imputed data for the intervention condition (which would have given a sample of 3,341) due to difficulty with the R package (semTools) in pooling model fit statistics across imputed datasets.

Participant selection

Schools were originally recruited by UCL and the LSHTM during the period from March to June 2014 from Greater London and the surrounding counties (Surrey, Kent, Essex, Hertfordshire, Buckinghamshire, and Berkshire). Approximately 500 schools were approached, initially by letter and email, and with a telephone follow-up. The final sample of 40 schools agreeing to participate were reported not to differ from 450 non-recruited schools in terms of school size, population, deprivation, student attainment or value-added education. However, participating schools were more likely to have an Ofsted rating of 'good' or 'outstanding'.

All pupils from participating schools who were at the end of Year 7 (aged 11 to 12) at baseline were included in the trial; there was no exclusion criteria for pupils. At the school level, schools had to be mainstream secondary schools that were in the state education system. For a school to be eligible, their latest Ofsted quality rating was required to be 'requires improvement', 'satisfactory', 'good', or 'outstanding'. Schools that had received an 'inadequate/poor' rating were excluded due to the special measures imposed on these schools, which may have negatively impacted the delivery of INCLUSIVE. Private schools, special educational needs schools, and pupil referral units were also excluded. Eligible schools from Greater London, Surrey, Kent, Essex, Hertfordshire, Buckinghamshire, and Berkshire were identified and contacted between March and June 2014. In total, 40 schools took part in the INCLUSIVE trial.

Outcome measures

Baseline measures

KS2 attainment for the 2012/2013 school year was used as a baseline measure; precise variables are Key Stage 2 reading (KS2_KS2READAPS) and KS2 mathematics attainment (KS2_KS2MATPS), as provided by the NPD.

Primary outcome

The primary outcome of this independent evaluation is pupil attainment at GCSE, specifically, KS4 Attainment 8 raw scores (KS4_ATT8) for July 2018 GCSE examinations, as provided by the NPD. [2]

Secondary outcomes

Secondary outcomes were attainment in the English (KS4_APENG) and maths (KS4_APMAT) GCSE examinations in July 2018, as provided by the NPD.

Bullying

'Bullying scores' were used to create subgroups. In the UCL/LSHTM trial, the Gatehouse Bullying Scale was used as a measure of self-reported bullying victimisation. It contains 12 items relating to the domains of (1) teasing or name-calling, (2) rumour spreading, (3) being left out or excluded from things, and (4) being threatened physically or physically hurt by another student (Bond et al., 2007). When completing this measure, students are asked to reflect on the previous three months and report the frequency of such bullying and how upset they were as a result. A summative score is calculated to give a total bullying score for each domain, ranging 0 to 3. In the INCLUSIVE dataset, total scores for each domain were summed to give an overall Gatehouse Bullying Score, which had a range of 0 to 12 (Bonell et al., 2018). Pupils completed this bullying measure at baseline (wave 1), at the second follow up when they were at the end of Year 9 (wave 2), and in the third follow up when they were at the end of Year 10 (wave 3).

In this evaluation report, we took this overall Gatehouse Bullying Score and divided it by four to represent better the original four subscales of severity, resulting in a score range of 0 to 3. This score was used to derive a categorical variable to indicate the extent of bullying among pupils. Due to small cell counts in the highest category of scoring, the range was collapsed into three categories: '0', not bullied (score of 0); '1', bullied but not frequent and not upset (score of 1); and '2', bullied frequently and/or upset (score of 2).

Sample size

The UCL/LSHTM INCLUSIVE trial (see Bonell et al., 2018; Bonell, Allen, Warren, et al., 2019) was adequately powered to detect effects on the primary outcome measures for which it was originally designed (measures of bullying and aggression). However, measures of academic attainment presented a quandary. First, the estimated intra-cluster correlation (ICC) for attainment in secondary schools (approximately 0.21)[3] is much larger than for aggression or bullying (which had been specified at 0.04). Second, the expected effect size (ES) for attainment was to be treated conservatively, estimated as smaller than proximal outcomes (that is, aggression or bullying, which had been specified at approximately 0.22) given its role as an indirect, distal outcome of the main intervention.

Given this, the current evaluation was powered thus:

Assuming N = 190 per cluster, 40 clusters, *ICC=0.21,* pre-post-test correlation*=0.5,* power*=0.8 and* alpha*=0.05, an ES of 0.41 or larger would be detectable (using* G**Power (Faul, Erdfelder, Buchner, & Lang, 2007)).*

This ES is a useful benchmark as it corresponds directly to Hattie's (2009) 'hinge point' of ES = 0.4, at which, 'the effects of innovation enhance achievement in such a way that we can notice real-world differences' (p.17).

[2] Attainment 8 is a measure published annually showing the average academic performance of a secondary school. It is calculated by adding together pupils' highest scores across eight government approved school subjects. While these numbers are not made publicly available on a pupil-by-pupil basis, scores taken from across a school year group are averaged to produce a school's overall score.

[3] This figure was calculated using GCSE scores for English and maths in the NPD.

Post-hoc sample size calculations were carried out using an online minimum detectable effect size (MDES) calculator (https://patricio-troncoso.shinyapps.io/mdesapp/; Troncoso, 2020) to determine the MDES at the randomisation and analysis stages. The average cluster size was 166.5 at randomisation and 128.2 at the analysis stage. The ICC for KS4 Attainment 8 (the primary outcome measure) was calculated using the `merTools` package in R (Knowles and Frederick, 2020) and was determined to be 0.22 at randomisation and 0.14 at the analysis stage. Full details of the parameters used in MDES calculations, and the MDES at each stage, can be found in Table 5.

Randomisation

Schools were the unit of randomisation and were randomly allocated (1:1) to either the intervention or control arms. Randomisation was stratified by:

- single sex versus mixed sex school (dichotomous categorical);
- school-level deprivation as measured by percentage of students eligible for EVERFSM (low/moderate: 0% to 23%; high: more than 23%, with 23% being the median for England); and
- school contextual value-added attainment (CVA) in GCSE exams (above and below median for England of 1,000). *

* Value added (VA) score is a school-level measure of students' attainment in public exams adjusting for their attainment on entry to the school. VA rather than Ofsted ratings for schools was used as there is better evidence for VA being associated with violence rates (Tobler et al., 2011).

The random allocation sequence was generated by the Clinical Trials Unit at the London School of Hygiene and Tropical Medicine using the `ralloc` command in Stata. This was concealed from schools and the wider evaluation and intervention teams. The research team was made aware of the allocation of each school to either the intervention or control arm after the baseline surveys were completed; the team then passed this information to schools and the intervention team. It was not possible for schools, the intervention team, or process evaluators to be blinded to arm allocation. Of the main trial team, fieldwork staff, the outcome research team lead, and all staff who entered and analysed data were blinded to condition. Further, the original intention was for researchers at Manchester University to also conduct the analysis blinded to condition. However, upon receipt of data, condition was non-blinded (see 'changes to statistical analysis plan').

Statistical analysis

Primary analysis

An intention to treat (ITT) analysis—including all pupils in the groups to which they were randomised irrespective of noncompliance (Gupta, 2011)—was conducted for the primary outcome variable (GCSE KS4 Attainment 8 raw scores). A two-level (school, pupil), hierarchical, multilevel model was estimated (random effects at the school level, utilising robust standard errors) to account for the nested nature of data using the `lme4` package in Rstudio (Bates et al., 2015). The primary outcome variable was standardised KS4 Attainment 8 raw scores (that is, converted to *z* scores) of the i-th pupil in the j-th school, and the model included the following covariates: Key Stage 2 scores (reading and maths) to adjust for prior attainment, group allocation at the school level, and variables which were used in the design (the randomisation factors shown above).

The model has the following algebraic form:

$$y_{ij} = \beta_{0ij} + \beta_1 group2_{0j} + \beta_2 KS2_{ij} + \beta\ mixed + \beta\ \mathrm{hi}g\mathrm{h}depriv + \beta\ belowCVA$$

Secondary analysis

The above model was replicated twice, once with KS4 English and once with KS4 maths scores, to determine the influence of the INCLUSIVE intervention on these secondary outcome variables.

Missing data analysis

Differences between complete and missing cases were examined to establish any pattern to the missingness. Logistic regression was used to predict missingness, whereby each child was coded as providing complete (0) or incomplete (1) outcome data, with treatment allocation, pre-test data, and demographic variables as explanatory variables. At the pupil level, KS2 maths scores and FSM eligibility were predictors of missingness in the KS4 Attainment 8 variable (see Appendix D). Multiple imputation was used as a technique to account for missing data, having established minimal threat to bias. This technique provided as large an analytical sample size as possible. This was especially important given that the nature and design of the evaluation meant statistical power was a concern—the trial by UCL and the LSHTM was powered to detected proximal effects on behaviour rather than educational attainment.

We performed a complete case analysis and a sensitivity analysis using multiple imputation, using the `mitml` package in Rstudio. The `jomoImpute` function was used to implement multiple imputation under the missing at random assumption (Grund et al., 2021). This enabled us to include both partially and completely observed cases of all schools and pupils in the analysis, thereby reducing the bias associated with attrition.

The imputation model was built using the logistic regression procedure described above and bearing in mind the main model of interest (primary ITT analysis). Therefore, the variables that were included in the imputation models were: treatment allocation, demographic variables (specifically, gender and FSM eligibility), prior attainment (KS2 scores), and the outcome variable (for example, KS4 scores). These variables were entered as auxiliary variables and used to impute missing values. Following general guidelines about multilevel multiple imputation (Carpenter et al., 2011), `jomoImpute` was set to run for 5,000 iterations, with a burn-in period of 500 iterations. We stored ten imputed datasets, allowing for 500 iterations to run between them, to ensure that they were independent. Results from the imputed datasets were pooled in Rstudio to obtain the final model coefficients. Results using complete-case analysis were subsequently compared to the results using the multiply-imputed (MI) datasets (see Appendices F and G). We considered that the data was missing at random, meaning minor variation was indicative of sampling error rather than an inherent bias in the dataset (for example, necessarily overestimating or underestimating an effect) and, importantly, we considered the fact that the trial was not otherwise optimally powered for effects. Therefore, MI was considered an acceptable approach for analysis.

Subgroup analyses

A series of analyses were conducted to examine specific subgroup effects. Models for each subgroup were estimated and included the specific variable as a cross-level interaction term (for example, FSM*Allocation group). The subgroups were:

- FSM eligibility (yes/no);
- victims of bullying at baseline: not bullied (score of 0); bullied (score of 1 or 2); and
- sex (male/female).

Additional analyses and robustness checks

Structural equation modelling (SEM) was used to test the logic model (see Figure 1). We explored whether proximal changes in behaviour, specifically bullying, explain distal changes in academic attainment. The proximal behaviours were selected based on findings from the main trial (Bonell et al., 2018), which found INCLUSIVE had a significant impact on bullying but not on aggression. We therefore only consider bullying as a proximal behaviour when testing the logic model. This analysis consisted of a two-level SEM, with pupils nested in schools, conducted using the r package `lavaan` (Rosseel, 2012). This analysis considered only those in intervention schools and consisted of a complete case analysis. Model fit was assessed using the comparative fit index (CFI; Hu and Bentler, 1999), root mean square error of approximation (RMSEA; MacCallum et al., 1996), standardized root mean square residual (SRMR), and Tucker Lewis Index (TLI; Hu and Bentler, 1999). CFI and TLI values greater than 0.9 indicate an acceptable fit; values greater than 0.95 indicate a good model fit (Hu and Bentler, 1999). RMSEA values of 0.01 indicate an excellent model fit, 0.05 indicates a good fit, and 0.08 indicates an acceptable model fit (MacCallum et al., 1996). Finally, SRMR values less than 0.08 are indicative of a good fit (Hu and Bentler, 1999; see Table 11: SEM model fit statistics).

Estimation of effect sizes

In all cases, effect sizes are reported using Hedges' *g* and accompanied by 95% confidence intervals as per EEF specifications. Hedges' *g* was calculated by taking the coefficient of the trial arm allocation variable from a model with covariates and dividing this by the pooled standard deviation for the outcome variable (the square root of pooled pupil level variance (the within-group variance) from an empty model. Confidence intervals for these effect sizes were calculated as the effect size ± 1.96*SE of the trial arm allocation variable.

Estimation of ICC

The ICC was estimated using the `merTools` package in R (Knowles and Frederick, 2020).

Timeline

Table 4: Project timeline

Dates	**Activity**	**Organisation responsible**
INCLUSIVE project delivery		
January–April 2014	Recruitment of schools, fieldworkers and consultants; instrument preparation	UCL, LSHTM
May–June 2014	Baseline (T1) outcome measures	UCL, LSHTM
July–August 2014	Randomisation	UCL, LSHTM
September–October 2014	Facilitated intervention begins; start of efficacy period	UCL, LSHTM
May–June 2016	Interim (T2) outcome measures	UCL, LSHTM
September–October 2016	Non-facilitated intervention begins	UCL, LSHTM
May–June 2017	Final (T3) outcome measures	UCL, LSHTM
Cohort examinations		
March–April 2014	NPD extraction: cohort data	NPD, UoM
January–February 2017	NPD extraction: KS3 data	NPD, UoM
May–June 2017	Year 10 teacher assessment judgements; extraction of teacher assessment judgements	Individual schools
May–June 2018	Cohort sit GCSE examinations	Individual schools
January–February 2019	Began NPD extraction	UoM
Project delays		
May 2018	GDPR introduced, invalidating original data sharing agreement between UoM and UCL/LSHTM	N/A
April 2020	NPD respond to proposed data sharing plan	NPD, UoM
September 2020	UCL/LSHTM provided NPD with data	UCL, LSHTM, NPD
November 2020	Data process by NPD	NPD
November 2020–March 2021	NPD undertook data matching to make data accessible for UoM	NPD
March 2021	RA Staff changes at UoM	UoM
August 2022	Draft report submitted to the EEF	UoM

Following completion of the primary trial, the project encountered a series of delays that led to an amended timescale to that proposed in Table 4. The following difficulties were encountered.

The introduction of GDPR invalidated the original data-sharing plan

The introduction of GDPR in May 2018 and the takeover of the NPD by the ONS invalidated the original data-sharing plan between UoM and UCL/LSHTM. Data sharing was direct between the two partners, but now required additional retooling, including:

- a revised protocol that included the NPD for data matching;
- construction of an ONS safe room; and
- the training of ONS accredited researchers.

