EVALUATION REPORT

Nurturing Empathy Before Transition

Efficacy Trial Report

Dr Sarah Reaney-Wood, Sean Demack, Jessica Benson-Eggleton, Dr Josephine Booth and Eleanor Byrne

January 2026

Sheffield Hallam University

YOUTH ENDOWMENT FUND

About the Youth Endowment Fund

The Youth Endowment Fund (YEF) is a charity with a mission that matters. We exist to prevent children and young people from becoming involved in violence. We do this by finding out what works and building a movement to put this knowledge into practice.

Children and young people at risk of becoming involved in violence deserve services that give them the best chance of a positive future. To make sure that happens, we'll fund promising projects and then use the very best evaluation to find out what works. Just as we benefit from robust trials in medicine, young people deserve support grounded in the evidence. We'll build that knowledge through our various grant rounds and funding activities.

And just as important, is understanding children and young people's lives. Through our Youth Advisory Board and national network of peer researchers, we'll ensure that they influence our work and that we understand and are addressing their needs. But none of this will make a difference if all we do is produce reports that stay on a shelf.

Together, we need to look at the evidence and agree on what works, then build a movement to make sure that young people get the very best support possible. Our strategy sets out how we'll do it. At its heart, it says that we'll fund good work, find what works and work for change. You can read it here.

For more information about the YEF or this report, please contact:

Youth Endowment Fund
C/O Impetus
10 Queen Street Place
London
EC4R 1AG

www.youthendowmentfund.org.uk
hello@youthendowmentfund.org.uk

Registered Charity Number: 1185413

Contents

About the evaluator

The evaluation team at the Sheffield Institute of Education includes experts in trial design, analysis, implementation and process evaluation. The team has extensive experience conducting evaluation studies as part of the Youth Endowment Fund evaluator panel and for other organisations, including the Education Endowment Foundation, government departments, charities and other policy makers.

For more information, please visit: https://www.shu.ac.uk/sheffield-institute-education-research.

Dr Sarah Reaney-Wood – principal investigator, trial manager and trial statistician – s.j.reaney-wood@shu.ac.uk

Sean Demack – principal investigator (during Dr Reaney-Wood's maternity leave) and trial statistician – s.demack@shu.ac.uk

Dr Josephine Booth – implementation and process evaluation (IPE) lead and fieldwork, analysis and write up – josephine.booth@shu.ac.uk

Jessica Benson-Egglenton – IPE analysis and write up – jessica.benson-egglenton@shu.ac.uk

Eleanor Byrne – IPE fieldwork – e.byrne@shu.ac.uk

Ben Willis – IPE fieldwork – b.willis@shu.ac.uk

Bernadette Stiell – IPE fieldwork

Acknowledgements

Thanks to Lucy Clague and Claire Wolstenholme who were involved in the project set up, to Giota Blouchou for her work on impact data collection in the second year of the Roots of Empathy trial and to Linda Bray and Louise Glossop for their role in coordinating control school incentive payments and their role in checking baseline questionnaire responses.

Executive summary

The project

The Nurturing Empathy Before Transition (NEBT) programme aims to increase empathy, improve social and emotional skills, and reduce aggression and bullying amongst Year 5 children. In the long term, the programme aims to reduce violent behaviour and offending. Delivered by the charity Roots of Empathy (ROE), the intervention places a parent and their baby in Year 5 classrooms as part of a structured programme of lessons designed to develop empathy. Children receive 27 45-minute sessions that cover nine themes (with three sessions delivered each theme). Examples of themes include meeting the baby, relationships, and communicating. Nine of the sessions are pre-family visit sessions, where children discuss what to expect; nine sessions are family visits, where children observe the baby's feelings, intentions and attachment to their parent; and nine sessions are post-family visits, where children reflect on their own feelings and the feelings of others. Sessions are delivered by an instructor who is trained by ROE. While previous research on ROE has been delivered by certified instructors who have completed training, in this project sessions were delivered by school-based teaching assistants (TAs) who were not certified at the start of the programme. Delivery in this project took place in five regions (Yorkshire, Merseyside, East and West Midlands, Greater London and Wales).

YEF funded a randomised controlled trial of NEBT. The trial aimed to assess the impact of NEBT on self-reported behavioural difficulties (as measured by the self-report Me and My Feelings (M&MF) questionnaire behavioural difficulties sub-scale). It also aimed to assess the impact on self-reported emotional difficulties and empathy, as well as on teacher-reported child behaviour. The evaluation was undertaken in two cohorts. Cohort 1 was delivered in 2022/3, with 16 intervention schools (that received NEBT) and 17 control schools (that continued with business as usual). Cohort 2 was delivered in 2023/4 to 30 intervention schools and 24 control schools (resulting in a total of 87 trial schools across both cohorts, 46 intervention [including 910 children] and 41 control schools [including 752 children]). The evaluation also included an implementation and process evaluation (IPE) that examined the key factors influencing the delivery of NEBT and explored the perceptions of children, teachers and deliverers. The IPE used eight case studies, observations of training sessions and the delivery of NEBT in classrooms, eight interviews with instructors and eight interviews with class teachers. It also conducted four interviews with senior school leaders, an interview with ROE and seven focus groups with children. The evaluation of NEBT in this project was impacted by the COVID-19 pandemic. Preparatory work began before the pandemic, and school closures led to pauses and delays. The impact of COVID-19 on schools and children and the heightened pressure and demand on these schools may also have impacted delivery.

Key conclusions
NEBT had a **small impact** on reducing children's self-reported behavioural difficulties. After the programme, children in NEBT schools reported slightly lower levels of behavioural difficulty compared to their counterparts in schools that did not receive NEBT. This result has an **extremely low** security rating.
NEBT had no impact on children's self-reported emotional difficulties or their cognitive empathy (understanding others' thoughts). It had a moderate impact on their self-reported affective empathy (empathy with others' emotions). NEBT showed a large impact on reducing teacher-reported behavioural difficulties, and this impact was driven by large impacts on reducing peer problems and hyperactivity and moderate impacts on reducing conduct and emotional problems. NEBT also showed a large impact on improving teacher-reported pro-social behaviour. These are the secondary outcomes, which should be interpreted with even more caution.
A very high level of attrition from the evaluation significantly weakens our confidence in the findings. 61% of children who started the trial were not included in the final analysis. 23/46 NEBT schools dropped out shortly after randomisation. School concerns regarding the time taken to deliver NEBT and measurement burden may have contributed to attrition.
61% of intervention schools delivered 8 out of 9 themes from the NEBT curriculum. TAs often made amendments to session scheduling to ensure that the content could be covered in time.

Positive relationships between teachers and the teaching assistants delivering NEBT, physical space for the sessions, and flexibility from mothers and school settings supported delivery. Barriers to delivery included challenges in recruiting mothers for some schools and insufficient time for TAs to prepare for sessions.

YEF security rating

These findings have an **extremely low** security rating. While the trial was set up as a well-designed efficacy randomised controlled trial that was large enough to detect meaningful impacts, a very high level of attrition substantially reduces the confidence we can have in the findings. 61% of the children who started the trial were not included in the final analysis because they did not participate in endline testing. We do not know if the effect found for NEBT would be the same if the children missing from the final analysis were included.

Interpretation

This result has an extremely low security rating, and all outcomes should be treated with significant caution. NEBT had a small impact on reducing children's self-reported behavioural difficulties. With regard to secondary outcomes, NEBT had no impact on children's self-reported emotional difficulties or their cognitive empathy. It had a moderate impact on their self-reported affective empathy and showed a large impact on reducing teacher-reported behavioural difficulties and improving teacher-reported pro-social behaviour. All outcomes should be treated with considerable caution, given the high level of attrition. Schools gave a variety of reasons for dropping out of the evaluation, including a lack of time to deliver the intervention and a change in the teaching assistant's circumstances. Using a community-based volunteer approach to deliver the programme (rather than solely using TAs) may have reduced the burden and time required from school staff.

Schools that successfully implemented the programme demonstrated flexibility in integrating NEBT within their existing operations. This included accommodating extended family visits, allocating preparation time for instructors, and prioritising sessions in the weekly schedule. The quality of classroom dynamics between teachers and the TAs who delivered the sessions also significantly influenced programme delivery. The flexibility of mothers and the availability of adequate physical space were key facilitators too. Barriers to delivery included timetabling challenges, insufficient preparation time for TAs to deliver the sessions, and the limited prior experience of some TAs. Some schools also faced challenges in recruiting mothers and babies.

Seven out of eight case study schools reported a positive perception of the programme, with active student engagement a common theme. In general, teachers and instructors felt that participating in the programme had positively impacted children's empathy and behaviour. In the pupil focus groups, children were very positive about the family visits, stating that it had been fun and exciting to see the baby grow, develop and start interacting with the world around them. One unintended consequence that some schools identified was that children with complex home situations (such as looked after children and young carers) could become distressed by the content (as the mother-baby bond may not have reflected their own experience).

Previous casual evaluations of ROE from beyond England and Wales have suggested that the programme may lead to an increase in teacher-reported pro-social behaviour. We also know from the wider evidence on social and emotional learning programmes that they can support reductions in children's involvement in violence. However, beyond the very useful reflections on implementation, the significant limitations in this study (caused by the high attrition rate) limit the contribution of the study to the wider evidence base. This report and the primary and secondary outcome findings only present the findings of one study. When considering implications, frontline professionals, policymakers and service commissioners should carefully consider the process evaluation, the wider evidence base and their own professional judgement. YEF has no plans for further evaluation of NEBT.

Summary of impact

Outcome	Effect size (95% confidence interval)	Impact	Evidence security	No. of children	P -value
M&MF behavioural difficulties	-0.06 (-0.22; 0.10)	Small	0/5 magnifying glasses	644	0.5

Introduction

Background

Youth violence has been increasing around the world in recent years at a serious cost to society (Haylock et al., 2020). The United Kingdom has consistently seen an increase in the incidence of youth violence since 2012/13 (Haylock et al., 2020). Specifically, poor mental health has been associated with violence among youth, including gang violence. During a time when the Covid-19 pandemic has negatively impacted child and youth mental health and wellbeing (Office for Health Improvement & Disparities, 2022), programmes that mitigate this negative impact and that support positive mental health are increasingly important.

There is extensive evidence for the benefits of and need for well-designed school-based interventions that focus on developing pupils' social and emotional wellbeing (Browne et al., 2004; Durlak et al., 2011; Tome et al., 2021). School-based interventions have been demonstrated to lead to improvements in behaviour and learning (Panayiotou et al., 2019), as well as academic success, better health outcomes and later life success. The need for well-designed and implemented interventions is thought to have increased since the COVID-19 pandemic (Lee et al., 2020; Hamoda et al., 2021), as this disruption has negatively impacted child and youth mental health and wellbeing (Office for Health Improvement & Disparities, 2022).

The Roots of Empathy (ROE) programme, which is only called Nurturing Empathy before Transition (NEBT) for the purposes of this trial, is a universal school-based intervention developed in Canada. Previous evaluations suggest that the ROE programme has the potential to reduce youth violence through targeting younger pupils, increasing prosocial behaviours, including empathy, and decreasing negative behaviours, including aggression (e.g. Santos et al., 2011; Latsch, Stauffer and Bollinger, 2017) and increased empathy (Wrigley, Makara and Elliot, 2015; Latsch, Stauffer and Bollinger, 2017)). Empathy is defined as both cognitive empathy and affective empathy, which are developed throughout the NEBT programme in tasks around perspective taking and emotional literacy.