There was four-month delay in obtaining the details necessary for the NPD application. There followed a three-month delay while waiting for the DfE to respond to our proposed data sharing plan, taking the project to April 2020.

Statutory changes to Year 10 teacher assessments

The protocol for the current study planned to use the compulsory return of Year 10 teacher assessments to assess interim impact (aligning with the point in the trial when external support was withdrawn from schools, effecting a change from ‘efficacy’ to ‘effectiveness’ conditions), however, this was not possible due to a subsequent change in government policy. During 2017/2018 (after the publication of the study protocol), Year 10 teacher assessments became optional for schools, drastically reducing the amount of data that was available. Stop/go criteria were introduced (see ‘Changes to Statistical Analysis Plan’) and subsequently there was insufficient data to enact this part of the protocol.

Delays in data being entered into to the NPD

The NPD was ready to receive data from UCL/LSHTM in June 2020 as per the amended data-sharing agreement to ensure compliance of new ONS regulation. The NPD did not receive the data until late September 2020. Further delays meant that the data was not ready to be processed by the DfE until November 2020.

NPD data processing lag

After receiving the data from UCL/LSHTM, the ONS had to data-match in order to make the file accessible for UoM. The file was not ready until March 2021.

Research associate moved from project

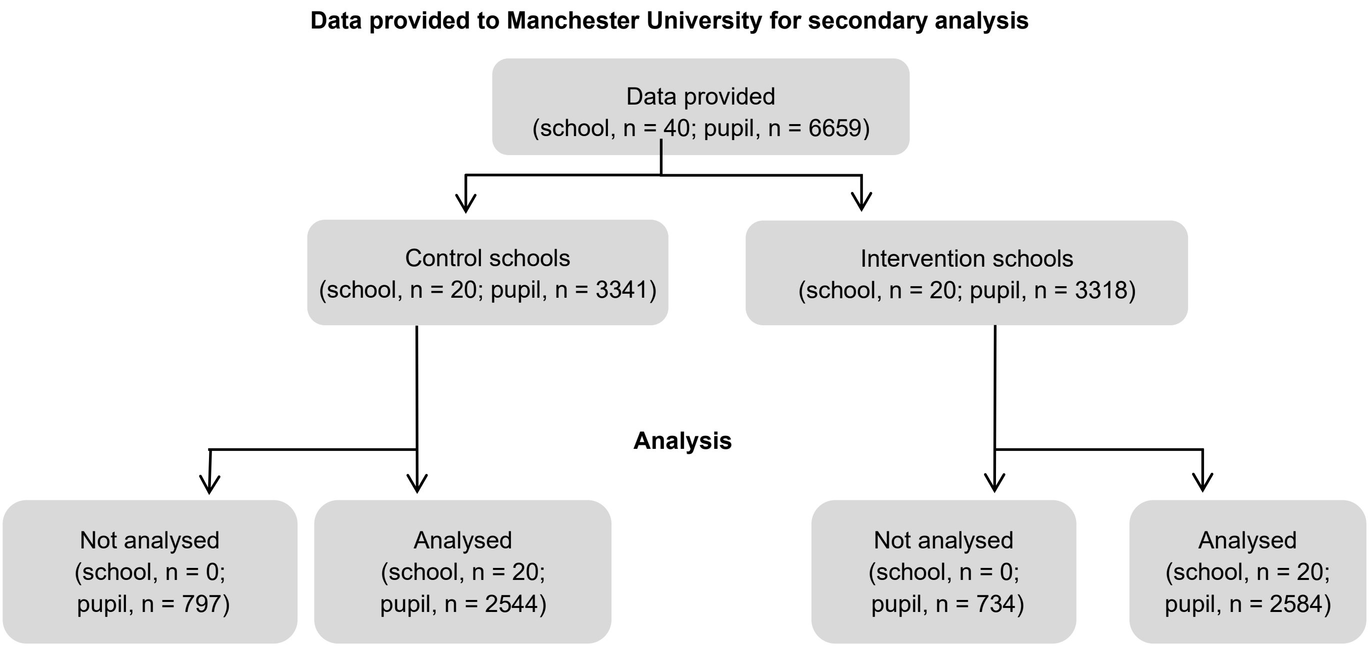
By March 2021, UoM’s research associate had moved institutions and the appointment of a new RA was hampered by the Covid-19 pandemic. Travel restrictions imposed in Portugal in 2021 prevented the new RA being able to take up the vacant position. Although between December 2020 and March 2021 the SRS was developing protocols for allowing international access to data, these were effectively abandoned in April 2021, curtailing our capacity to work on the evaluation.

Impact evaluation

Participant flow including losses and exclusions

Figure 2 shows the flow of participants and schools through the trial. In total, 40 schools (20 intervention; 20 control) took part in the original RCT upon which this evaluation is based. Of the 6,667 pupils that were recruited in the original trial, data from 6,659 was provided to the UoM by UCL/LSHTM. Of these, 3,341 were in the control condition and 3,318 received the INCLUSIVE intervention. The data provided by UCL/LSHTM was matched to the NPD by the ONS on behalf of the Department for Education to provide KS2 and KS4 attainment along with pupil characteristics (sex and FSM eligibility). The matching process was successful for 5,128 pupils (77%: 2,584 intervention, 2,544 control). Missing data in this instance (for NPD variables) was a result of fuzzy matching with the NPD data (see Table 6 for attrition).

Figure 2: Participant flow diagram (two arms)


Data provided to Manchester University for secondary analysis
Data provided (school, n = 40; pupil, n = 6659)
Control schools (school, n = 20; pupil, n = 3341)
Intervention schools (school, n = 20; pupil, n = 3318)
Analysis
Not analysed (school, n = 0; pupil, n = 797)
Analysed (school, n = 20; pupil, n = 2544)
Not analysed (school, n = 0; pupil, n = 734)
Analysed (school, n = 20; pupil, n = 2584)

The discrepancy between the total sample size at baseline for the main trial (N = 6,667) and with the total sample size in this report (N = 6,659) is due to the data provided to the University of Manchester team from the NPD and UCL. The participant flow diagram for the original INCLUSIVE trial can be found in Appendix C.

Table 5: Minimum detectable effect size at different stages

		Randomisation		Analysis	
		Overall	FSM	Overall	FSM
MDES		0.22	0.16	0.19	0.16
Pre-test/post-test correlations	Reading	0.55	0.53	0.55	0.62
	Mathematics	0.63	0.62	0.63	0.53
R^2	Level 1 (pupil)	0.41	0.41	0.41	0.41
	Level 2 (school)	0.71	0.71	0.71	0.71
Intracluster correlations (ICCs)	Level 2 (school)	0.21	0.07	0.14	0.07
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?					
Average cluster size		166.5	41.7	128.2	40.25
Number of schools	Intervention	20	20	20	20
	Control	20	20	20	20
	Total:	40	40	40	40
Number of pupils	Intervention	3318	892	2584	859
	Control	3341	775	2544	751
	Total:	6659	1667	5128	1610

Where it would be typical to report minimum detectable effects at protocol, as this was a secondary analysis of an existing trial dataset, authors of this report did not have access to protocol-stage statistics. The analysis was proposed as an opportunistic sample, based on the resultant recruitment figures from UCL and the LSHTM.

Attrition

Pupil-level attrition is reported in Table 6. For the primary ITT analysis (KS4 Attainment 8 as the outcome variable), data from a total of 5,128 pupils was analysed, with 23% missing data overall. This discrepancy is attributed to difficulties with fuzzy matching between UCL/LSHTM and NPD data, as no compatible unique identifier was present in the UCL/LSHTM dataset. Instead, matching was through school name, pupil forename and surname, and date of birth. Details of missing data in the UCL/LSHTM trial can be found on page 35 of the supplementary materials of Bonell et al. (2018). Overall, attrition is slightly higher in intervention schools compared to control schools across a range of outcomes, however, no formal analyses were conducted so we cannot discuss this here.

Table 6: Pupil-level attrition from the trial (primary outcome)

		Intervention	Control	Total
Number of pupils	Randomised	3,318	3,341	6,659

	Analysed	2,584	2,544	5,128
Pupil attrition (from randomisation to analysis)	Number	734	797	1,531
	Percentage	22.12%	23.86%	23%

Pupil and school characteristics

Descriptive statistics for school- and-pupil level variables at baseline can be found in Table 7. No data was missing for sociodemographic school-level variables, and control and intervention schools were similar in terms of whether they were a mixed or single sex school, the proportion of their pupils eligible for FSM, and the Contextual Value Added (CVA) score of the school. CVA is a school-level measure of students' attainment in public exams adjusting for their attainment on entry to the school. Pupil attainment at KS2 was similarly distributed in the intervention and control conditions, with similar means in each condition (see Table 7 and histograms in Appendix E). Pupils who received the INCLUSIVE intervention had higher average KS4 attainment compared to those in control schools. In respect to information available in order to compare the sample with national averages, the sample group was seen to have a higher than average number of students eligible for free school meals, —approximately twice the national average. There was also an imbalance in gender with less than the average number of males—15% to 20% less than the national average.

Table 7: Baseline characteristics of groups as randomised

School level (categorical)	**National-level mean**	**Intervention group**		**Control group**		**Balance at randomisation (effect size; * 95% CI)**
		n/N (missing)	**Count (%)**	**n/N (missing)**	**Count (%)**	
Single sex/mixed[1]		0	Mixed = 15 (75%) Single sex = 5 (25%)	0	Mixed = 15 (75%) Single sex = 5 (25%)	0* (-0.71; 0.71)
% of pupils eligible for FSM[2]		0	Low/mod = 8 (40%) High = 12 (60%)	0	Low/mod = 7 (35%) High = 13 (65%)	0.11* (-0.59; 0.83)
Contextual value added score[3]		0	Above = 12 (60%) Below = 8 (40%)	0	Above = 12 (60%) Below = 8 (40%)	0* (-0.71; 0.71)
Pupil level (categorical)		**n/N (missing)**	**Count (%)**		**Count (%)**	
Sex	49% Female[4]	654 (19.71%)	Male: 1180 (35.56%) Female: 1484 (44.73%)	766 (22.93%)	Male: 1242 (37.17%) Female: 1333 (39.9%)	0.09* (0.03; 0.15)
Ever FSM	16%[5]	654 (19.71%)	Yes: 1040 (31.34%) No: 1624 (48.95%)	767 (22.96%)	Yes: 914 (27.36%) No: 1660 (49.69%)	0.08* (0.02; 0.15)

Bullying (baseline)		312 (9.4%)	Not Bullied: 1717 (51.75%) Bullied (not frequently/not upset): 999 (30.11%) Bullied (frequently and/or upset): 290 (8.74%)	309 (9.25%)	Not Bullied: 1683 (50.37%) Bullied (not frequently/not upset): 1018 (30.47%) Bullied (frequently and/or upset): 331 (9.91%)	Since this variable is derived from T1 GBS total (continuous), equivalence is demonstrated using the continuous score.
Pupil level (continuous)		**n/N (missing)**	**Mean (SD)**	**n/N (missing)**	**Mean (SD)**	**Effect size**
KS2 Reading (Range: 0-39)	87% of pupils achieved Level 4 or above in KS2 Reading (2012)[6]	663 (19.98%)	29.2 (4.83)	749 (22.42%)	28.7 (4.76)	-0.09 (-0.15; -0.03)
KS2 Maths (Range: 0-39)	84% of pupils achieved Level 4 or above in KS2 maths (2012)[6]	663 (19.98%)	29.7 (5.54)	749 (22.42%)	29.0 (5.14)	-0.14 (-0.19; -0.09)
KS4 Attainment 8 (Range: 0–92)	44.5[7]	620 (18.69%)	52.9 (21.0)	730 (21.85%)	48.2 (19.5)	NA at randomisation
KS4 English (Range: 0-9)	70% achieved Grade 4/ C or above in GCSE English (2018)[8]	695 (20.95%)	5.24 (1.87)	789 (23.62%)	4.88 (1.79)	NA at randomisation
KS4 Maths (Range: 0-9)	71% achieved Grade 4/C or above in GCSE maths (2018)[8]	689 (20.77%)	5.25 (2.23)	790 (23.65%)	4.75 (2.06)	NA at randomisation
T1 GBS total9 (Range: 0-12)		312 (9.40%)	1.9(2.4)	309 (9.25%)	2.03 (2.52)	0.05 [0; 0.01]
T2 GBS total (Range: 0-12)		914 (27.55%)	1.45(2.06)	800 (23.94%)	1.64 (2.22)	NA at randomisation

T3 GBS total (Range: 0-12)		1208 (36.41%)	1.16(1.87)	1048 (31.37%)	1.32 (2.01)	NA at randomisation

[1] Single-sex schools (boys and girls) have been aggregated for SDC purposes.
[2] Low/moderate: 0%–23% of pupils eligible; high: >23% of pupils eligible.
[3] Score above or below 1,000 (the median score for England).
[4] Proportion of pupils aged 11–12 in state funded schools in 2014 (DfE, 2014).
[5] 16% of pupils in state funded secondary schools eligible for and claiming free school meals (DfE, 2014).
[6] Due to availability of national data, national KS2 attainment reported as percentage achieving level 4 or above (DfE, 2012).
[7] DfE, 2018c.
[8] https://www.gov.uk/government/news/guide-to-gcse-results-for-england-2018
[9] GBS: Gatehouse bullying scale.

* For continuous variables, Hedges' *g* has been calculated directly using measures of standard deviation. In the instance of categorical variables, comparable effects sizes have been calculated using formula provided by Chinn (2000).

Outcomes and analysis

Primary analysis

As can be seen from Table 8, an intervention effect was apparent for the primary outcome: those who were allocated to the intervention had better overall attainment at GCSE (KS4 Attainment 8 scores; ES: 0.14, 95% CI: 0.05; 0.23). Results from missing data analyses can be found in Table 3, Appendix F. These analyses revealed that for the ITT analyses, missing data did not impact the results: similar effect sizes were reported for the same analysis conducted on imputed data. Minor differences were identified between the ITT and MI datasets, notably the MI set showed a reduced trial effect for KS2 Attainment 8 by 0.01. Although overall there were identifiable differences, these were minor.

Secondary analyses

This intervention effect was also present for KS4 English attainment (ES: 0.13, 95% CI: 0.01; 0.25), however, a smaller effect was found for KS4 maths attainment (ES: 0.09, 95% CI: 0.00; 0.19). Full results for these ITT models can be found in Table 2, Appendix F. Results from missing data analyses can be found in Table 3, Appendix F. These analyses revealed that for the ITT analyses, missing data did not impact the results: similar effect sizes were reported for the same analysis conducted on imputed data. Minor differences were identified between the ITT and MI datasets, notably the MI set showed a 0.02 reduction in English. Although overall there were identifiable differences, these were minor.

Subgroup analyses

We investigated whether the INCLUSIVE intervention was more effective for some groups than others. Specifically, we looked at FSM eligibility, sex, and bullying victimisation in three subgroup analyses (see Table 9). A subgroup*trial arm interaction term was included in the original ITT model. If a significant interaction term is observed, this would indicate subgroup effects whereby the intervention was more effective for that specific subgroup. However, findings revealed no significant subgroup effects (FSM subgroup ES: 0 95% CI: -0.09; 0.08; gender subgroup ES: -0.01, 95% CI: -0.01; 0.07; bullying subgroup ES: 0.02, 95% CI: -0.07; 0.10), suggesting the INCLUSIVE intervention is not particularly effective for any specific subgroup, rather all participants appear to benefit similarly from the intervention. We also conducted the subgroup analyses on ten imputed datasets and the pooled results from these analyses revealed the same pattern of results, indicating no subgroup effects (see Tables 4 to 9 in Appendix G). Reductions in effect between 0.01 and 0.02 were also shown in subgroup differences for bullying. There was no identifiable difference between ITT and MI data for the FSM subgroup analysis. Although overall there were identifiable differences, these were minor.

Table 8: Intention to Treat analyses and parameters used to calculate effect sizes—primary outcome

	Unadjusted means				**Effect size**		
	Intervention group		**Control group**				
Outcome	**n (missing)**	**Mean (95% CI)**	**n (missing)**	**Mean (95% CI)**	**Total n (intervention; control)**	**Hedges' *g* (95% CI)**	**p-value**
KS4 Attainment 8	2584 (734)	52.93 (52.13; 53.73)	2544 (797)	48.25 (47.49; 49.01)	5128 (2584; 2544)	0.14 (0.05; 0.23)	0.004**

* Significant at < 0.05; ** significant at < 0.01.

Table 9: Intention to Treat analyses and parameters used to calculate effect sizes—secondary outcomes

	Unadjusted means				**Effect size**		
	Intervention group		**Control group**				
Outcome	**n (missing)**	**Mean (95% CI)**	**n (missing)**	**Mean (95% CI)**	**Total n (intervention; control)**	**Hedges' *g* (95% CI)**	**p-value**
KS4 English	2519 (799)	5.24 (5.16; 5.31)	2487 (854)	4.89 (4.82; 4.96)	5006 (2519; 2487)	0.13 (0.01; 0.25)	0.039 *
KS4 maths	2526 (792)	5.23 (5.15; 5.32)	2486 (8550	4.76 (4.68; 4.84)	5012 (2526; 2486)	0.09 (0.0; 0.19)	0.059

* Significant at < 0.05; ** significant at < 0.01.

Table 10: Subgroup analyses

Unadjusted means					**Effect size**		
	Intervention group		**Control group**				
Outcome	**n (missing)**	**Mean (95% CI)**	**n (missing)**	**Mean (95% CI)**	**Total n (intervention; control)**	**Hedges' *g* (95% CI)**	**p-value**
Free school meals							
Not eligible (reference)							
KS4 Attainment 8	1622 (<10)	58.55 (57.62; 59.47)	1656 (<10)	51.99 (51.12; 52.85)	3278 (<10)		
KS4 English	1607 (17)	5.62 (5.53; 5.71)	1637 (23)	5.11 (5.02; 5.19)	3244 (40)		
KS4 maths	1608 (16)	5.77 (5.67; 5.87)	1640 (20)	5.09 (4.99; 5.18)	3248 (36)		
Eligible							
KS4 Attainment 8	1033 (<10)	45.55 (44.31; 46.79)	909 (<10)	42.94 (41.68; 44.20)	1942 (12)	0.00 (-0.09; 0.08)	0.956

KS4 English	1000 (40)	4.65 (4.53; 4.76)	892 (220)	4.50 (4.38; 4.61)	1892 (260)	0.00 (-0.09; 0.10)	0.962
KS4 maths	1007 (33)	4.42 (4.28; 4.55)	884 (30)	4.18 (4.04; 4.31)	1891 (63)	-0.01 (-0.09; 0.07)	0.825
Gender subgroup							
Male (reference)							
KS4 Attainment 8	1178 (<10)	49.65 (48.47; 50.83)	1238 (<10)	46.81 (45.72; 47.90)	2416 (<10)		
KS4 English	1147 (33)	4.73 (4.63; 4.84)	1213 (29)	4.57 (4.47; 4.67)	2360 (62)		
KS4 Maths	1153 (27)	5.14 (5.01; 5.27)	1212 (30)	4.86 (4.74; 4.98)	2365 (57)		
Female							
KS4 Attainment 8	1477 (<10)	56.55 (55.54; 57.56)	1327 (<10)	50.62 (49.64; 51.60)	2804 (13)	-0.01 (-0.01; 0.07)	0.782
KS4 English	1460 (24)	5.64 (5.55; 5.74)	1316 (17)	5.19 (5.09; 5.29)	2776 (41)	-0.06 (-0.16; 0.03)	0.186
KS4 maths	1462 (22)	5.33 (5.22; 5.45)	1312 (21)	4.69 (4.58; 4.79)	2774 (43)	-0.01 (-0.09; 0.08)	0.870
Bullying							
Not bullied (reference)							
KS4 Attainment 8	1387 (330)	55.36 (54.29; 56.43)	1345 (338)	49.90 (48.86; 50.94)	2732 (668)		
KS4 English	1359 (358)	5.39 (5.29; 5.49)	1317 (366)	4.96 (4.87; 5.06)	2676 (724)		
KS4 maths	1361 (356)	5.47 (5.35; 5.58)	1317 (366)	4.95 (4.84; 5.06)	2678 (722)		
Bullied not frequently and not upset							
KS4 Attainment 8	840 (159)	53.84 (52.47; 55.22)	790 (228)	48.65 (47.33; 49.97)	1630 (387)	0.02 (-0.07; 0.10)	0.740
KS4 English	820 (179)	5.32 (5.19; 5.45)	776 (242)	4.94 (4.81; 5.06)	1596 (421)	-0.02 (-0.12; 0.08)	0.723
KS4 maths	824 (175)	5.33 (5.18; 5.48)	776 (242)	4.77 (4.63; 4.91)	1600 (417)	0.02 (-0.07; 0.11)	0.658
Bullied frequently and/or upset							
KS4 Attainment 8	231 (59)	45.12 (42.28; 47.97)	241 (90)	44.90 (42.37; 47.44)	472 (149)	0.01 (-0.14; 0.15)	0.925
KS4 English	217 (73)	4.81 (4.56; 5.06)	232 (99)	4.71 (4.48; 4.95)	449 (172)	0.02 (-0.14; 0.18)	0.770
KS4 maths	216 (74)	4.55 (4.25; 4.84)	233 (98)	4.42 (4.16; 4.67)	449 (172)	0 (-0.14; 0.14)	0.951

Where the number of missing responses is less than the threshold of 10, this has been replaced with <10, for statistical disclosure control purposes as a result of small ns.

Additional analyses and robustness checks

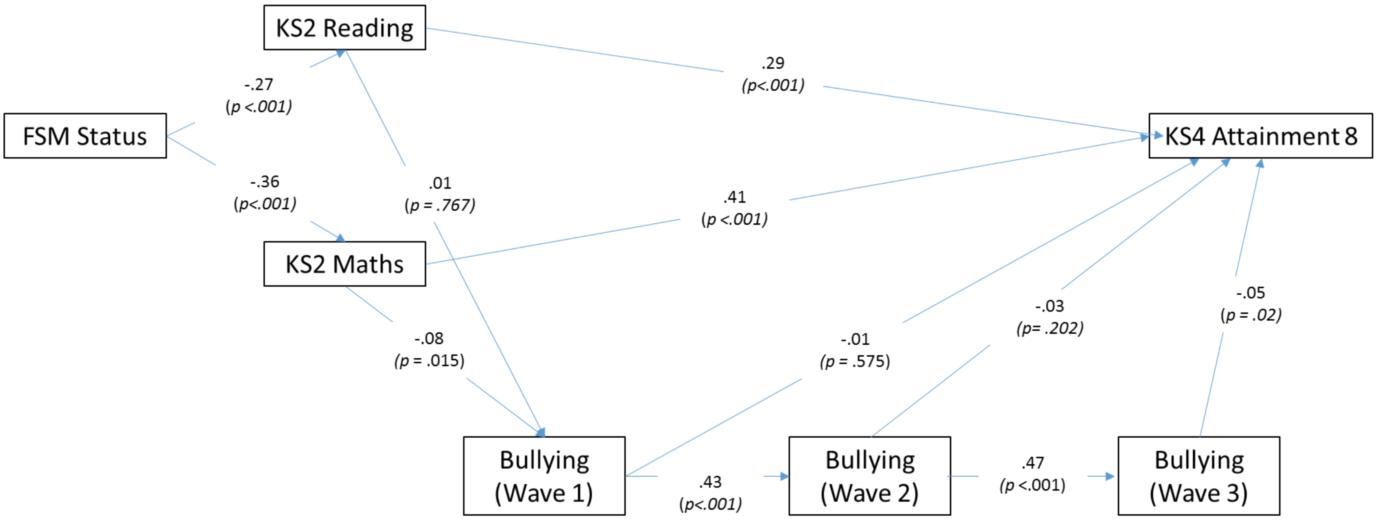
Structural equation modelling (SEM) was used to test the logic model (Figure 1). We explored whether proximal changes in behaviour, specifically bullying, explain distal changes in academic attainment. SEM was conducted to test the logic model and fit indices indicated that the model fit was poor (CFI: 0.72; TLI: 0.59; SRMR: 0.68; RMSEA: 0.13; see Table 11). Coefficients for the paths tested in the SEM can be found in Figure 3. As can be seen from Figure 3, bullying experiences at wave 3 were predictive of KS4 Attainment 8 scores (β = -0.05) with those who were bullied more having poorer KS4 Attainment 8 scores. Although earlier bullying experiences predicted later bullying experiences, bullying at wave 1 and wave 2 were not directly predictive of KS4 Attainment 8 scores. However, any conclusions drawn are spurious and should be interpreted with caution due to the poor model fit to the data.

The logic model indicated that the INCLUSIVE intervention would affect educational outcomes as a result of the impact this intervention has been reported to have on bullying (Bonell et al., 2018). However, the lack of a subgroup effect for those who were victims of bullying suggests that these pupils did not experience a particular benefit of the intervention in terms of their educational attainment compared to those who were not bullied. There was a significant main effect of being a victim of bullying on KS4 Attainment 8 scores such that those who were 'bullied frequently and/or were upset' as a result had poorer KS4 Attainment 8scores compared to those who were 'not bullied at all' or 'bullied but not frequently or upset'. There was also a significant main effect of receiving the intervention compared to control schools such that those who received the intervention had higher KS4 Attainment 8 scores compared to those who continued with usual practice (control schools). However, this intervention effect did not differ as a function of bullying exposure. This is inconsistent with the idea that INCLUSIVE impacts educational attainment through reduced bullying victimisation.

Table 11: SEM model fit statistics

Fit statistic	Threshold	Model results
Comparative fit index (CFI)	Good fit > 0.95, Acceptable fit > 0.9	0.72
Root mean square error of approximation (RMSEA)	Excellent fit = 0.01 Good fit = 0.05 Acceptable fit = 0.08	0.13
Standardized root mean square residual (SRMR)	Good fit < 0.08	0.68
Tucker Lewis Index (TLI)	Good fit > 0.95 Acceptable fit > 0.9	0.59

Figure 3: Structural equation model testing the logic model (shown in Figure 1)



Cost

The average cost of INCLUSIVE for one intervention school was around £50,244, or £58 per pupil per year when averaged over three years. This is an estimate of the costs incurred by schools in the intervention group above the costs that control schools incurred as part of their usual practice dealing with bullying in school. This cost estimates were calculated based on NIHR protocol rather than EEF's cost analysis guidelines so caution should be taken when comparing costs across EEF evaluations. A full costing analysis is available as part of UCL's primary report to the NIHR (Bonell, Allen, Warren, et al., 2019).

Conclusion

Table 12: Key conclusions

Key conclusions
1. Children in INCLUSIVE schools made the equivalent of two months' progress in KS4 Attainment 8 scores, on average, compared to children in other schools. This finding has not been assigned an EEF security rating, although it should be noted that there is uncertainty around the results.
2. There is some evidence that INCLUSIVE had a positive impact on pupils' maths and English GCSE results, with pupils in INCLUSIVE schools making an additional month of progress in maths and an additional two months' progress in English, on average.
3. There was no notable difference in the impact of the programme for pupils eligible for free school meals (FSM) compared to other pupils.
4. While the evaluation led by UCL and the LSHTM found that INCLUSIVE reduced bullying in schools and being bullied has been found to be associated with lower attainment, this report found no evidence that INCLUSIVE had a greater impact on the attainment of pupils who had reported being bullied compared to other pupils.
5. Although the evaluation led by UCL and the LSHTM found that INCLUSIVE had greater effects for boys than girls, there was no evidence in this evaluation that INCLUSIVE impacted children's attainment differently based on gender.

Interpretation

Main programme effects

The current study is first to examine the impact of INCLUSIVE in respect to academic attainment as a distal outcome of the intervention, following the examination of primary behavioural and psycho-social outcomes in the main trial conducted by UCL and the LSHTM (for example, smoking, under-age drinking, bullying, and aggression). The study showed a positive effect in respect to pupils' KS4 attainment scores, consistent with two month's additional progress. The study did not identify any subgroup effects in respect to academic attainment, even though subgroups were drawn from significant findings reported by the main trial conducted by UCL and the LSHTM (specifically, pupils reporting being bullied and gender differences). At first glance, the results suggest that INCLUSIVE is an effective intervention for improving pupil's attainment but not for the reasons hypothesised in respect to the proposed logic model.