When considering the evidence base for programmes like NEBT for reducing violence and crime, the Youth Endowment Fund (YEF) toolkits on anti-bullying and social skills are the closest in topic to the NEBT programme. These toolkits summarise the evidence base to date on the success of these types of programmes for violence and crime reduction and detail the tenets of successful programmes. For social skills programmes, evidence suggests that, on average, they are likely to reduce the number of children involved in crime by 32%. However, whilst universal programmes exist (that include young people regardless of their risk of involvement in violence and crime), targeted programmes (targeted to youth thought to be at risk of violence) tend to be more effective (Gaffney, Harrington and White, 2021). The evidence of the impact of anti-bullying programmes on crime and behaviour is less clear. What is known is that anti-bullying programmes can be effective in reducing bullying in school and that bullying in school is associated with later involvement in violence in subsequent years (Gaffney, Harrington and White, 2021). However, anti-bullying programmes tend to be most successful if they utilise a whole-school approach. Due to methodological issues impacting the quality of the existing evidence and the limited evidence available, the security of the findings for social skills interventions is higher (score of 4) than for anti-bullying programmes (score of 1) (see the YEF technical guide for further information on how security ratings are created).

The ROE programme (https://rootsofempathy.org/programs/roots-of-empathy/) was established in 1996 in Canada and has since been implemented in other countries, including the USA, New Zealand, Ireland, Wales, Northern Ireland, Scotland, Norway, Switzerland, Australia, Isle of Man, Japan, Korea, Germany, Costa Rica and the Netherlands. At the core of ROE is the assumption that empathy is innate and that the extent to which it develops is dependent upon the attachment relationships children build. ROE aims to facilitate this through repeated classroom visits with a parent and baby, thus providing a model of a secure infant and parent attachment.

Current evaluation in the context of previous evaluations of the ROE programme

There have been several previous evaluations of the ROE programme, with the majority employing a quasi-experimental design and with two (including in Northern Ireland) employing a randomised controlled trial (RCT) design (Santos et al., 2011; Wrigley, Makara and Elliot, 2015; Latsch, Stauffer and Bollinger, 2017; Connoly et al, 2018). These have been conducted in several countries, with the current evaluation being the first in England and Wales. Previous evaluations of the ROE programme have spanned a range of age groups and collectively demonstrated that the programme may lead to an immediate increase in teacher perceptions of pupils' prosocial behaviour and their understanding of infant development and a reduction in teacher-reported problem behaviour (Santos et al., 2011; Wrigley, Makara and Elliot, 2015; Latsch, Stauffer and Bollinger, 2017; Connolly et al, 2018).

The current evaluation of the NEBT programme (identical to the ROE programme but called NEBT only for the purposes of this trial) focuses nine- to 10-year-olds before their transition to secondary school. This evaluation builds on previous evaluations, including improving on previous methodological weaknesses (e.g. an over-reliance on teacher reports, which cannot be blinded due to the nature of the programme, cf. using direct observations and pupil reports; Schonert-Reichl et al., 2012; Latsch et al., 2017) and expanding on the geographical locations where robust evaluations of the ROE programme have taken place. The causal impact of NEBT on the social and emotional development of year five pupils (ages nine to 10 years) was estimated using a split-cohort clustered RCT design. Participating primary schools were drawn from throughout England and Wales, with a specific focus on areas of disadvantage.

The NEBT programme is often delivered in the timetable space allocated for the delivery of Personal, Social, Health and Economic education (PSHE) lessons. PSHE is a non-statutory subject in England, with schools having some flexibility in tailoring their curriculum to meet the needs of their pupils, with government-funded support from the PSHE Association (Department for Education [DfE], 2024). Whilst the content of the NEBT programme is thorough, it does not fully capture the entirety of PSHE needs, but it could complement existing curricula. Importantly, whilst schools often choose to use PSHE lessons to deliver the ROE programme, this isn't mandated by ROE. Schools are free to choose which sessions they deliver the NEBT programme in.

Changes to the evaluation and delivery schedule due to COVID-19

This evaluation is one of YEF's Launch Grant Round projects commissioned in 2019. Prior to the COVID-19 pandemic, YEF commissioned both the project and evaluation teams and did some early collaborative work on evaluation design. For NEBT, it was planned that 140 schools would be recruited to the trial using a community-based model. As the pandemic developed, many of the practical steps of setup had to be put on hold. The YEF adopted and agreed upon a range of responses to support each Launch Grant Round project, including pausing during school closure phases and restarting at a later time (DARE25 and LNK Educate), switching to online delivery (Transition Hub for Children Looked After), and reducing the quantity

of data collected from participants (Functional Family Therapy Gangs). Specifically for this project, changes agreed between the YEF, Sheffield Hallam University (SHU) and ROE included pausing the project until ROE could return to schools to commence delivery, changing to school-based instructors and splitting the cohort trial (rather than including 140 schools in one year). With regard to the last change, this meant that the initial plans for a two-year project and one-year evaluation in 2020/21 (recruitment in 2019/20) expanded to a five-year project with two one-year evaluation cohorts: cohort 1 in 2022/23 (recruitment in 2021/22) and cohort 2 in 2023/24 (recruitment in 2022/23). It is important to note that the pandemic had an impact on all three organisations involved in this work, the ROE, SHU and YEF. The length of the project meant that there have been changes in team members on all sides, with key individuals from the ROE and SHU present across all five years. Whilst every effort has been made to minimise the disruption this may have caused to the trial, the possibility of disruption needs to be acknowledged.

Although recruitment, delivery and evaluation of this project were postponed until after the peak of the pandemic, and the evaluation was not directly aimed at assessing the impact of the pandemic, it is clear that delivery of most school-based interventions was affected by the pandemic, which created methodological challenges for their evaluations.[1] It is difficult to quantify the full impact of the pandemic on the delivery of this intervention – beyond the general departure from the typical ROE model,[2] pupils, instructors, teachers and schools as a whole were experiencing a heightened state of stress during the pandemic and its aftermath. Increased levels of stress, alongside isolation during lockdown, have negatively impacted young people, evidenced by increased reports of mental ill-health in young people since the pandemic. Alongside this, pupil absence has remained problematic in schools, as parents' and pupils' attitudes towards attendance have changed, with this change thought to have had a greater impact in areas of socioeconomic disadvantage (Gibbons, McNally and Montebruno, 2025). It is important that the findings and conclusions of this evaluation are considered within this context.

Intervention

1. Named

Nurturing Empathy before Transition (NEBT)

2. Why

ROE designed the NEBT programme (just called the ROE programme outside the context of this trial), which aims to increase empathy and prosocial behaviour among school children in Year 5. This well-established programme involves bringing a parent and baby into the classroom as part of a structured programme of lessons focused on building empathy. It is described by the delivery partners as "an evidence-based, preventative intervention for primary school children that aims to reduce aggression, including bullying, and [increase] children's social and emotional competence". The programme is underpinned by the assumption that empathy is innate and that the extent to which it develops is

[1] https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/reporting-templates/Reporting_checklist_for_Covid-affected_evaluations_May2021.pdf?v=1749734292

[2] https://foundations.org.uk/toolkit/guidebook/roots-of-empathy/

dependent upon the attachment relationships children build. The programme teaches both the cognitive and emotional elements of empathy by encouraging pupils to identify the baby's feelings whilst they also reflect on their own feelings and the feelings of others, thus improving their emotional literacy. This improved emotional literacy, alongside witnessing the parent regulating the baby's emotions, enables children to better regulate their own emotions, leading to improvements in emotional regulation, resilience and wellbeing. This, in turn, leads to reduced aggression and an increase in prosocial behaviour. Please see the finalised logic model on page 17 and appendix D.

In line with the above theoretical assumptions, a previous RCT of the ROE programme in Northern Ireland found that the programme was well received in schools and that a positive effect on teacher-rated prosocial behaviour could be observed (Connoly et al., 2018). Furthermore, previous studies (mainly quasi-experimental designs) have highlighted that the programme may lead to decreased aggression (e.g. Santos et al., 2011; Latsch, Stauffer and Bollinger, 2017) and increased empathy (Wrigley, Makara and Elliot, 2015; Latsch, Stauffer and Bollinger, 2017). However, an evaluation of the ROE programme and its effects has not yet been conducted in England. This evaluation builds on previous research, including improving upon previous methodological weaknesses (e.g. an over-reliance on teacher reports) as well as expanding the geographical locations where robust evaluations of NEBT have taken place.

3. What (materials)

NEBT instructors were provided with instructor topic guides and supporting materials, which should have enabled them to teach the 27-session programme of NEBT in one academic school year. These had not been adapted and still contained language specific to Canada (adaptations were left to the instructors' initiative, as detailed later in the report). These were accessed via an iPad, and materials could also be downloaded and printed when required.

4. What (procedures)

Instructor training

Members of the ROE team trained instructors from participating schools, e.g. teaching assistants (TAs). Instructors received specialist training in how to deliver the NEBT programme through four days of face-to-face training (three initial days and one mid-year day).

NEBT delivery

For cohort 1, the NEBT intervention took place within four geographical regions (Yorkshire, Merseyside, East and West Midlands, and Greater London). To assist with meeting recruitment targets, this was later expanded to five areas, with schools in Wales included.

Following training, instructors delivered the NEBT programme in the nominated Year 5 classes at each intervention school. The NEBT programme consists of 27 sessions split into nine themes, with three sessions per month. The parent and baby attended one session per month, with a preparation and a debrief session on either side. Sessions are based around the following nine broad themes:

Theme 1 – Meeting the baby

Theme 2 – Crying

Theme 3 – Caring and planning for the baby

Theme 4 – Relationships

Theme 5 – Sleep

Theme 6 – Safety

Theme 7 – Communicating

Theme 8 – Who am I?

Theme 9 – Goodbye and good wishes

Sessions last for approximately 40–45 minutes. However, as the welfare of the parent and baby is of paramount importance, it is feasible that the parent and baby session may be shorter if needed. In addition, there are a number of safety messages that are included in sessions with pupils aged nine and older, and leaflets are sent home to families on topics such as smoking during pregnancy, never shaking a baby and Fetal Alcohol Spectrum Disorders.

5. Who (provider)

ROE is a children's charity whose mission is to build caring, peaceful and civil societies through the development of empathy in children and adults. ROE instructors were members of staff based in participating schools, mostly TAs or Special Educational Needs Coordinators (SENCOs), who were trained by ROE prior to the start of the intervention.[3]

6. How

The NEBT intervention is designed to be delivered face-to-face in the school setting to a whole class. Sessions took place when the schools felt it was appropriate within the timetable, but they usually replaced a PSHE lesson. ROE considers the programme (occasionally accompanied by a father) to contribute to many of the same learning objectives as PSHE. Each NEBT session has a specific lesson plan that the instructor follows step by step. The family visit focuses on guided observation, discussion and interaction with the parent and baby and, as such, is different each time. The pre- and post-visit sessions include questions and discussion, group work, art and storytelling.

7. Where

For cohort 1, NEBT ran in four geographical locations in the UK: Yorkshire, Merseyside, East and West Midlands, and Greater London.[4] For cohort 2, delivery was expanded to Wales. Schools were to be recruited from areas of social disadvantage, which ROE have classified as schools with over 21% pupil premium/Free School Meals (FSM) in any of the five broader geographical areas.

4 Recruitment originally focused on five smaller geographical areas (Doncaster, Birmingham, Northamptonshire, Nottingham and London) but this was changed to four larger areas (Yorkshire, Merseyside, the Midlands and Greater London) following initial recruitment difficulties.

8. When and how much

The NEBT intervention ran for one academic year but involved two Year 5 pupil cohorts: cohort 1 in 2022/23 and cohort 2 in 2023/24. In both cohorts, the intervention ran for around nine months of the year, with programmes typically starting in October and finishing in May/June. The original evaluation design planned an overall sample size of 140 from the combined cohorts (70 intervention and 70 control). The first cohort, from Autumn 2022 to Summer 2023, aimed to involve 60 schools (30 intervention and 30 control), and the second cohort, from Autumn 2023 to Summer 2024, aimed to involve 80 schools (40 intervention and 40 control). In each intervention school, one Year 5 class (approximately 30 pupils) received the intervention. Three sessions were delivered to the class each month. Recruitment was challenging, and 33 schools were randomised for cohort 1 (16 intervention and 17 control), and 54 were randomised for cohort 2 (30 intervention and 24 control), resulting in a total of 87 schools before attrition (46 intervention and 41 control).

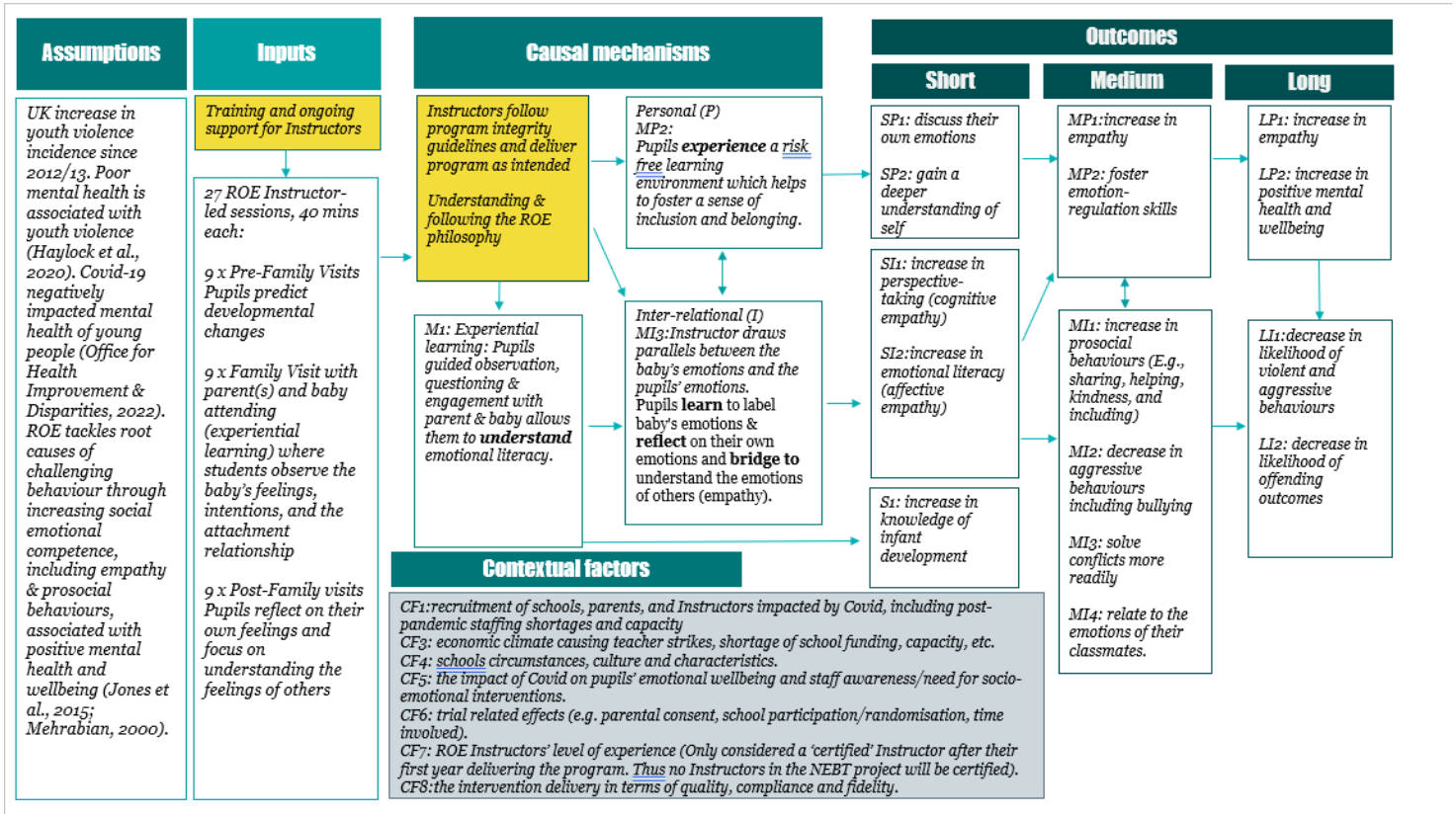


Figure 1 Logic model

The original ROE NEBT theory of change (ToC) was developed across several iterations with the evaluation team and the developers. The model details the programme inputs for both instructors (instructor inputs and mechanisms are indicated by the yellow boxes in the logic model) and pupils, causal mechanisms for change, and the short-, medium- and longer-term outcomes. The ROE NEBT programme has three causal mechanisms: experiential learning, inter-relational experiences and personal experiences. During the parent–baby sessions, guided observations and questioning help pupils understand the emotions of the baby and parent, thereby helping them develop emotional literacy. Risk-free learning allows pupils to feel safe to participate in an inclusive environment that fosters belonging, and the instructor is uniquely placed to draw parallels between the baby's emotions and those of pupils. This then enables pupils to label the baby's and their own emotions, helping them understand the emotions of others (empathy). Several short-term outcomes need to be realised to lead to the medium-term outcomes of increased empathy, emotional regulation and prosocial behaviour.

The long-term outcomes may not be met across the course of the programme, but they relate to sustained improvements in empathy, subsequent improvements in mental wellbeing, and reductions in the likelihood of violent and aggressive behaviours and offending outcomes. This ToC is revisited throughout the findings and in the conclusion and is updated with knowledge gained across the course of the evaluation.

Evaluation objectives

Primary research question

1. (Impact evaluation – primary outcome) What is the impact of the NEBT programme on self-reported behavioural difficulties (BD) of primary school–aged children when compared to a 'business as usual' control?

Secondary research questions

2. (Impact evaluation – secondary outcome) What is the impact of the NEBT programme on self-reported emotional difficulties of primary school–aged children when compared to a 'business as usual' control?
3. (Impact evaluation – secondary outcome) What is the impact of the NEBT programme on self-reported affective empathy of primary school–aged children when compared to a 'business as usual' control? (secondary outcome)?
4. (Impact evaluation – secondary outcome) What is the impact of the NEBT programme on self-reported cognitive empathy of primary school–aged children when compared to a 'business as usual' control? (secondary outcome)?
5. (Impact evaluation – secondary outcome) What is the impact of the NEBT programme on teacher-reported prosocial behaviour of primary school–aged children when compared to a 'business as usual' control? (secondary outcome)?
6. (Impact evaluation – secondary outcome) What is the impact of the NEBT programme on teacher-reported school behaviour of primary school–aged children when compared to a 'business as usual' control? (secondary outcome)?

Exploratory

7. What is the difference in teacher-reported emotional problems (Strengths and Difficulties Questionnaire [SDQ] subscale) between the intervention group and a 'business as usual' control?
8. What is the difference in conduct problems (SDQ subscale) between the intervention group and a 'business as usual' control?
9. What is the difference in peer relationship problems (SDQ subscale) between the intervention group and a 'business as usual' control?
10. What is the difference in hyperactivity (SDQ subscale) between the intervention group and a 'business as usual' control?

Implementation and Process Evaluation

The implementation and process evaluation (IPE) aimed to answer the following research questions:

11. What are the key factors which influence the successful delivery of the NEBT programme in years 1 and 2?
12. What are the perceptions of pupils, teachers, deliverers and instructors about the effectiveness of the programme in years 1 and 2?
13. What fidelity issues are observed during years 1 and 2 of the trial?
14. What does the trial indicate about scalability?

The ROE NEBT protocol can be found here, and the statistical analysis plan here.

Ethics and trial registration

The ROE NEBT evaluation underwent a full independent review and approval through the university ethics committee (Ref ER19810112). All researchers visiting schools were experienced in this role, held a current enhanced Disclosure and Barring Service (DBS) certificate and had completed additional National Society for the Prevention of Cruelty to Children (NSPCC) safeguarding training[5].

Agreement to participate in the trial was obtained in stages. Firstly, interested schools were given a memorandum of understanding (MoU), which they were asked to read, sign and return as confirmation of their agreement to take part. Signatures were gained from the head teacher and a school-based lead (volunteered by the school to coordinate all evaluation activities). In 2023/24, for the second cohort, a checklist was also sent with the MoU to check understanding. This was agreed in a lessons learnt meeting at the end of the first cohort because of high school-level attrition. Once schools had signed up, parents were given the opportunity to opt their child out of the evaluation activities, such as surveys. For those

[5]https://learning.nspcc.org.uk/training/introduction-safeguarding-child-protection Child protection: an introduction (3.0) - elearning

pupils who did not opt out, consent for participation was sought from pupils themselves when collecting baseline and endpoint outcome data and for any qualitative field work.

The ROE NEBT is registered on the ISRCTN registry (ISRCTN98490275).

Right to withdraw and consent

SHU provided information sheets to pupils, staff and ROE parents to make them aware of the expectations underpinning their involvement. The sheets made clear that participating pupils/school staff were free to withdraw from data processing as part of the evaluation. Parents were able to withdraw their child from data processing either by visiting an online site or by returning a slip to school. Pupils and teachers were asked to provide consent at the start of their questionnaires at both baseline and endpoint. For the parent–baby visits to the school, an information sheet was supplied. The parent was advised that they were not part of the evaluation and that no data was being collected or processed. ROE was responsible for the safety and wellbeing of the parent–baby dyad.

Given ethical considerations, separate permission was sought from school staff and pupils for additional data collection, e.g. primary and secondary outcomes. For any qualitative data, verbal consent was also taken before proceeding with any interview or focus group.

Data protection

SHU strictly complies with all current legislation in relation to data processing and storage.

Materials relating to consent and data processing (including the MoU) can be found in Appendix C.

SHU was the data controller for the data collected as part of the NEBT programme evaluation until the evaluation finished. Following completion of the trial and submission of the report, the trial data was sent to the DfE (at which point SHU ceased to be responsible for the data). At the DfE, it was pseudonymised and transferred to the secure archive, held by the Office for National Statistics in their Secure Research Service (SRS). Once the data was transferred to the SRS, the YEF became responsible for the data. No pupils are individually identifiable in the archived data, and the archived data will be kept indefinitely. Further information on YEF's data archive can be found below.

The processing of personal data through the NEBT project evaluation was defined under General Data Protection Regulations (GDPR) as a specific task in the public interest. The legal basis for processing your personal data is 'Public Task' (Article 6 (1) (e)). https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/public-task/

SHU was responsible for retrieving and processing data, including management information, such as pupil attendance and demographics. For the retrieval and transmission of data, SHU Zend To (a secure transfer method) was used. The nominated school-based leads were sent a standard pro forma to complete with pupils' names, dates of birth and unique pupil reference numbers. Leads were then required to send back the complete pro forma to SHU via SHU ZendTo. Pupil names, dates of birth and unique pupil reference numbers were used for the matching of data only. Once the data was matched and ready for analysis, these identifiers were removed. No pupil/staff/school names were used or will be used in any report arising from the research. The data was stored securely in a password-protected folder accessible only to members of the evaluation team. SHU was responsible for the qualitative and quantitative analysis of all the data collected.

All researchers visiting schools held a current enhanced DBS certificate. SHU researchers undertook case study visits at times convenient for the school.

Communication

SHU was the point of liaison for schools on anything related to the evaluation throughout the course of the evaluation. SHU liaised with ROE and the YEF throughout the course of the evaluation.

Project team/stakeholders

Dr Sarah Jane Reaney-Wood – principal investigator (PI), trial statistician and manager

Sean Demack – senior statistical advisor, principal investigator (PI) and trial statistician (cover for parental leave)

Dr Pangiota Blouchou – data collection researcher (maternity cover)

Dr Josephine Booth – IPE lead and quality assurance

Jessica Benson-Egglenton – IPE researcher

Eleanor Byrne – IPE researcher

Due to the length of this project and the interruption due to COVID-19, there were several changes to the evaluation team. Dr Sarah Reaney-Wood has been part of the evaluation team since inception (with a quality assurance role during parental leave) and has been the principal investigator, trial statistician and manager since 2022; prior to 2022, Sarah was the trial statistician and manager. Following her return from parental leave, Sarah continued as the PI, joined Sean Demack in conducting the analysis and led the write-up of the impact evaluation and this report.