With respect to the evaluation objectives of the current study, INCLUSIVE was associated with effects on attainment that are broadly consistent with those of existing SEL programmes. For instance, a number of meta-analyses of universal programmes that aim to address student behaviour and provide social and emotional skills education are seen to produce an average effect size of between 0.19 and 0.28 (Corcoran et al., 2018; Sklad et al., 2012; Wigelsworth et al., 2016). Although the result of the current trial shows INCLUSIVE to produce more conservative effects (an ES of 0.14) when compared to the wider literature, it is worth noting several mitigating factors.

First, it is worth considering the high degree of heterogeneity typically associated with meta-analyses for these types of programme. Impact varies widely, with 95% confidence intervals reported between 0.1 to 0.4 in some instances (Wigelsworth et al., 2016). Some estimates are wider, with Tanner-Smith and colleague's (2018) meta-analytic review indicating a range between 0.01 and 0.34 for universal prevention programmes similar to INCLUSIVE. Further to this point, there are concerns that estimated effects in the literature are inflated as programmes with more rigorous randomised studies with large samples sizes are generally seen to have lower effect sizes (Corcoran et al., 2018). Given that the strengths of the current study include random allocation and comparatively large sample size, it would generally be expected for INCLUSIVE to produce more modest effects in comparison to less well controlled studies that currently inform a portion of the meta-analytic evidence base. Second, although whole-school approaches are seen to be generally effective, they also yield lower effect sizes in comparison to interventions with a focus on providing class-based curricula (Goldberg et al., 2018). On this basis, the reported effect is broadly consistent with the expectations of a programme of this type.

The points above help to critically interpret the second of the current study's objectives, namely whether INCLUSIVE's effects are 'meaningful' within the specific frame of Hattie's (2009) 'hinge point' of an ES of 0.4 or above. Derived from a meta-synthesis of over 1,200 meta-analyses examining influences on school-based attainment, an average effect of 0.4 was found through Hattie's analyses (broadly consistent with one year's progress), cited to be the point at which educational intervention is considered to be effective. This value has often been used to judge the comparative efficacy of approaches in order to critically consider the 'opportunity cost' of intervention (for example, why implement 'A' when 'B' may produce a larger effect?). However, it is important to highlight some significant critique of Hattie's approach as there is a real risk of misinterpretation of the use of the hinge point highlighted in work updated since the publication of Hattie's meta-synthesis (and, indeed, also since publication of the current study's statistical analysis plan).

Kraft (2020) notes important considerations in the critical interpretation of ES including assessing what outcomes are measured. Bespoke instruments (for example, researcher-administered competency assessments) are seen to produce effects two to four times larger than that standardised testing (Lynch et al., 2019). As INCLUSIVE's outcomes were measured through national standardised testing (Key Stage 4 results) we would expect a more conservate interpretation of effect in comparison to studies included in Hattie's meta-synthesis, a portion of which utilised bespoke competency assessments. Kraft also notes concerns regarding the treatment-control contrast. Experiences of the experimental and comparison groups plays an important role in determining ES. In contexts where control groups are otherwise isolated from comparable or alternative intervention effects, effects are seen to be higher in comparison to those where comparison groups are less controlled (Kline and Walters, 2015). There is little in the UCL/LSHTM study design to allow for controllable conditions for those not allocated to the intervention condition, befitting the naturalistic context and nature of INCLUSIVE as a whole-school approach. A limitation to the current study is the relative absence of direct information drawn from schools allocated to the comparison condition (see 'Limitations and Lessons Learned'). However, findings from the UCL/LSHTM trial showed that although most control schools did not report addressing bullying or aggression as a main priority, six did deliver some form of restorative practice or social and emotional skills education after starting the trial and 15 reported using restorative practice as part of school practice. Similar reports show a number of schools utilising student action groups which also form a part of the INCLUSIVE approach. Therefore, there is reasonable evidence that school practices in the control schools were implementing activities similar to INCLUSIVE and some indicative evidence that allocation to the comparison condition itself may have accelerated activity in this area (McMillan, 2007). As such, this is an additional reason to treat an otherwise conservative effect size as a pessimistic (albeit demonstrable) interpretation of effect.

As a result of these considerations, revised estimates drawn from an updated meta-synthesis, notably both Kraft (2020) and Tanner-Smith et al. (2018), suggest effects on attainment as low as 0.05 may be considered 'worthwhile', with effects for INCLUSIVE to be close to the average expectation of impact for a programme of this type. Therefore, benchmarks updated from Hattie's original meta-synthesis suggest that INCLUSIVE may be 'worthwhile' as an intervention. This is independent of considering the relatively low cost of the intervention—UCL's and the LSHTM's costing suggests an implementation cost of £47 to £58 per pupil (Bonell et al., 2018). Notably, recent authors note the importance of individual settings judging their own context in respect of what they want to achieve through intervention. Accordingly, noting that INCLUSIVE is primarily a psycho-social intervention aimed at addressing behaviours and attitudes, with academic attainment being a distal outcome to this activity, is arguably an important factor in deciding whether to implement this programme.

Subgroup effects and programme theory

The current study failed to identify any subgroup effects with specific reference to eligibility for FSM, gender, and those identified as being bullied at baseline. In terms of the former, although socioeconomic disadvantage is an established predictor of both poor mental health and impaired academic attainment (Bradley and Corwyn, 2002), the factors underpinning this relationship are complex. For instance, socioeconomic status (SES) is a proxy of several factors, including access to community resources and exposure to stressors (Hetzner, Johnson and Brooks-Gunn, 2010). There is little direct literature exploring differential treatment effects for this subgroup, especially regarding whole-school social and emotional learning interventions, and this was not explicitly part of the proposed logic model explaining intervention effects. Similarly, differential gender effects are not noted in INCLUSIVE's logic model and, therefore, there was not strong a priori justification for examining this effect, beyond noting improvements for boys in the UCL/LSHTM trial. Conversely, a link between bullying and attainment is noted both in the wider literature base (Brown and Taylor, 2008) and is present in UCL's and the LSHTM's logical model, meaning results are contrary to expectations. However, further examination of INCLUSIVE's approach may provide an explanation for a lack of distal impact for these groups despite UCL and LSHTM finding immediate effects.

As a whole-school approach, INCLUSIVE does not contain targeted elements, instead seeking change at an institutional and staff level (Bonell et al., 2019). Whole-school approaches are not necessarily considered the optimal delivery strategy for addressing at-risk groups unless tiered-type approaches (for example, indicated and targeted) are offered as part of an integrated package. Consistent with INCLUSIVE's logic model, conceptual models of SEL (see, for example, Jennings and Greenberg, 2009) associate improved attitudes about school, self, and others with subsequent reductions in aggressive and bullying-type behaviours, the latter of which is seen to contribute to improved academic performance (CASEL, 2015). However, alternative models suggest that it is the change in teaching practices that allow for a more engaging classroom environment, which in turn impacts upon more performance related skills such as attention and emotional regulation (Duckworth and Yeager, 2015). Such skills are theorised to have a larger impact on academic outcomes in comparison to interpersonal skills such as prosocial behaviour. Given INCLUSIVE's focus on training teachers, it is possible that any indirect improvement in prosocial behaviours explaining a reduction in reported bullying that otherwise leads to improved attainment is otherwise subsumed by a more direct and universal impact of performance-related SEL skills such as the ability to maintain and regulate attention.

Another fundamental aspect of INCLUSIVE's approach, knowledge around the mechanisms of restorative practice (Duckworth and Yeager, 2015), may explain current findings. The aim of restorative approaches is to reduce bullying and aggression through engaging students in shared decision making and, as such, teacher's actions and approaches are a key aspect (Lodi et al., 2021). In discussing restorative practices in schools, Weaver et al. (2020) note that increases in academic attainment may be due to the creation of an equitable, safe, and inclusive classroom climate, meaning students are more able to engage with the learning environment. Such an explanation is consistent with Duckworth and Yeager's (2015) pupil-level skill-building as restorative practice approaches may create a constructive environment for these skills to be effectively taught and practiced. Such an explanation is also consistent with the findings of the empirical logic model as it is not through the reduction of perceived bullying that learning (and therefore attainment) is supported. This explanation is also consistent with the identified impact of gender as it has been theorised that restorative practice techniques may be particularly useful for female adolescents on the basis that a focus on relationships and connections is consistent with gender theories of self-identify and growth (Londi et al., 2022).

Limitations and lessons learned

The current study demonstrates a number of strengths, supporting confidence in the rigour of the findings. The examination of attainment explored in the current study is supported by the rigour of the original trial, namely a large-scale RCT. Original sampling and recruitment efforts by UCL/LSHTM were seen to produce a sample representative of the approximately 500 schools initially approached to participate in the trial. Randomisation was not optimal, with some imbalance between trial arms, however, as, overall, baseline imbalances were small—specifically, KS2 reading (ES: -0.09) and maths (ES: -0.14) and were controlled for in the analytical model—these potentially increase the rigour and security of the findings (EEF, 2019). Accordingly, we do not judge the identification of balance at baseline to have impaired the rigour of the analysis or results.

Ecological validity is established through the use of standardised attainment conducted as part of national assessments. Our principal findings relating to the impact of INCLUSIVE on pupil-level outcomes at the ITT level were partly sensitive to any changes in our modelling parameters, with MI models showing differences of between 0.01 and 0.02 on some parameters (see Appendices F to G).

Limitations to the study pertain to the nature of the approach, namely the difficulties in working with a pre-established trial design that restrict certain parameters. In this instance, the current evaluation was powered to detect proximal impacts on behaviour, which were anticipated to be higher in terms of effect size when compared to the distal outcome of attainment. However, power is only a principal issue in respect to null hypothesis significance testing (NHST; Crutzen and Peters, 2017). However, as the current study did not rely on NHST in interpretation of the impact of INCLUSIVE, instead relying on the interpretation of effect size through comparative benchmarking, this concern is arguably mitigated.

With a post-hoc dataset, there were no opportunities to consider data pertaining to the comparison group. In the original trial, control data was available and was used to exclude five control schools that implemented activities that closely resembled elements of INCLUSIVE (restorative practice, social and emotional skills education, and student participation in decision-making). Although per-protocol analysis from the original trial showed similar intervention effects (suggesting that the control school activities did not make an observable difference to the primary behaviour outcomes), it is not known whether this would be the case for the distal outcome of attainment, especially as there is data to suggest that the intervention effects are not necessarily directly ascribable to the proposed logic model (see above). Although the interpretation of findings with respect to INCLUSIVE are, in effect, 'in comparison to normal practice', we cannot be entirely confident that 'normal practice' did occur, nor be able to define what activity this might or might not entail in future.

A second source of limitations come from a number of external events that led to alterations to the original plans, as detailed above (see Changes to Statistical Analysis Plan). Notably, rigour is impacted by a lack of wider data from which to further examine INCLUSIVE, namely a lack of interim Key Stage 4 teacher estimates and lack of implementation data. With respect to a lack of interim assessment, this prevented an opportunity to consider potential 'wash out' effects. As INCLUSIVE was associated with a positive impact at post-test this proved unnecessary. Though there was arguably an opportunity to consider whether effects changed over time—for instance, diminishing results as the researchers stopped supporting implementation or increasing impact as the intervention effects took hold—the rigour of this analysis would have been mitigated by the variable of choice: teacher assessment. Recent literature has taken note of evidence of systemic divergence between teacher assessment and standardised testing (Lee and Newton, 2021) and therefore any comparison of interim assessment using estimates and a post score derived from standardised testing would have needed to be treated cautiously given the possibility of bias. In a similar vein, further opportunities to explore the impact of INCLUSIVE were restricted by the lack of robust implementation data, meaning a CACE analysis could not be conducted. This omission prevents a consideration as to whether compliance with the intervention protocol was related to variation in impact. In the UCL/LSHTM trial, positive results for behaviour were found despite variability in fidelity to the intervention (Bonell, 2019). Given the principal aim of INCLUSIVE to enable school practice (rather than just the delivery of specific intervention components as per the developer's intent), an examination of fidelity to 'form' (that is, delivery of the teacher training and SEL intervention components) was argued by UCL/LSHTM as less important than the overall fidelity of function (that is, whether the intervention triggered intended change in locally appropriate ways). CACE analysis is not a sufficient tool to examine functional fidelity, and in this way, the proposed mechanism by which implementation was to be considered is now dated given UCL's and LSHTM's subsequent findings (released after the current evaluation began). That said, there arguably remains a significant opportunity to further consider the mechanisms behind organisational change in respect of the implementation of INCLUSIVE.

Future research

Findings suggest that INCLUSIVE is a promising approach that may offer positive impacts on attainment in addition to previously reported improvements in student heath and behaviour (for example, Bonell et al., 2019). There remain, however, further opportunities to peruse key lines of enquiry in order to understand better, and potentially optimise, the impact of INCLUSIVE.

Given our inability to consider implementation, this remains a line of current enquiry. Beyond the omission of CACE as originally planned, further opportunities are also present. Recent publications from the UCL/LSHTM trial offer further insight into optimal modes in doing so, suggesting that CACE analysis, though offering potential insight into functional implementation and any relationship to impact, other methods may be needed to consider form-based implementation. This would involve examining the triggers and subsequent changes in staff behaviours resulting from the delivery of material and training. For instance, attitudinal surveys from staff or comparative qualitative exploration may be needed to capture the intermediate mechanism of a change in values or beliefs that form part of UCL's and the LSHTM's hypothesised mechanism of change. Further to this, as a multi-component intervention, there is further work in examining more closely the 'critical components' that are required for this change to occur. This might feasibly be achieved using factorial based approaches by which different components can be systematically tested. For instance, health sciences have recently adopted Multiphase Optimisation Strategies (MOST) as a research design that incorporates a screening phase by which suitable core components are first matched against the needs of the study participants. Subsequent trial conditions are arranged in factorial design by which different combinations of the approach are trialled (known as SMART—Sequential Multiple Assignment Randomized Trial). Later phases allow for fine tuning by identifying the most optimal components by subsequently using factorial designs (Collins et al., 2008). This may be particularly suited to interventions such as INCLUSIVE given the original developers' note in respect to schools adopting locally decided actions, suitable to individual contexts.