Bernadette Stiell was the initial PI on the project and left the team when she left SHU at the end of February 2024. Sean Demack took on the PI role while Sarah was on parental leave, and Dr Josephine Booth and Eleanor Byrne joined the team to lead on the IPE, and they carried out fieldwork, analysis and report write-up. Ben Willis also undertook fieldwork, and Jessica Benson-Egglenton undertook IPE analysis and write-up. Dr Giota Blouchou joined the team in April 2024 to help with endpoint data collection.

The extent of change to the evaluation team has been notable; a larger-than-usual number of SHU academics has been involved during the five years of the project, with two of these people (Sean Demack and Sarah Reaney-Wood) having consistent involvement. Some of this can be accounted for by the timescale of the project and the impact of Covid-19.

The design of the trial was conducted by SHU. However, during the process of designing the trial, several meetings were held with the YEF and ROE to ensure the design would be feasible for ROE. When designing the trial, SHU worked within the parameters laid out by the YEF, for example, when considering outcome measures.

Methods

Trial design

Table 1. Trial design

Trial design, including the number of arms		*Split-cohort, two-armed, cluster randomised controlled trial*
Unit of randomisation		*School*
Stratification variable(s) (if applicable)		*Geography*
Primary outcome	**Variable**	*Behavioural difficulties*
	Measure (instrument, scale, source)	*Behavioural difficulties as measured using the Self-report Me and My Feelings (M&MF) questionnaire behavioural difficulties subscale (Deighton et al., 2012) [0 to 12 scale]*
Secondary and exploratory outcome(s)	**Variable(s)**	*Emotional difficulties, cognitive empathy, affective empathy, teacher-reported behaviour difficulties, prosocial behaviour, hyperactivity, emotional problems, conduct problems and peer problems*
	Measure(s) (instrument, scale, source)	*M&MF emotional difficulties subscale [0 to 20 scale]* *Basic Empathy Scale (BES) cognitive empathy subscale [9 to 45 scale]* *BES affective empathy subscale [11 to 55 scale]* *Teacher Strengths and Difficulties Questionnaire (SDQ) total difficulties subscale [0 to 40 scale]* *Teacher SDQ prosocial score [0 to 10 scale]* *Teacher SDQ hyperactivity score [0 to 10 scale]* *Teacher SDQ emotional problems score [0 to 10 scale]* *Teacher SDQ conduct problems score [0 to 10 scale]* *Teacher SDQ peer problems score [0 to 10 scale]*
Baseline for primary outcome	**Variable**	*Behavioural difficulties at baseline*
	Measure (instrument, scale, source)	*M&MF behavioural difficulties subscale [0 to 12 scale]*
Baseline for secondary and exploratory outcome(s)	**Variable**	*Emotional difficulties, cognitive empathy, affective empathy, teacher-reported behaviour difficulties and prosocial behaviour*
	Measure (instrument, scale, source)	*M&MF emotional difficulties subscale [0 to 20 scale]* *BES cognitive empathy subscale [9 to 45 scale]* *BES affective empathy subscale [11 to 55 scale]* *Teacher SDQ total difficulties subscale [0 to 40 scale]* *Teacher SDQ prosocial score [0 to 10 scale]* *Teacher SDQ hyperactivity score [0 to 10 scale]* *Teacher SDQ emotional problems score [0 to 10 scale]* *Teacher SDQ conduct problems score [0 to 10 scale]* *Teacher SDQ peer problems score [0 to 10 scale]*

The NEBT trial is an efficacy trial that is formed by combining data from two cohorts that ran in 2022/23 and 2023/24 (a split-cohort design). This evaluation was commissioned prior to the Covid-19 pandemic in 2019 as part of the YEF's first grant round, but it was postponed until 2021/22. Since its inception, the YEF has developed a number of methodological expectations to ensure that equity and diversity are embedded throughout all evaluation processes. However, these were not in place during the first grant round, and as such, the focus was on collecting information directly pertaining to the research questions, with the overarching aims of minimising burden and cost. This is a limitation of this evaluation, and this is addressed further in the discussion section. Post-Covid-19 recruitment difficulties led to the decision to split the efficacy trial over two years. Both smaller evaluations adopted a two-armed design, with randomisation at the school level. Randomisation at the school level was chosen over randomisation at the class level due to

issues with spillover, which would be a significant risk to the validity of the trial. With class-level randomisation, there is a risk of spillover when pupils from a control class interact with those from an intervention class. This can occur indirectly, through control pupils benefiting from behavioural improvements in intervention participants or directly through talking about what has been learnt. This risk was deemed to be greater than the benefit of class-level randomisation (explanatory power). Prior to randomisation, schools identified a teacher and their Year 5 class of pupils who would receive the NEBT programme if the school was randomised to the intervention group. For schools with more than a single form entry, only one class could be included in the trial due to delivery capacity. Additionally, prior to randomisation, schools completed a baseline pupil survey to collect the following questionnaires: Me and My Feelings (M&MF; Deighton et al, 2013) and the Basic Empathy Scale (BES; Joliffe and Farrington, 2006) and a teacher survey to collect teacher SDQs (Goodman, 2001).

At the protocol stage, the aim was to recruit 60 schools for cohort 1 and 80 schools for cohort 2. For the minimal detectable effect size (MDES) calculations, a school was assumed to have a class size of approximately 30 pupils, resulting in a total sample of 4,200 pupils across 140 schools for the combined cohort. However, the number of recruited schools differed from these initial aims, and this is discussed in subsequent sections of this report. ROE was responsible for the recruitment of schools, with the evaluation team managing the randomisation process to maintain independence between the two processes and minimise imbalance or bias in treatment assignment.

Control schools operated under business as usual during the trial and received an incentive payment of £400 to recognise their commitment and effort towards participating in the trial. This payment was made in two smaller payments of £200. The first was after completion of the baseline data collection, and the second was after completion of the endpoint data collection. Where control schools did not provide the data required, the control payment was withheld.

Randomisation was conducted by the evaluation team at SHU. Schools were randomised following baseline testing and prior to the programme starting. Randomisation was stratified by geographical area so that around half of the schools in each area were randomised to the NEBT intervention or control groups. One difference between the two cohorts was the inclusion of primary schools in Wales for cohort 2. This was done following a discussion between the YEF, ROE and SHU, during which it was pragmatically agreed on to help maximise the recruited sample for cohort 2.

The primary outcome measure was the BD subscale from the M&MF questionnaire (Deighton et al., 2013). The secondary outcomes were the emotional difficulties subscale of the M&MF questionnaire, the subscales of affective empathy and cognitive empathy from the BES (Joliffe & Farrington, 2006) and the teacher SDQ (Goodman, 2001) total difficulties and prosocial subscales. An exploratory analysis of the teacher SDQ emotional problems, conduct problems, peer problems and hyperactivity subscales was also undertaken. Data for all measures was collected either directly from pupils (M&MF and BES) or directly from teachers (SDQ).

Participant selection

The ROE NEBT intervention was a whole-class intervention and was not targeted towards a particular demographic of pupils. However, to be eligible, schools had to be within an area of disadvantage (assessed by having greater than 21% FSM) within the following geographic regions: Yorkshire, Merseyside, East and West Midlands, and Greater London in cohort 1, with the inclusion of Wales in cohort 2. The recruitment and screening of schools and instructors was conducted by ROE.

A number of exclusion criteria were in place to ensure that certain types of schools were not recruited to the trial. These were:

- Schools with prior experience of the ROE programme (this is the name of the NEBT programme outside the context of this evaluation)
- Private schools, special schools or pupil referral units / alternative provision

Not all schools were one form entry. For those with multiple classes, the school chose which class to select; they were not guided to do so on the basis of pupil needs, but some schools may have done this. Classes were chosen prior to randomisation, reducing the risk of bias. The intervention was delivered by a member of school staff, in school, during a timetabled lesson. To be eligible to deliver the intervention, an individual had to work at the school but not be a class teacher. In addition, they had to be willing and able to attend the training course run by ROE to be trained to deliver the intervention. Typically, ROE works only with regions and schools that have actively requested the ROE programme and where the necessary supports are firmly in place. This was not the case for this trial, as ROE was actively recruiting schools for the evaluation rather than schools actively requesting.

Outcome measures

Baseline measures

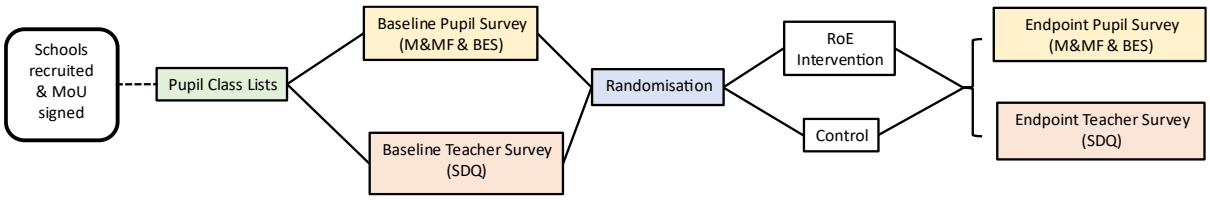


Figure 2. Data collection process for outcomes measures

At baseline, all of the primary, secondary and exploratory outcomes listed below were collected. Data collection took place online, with pupils completing the primary and secondary outcome measures in school in a computer suite or other appropriate room. An FAQ document was provided for teachers. It aimed to provide answers to commonly asked questions pertaining to the outcome measures, and details about the project PI were available should anyone need to make contact. Many studies use the same or similar outcome measures without invigilation or without a researcher being present in each class. Having a researcher present would be financially burdensome for evaluations. However, teachers were able to contact the project PI, who is experienced in these outcome measures, should clarification be needed for

themselves or the pupils. Previous YEF and other What Works Centre evaluations have utilised the same or a similar approach for outcome testing.

Where available, the outcome measures were provided in Welsh to meet the needs of pupils and teachers. This also included translating information sheets at the start of questionnaires into Welsh if required. None of the schools elected to use the Welsh translations.

Primary outcome

The primary outcome for the NEBT trial is self-reported BD as measured using the M&MF questionnaire (Deighton et al., 2013). M&MF is a 16-item school-based measure of child mental health suitable for children as young as eight years of age, covering two domains: emotional difficulties and BD (Deighton et al., 2013). The BD score for M&MF was used as the primary outcome, with the emotional difficulties subscale as a secondary outcome.

The self-report M&MF BD scale was selected as the primary outcome because previous studies have shown a reduction in BD (aggression) following the ROE programme (Schonert-Reichl et al., 2012). The NEBT programme aims to tackle the root causes of challenging behaviours by increasing social and emotional wellbeing. The evidence-based short- and medium-term outcomes detailed within the NEBT ToC relate to an increase in emotional literacy and a decrease in violent and aggressive behaviours. In addition, whilst previous studies have reported an increase in prosocial behaviour, the only way of assessing this was indirectly through teacher reports using the SDQ, as Year 5 pupils are too young to complete the self-report SDQ (which is only validated from age 10 to 16). As such, prosocial behaviour was considered unsuitable as a primary outcome. Previous research into the impact of the ROE programme has been dependent on teacher reports, and this has been cited as a limitation. Use of the M&MF scale, specifically BD, as the primary outcome keeps this evaluation comparable with other YEF evaluations (by using an age-appropriate YEF core measure) and in line with the YEF strategy to provide evidence to prevent youth crime whilst also overcoming the methodological limitations of previous evaluations.

The psychometric properties of the M&MF are broadly good. This is demonstrated by good internal consistency (Deighton et al., 2013; Patalay et al., 2014), construct validity (Deighton et al., 2013), convergent validity with the subscales of the SDQ (Deighton et al., 2013) and discriminant validity (Deighton et al., 2013).

Table 2. Me and My Feelings (M&MF) scoring

M&MF Statement	Response/Coding
I get very angry I lose my temper I hit out when I am angry I do things to hurt people I break things on purpose	Never=0, Sometimes=1, Always=2
I am calm [Reverse Coded]	Never=2, Sometimes=1, Always=0

M&MF behavioural difficulties scoring: calculated from summing six of the 16 M&MF items.