A final consideration is that of scalability. The original trial had the benefit of operating both an efficacy and effectiveness phase by which support for implementation was removed at the mid-point of the trial. Although both proximal findings around health and behaviour from the UCL/LSHTM trial and attainment from the current study both suggest continuation of effects under effectiveness conditions, there is not yet data to indicate effects when support is not initially provided. Such a consideration would be important for the scalability of effect in any future deployment of the intervention.

References

Armitage, R. (2021) ‘Bullying in Children: Impact on Child Health’, *BMJ Paediatrics Open*, 5 (1), e000939. https://doi.org/10.1136/bmjpo-2020-000939

Arseneault, L. (2017) ‘Annual Research Review: The Persistent and Pervasive Impact of Being Bullied in Childhood and Adolescence: Implications for Policy and Practice’, *Child Psychology and Psychiatry*, 59 (4), pp. 405–421. https://doi.org/10.1111/jcpp.12841

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015) ‘Fitting Linear Mixed-Effects Models Using Ime4’, *Statistical Software*, 67 (1), pp. 1–48. https://doi.org/10.18637/jss.v067.i01

Bellis, M. A., Hughes, K., Perkins, C. and Bennett, A. (2012) ‘Protecting people, Promoting health: A Public Health Approach to Violence Prevention for England’, North West Public Health Observatory, Liverpool John Moores University: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/216977/Violence-prevention.pdf

Blank, L., Baxter, S., Goyder, E., Naylor, P. B., Guillaume, L., Wilkinson, A., Hummel, S. and Chilcott, J. (2010) ‘Promoting Well-Being by Changing Behaviour: A Systematic Review and Narrative Synthesis of the Effectiveness of Whole Secondary School Behavioural Interventions’, *Mental Health Review*, 15, pp. 43–53.

Bond, L., Wolfe, S., Tollit, M., Butler, H. and Patton, G. (2007) ‘A Comparison of the Gatehouse Bullying Scale and the Peer Relations Questionnaire for Students in Secondary School’, *School Health*, 77 (2).

Bonell, C., Allen, E., Christie, D., Elbourne, D., Fletcher, A., Grieve, R., LeGood, R., Mathiot, A., Scott, S., Wiggins, M. and Viner, R. M. (2014) ‘Initiating Change Locally in Bullying and Aggression Through the School Environment (INCLUSIVE): Study Protocol for a Cluster Randomised Controlled Trial’, *Trials*, 15, p. 381. https://doi.org/10.1186/1745-6215-15-381

Bonell, C., Allen, E., Opondo, C., Warren, E., Elbourne, D. R., Sturgess, J., Bevilacqua, L., McGowan, J., Mathiot, A. and Viner, R. M. (2019) ‘Examining Intervention Mechanisms of Action Using Mediation Analysis Within a Randomised Trial of a Whole-School Health Intervention’, *Epidemiology and Community Health*, 73 (5), pp. 455–464. https://doi.org/10.1136/jech-2018-211443

Bonell, C., Allen, E., Warren, E., McGowan, J., Bevilacqua, L., Jamal, F., Legood, R., Wiggins, M., Opondo, C., Mathiot, A., Sturgess, J., Fletcher, A., Sadique, Z., Elbourne, D., Christie, D., Bond, L., Scott, S. and Viner, R. M. (2018) ‘Effects of the Learning Together Intervention on Bullying and Aggression in English Secondary Schools (INCLUSIVE): A Cluster Randomised Controlled Trial’, *The Lancet*, 392 (10163), pp. 2452–464. https://doi.org/10.1016/s0140-6736(18)31782-3

Bonell, C., Allen, E., Warren, E., McGowan, J., Bevilacqua, L., Jamal, F., Sadique, Z., Legood, R., Wiggins, M., Opondo, C., Mathiot, A., Sturgess, J., Paparini, S., Fletcher, A., Perry, M., West, G., Tancred, T., Scott, S., Elbourne, D., . . . Viner, R. M. (2019) ‘Modifying the Secondary School Environment to Reduce Bullying and Aggression: The INCLUSIVE Cluster RCT’: https://www.ncbi.nlm.nih.gov/pubmed/31682394

Bradley, R. H. and Corwyn, R. F. (2002) ‘Socioeconomic Status and Child Development’, *Annual Review of Psychology*, 53 (1), pp. 371–399.

Brooks, F., Magnusson, J., Klemera, E., Chester, K., Spencer, N. and Smeeton, N. (2015) ‘HBSC England National Report: Health Behaviour in School-aged Children (HBSC): World Health Organization Collaborative Cross National Study’, University of Hertfordshire: http://www.hbsc.org/news/index.aspx?ni=3256

Brown, S. and Taylor, K. (2008) ‘Bullying, Education and Earnings: Evidence from the National Child Development Study’, *Economics of Education Review*, 27 (4), pp. 387–401. https://doi.org/10.1016/j.econedurev.2007.03.003

Carpenter, J. R., Goldstein, H. and Kenward, M. G. (2011) ‘REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types’, *Statistical Software*, 45, pp. 1–14.

Chinn, S. (2000) ‘A Simple Method for Converting an Odds Ratio to Effect Size for Use in Meta-Analysis’, *Statistics in Medicine*, 19, pp. 3127–131.

Collins, L., Murphy, S. and Strecher, V. (2008) ‘The Multiphase Optimization Strategy (MOST) and the Sequential Multiple Assignment Randomized Trial (SMART): New Methods for More Potent eHealth Interventions’, *Preventitative Medicine*, 32 (5), pp. 112–118. https://doi.org/https://doi.org/10.1016/j.amepre.2007.01.022

Corcoran, R. P., Cheung, A. C. K., Kim, E. and Xie, C. (2018) ‘Effective Universal School-Based Social and Emotional Learning Programs for Improving Academic Achievement: A Systematic Review and Meta-Analysis of 50 Years of Research’, *Educational Research Review*, 25, pp. 56–72. https://doi.org/10.1016/j.edurev.2017.12.001

Crutzen, R. and Peters, G. Y. (2017) ‘Targeting Next Generations to Change the Common Practice of Underpowered Research’, *Frontiers in Psychology*, 8, p. 1184. https://doi.org/10.3389/fpsyg.2017.01184

DfE (2012) ‘National Curriculum Assessments at Key Stage 2 in England, 2011/2012’ (revised), Department for Education.

DfE (2014) ‘Schools, Pupils and Their Characteristics: January 2014’, Department for Education.

DfE (2017) ‘Preventing and Tackling Bullying. Advice for Headteachers, Staff and Governing Bodies’, Department for Education.

DfE (2018a) ‘Bullying in England, April 2013 to March 2018’, Department for Education.

DfE (2018b) ‘Bullying: Evidence from LSYPE2, Wave 3’, Department for Education.

DfE (2018c) ‘Key stage 4 Including Multi- Academy Trust Performance, 2018’ (revised), Department for Education.

Duckworth, A. L. and Yeager, D. S. (2015) ‘Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes’, *Educational Researcher*, 44 (4), pp. 237–251. https://doi.org/10.3102/0013189X15584327

EEF (2019) ‘Classification of the security of findings from EEF evaluations’, London: Education Endowment Foundation: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Carrying_out_a_Peer_Review/Classifying_the_security_of_EEF_findings_2019.pdf

Glew, G. M., Fan, M.-Y., Katon, W., Rivara, F. P. and Kernic, M. A. (2005) ‘Bullying, Psychosocial Adjustment, and Academic Performance in Elementary School’, *Archives of Pediatrics and Adolescent Medicine*, 159, pp. 1026–031. https://jamanetwork.com/journals/jamapediatrics/articlepdf/486162/poa50054.pdf

Goldberg, J. M., Sklad, M., Elfrink, T. R., Schreurs, K. M. G., Bohlmeijer, E. T. and Clarke, A. M. (2018) ‘Effectiveness of Interventions Adopting a Whole School Approach to Enhancing Social and Emotional Development: A Meta-Analysis’, *Psychology of Education*, 34 (4), pp. 755–782. https://doi.org/10.1007/s10212-018-0406-9

Greenberg, M. T. (2010) ‘School-Based Prevention: Current Status and Future Challenges’, *Effective Education*, 2 (1), pp. 27–52. https://doi.org/10.1080/19415531003616862

Grund, S., Robitzsch, A. and Luedtke, O. (2021) ‘mitml: Tools for Multiple Imputation in Multilevel Modelling’, R package version 0.4-3. https://cran.r-project.org/web/packages/mitml/mitml.pdf

Gupta, S. K. (2011) ‘Intention-to-Treat Concept: A Review’, *Perspectives in Clinical Research*, 2 (3), pp. 109–112. https://doi.org/10.4103/2229-3485.83221

Hattie, J. A. C. (2009) *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*, London: Routledge.

Hu, L.,. and Bentler, P. M. (1999) ‘Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives’, *Structural Equation Modelling*, 6 (1), pp. 1–55. https://doi.org/10.1080/10705519909540118

Humphrey, N. (2013) *Social and Emotional Learning: A Critical Appraisal*, London: Sage. https://doi.org/10.4135/9781446288603

Jennings, P. A. and Greenberg, M. T. (2009) ‘The Prosocial Classroom: Teacher Social and Emotional Competence in Relation to Student and Classroom Outcomes’, *Review of Educational Research*, 79 (1), pp. 491–525. https://doi.org/10.3102/0034654308325693

Kelly, G., Coleman, N., Hickman, M., and Word of Mouth. (2010). *TellUs4 Evaluation*. Department for Education.

Kline, P. and Walters, C. (2015) ‘Evaluating Public Programs with Close Substitutes: The Case of Head Start’, NBER Working Paper Series, Cambridge MA. http://www.nber.org/papers/w21658

Knowles, J. E. and Frederick, C. (2020) ‘merTools: Tools for Analysing Mixed Effect Regression Models’, R package version 0.5.2: https://CRAN.R-project.org/package=merTools

Kraft, M. A. (2020) ‘Interpreting Effect Sizes of Education Interventions’, *Educational Researcher*, 49 (4), pp. 241–253.

Langford, R., Bonell, C. P., Jones, H. E., Pouliou, T., Murphy, S. M., Waters, E., Komro, K. A., Gibbs, L. F., Magnus, D. and Campbell, R. (2014) 'The WHO Health Promoting School Framework for Improving the Health and Well-Being of Students and Their Academic Achievement', *Cochrane Database Systematic Reviews*, (4), CD008958. https://doi.org/10.1002/14651858.CD008958.pub2

Lee, M. W. and Newton, P. (2021) 'Systematic Divergence Between Teacher and Test-Based Assessment: Literature Review', Ofqual.

Lendrum, A., Humphrey, N. and Wigelsworth, M. (2013) 'Social and Emotional Aspects of Learning (SEAL) for Secondary Schools: Implementation Difficulties and Their Implications for School-Based Mental Health Promotion', *Child and Adolescent Mental Health*, 18 (3), pp. 158–164. https://doi.org/10.1111/camh.12006

Lloyd, G., Kane, J., McCluskey, G., Stead, J. and Riddell, S. (2006) 'Restorative Approaches in Scottish Schools: Transformations and Challenges', Scottish Executive Education Department.

Lodi, E., Perrella, L., Lepri, G. L., Scarpa, M. L. and Patrizi, P. (2021) 'Use of Restorative Justice and Restorative Practices at School: A Systematic Literature Review', *Environmental Research and Public Health*, 19 (1). https://doi.org/10.3390/ijerph19010096

Lynch, K., Hill, H. C., Gonzalez, K. E. and Pollard, C. (2019) 'Strengthening the Research Base That Informs STEM Instructional Improvement Efforts: A Meta-Analysis', *Educational Evaluation and Policy Analysis*, 41 (3), pp. 260–293. https://doi.org/10.3102/0162373719849044

MacCallum, R. C., Browne, M. W. and Sugawara, H. M. (1996) 'Power Analysis and Determination of Sample Size for Covariance Structure Modeling', *Psychological Methods*, 1 (2), pp. 130–149.

McMillan, J. H. (2007) 'Randomized Field Trials and Internal Validity: Not So Fast My Friend', *Practical Assessment, Research, and Evaluation*, 12. https://doi.org/10.7275/3vh7-m792

Merrell, K. W. and Gueldner, B. A. (2010) *Social and Emotional Learning in the Classroom: Promoting Mental Health and Academic Success*, New York: Guilford.

Morrison, B. (2005) 'Restorative Justice in Schools', in Elliot. E and Gordon, R. M. (eds), *New Directions in Restorative Justice: Issues, Practice, Evaluation*, Portland OR: Willan.

OECD (2019) *PISA 2018 Results (Volume III): What School Life Means for Students' Lives*, OECD Programme for International Student Assessment (PISA).

Olweus, D. (2013) 'School Bullying: Development and Some Important Challenges', *Annual Review of Clinical Psychology*, 9, pp. 751–780. https://doi.org/10.1146/annurev-clinpsy-050212-185516

Risser, S. D. (2013) 'Relational Aggression and Academic Performance in Elementary School', *Psychology in the Schools*, 50 (1), pp. 13–26. https://doi.org/10.1002/pits.21655

Rosseel, Y. (2012) 'lavaan: An R Package for Structural Equation Modeling', *Statistical Software*, 48 (2), pp. 1–36. https://doi.org/10.18637/jss.v048.i02

RStudio Team (2022). *RStudio: Integrated Development Environment for R*. RStudio, PBC.

Skinns, L., Du Rose, N. and Hough, M. (2009) 'An Evaluation of Bristol RAiS [Restorative Approaches in Schools]', ICPR, King's College London. https://transformingconflict.org/wp-content/uploads/2017/09/Bristol-RAiS-Report-2009.pdf

Sklad, M., Diekstra, R., Ritter, M. D., Ben, J. and Gravesteijn, C. (2012) 'Effectiveness of School-Based Universal Social, Emotional, and Behavioural Programs: Do They Enhance Students' Development in the Area of Skill, Behaviour, and Adjustment?', *Psychology in the Schools*, 49 (9), pp. 892–909. https://doi.org/10.1002/pits.21641

Smith, J. D., Schneider, B. H., Smith, P. K. and Ananiadou, K. (2004) 'The Effectiveness of Whole-School Antibullying Programs: A Synthesis of Evaluation Research', *School Psychology Review*, 33 (4), pp. 547–560.

Steer, A. (2009) 'Learning Behaviour: Lessons Learned: A Review of Behaviour Standards and Practices in Our Schools', Department for Children, Schools and Families.

Takizawa, R., Maughan, B. and Arseneault, L. (2014) 'Adult Health Outcomes of Childhood Bullying Victimization: Evidence From a Five-Decade Longitudinal British Birth Cohort', *The American Journal of Psychiatry*, 171 (7), pp. 777–784. https://doi.org/10.1176/appi.ajp.2014.13101401

Tanner-Smith, E. E., Durlak, J. A. and Marx, R. A. (2018) ‘Empirically Based Mean Effect Size Distributions for Universal Prevention Programs Targeting School-Aged Youth: A Review of Meta-Analyses’, *Prevention Science*, 19 (8), pp. 1091–101. https://doi.org/10.1007/s11121-018-0942-1

OECD (2017) ‘PISA 2015 Results (Volume III): Students’ Well-Being’, Organization for Economic Cooperation and Development, Programme for International Student Assessment (PISA).

Tobler, A. L., Komro, K. A., Dabroski, A., Aveyard, P. and Markham, W. A. (2011) ‘Preventing the Link Between SES and High-Risk Behaviours: “Value-Added” Education, Drug Use and Delinquency in High-Risk, Urban Schools’, *Prevention Science*, 12 (2), pp. 211–221. https://doi.org/10.1007/s11121-011-0206-9

Troncoso, P. (2020) ‘Minimum Detectable Effect Size Calculator’: https://patricio-troncoso.shinyapps.io/mdesapp/

Vreeman, R. C. and Carroll, A. E. (2007) ‘A Systematic Review of School-Based Interventions to Prevent Bullying’, *American Medical Association*, 161, pp. 78–88.

Vuoksimaa, E., Rose, R. J., Pulkkinen, L., Palviainen, T., Rimfeld, K., Lundstrom, S., Bartels, M., van Beijsterveldt, C., Hendriks, A., de Zeeuw, E. L., Plomin, R., Lichtenstein, P., Boomsma, D. I. and Kaprio, J. (2021) ‘Higher Aggression is Related to Poorer Academic Performance in Compulsory Education’, *Child Psychology and Psychiatry*, 62 (3), pp. 327–338. https://doi.org/10.1111/jcpp.13273

Weaver, J. L. and Swank, J. M. (2020) ‘A Case Study of the Implementation of Restorative Justice in a Middle School’, *RMLE Online*, 43 (4), pp. 1–9. https://doi.org/10.1080/19404476.2020.1733912

WHO (2014) ‘Global Status Report on Violence Prevention 2014’, World Health Organisation.

Wigelsworth, M., Lendrum, A., Oldfield, J., Scott, A., ten Bokkel, I., Tate, K. and Emery, C. (2016) ‘The Impact of Trial Stage, Developer Involvement and International Transferability on Universal Social and Emotional Learning Programme Outcomes: A Meta-Analysis’, *Cambridge Journal of Education*, 46 (3), pp. 347–376. https://doi.org/10.1080/0305764X.2016.1195791

Wigelsworth, M., Verity, L., Mason, C., Qualter, P. and Humphrey, N. (2021) ‘Social and Emotional Learning in Primary Schools: A Review of the Current State of Evidence’, *Educational Psychology*, e12480. https://doi.org/10.1111/bjep.12480

Wolke, D. and Lereya, S. T. (2015) ‘Long-Term Effects of Bullying’, *Archives of Disease in Childhood*, 100 (9), pp. 879–885. https://doi.org/10.1136/archdischild-2014-306667

Woods, S. and Wolke, D. (2004) ‘Direct and Relational Bullying Among Primary School Children and Academic Achievement’, *School Psychology*, 42 (2), pp. 135–155. https://doi.org/10.1016/j.jsp.2003.12.002

Wright, M. (1999) *Restoring Respect for Justice*, Hook, Hants: Waterside.

YJB (2004) ‘National Evaluation of the Restorative Justice in Schools Programme’, Youth Justice Board for England and Wales.

Appendix A: EEF cost rating

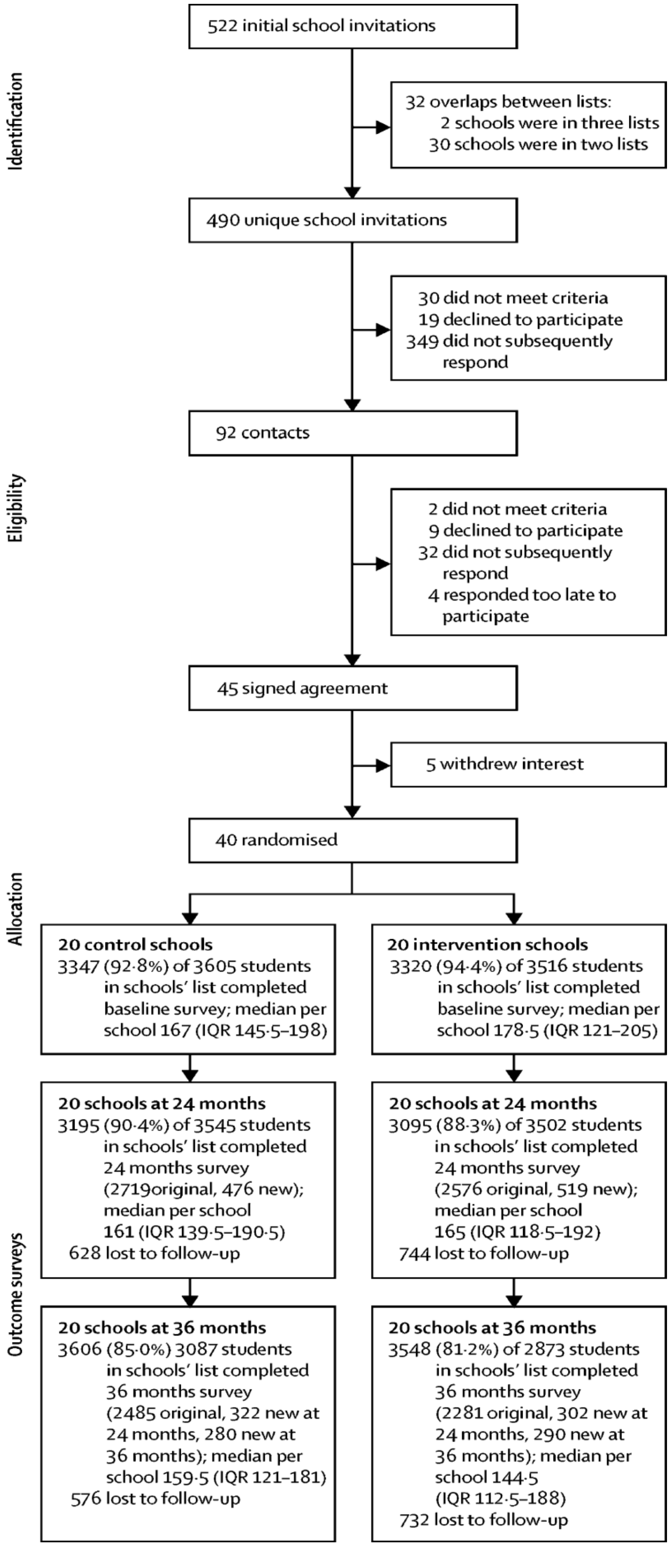
Appendix Figure 1: Cost rating

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

Appendix B: Threats to the validity of findings

Threats to validity of findings	Comments
Measurement of outcomes	The selected outcome, Attainment 8 scores, are a valid and reliable measure derived from the NPD.
Power and sample size	The UCL/LSHTM trial was a well-designed randomized controlled trial. However, the trial was powered to detect proximal effects on behaviour, rather than educational attainment. Although ex ante power considerations are outside the control of the research team for the purpose of this evaluation on attainment, it is an important aspect to consider when interpreting the results in this report.
Confounding	There was evidence of small pre-test imbalances in KS2 reading (ES = 0.09) and maths (ES = 0.14). However, the research team controlled for both variable in regression model and deemed that these baseline imbalances did not diminish the rigour of the results.
Missing data	Due to issues with pupil data matching, 23% of pupils who started the trial were not included in the final attainment analyses (a relatively high figure). The research team took this into account by performing multiple imputation as a sensitivity analysis and found negligible identifiable differences in the results.
Implementation and process evaluation (IPE)	A constraint of the current UoM evaluation on attainment was the absence if IPE data. Therefore, implementations dimension, such as fidelity, could not be fully assessed in the context of this study.
Concurrent interventions and experimental effects	There was some evidence of control schools implementing activities similar INCLUSIVE, which could not be controlled for analytically. Moreover, there is also some indicative evidence that allocation to the comparison condition itself may have accelerated activity in this area. Although the original trial showed analytically that control schools partaking in similar activities did not make an observable difference to the primary behaviour outcomes, it is not known whether this would be the case for the distal outcome of attainment.
Selective reporting	No risks in terms of selected reporting. The original study has been registered and protocol and SAP published. The UoM team worked with limited data which ultimately affected the reporting of the current study.

Appendix C: Participant flow diagram from original INCLUSIVE trial (Bonell et al., 2018)


Identification
Eligibility
Allocation
Outcome surveys
522 initial school invitations
32 overlaps between lists:
2 schools were in three lists
30 schools were in two lists
490 unique school invitations
30 did not meet criteria
19 declined to participate
349 did not subsequently respond
92 contacts
2 did not meet criteria
9 declined to participate
32 did not subsequently respond
4 responded too late to participate
45 signed agreement
5 withdrew interest
40 randomised
20 control schools
3347 (92·8%) of 3605 students in schools' list completed baseline survey; median per school 167 (IQR 145·5–198)
20 intervention schools
3320 (94·4%) of 3516 students in schools' list completed baseline survey; median per school 178·5 (IQR 121–205)
20 schools at 24 months
3195 (90·4%) of 3545 students in schools' list completed 24 months survey (2719original, 476 new); median per school 161 (IQR 139·5–190·5)
628 lost to follow-up
20 schools at 24 months
3095 (88·3%) of 3502 students in schools' list completed 24 months survey (2576 original, 519 new); median per school 165 (IQR 118·5–192)
744 lost to follow-up
20 schools at 36 months
3606 (85·0%) 3087 students in schools' list completed 36 months survey (2485 original, 322 new at 24 months, 280 new at 36 months); median per school 159·5 (IQR 121–181)
576 lost to follow-up
20 schools at 36 months
3548 (81·2%) of 2873 students in schools' list completed 36 months survey (2281 original, 302 new at 24 months, 290 new at 36 months); median per school 144·5 (IQR 112·5–188)
732 lost to follow-up

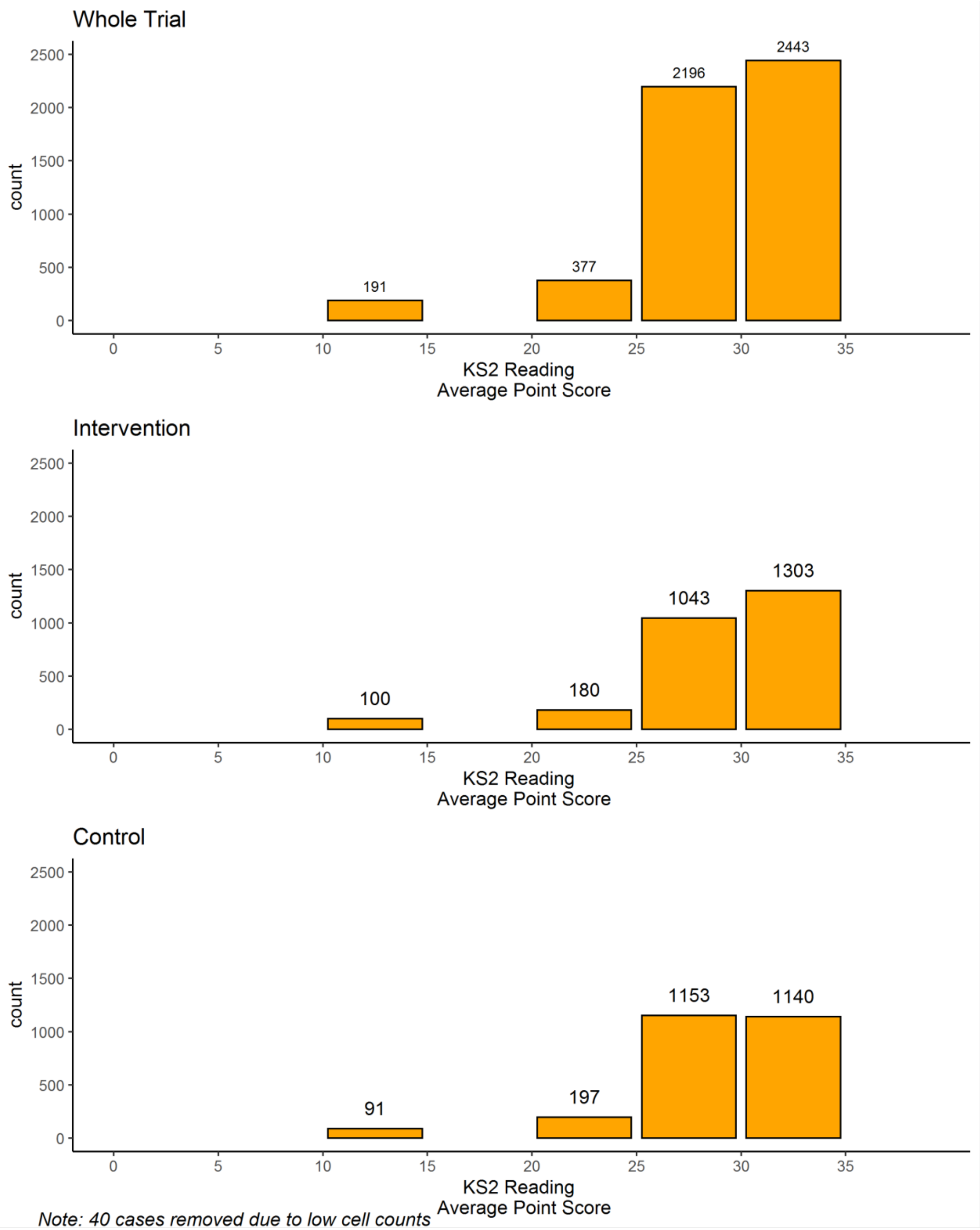
Appendix D: Predictors of missing data for KS4 Attainment 8 outcome