The M&MF BD subscale ranges from 0 to 12, where a higher score indicates higher (self-reported) BD. Scores were treated as continuous[6] in this evaluation.

Data for the primary outcome was collected at baseline, pre-randomisation (September/October 2022 for cohort 1 schools and September/October 2023 for cohort 2 schools) and at endpoint (June/July 2023 for cohort 1 schools and June/July 2024 for cohort 2 schools.

M&MF is publicly available at Me and My Feelings (corc.uk.net)

Secondary and exploratory outcomes

A total of nine secondary and exploratory outcomes were used: self-reported emotional difficulties (from the M&MF), self-reported affective and cognitive empathy (from the BES) and six teacher-reported measures (from the SDQ): total difficulties, prosocial, emotional problems, conduct problems, peer problems and hyperactivity.

Empathy

Empathy was assessed at baseline and endpoint using the BES (Joliffe et al., 2006). The BES is a pre-validated scale to assess empathy in young people aged 9–18, focusing on cognitive and affective empathy. Whilst other existing measures focus on the three elements of empathy (rather than two as with the BES), the age group of the young people in the NEBT trial makes the BES the most appropriate.

As per the ToC, the ROE programme aims to improve empathy, prosocial behaviour and wellbeing. As such, the BES is an age-appropriate tool for assessing this, using pupil self-report rather than parent or teacher perceptions.

Previous validation of the BES included exploratory factor analysis that demonstrated a two-factor structure, cognitive empathy and affective empathy, and this was confirmed using confirmatory factor analysis. The BES is a 20-item questionnaire with acceptable internal consistency (77–87). Confirmatory factor analysis showed adequate model fit and test–retest were between r=0.54 and r=0.70 (Jolliffe and Farrington, 2006; D'Ambrosio et al., 2009).

SDQ – behaviour

Pupil behaviour and mental health were also assessed using the teacher version of the SDQ (Goodman, 2001). The SDQ, teacher version, is a 25-item scale used to assess behaviour in the school context in 4–16-year-olds.

The total difficulties score for the SDQ is calculated by combining responses to 20 items, whilst the prosocial score is calculated by combining responses to the remaining five items. The total difficulties score can be unpacked into four SDQ subscales (conduct problems, emotional problems, peer problems and hyperactivity), each on a 0 to 10 scale.

[6] Further details on the cut-offs for the M&MF scale which indicate the severity of behavioural and/or emotional difficulties scores pertain to can be found here: https://www.corc.uk.net/outcome-experience-measures/me-and-my-feelings-mmf/

The SDQ is commonly used in clinical assessments and has become increasingly popular as an outcome measure in a variety of evaluations. Furthermore, at the time this trial was designed, the SDQ was a core measure for the YEF. From validation studies, the SDQ has good psychometric properties with good internal consistency (Cronbach's alpha = 0.63–0.87) and test–retest reliability of r=0.71–0.81. In addition, the SDQ is a multi-respondent measure with inter-rater reliability of 0.37–0.58 for self-report and parent-report and 0.24–0.39 for self-report and teacher-report measures (Goodman, 1997; Goodman, Meltzer and Bailey, 1998; Goodman, 2001).

The teacher who completes the SDQ needs to be familiar with the pupil they are completing the SDQ for, as they are being asked to assess that pupil's regular behaviour. Additionally, it is important that the same teacher completes the SDQ at baseline and endpoint. The trial was designed to facilitate this – that is, data collection was organised at the start of the school year to ensure the same teacher completed at both baseline and endpoint.

Sample size

Sample size was determined prior to recruitment taking place by calculating MDESs and was undertaken by the evaluation team. Specifically, the MDES is the estimated smallest difference (between the intervention and control groups) in the primary outcome that the design could detect as being statistically significant ($p<0.05$, two-tailed) with a statistical power of 0.80 or greater. The MDES is presented as a standardised Hedges g effect size in units of standard deviations.

The power analyses drew on data available from a previous RCT (Connolley et al., 2018), which reported effect sizes of +0.20, p=0.05 for prosocial behaviour, meaning the intervention group was rated as more prosocial by their teachers, and −0.16, p=0.06 for difficult behaviour, meaning the intervention group exhibited less difficult behaviour than the control group. We therefore looked to design an evaluation with enough sensitivity to detect similar effect sizes. Discussions were then held between the evaluator, delivery team and funder team to ensure that the MDES calculations were based on feasible estimates that accounted for practical constraints and delivery capacity for NEBT.

MDES calculations were estimated using the formula set out by Bloom et al. (2007) below and cross-checked using the Powerup! Software (Dong et al., 2013; sheet CRA2_2r).

$$MDES \sim \left(\frac{P}{1-P}\right) M_{(J-m-2)} \sqrt{\frac{ICC_2(1-R_C^2)}{J} + \frac{(1-ICC_2)(1-R_R^2)}{Jn}}$$

Where:

- P is the proportion of schools/clusters allocated to the intervention group, set at 0.50 for protocol and 0.53 at the randomisation stage (46 of the 87 schools). Across both cohorts, 48 of the 87 schools randomised provided baseline and outcome data for the primary outcome, 23 of which were in the NEBT intervention group (P=0.48 at analysis stage).
- ICC_2 – Cluster (school)-level Intra-cluster Correlation Coefficient (ICC) (proportion of variance in the outcomes between schools). This was set between 0.10 and 0.20 at the protocol stage and at 0.03

at the randomisation stage, drawing on cohort 1 for this estimate. At the analysis stage, with the combined data from cohorts 1 and 2, the unconditional school ICC was observed as 0.04.

- R_C^2 is the covariate explanatory power at the cluster (school) level. This was set at 0.06 at the protocol stage and 0.58 at the randomisation stage (drawing on cohort 1) and was observed as 0.53 in the final intention to treat (ITT) analysis of the combined data.
- R_R^2 is the residual (within-school, between-pupils) covariate explanatory power. This was set at 0.25 at the protocol stage and 0.27 at the randomisation stage (drawing on cohort 1) and was observed as 0.27 in the ITT analysis of the combined data.
- J is the total number of schools in the evaluation. This was set at 140 at the protocol stage, 87 at randomisation and 48 at the analysis stage.
- n is the number of pupils per school. This was set at 30 at the protocol stage, 19 at randomisation and 13 at the analysis stage in response to average pupil numbers at these stages.
- m is the number of cluster-level covariates included in the impact analyses. This was set at 7 (group membership, baseline measure at school and pupil levels, four dummy variables for the five geographical areas).
- M is the t-distribution multiplier, and this has (J-m-2) degrees of freedom.

Table 3. Sample and minimum detectable effect sizes (MDES)

Whole trial (cohorts 1 and 2 combined)		Protocol: planned sample size at the start of the evaluation	Randomisation: Drawing on cohort 1 for ICC and correlation estimates.	Intention to treat analysis
MDES		0.16–0.22	0.14	0.22
Pre-test/post-test correlations	Level 1 (participant)	0.50 (R^2=0.25)	0.52	0.52
	Level 2 (cluster)	0.25 (R^2=0.06)	0.76	0.73
Intracluster correlations (ICCs)	Level 2 (cluster)	Low=0.10 High=0.20	0.03	0.04
Alpha[7]		0.05	0.05	0.05
Power		0.80	0.80	0.80
One-sided or two-sided?		Two-sided	Two-sided	Two-sided
Number of clusters[8]	Intervention	70 schools	46 schools	23 schools
	Control	70	41	25
	Total	140	87	48
Number of participants	Intervention	2,100	910	314
	Control	2,100	752	330
	Total	4,200	1,662	644

Estimates of the ICC and pre-/post-test correlations were based on previous research using the M&MF scale (Deighton et al., 2018; Humphrey & Panayiotou, 2022)[9]. Post-hoc observed ITT MDESs were in the range of

[8] Please adjust as necessary e.g., for trials that are randomised at the setting, practitioner or participant level.

[9] past ICCs may not be comparable given some focused on targeted cohorts rather than universal cohorts like in this evaluation

those reported in the protocol, despite having 92 fewer schools and randomisation being undermined. The randomisation and analysis MDES estimates are illustrative only (as they assume missingness is random). Gains in sensitivity were bought by a school ICC that was much lower than originally estimated and a higher school-level explanatory power.

A two-level clustered design was considered appropriate due to the assumption that the NEBT intervention would be delivered in the same way across all geographical areas, and the decision to block by geographical area was only done to aid intervention delivery rather than because of suspected distinct differences between the areas. Geographical areas will be included in the analysis as school-level covariates.

No corrections for multiple testing have been made, as the trial is a two-armed RCT powered only for the primary outcome, with only one primary outcome for the pooled data for cohorts 1 and 2. As such, correction for multiple testing is not needed. Follow-on sensitivity analyses will explore the impacts in cohorts 1 and 2 separately.

Randomisation

Randomisation was conducted by the principal investigator. Schools were randomised following baseline testing and prior to the programme starting. Randomisation was stratified by geographical area so that around half of the schools in each area were in the intervention group and half in the control groups. One difference between the two cohorts was the inclusion of primary schools in Wales for cohort 2. This was done following a discussion between the YEF, ROE and the evaluator and was pragmatically agreed to help maximise the recruited sample for cohort 2. In addition, the allocation ratio throughout cohort 1 and cohort 2 randomisation had been one-to-one; towards the end of the randomisation period for cohort 2, the evaluation team ended up facilitating the rolling recruitment and randomisation of schools in very small batches. As a consequence, we could not ensure the one-to-one balance between groups.

Randomisation was conducted in batches for both years of the trial. As there was such a tight timeframe between randomisation and training starting (for those schools randomised to the intervention), randomisation in batches maximised the time available, meaning that those schools that had returned all the evaluation data needed could be informed of their allocation more quickly.

Schools were listed in an Excel spreadsheet within their geographical locations. Schools were assigned a random number using the RAND() function in Excel and then sorted in ascending numerical order. The cells were then assigned either a ‘1’ or a ‘2’ by alternating (1=ROE intervention, 2=control). Dr Sarah Reaney-Wood informed all schools of their allocation, and the ROE then picked up communication with intervention schools. The process effectively blinded the delivery team to the allocation of schools until schools had been informed.

Statistical analysis

Primary analysis

The M&MF questionnaire (Deighton et al., 2013) is a 16-item school-based measure of child mental health, suitable for children aged 8–11. The BD subscale was used as the primary outcome.

The primary analysis answered research question 1 (RQ1); multi-level linear regression models were constructed that acknowledged that pupils are clustered in schools, as specified in equation P1 below.

$$[P1] \quad Y_{ij} = \beta_0 + \beta_1 Group_j + \beta_2(X_{ij} - \bar{X}_j) + \beta_3(\bar{X}_j - \bar{X}) + \beta_{4..7}[Region]_j + \varepsilon_{ij}$$

Where

- Y_{ij} is the primary outcome (pupil-level endpoint M&MF BD score).
- $Group_j$ is a binary variable that identifies trial arm (1=NEBT or 0=control), $\beta_1$is the regression coefficient that is used to estimate the effect size (see Appendix A).
- X_{ij} is the pupil-level baseline M&MF BD score (which is centred around the mean score for their school), β_2 is the regression coefficient.
- $\bar{X}_j$ is the school-level mean (which is centred around the overall school-level grand mean), β_3 is the regression coefficient.
- $[Region]_j$ represents a collection of four school-level binary dummy variables used to identify the five regions in which schools were clustered, $\beta_{4...7}$ are the four regression coefficients for the dummy variables.

The main analysis combined data for the two cohorts, whilst follow-on sensitivity analyses examined the impact for cohorts 1 and 2 separately.

Secondary analysis

The secondary outcomes are the M&MF emotional difficulties (RQ2), affective empathy (RQ3) and cognitive empathy (RQ4) subscales of the BES (pupil self-report) and six scales from the SDQ (teacher report): prosocial behaviour (RQ5), total difficulties (RQ6), emotional problems (RQ7), conduct problems (RQ8), peer problems (RQ9) and hyperactivity (RQ10).

<u>BES</u>

The BES (Jolliffe & Farrington, 2006) is a 20-item questionnaire that assesses cognitive and affective elements of empathy. Both affective empathy and cognitive empathy are included as secondary outcomes. Models were conducted as detailed for the primary outcome analysis above to answer RQ3 and RQ4. Within these models, the baseline affective empathy score OR cognitive empathy score was included as a covariate as appropriate.

<u>The SDQ</u>

The teacher SDQ (Goodman, 2001) is a 25-item behavioural screening questionnaire for 4–16-year-olds. The SDQ is being used by the YEF across its projects to create consistency and comparability between different evaluations. A similar approach to constructing models for the two teacher SDQ outcomes was taken as specified for the primary outcome to answer RQ5, RQ6 and exploratory RQ7 to RQ10. Within these models, baseline total SDQ and baseline SDQ subscales were used as covariates (as appropriate). The exploratory analyses of the SDQ total difficulties subscales adopted the same approach.

Analysis in the presence of non-compliance

In experimental design, compliance describes whether the actual intervention (NEBT) coincides with the assigned group. Full compliance describes when all participants in the intervention group and none of those

in the control group receive the intervention. Noncompliance is when some participants in the intervention group do not receive the intervention and/or when some participants in the control group receive the intervention. To evaluate the impact of NEBT, we have assumed compliance to be one-sided. Specifically, the clustered RCT design means that it is reasonable to assume that none of the control group received the intervention, so compliance relates solely to pupils in the schools randomised to the NEBT intervention.

As specified in the Statistical Analysis Plan (SAP), compliance with the NEBT programme was defined by ROE using three dimensions: instructor training, schools delivering the NEBT curriculum and pupils attending the NEBT lessons. Specifically, to be considered compliant, the following conditions need to be met:

- Instructors attend all four NEBT training sessions (=1) or not (=0).
- Schools deliver at least eight of the nine NEBT themes during the trial period (=1) or not (=0).
- A pupil attends all three lessons in at least eight of the nine NEBT themes (=1) or not (=0).

Overall compliance is when all three conditions are met, and this is operationalised at the pupil level to identify pupils in NEBT schools who are compliant (= 1) or noncompliant (= 0). We also undertook some descriptive analyses to examine how the three separate compliance conditions were associated with the primary outcome (M&MF BD).

The ITT analysis of the primary outcome estimates the impact (on M&MF BD) of being assigned to the NEBT intervention rather than the control group. However, because there may have been different levels of engagement with the NEBT intervention for instructors, schools and pupils (not as intended), the ITT estimate could potentially be different to the impact for pupils who received the NEBT programme as intended.

The Complier Average Causal Effect (CACE) analysis has two assumptions. First, both treatment and control groups have the same probability of non-compliance. Second, the exclusion restriction assumes that being offered the intervention will have had no direct impact on the primary outcome; instead, the impact is fully mediated by compliance with the intervention (Raudenbush and Bloom 2015). The first assumption should theoretically be met through randomisation. However, given the notable attrition experienced, the integrity of randomisation (and hence the first assumption of CACE) is undermined. The second assumption splits the ITT sample into two: one which includes pupils identified as compliant and the second which includes pupils who are not compliant. The CACE analysis aims to reveal the impact of the NEBT programme for the subsample of pupils identified as meeting the three compliance conditions. However, it remains possible that some engagement, even if below the thresholds for compliance, may be sufficient to achieve an impact.

We applied CACE analysis (Gerber and Green, 2012) to estimate the impact of the NEBT programme for pupils identified as compliant. A two-stage least squares regression (2SLS) CACE analysis was used: the first stage estimated compliance as described in equation X.1

$$[X.1] \quad C_{ij} = \beta_0 + \beta_1 Group_j + \beta_2 (X_{ij} - \bar{X}_j) + \beta_3 \left(\bar{X}_j - \bar{X}\right) + \beta_{4..7} [Region]_j + \varepsilon_{ij}$$

Where C_{ij} is a pupil-level binary variable that distinguishes between pupils identified as compliant (=1) and pupils identified as non-compliant (=0). β_0 is the constant term. $Group_j$ is the school-level group membership variable (NBT=1; control=0) and $\beta_1$is the coefficient for group membership. $(X_{ij} - \bar{X}_j)$ is the pupil-level baseline M&MF BD score centred around their school's mean and β_2 is the coefficient. $\left(\bar{X}_j - \bar{X}\right)$ is the mean school-level baseline M&MF BD score centred around the school-level grand mean, and β_3 is the coefficient. $[Region]_j$ is the collection of four school-level binary dummy variables to account for the

geographical stratification across five regions (London, Merseyside, Midlands, Yorkshire and Wales). ε_{ij} is the error term.

The predicted values of C_{ij} from the first stage model are used in place of group membership in the second stage model to obtain the CACE estimate, as described in equation X.2:

$$[X.2] \quad Y_{ij} = \beta_0 + \beta_1 C'_{ij} + \beta_2 (X_{ij} - \bar{X}_j) + \beta_3 \left(\bar{X}_j - \bar{X}\right) + \beta_{4..7}[Region]_j + \varepsilon_{ij}$$

Where Y_{ij} is the M&MF BD primary outcome, and C'_{ij} is predicted compliance. CACE will be estimated using β_1, standardised into an effect size by dividing by the standard deviation of the M&MF BD primary outcome.

These two stages are included in the STATA 'ivregress' command. Standard errors of estimates will be adjusted for clustering of pupils in schools using the 'vce (cluster robust)' subcommand. The use of instrumental variable models to estimate CACE assumes compliance to be an endogenous variable. This assumption was examined by running the 'estat endogenous' postestimation command for STATA 'ivregress' to obtain a test for endogeneity using an F-test[10]. If compliance is found to be endogenous, this supports the need for instrumental variables with 2SLS, as specified above. However, if compliance is found to be exogenous, this approach is not necessary. The alternative approach to estimating CACE if compliance is found to be exogenous is specified in the SAP and shown in equation Z.1

$$[Z.1] \; CACE\ estimate = \frac{\text{ITT estimate}}{\text{proportion of pupils identified as compliant}}$$

Following the completion of evaluation data collection activities in cohort 2, the paucity of data on pupil attendance was confirmed[11]. The compliance variable specified above results in a further 155 pupils (in 13 NEBT schools) being classed as missing (SAP-specified compliance estimate based on a sample of 159 pupils in 10 NEBT schools). All of this additional missing data is related to missing pupil attendance data.

In response to this, we specify two additional follow-on sensitivity analyses here that were not specified in the SAP. First, the headline ITT analysis is based on a sample of 644 pupils (314 NEBT, 330 control) located in 48 schools (23 NEBT, 25 control). The SAP-specified CACE analysis is based on a smaller sample of 489 pupils (159 NEBT, 330 control) located in 35 schools (10 NEBT, 25 control). Estimates of impact from the SAP-specified CACE analysis and the headline ITT analysis of the primary outcome are therefore based on different sample sizes. To address this, the first follow-on sensitivity analysis ran the ITT analysis using the smaller SAP-specified CACE sample. This enables the ITT and SAP-specified CACE estimates to be directly compared (although both have a caveat of suffering from extremely high attrition). The second additional

10 Null hypothesis assumes that compliance is exogenous and if this is found to be unlikely ($p<0.05$), the alternative hypothesis that compliance is endogenous is assumed. The 'estat endogenous' postestimation for the 'ivregress' STATA command with robust VCE provides a robust regression-based test for endogeneity.

11 Pupil attendance data in cohort 1 was obtained for just four of the 16 schools randomised to NEBT (eight of which had withdrawn from the programme and evaluation mid-trial). We expected similar difficulties for cohort 2 and targeted the collection at schools that supplied baseline and outcome M&MF BD data to try and maximise response. However, in cohort 2, pupil attendance data was obtained for just six of the 30 schools randomised to NEBT (15 of which had withdrawn from the programme and evaluation mid-trial).

analysis revisited the compliance variable and was based on the assumption that in schools known to have met the instructor-level compliance conditions, pupil attendance was 100%[12]. This sensitivity CACE analysis was based on the same sample as the headline ITT analysis, so it can be directly compared (but again with the caveat of very high attrition).

Missing data analysis

The baseline and ITT samples were compared to help illustrate the impact of missing data on the primary outcome variable only: the M&MF BD score. This was initially done descriptively by tabulating missing cases across the categories of variables included in the ITT analysis (M&MF baseline). Reasons for any missingness were summarised, and a multi-level logistic regression model (1=in the ITT model; 0=not in the ITT model) was conducted to examine whether missingness was associated with school-, intervention- and/or pupil-level covariates. School-level covariates included geographical location, % FSM, % English as an Additional Language (EAL) and Ofsted. Pupil-level covariates included baseline M&MF (behavioural and emotional difficulties scales), baseline teacher SDQ (prosocial and total difficulties scales) and baseline BEC (cognitive and affective empathy scales). As specified in the SAP, missing data for the primary outcome was not imputed, given that no pupil-level data was collected at baseline, except for the specified outcomes. Patterns of missingness were examined to inform the interpretation of the ITT impact analysis, rather than preceding the imputation of missing cases.

Sub-group analyses

This trial was commissioned as part of the first funding round for the YEF after its set-up. As such, requirements for subgroup analyses were not set out. Discussions were held between all parties (the YEF, SHU and ROE), and it was decided as part of this trial that the collection of additional information from schools would be kept to a minimum and that a follow-on analysis could be undertaken at a later stage as part of the YEF's archiving process. This means that no data on participant gender, ethnicity, FSM status or pupil-level details, other than the specified outcomes, was collected.

Additional analyses and robustness checks

A sensitivity analysis was conducted to look at the impact of cohort 1 and cohort 2 separately. This was done descriptively by replicating the multilevel model process outlined above for the primary outcome.

Imbalance at baseline

Imbalance at baseline focused on the specified pupil-level self-reported outcomes (M&MF and BES), pupil-level teacher-reported outcomes (SDQ) and school-level statistics (% FSM, Ofsted, geographical area, etc.). These analyses provide an indication of an imbalance at baseline following randomisation. We also examined how these pupil-level outcomes and school-level details are associated with the M&MF BD primary outcome.

12 In the 10 schools where pupil attendance was obtained, 86% of pupils are recorded as attending 8+ of the NEBT themed lessons. To include the missing 155 pupils, they could only be classed as all compliant or all non-compliant. Neither are perfect but assuming all to be compliant (i.e. 100%) seems to be the closer approximation of the observed pattern in schools where pupil attendance data was collected.

Estimation of effect sizes

Effect sizes were calculated using Hedges' g, as specified in the following equation, where T is the treatment mean, C is the control mean, δ^2_{sch} is the school-level variance and δ^2_{pup} is the pupil-level variance for the empty/null model:

$$ES = \frac{(T - C)_{adjusted}}{\sqrt{\delta^2_{sch} + \delta^2_{pup}}}$$

The headline effect size was calculated from the group allocation (intervention/control) coefficient in the full analysis model (including geographical area and cohort), with the unconditional variance used as the denominator. The effect sizes are reported along with 95% confidence intervals and p-values to reflect statistical uncertainty.

Estimation of ICC

School-level Intracluster Correlation (ICC) was estimated using a null (empty) two-level multilevel variance components model. Variance decomposition for the two levels (school and student) is presented below, along with the ICC estimates for the ITT student sample for the primary outcome.

Table 4. ICC estimates for Me and My Feelings behavioural difficulties

	Intervention vs control Estimate (95% CI)
School-level	0.26 (0.09–0.77)
Residual	5.54 (4.94–6.21)
School-level ICC	**0.04 (0.015–0.125)**

Longitudinal analysis

No longitudinal follow-ups will be undertaken as part of the NEBT RCT evaluation. However, unique pupil identifiers (UPNs) have been collected by the evaluation team for pupils in England to enable long-term follow-up by others as part of the YEF data archiving process.