Appendix Table 1: Logistic regression to predict missingness

Intercept (SE): -6.05 (0.05)		
	Coefficient (SE)	*p*
School level		
Intervention	-0.03 (0.48)	.957
Pupil level		
KS2 Reading	0.23 (0.26)	.358
KS2 Maths	-0.57 (0.23)	.013
Gender	-0.72 (0.53)	.177
Free school meal eligibility	1.06 (0.51)	.038

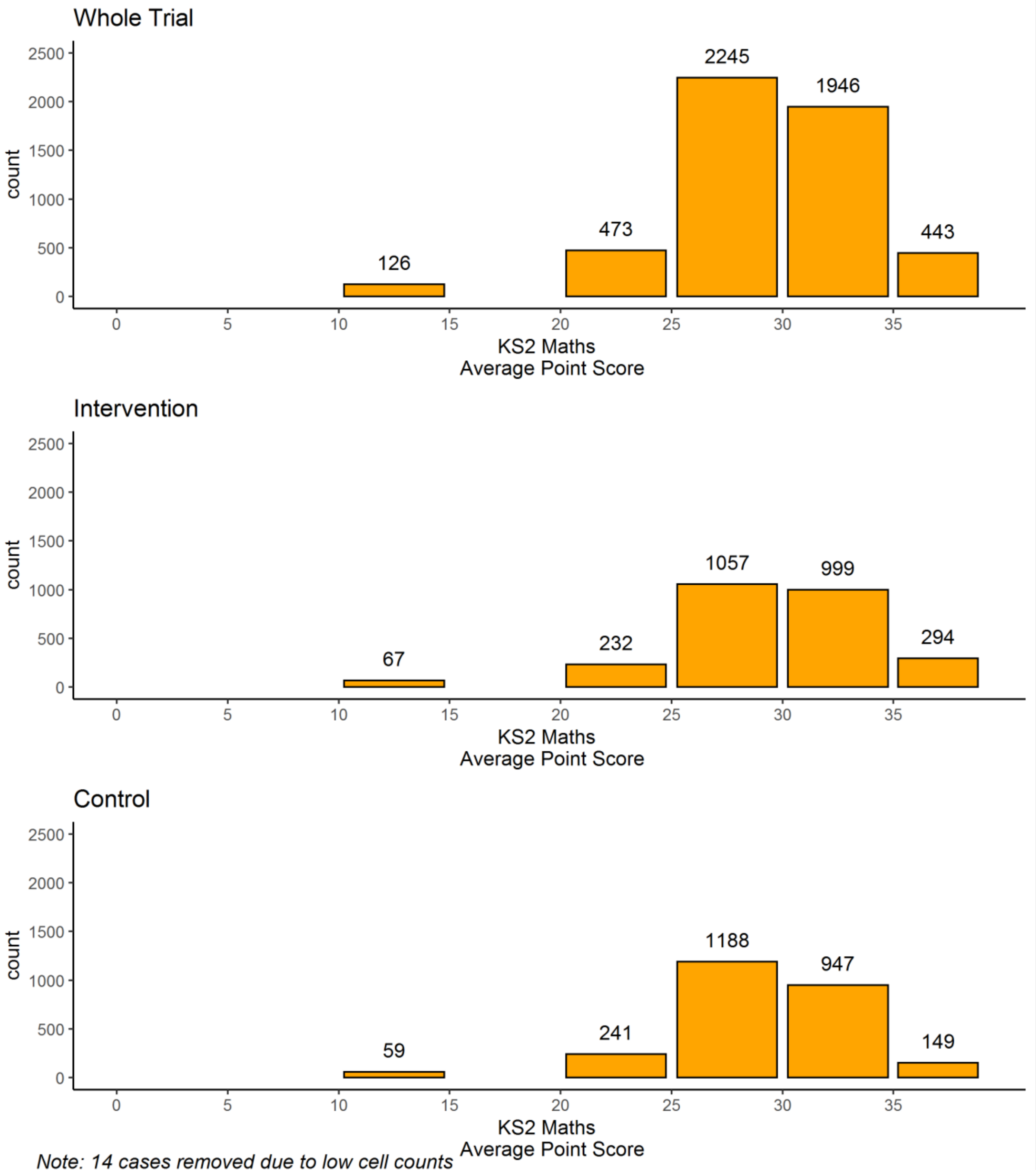
Appendix E: Pre- and post-test histograms

Appendix Figure 2: Histogram displaying the distribution of pre-test (KS2) reading scores



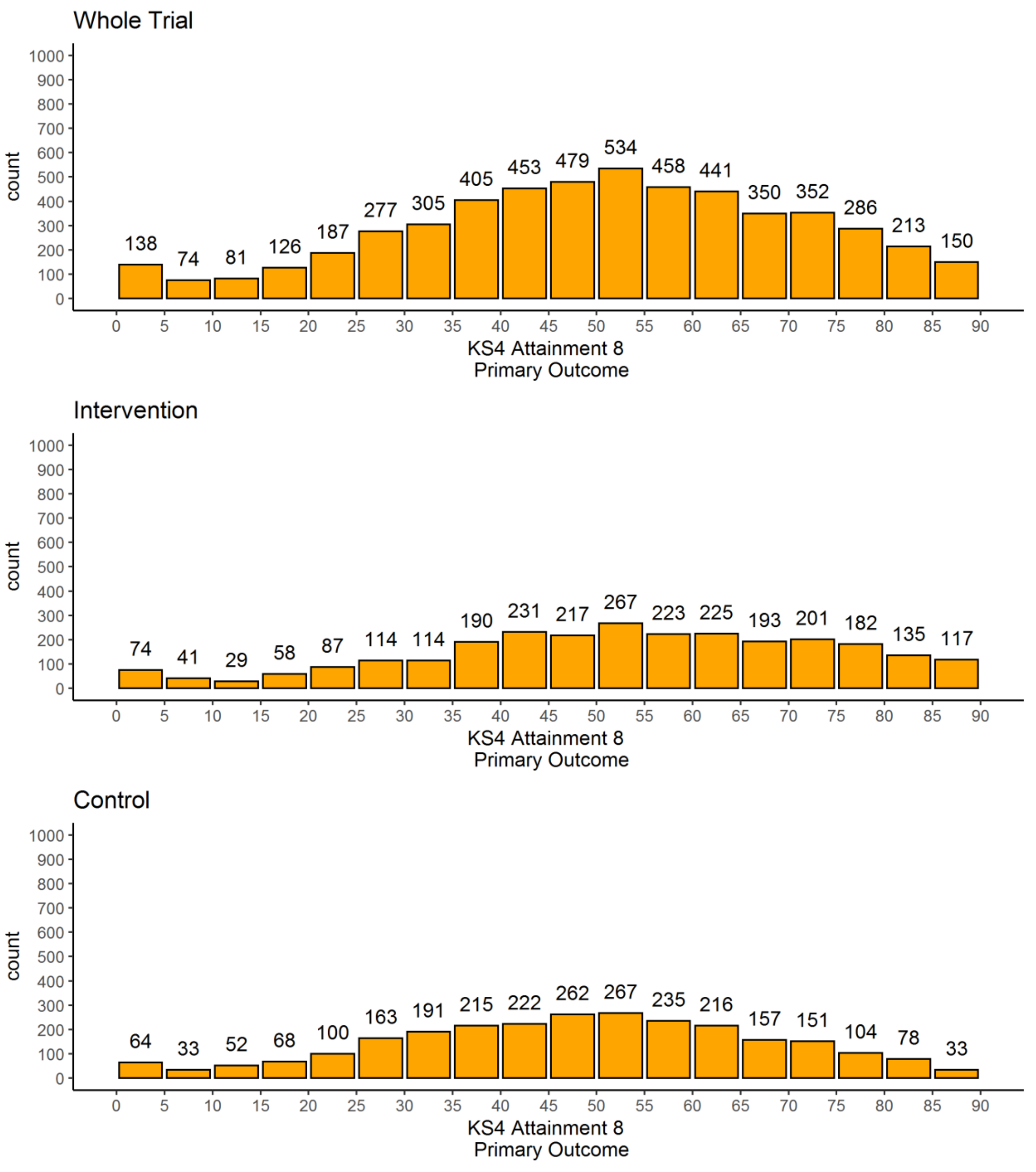
Note: 40 cases removed due to low cell counts

Appendix Figure 3: Histogram displaying the distribution of pre-test (KS2 maths) scores

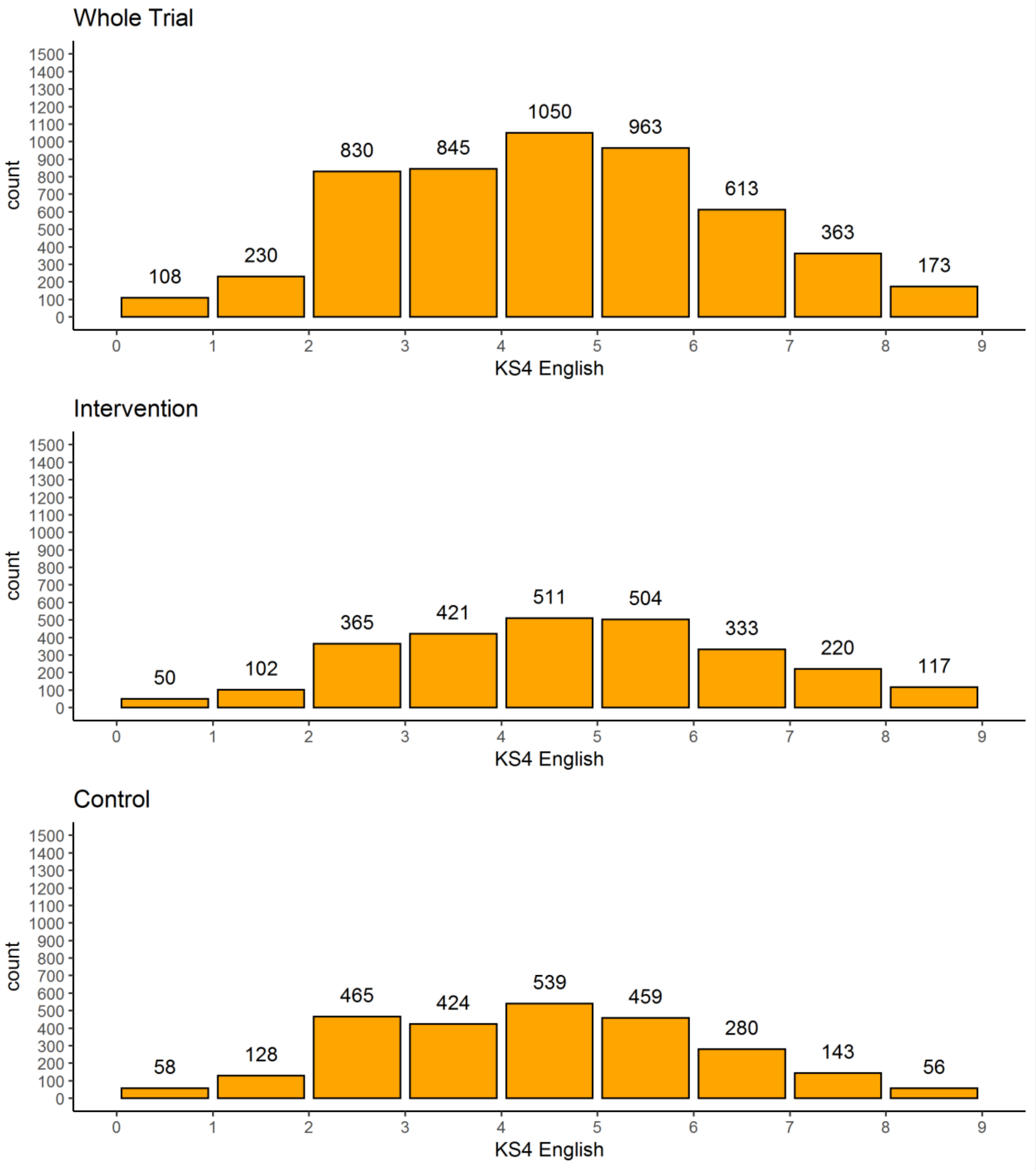


Note: 14 cases removed due to low cell counts

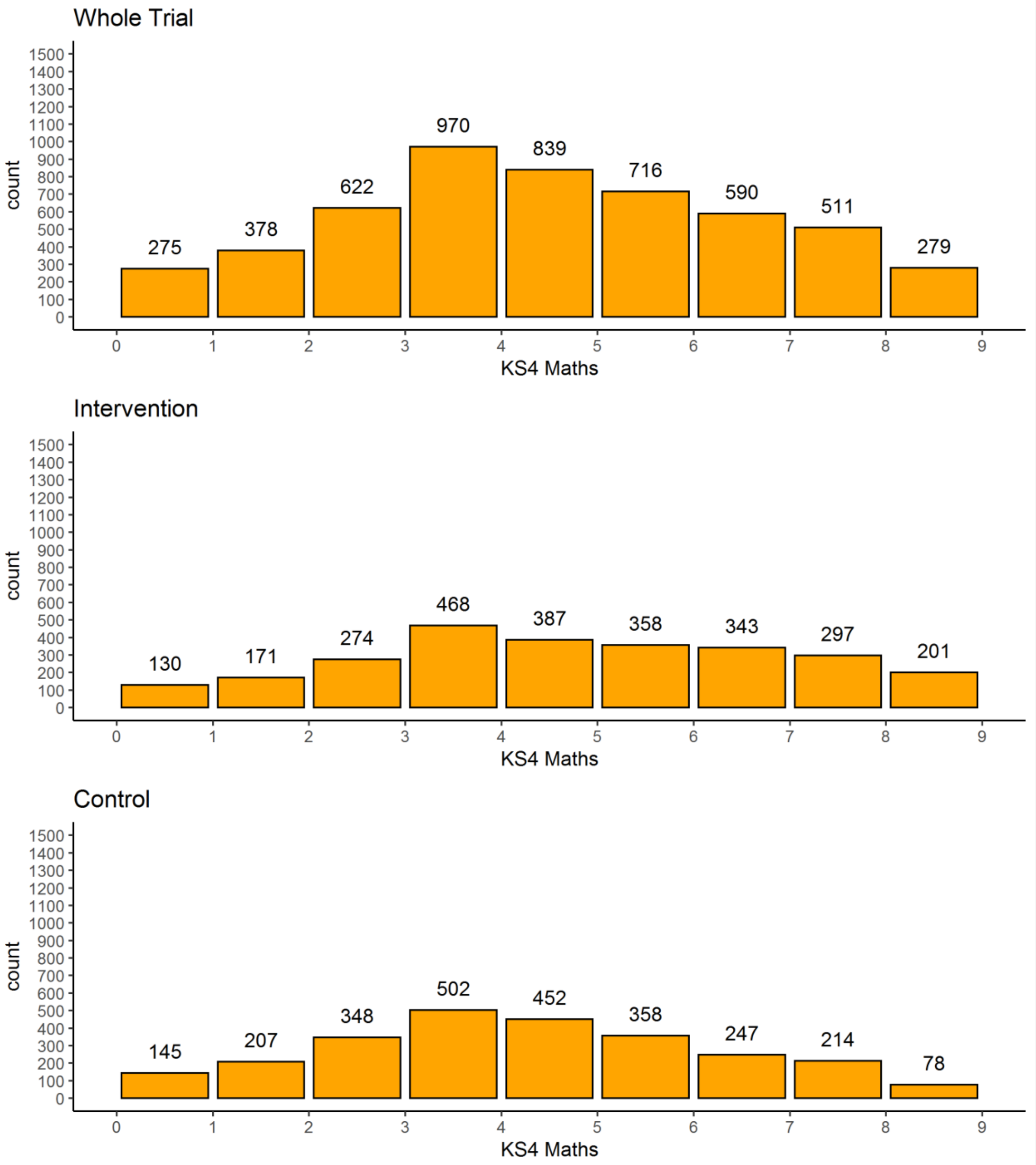
Appendix Figure 4: Histogram showing distribution of Attainment 8 scores