IPE

Research methods

At all stages of the evaluation, we endeavoured to ensure that our approaches, including analysis, were inclusive and cognisant of all aspects of diversity. As detailed in the earlier discussion of subgroup analysis, we acknowledge that due to this project being part of the YEF launch round, the equality and diversity aspects are not in line with the current requirements of YEF work. This is also addressed in the limitations section.

We used observations and case study qualitative data collection (interviews with staff and focus groups with pupils) as opportunities to understand and critically consider issues of inclusivity. For example, observations of the instructor training and NEBT delivery in the classrooms were used to note the cultural appropriateness of the materials, language and approaches. More widely, this included questions around:

- The appropriate adaptation of generic ROE delivery for English and Welsh classroom contexts, given the Canadian genesis, development and tone of the programme
- The appropriate recognition and sensitivity of the content and delivery for a post-Covid classroom and school context, e.g. increased concerns about pupils' socio-emotional wellbeing and behavioural issues in the ongoing Covid recovery period
- The language, materials and delivery approaches, noting any sensitive issues; appropriately acknowledging the impact the pandemic, poverty, special education needs and disabilities (SEND), mental health difficulties and diverse family contexts may have on pupils, their learning, attachments and behaviours; noting explicit and implicit assumptions about family structure and relationships, e.g. language pertaining to traditional white nuclear family structures, heteronormativity or ableism; and noting whether the language, delivery and resources indicate awareness and inclusion of other cultures, ethnicities and diverse family contexts and dynamics.

As a result of the issues with retention and attrition already described, the IPE fieldwork took place over two school years. The year 1 fieldwork included three small-scale case studies, one of which was limited to a remote interview with the instructor. In year 1, attrition had been high, and a concern was raised that additional IPE evaluation burden may further increase attrition. As such, the fieldwork conducted in year 1 was far less extensive than in year 2.

In year 2, random stratified sampling of schools was undertaken based on geography, with the hope of being able to visit two schools in Wales, with the remaining visits being spread across Merseyside, London and the Midlands/Yorkshire. Schools were then contacted by email, starting in March 2024. When schools did not reply or were unable to accommodate a visit, the next school on the list was contacted. Using this method, all intervention schools were eventually asked to take part in a focus group. Five case study visits were undertaken, two in Wales, one in London, one in Merseyside and one in Yorkshire. Table 7 indicates the sessions observed. It was the intention to observe two of each session type (i.e. pre-visit, visit and post-visit). However, as a sixth school could not be found despite all efforts, only one post-family-visit session was observed.

Table 5. Implementation and process evaluation methods overview

Data collection method	Outcome/measure	Research question[13]	Logic model code
Case studies (overall)	3 case studies in Y1 (summer 2023) 6 case studies in Y2 (3 in Wales and 3 in England; summer 2024)	RQ11–14	CM1, CMP2, CMI3 SP1, SP2, SI1, SI2, S1 MP1, MP2, MI1, MI2, MI3, MI4 CF1–8
Observations	Instructor training observations: Y1: 1 visit (Oct 2022) Y2: 1 mid-year visit (Feb 2024; no data observation) Classroom observations of Roots of Empathy (ROE) sessions: Y1: 2 (summer 2023) Y2: 5 (summer 2023)	RQ13 RQ11,13	(yellow inputs and yellow causal mechanisms in the logic model) CM1, CMP2, CMI3 SP1, SP2, SI1, SI2, S1, CF8
Amended Nurturing Empathy Before Transition instructor interviews	Y1: 3 interviews Y2: 5 interviews	RQ11-13	(yellow inputs and yellow causal mechanisms) CM1, CMP2, CMI3 SP1, SP2, SI1, SI2, S1 MP1, MP2, MI1, MI2, MI3, MI4 CF1–8
Class teacher interviews	Y1: 3 interviews Y2: 5 interviews	RQ11–13,	CM1, CMP2, CMI3 SP1, SP2, SI1, SI2, S1 CF3–6, CF8 MP1, MP2, MI1, MI2, MI3, MI4
Senior Leadership Team (SLT) interviews	Y1: 2 interviews Y2: 2 interviews	RQ11–13,	CF1–8 SP1, SP2, SI1, SI2, S1 MP1, MP2, MI1, MI2, MI3, MI4
Pupil focus groups	Y1: 2 focus groups Y2: 5 focus groups Approximately 8 pupils in each group	RQ11–13	CM1, CMP2, CMI3 SP1, SP2, SI1, SI2, S1 MP1, MP2, MI1, MI2, MI3, MI4
Deliverer interviews	Y1: 1 ROE interview Y2: ROE interview – unable to complete due to Katie leaving	RQ11–14	Yellow inputs; CF1–8 (and their perspectives on CMs and outcomes)

Note: Y=year, CM1=Causal mechanism experiential learning, CMI=Causal mechanism inter-relational, CMP=Causal mechanism, personal, CF=Contextual factors, SP-Short-term outcome-personal,

[13] The implementation and process evaluation aimed to answer the following research questions:
RQ11. What are the key factors which influence successful delivery of the NEBT programme in years 1 and 2?
RQ12. What are the perceptions of pupils, teachers, deliverers and instructors about the effectiveness of the programme in years 1 and 2?
RQ13. What fidelity issues are observed during years 1 and 2 of the trial?
RQ14. What does the trial indicate about scalability?

SI=short term outcome-inter-relational, MP=Medium Outcome-Personal, MI=Medium outcome-inter-relational

Data collection

Cohort 1 IPE data was collected by a single member of staff. As a result of staff turnover in the intervening time, the IPE data collection for cohort 2 was carried out by three researchers using the same research tools as for cohort 1 in order to ensure consistency of approach.

Data collection instruments were developed by the core evaluation team with close reference to the ToC, logic model and IPE research questions. The instruments went through several iterations before a final version was agreed upon by all team members. Prior to data collection, the team conducted calibration meetings to ensure a shared understanding of the research tools and consistent application across cohorts. Detailed guidance documents and standardised interview guides were used to minimise variations in data collection. Regular check-ins and reflective sessions were held throughout cohort 2 fieldwork to monitor compliance and address any deviations from the agreed-upon methods. This approach ensured that, despite multiple researchers being involved in cohort 2, data collection remained consistent with that in cohort 1, supporting the reliability and validity of the findings.

Data analysis

IPE data was analysed using NVivo 12. Transcripts from the instructor, teacher and headteacher interviews and pupil focus groups were coded deductively at first, informed by the research questions and logic model. Following this, a second round of inductive coding was undertaken to capture any unexpected patterns or themes that went beyond the initial research questions. Session observation notes were descriptively summarised and used to supplement the main analysis (see Appendix F for case study summaries). Dedicated meetings were held throughout the analysis to share emerging findings, allowing the team to collectively explore patterns, similarities and differences across the data. This collaborative approach supported a deeper interpretation of the findings and strengthened analytical rigour.

Timeline

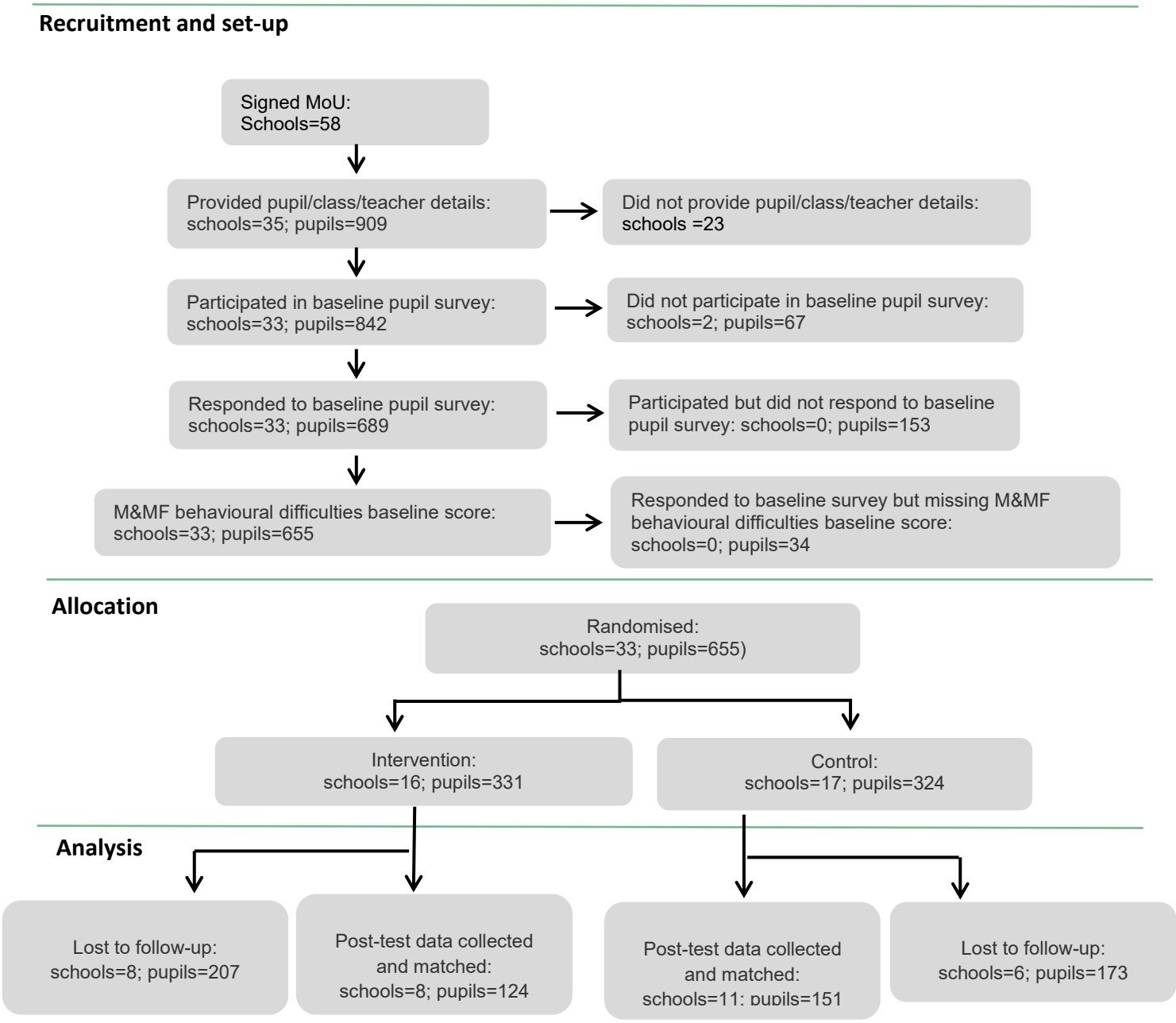
Table 6. Timeline

Dates (WHEN)	Activity (WHAT)	Staff responsible /leading (WHO)
April–July 2022	School recruitment for cohort 1	ROE
September–October 2022	Baseline data collection for cohort 1 schools (class lists, pupil survey, teacher survey)	SHU
September–October 2022	Randomisation for cohort 1	SHU
October 2022	Cohort 1 intervention delivery starts	ROE
October 2022	Control schools receive part of their incentive payment	SHU
October 2022–May 2023	IPE school visits for cohort 1	SHU
May–June 2023	Intervention delivery ends	ROE

June 2023	Cohort 1 endpoint data collection (pupil survey and teacher survey)	SHU
Cohort 2		
Septemer 2022–July 2023	School recruitment for cohort 2	ROE
September 2022	Evaluation protocol published	SHU
September–October 2023	Baseline data collection for cohort 2 schools (class lists, pupil survey, teacher survey)	SHU
September–October 2023	Randomisation for cohort 2	SHU
October 2023	Cohort 2 intervention delivery starts	ROE
February–May 2023	IPE school visits for cohort 2	SHU
March 2024	Evaluation statistical analysis plan published	SHU
April 2024	Control schools receive their first incentive payment	SHU
May–June 2024	Intervention delivery ends in cohort 2 schools	ROE
June–July 2023	Endpoint data collection in cohort 2 schools	SHU
July–September 2024	Analysis	SHU
August 2024–November 2024	Write-up	SHU
September 2024	Control schools receive their second incentive payment	SHU
February 2024–April 2025	Peer review of final report and revisions	YEF and SHU
April–May 2025	Report shared with ROE and updates/revisions made	YEF, SHU and ROE
August 2025	Final report published	YEF