Appendix Figure 5: Histogram showing distribution of KS4 English scores



Appendix Figure 6: Histogram showing distribution of KS4 maths scores



Appendix F: MLM for ITT and MI analyses

Appendix Table 2: Fixed effects for ITT analyses (primary and secondary outcomes—complete cases)

	KS4 Attainment 8 (n = 5128)		KS4 English (n = 5006)		KS4 maths (n = 5012)	
	Intercept = -0.14 (0.04)		**Intercept** =-0.15 (0.06)		**Intercept** -0.14 (0.05)	
	Coefficient β (SE)	***p***	**Coefficient β (SE)**	***p***	**Coefficient β (SE)**	***p***
School						
School type (single sex)	0.25 (0.05)	<.001	0.25 (0.07)	.001	0.14 (0.06)	.012
FSM eligibility (low/medium)	0.09 (0.05)	<.001	0.06 (0.06)	.348	0.15(0.05)	.004
CVA (below median)	-0.12 (0.05)	.012	-0.06 (0.06)	.347	-0.11(0.05)	.037
Trial arm (intervention)	**0.14 (0.04)**	**.004**	**0.13 (0.06)**	**.039**	**0.09 (0.05)**	**.059**
Pupil						
KS2 Reading	0.25 (0.01)	<.001	0.34 (0.01)	<.001	0.13 (0.01)	<.001
KS2 maths	0.45 (0.01)	<.001	0.29 (0.01)	<.001	0.61 (0.01)	<.001

Appendix Table 3: Fixed effects for ITT analyses (primary and secondary outcomes—MI analyses)

	KS4 Attainment 8 (n = 6659)		KS4 English (n=6659)		KS4 maths (n=6659)	
	Intercept = -0.13 (0.04)		**Intercept** = -0.14 (0.06)		**Intercept**= -0.12(0.05)	
	Coefficient β (SE)	***p***	**Coefficient β (SE)**	***p***	**Coefficient β (SE)**	***p***
School						
School type (single sex)	0.26 (0.05)	<.001	0.25 (0.07)	<.001	0.14 (0.06)	.016
FSM eligibility (low/medium)	0.09 (0.04)	.040	0.08 (0.06)	.176	0.16 (0.05)	.001
CVA (below median)	-0.12 (0.05)	.010	-0.04 (0.06)	.497	-0.08 (0.05)	.088
Trial arm (intervention)	**0.13 (0.04)**	**.003**	**0.11 (0.06)**	**.056**	**.08 (0.05)**	**.095**
Pupil						
KS2 Reading	0.24 (0.01)	<.001	0.32 (0.02)	<.001	0.12 (0.01)	<.001
KS2 maths	0.45 (0.02)	<.001	0.29 (0.02)	<.001	0.60 (0.01)	<.001

Appendix G: MLM for Subgroup Analyses (IT and MI)

Appendix Table 4: Fixed effects for FSM subgroup analysis (primary and secondary outcomes—complete cases)

	KS4 Attainment 8 (N=5048)		KS4 English (N= 4969)		KS4 maths (N= 4972)	
	Intercept = 0.04(.04)		**Intercept** = -0.04 (.06)		**Intercept** = -0.03 (0.05)	
	Coefficient β (SE)	***p***	**Coefficient β (SE)**	***p***	**Coefficient β (SE)**	***p***
FSM subgroup analysis						
School						
School type (single sex)	0.24(0.05)	<.001	0.24(0.07)	.002	0.13(0.06)	.028
FSM eligibility (low/medium)	-0.02(0.05)	.671	-0.01 (0.06)	.889	0.08(0.05)	.131
CVA (below median)	-0.14 (0.05)	.005	-0.05 (0.06)	.409	-0.1 (0.05)	.046
Trial arm (intervention)	0.15 (0.05)	.002	**0.14 (0.06)**	**.036**	**0.11(0.05)**	**.034**
Pupil						
KS2 Reading	0.25 (0.01)	<.001	0.33 (0.01)	<.001	0.12 (0.01)	<.001
KS2 maths	0.43 (0.01)	<.001	0.27 (0.01)	<.001	0.60 (0.01)	<.001
Ever been eligible for FSM (Yes)	-0.28 (0.03)	<.001	0.22 (0.03)	<.001	-0.22 (0.03)	<.001
FSM (yes)*Trial Arm (intervention)	0 (0.04)	.956	0 (0.05)	.962	-0.01 (0.04)	.825

Appendix Table 5: Fixed effects for FSM subgroup analysis (primary and secondary outcomes—imputed data)

	KS4 Attainment 8 (N=6659)		KS4 English (N= 6659)		KS4 maths (N= 6659)	
	Intercept = 0 (0.05)		**Intercept** = -0.04 (0.06)		**Intercept** = -0.01 (0.05)	
	Coefficient β (SE)	***p***	**Coefficient β (SE)**	***p***	**Coefficient β (SE)**	***p***
FSM subgroup analysis						
School						
School type (single sex)	0.25 (0.05)	<.001	0.24 (0.07)	.001	0.13 (0.06)	.018
FSM eligibility (low/medium)	0.01(0.05)	.887	0.01 (0.06)	.871	0.09 (0.05)	.066
CVA (below median)	-0.12 (0.05)	.012	-0.04 (0.06)	.569	-0.08 (0.05)	.082
Trial arm (intervention)	0.16 (0.05)	.001	**0.14 (0.06)**	**.027**	**0.11 (0.05)**	**.018**
Pupil						

KS2 Reading	0.25(0.02)	<.001	0.33 (0.02)	<.001	0.12 (0.02)	<.001
KS2 maths	0.43 (0.02)	<.001	0.26 (0.02)	<.001	0.58(0.01)	<.001
Ever been eligible for FSM (Yes)	-0.29 (0.03)	<.001	-0.22 (0.04)	<.001	-0.23 (0.03)	<.001
FSM (yes)*Trial Arm (intervention)	-0.01 (0.04)	.840	-0.01 (0.05)	.805	-0.02 (0.04)	.604

Appendix Table 6: Fixed effects for gender subgroup analysis (primary and secondary outcomes—complete cases)

	KS4 Attainment 8 (N=5048)		KS4 English (N= 4969)		KS4 maths (N=4972)	
	Intercept = 0.03 (0.04)		**Intercept** = 0.03 (0.05)		**Intercept** = -0.14 (0.05)	
	Coefficient β (SE)	***p***	**Coefficient β (SE)**	***p***	**Coefficient β (SE)**	***p***
Gender subgroup analysis						
School						
School type (single sex)	0.20(0.05)	<.001	0.16 (0.06)	.013	0.14(0.06)	.018
FSM eligibility (low/medium)	0.07 (0.04)	.100	0.07 (0.06)	.219	0.14(0.05)	.007
CVA (below median)	-0.13(0.04)	.007	-0.03 (0.06)	.595	-0.11(0.05)	.037
Trial arm (intervention)	**0.13(0.05)**	**.009**	**0.14 (0.06)**	**.020**	**0.09 (0.05)**	**.083**
Pupil						
KS2 Reading	0.23 (0.01)	<.001	0.31 (0.01)	<.001	0.13 (0.01)	<.001
KS2 maths	0.46(0.01)	<.011	0.32 (0.01)	<.001	0.61(0.01)	<.001
Gender (Female)	-0.23 (0.03)	<.001	-0.34 (0.03)	<.001	0.02(0.03)	.614
Gender (female)* Trial Arm (intervention)	-0.01(0.04)	.782	-0.06(0.05)	.186	-0.01 (0.04)	.87

Appendix Table 7: Fixed effects for gender subgroup analysis (primary and secondary outcomes—imputed data)

	KS4 Attainment 8 (N=6659)		KS4 English (N= 6659)		KS4 maths (N=6659)	
	Intercept = -0.01 (0.05)		**Intercept** = .04 (.05)		**Intercept** = -.12 (.05)	
	Coefficient β (SE)	***p***	**Coefficient β (SE)**	***P***	**Coefficient β (SE)**	***p***
Gender subgroup analysis						
School						
School type (single sex)	0.21 (0.05)	<.001	0.17 (0.06)	.005	0.14 (0.06)	.013
FSM eligibility (low/medium)	0.1 (0.05)	.025	0.09 (0.05)	.087	0.16 (0.05)	.001
CVA (below median)	-0.11 (0.05)	.021	-0.01 (0.05)	.840	-0.09 (0.05)	.070

Trial arm (intervention)	**0.13 (0.05)**	**.004**	**0.12 (0.06)**	**.026**	**0.1 (0.05)**	**.046**
Pupil						
KS2 Reading	0.23 (0.02)	<.001	0.30 (0.02)	<.001	0.12(0.04)	<.001
KS2 maths	0.46 (0.02)	<.001	0.31 (0.02)	<.001	0.59 (0.02)	<.001
Gender (Female)	-0.25 (0.03)	<.001	-0.36 (0.03)	<.001	0 (0.03)	.900
Gender (female)* Trial Arm (intervention)	-0.01 (0.04)	.764	-0.05 (0.05)	.285	-0.02 (0.04)	.632

Appendix Table 8: Fixed effects for bullying subgroup analysis (primary and secondary outcomes—complete cases)

	KS4 Attainment 8 (N=4678)		KS4 English (N=4576)		KS4 maths (N= 4583)	
	Intercept = -0.09 (0.04)		**Intercept** = -0.14 (0.06)		**Intercept** = -0.10(0.05)	
	Coefficient β (SE)	***p***	**Coefficient β (SE)**	***p***	**Coefficient β (SE)**	***p***
Bullying subgroup analysis						
School						
School type (single sex)	0.25 (0.05)	<.001	0.25 (0.07)	.001	0.15(0.06)	.001
FSM eligibility (low/medium)	0.08 (0.04)	.097	0.05 (0.06)	.470	0.14(0.05)	.007
CVA (below median)	-0.14 (0.05)	.005	-0.07 (0.06)	.29	-0.12(0.05)	.024
Trial arm (intervention)	**0.14 (0.05)**	**.005**	**0.14 (0.06)**	**.031**	**0.08(0.05)**	**.103**
Pupil						
KS2 Reading	0.25 (0.01)	<.001	0.35 (0.02)	<.001	0.13(0.01)	<.001
KS2 maths	0.45 (0.01)	<.001	0.29 (0.02)	<.001	0.61(0.01)	<.001
Bullied (not upset or frequent)	-0.03 (0.03)	.286	0.02 (0.04)	.64	-0.04(0.03)	.191
Bullied (upset and/or frequently)	-0.17 (0.05)	.001	-0.06 (0.06)	.29	-0.13(0.05)	.010
Bullied (not upset or frequent) *Trial arm (intervention)	0.02 (0.05)	.740	-0.02 (0.05)	.724	0.02(0.04)	.658
Bullied (upset and/or frequently) *Trial arm (intervention	0.01 (0.07)	.925	0.02 (0.08)	.77	0(0.07)	.951

Appendix Table 9: Fixed effects for bullying subgroup analysis (primary and secondary outcomes—imputed data)

	KS4 Attainment 8 (N=6659)	KS4 English (N=6659)	KS4 maths (N= 6659)
	Intercept = -0.1 (0.05)	**Intercept** = -0.14 (0.06)	**Intercept** = -0.10 (0.05)

	Coefficient β (SE)	***p***	**Coefficient β (SE)**	***p***	**Coefficient β (SE)**	***p***
Bullying subgroup analysis						
School						
School type (single sex)	0.26 (0.05)	<.001	0.25 (0.07)	<.001	0.14 (0.06)	.010
FSM eligibility (low/medium)	0.09(0.05)	.048	0.07(0.06)	.217	0.16 (0.05)	.001
CVA (below median)	-0.12(0.05)	.008	-0.04 (0.06)	.507	-0.09 (0.05)	.068
Trial arm (intervention)	**0.14 (0.05)**	**.003**	**0.13(0.06)**	**.037**	**0.09 (0.05)**	**.071**
Pupil						
KS2 Reading	0.25 (0.02)	<.001	0.33(0.02)	<.001	0.13(0.02)	<.001
KS2 maths	0.44 (0.02)	<.001	0.28(0.02)	<.001	0.59 (0.02)	<.001
Bullied (not upset or frequent)	-0.04 (0.03)	.180	0.02 (0.03)	.591	-0.03 (0.03)	.251
Bullied (upset and/or frequently)	-0.16 (0.05)	.003	-0.07(0.05)	.172	-0.13 (0.05)	.010
Bullied (not upset or frequent) *Trial arm (intervention)	0.01 (0.04)	.912	-0.03 (0.05)	.544	0.01 (0.04)	.757
Bullied (upset and/or frequently) *Trial arm (intervention	-0.02 (0.07)	.810	-0.02 (0.07)	.784	-0.03 (0.07)	.714

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

Facebook.com/EducEndowFoundn