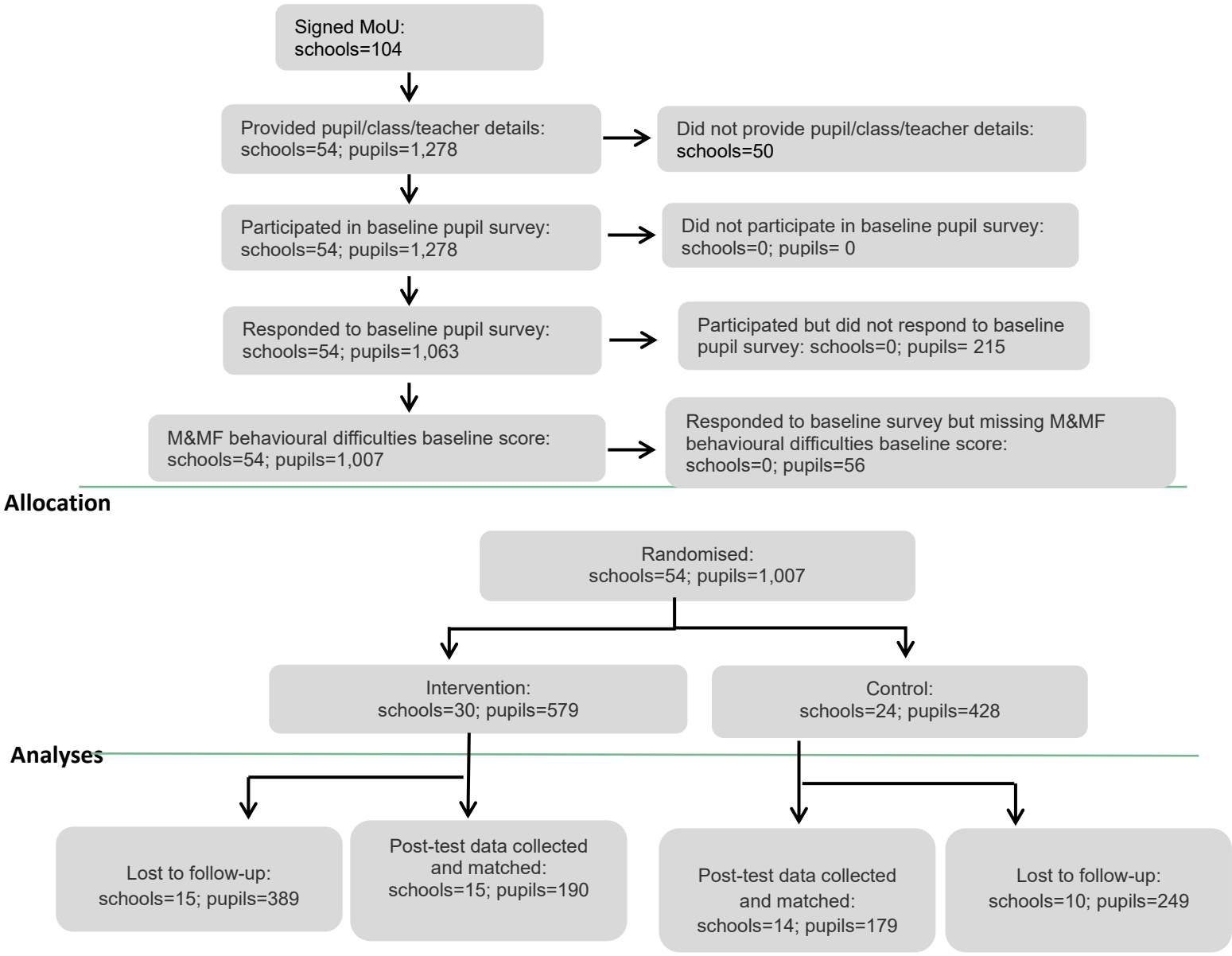
Note: ROE=Roots of Empathy, SHU=Sheffield Hallam University, YEF=the Youth Endowment Fund

Participant flow, including losses and exclusions



Note: MoU=memorandum of understanding, M&MF=Me and My Feelings

Figure 3. Consort flow diagram for primary outcome – cohort 1 (2022/23)



Note: MoU=memorandum of understanding, M&MF=Me and My Feelings

Figure 4. Consort flow diagram for primary outcome – cohort 2 (2023/24)

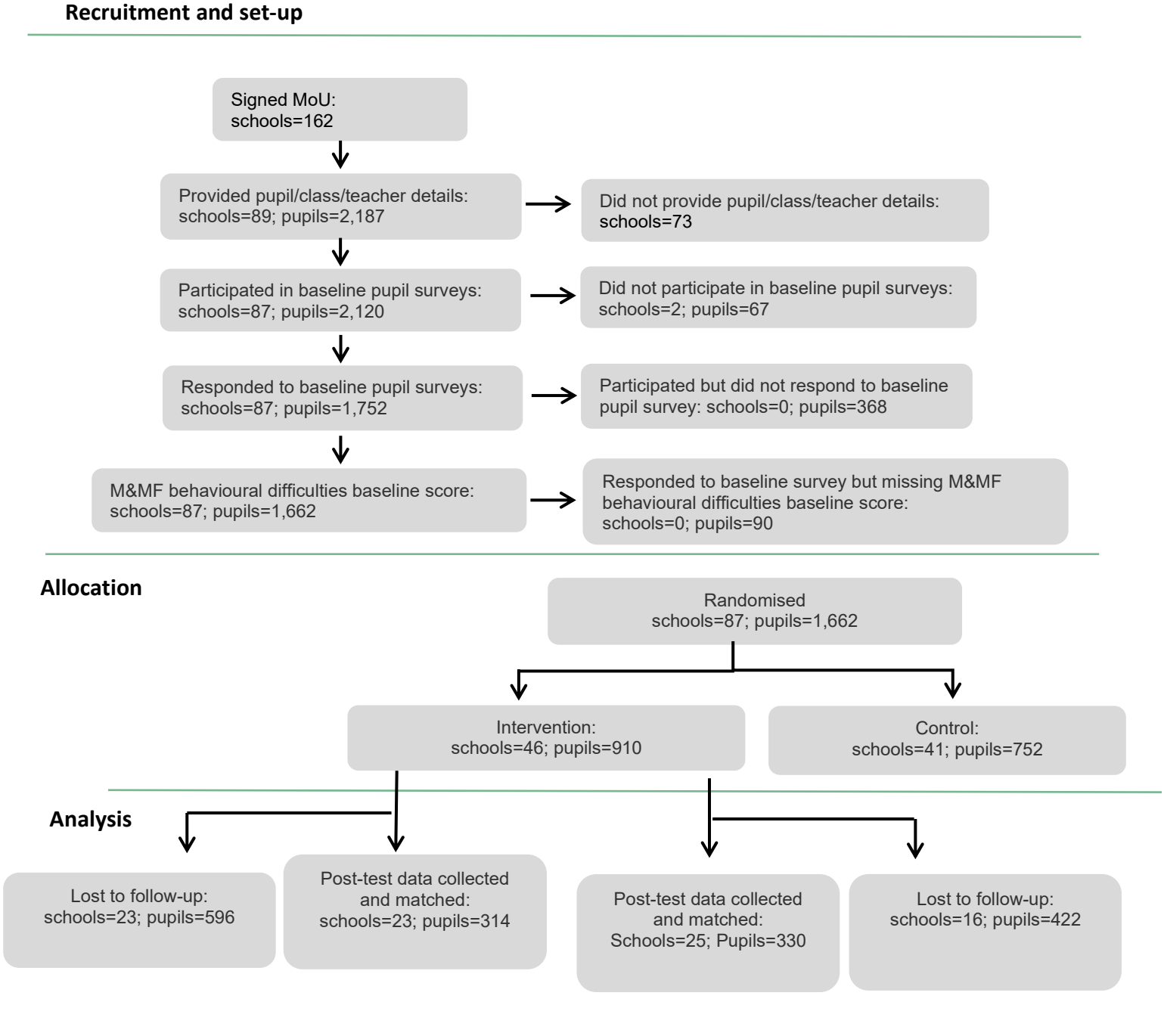


Figure 5. Consort flow diagram for primary outcome – cohorts 1 and 2 combined

Note: MoU=memorandum of understanding, M&MF=Me and My Feelings

Attrition

Attrition across the trial was notable, with the majority being at the school level, with whole schools dropping out of the programme and the evaluation data collection. Across both cohorts (Fig 5), M&MF baseline data was collected from pupils in 87 schools, but this dropped to just 48 schools at endpoint (45% school-level attrition overall, 50% amongst Intervention schools and 39% amongst control schools). In Figure 3-5 above, 'participated in' vs 'responded to' details the difference between schools and pupils participating and being invited to participate vs those responding. Schools may have participated in the survey, as indicated in the 'participated in' box in Figure 3-5, but not all pupils from that school who were on the trial responded, as detailed in the 'responded to' box in Figure 3-5.

As might be expected, pupil-level response was notably higher in the subsample of 48 schools that did participate in the baseline and endpoint pupil surveys, although even in this subsample of 48 schools, overall pupil-level attrition between the baseline and endpoint surveys was still high, at 31% (30% in intervention schools and 32% in control schools).

The reasons for attrition across the trial were varied. At the point of dropping out, some schools provided a reason, which could be grouped into lack of time to deliver the intervention in school, inability to attend instructor training and a change in the nominated instructor's circumstances. Given the number of primary and secondary outcomes in this trial (three questionnaires), measurement burden may have also impacted attrition levels. On average, the pupil questionnaires (M&MF and BES) took nine minutes for pupils to complete, and on average, each of the teacher SDQs took 2.5 minutes to complete. It is also likely that some pupils may have moved schools during the programme. Furthermore, where class lists weren't provided, schools were not included, adding to attrition. Responses from pupils (or teachers for pupils) who were not listed on class lists were also not included.

Table 7. Pupil-level attrition from the trial (primary outcome)

Combined cohorts		**Intervention**	**Control**	**Total**
Number of participants	Randomised	910	752	1662
	Analysed	314	330	644
Participant attrition	Number	596	422	1,018
(from randomisation to analysis)	Percentage	65.5%	56.1%	61.3%
Cohort 1		**Intervention**	**Control**	**Total**
Number of participants		331	324	655
		124	151	275
Participant attrition		207	173	380
(from randomisation to analysis)		62.5%	53.4%	58.0%
Cohort 2		**Intervention**	**Control**	**Total**
Number of participants	Randomised	579	428	1,007
	Analysed	190	179	369
Participant attrition	Number	389	249	638
(from randomisation to analysis)	Percentage	67.2%	58.2%	63.4%

Participant characteristics

Table 8. Baseline characteristics of groups as randomised

	Intention to treat sample randomised	
	Intervention n (%)	Control n (%)
Pupil		
London	232 (25.5%)	216 (28.7%)
Merseyside	279 (30.7%)	222 (29.5%)
Midlands	79 (8.7%)	75 (10%)
Yorkshire	152 (16.7%)	108 (14.4%)
Wales	168 (18.5%)	131 (17.4%)
Total	910 (100%)	752 (100%)
School		
London	12 (26.1%)	11 (26.8%)
Merseyside	13 (28.3%)	12 (29.3%)
Midlands	4 (8.7%)	4 (9.8%)
Yorkshire	7 (15.2%)	5 (12.2%)
Wales	10 (21.7%)	9 (22%)
Total	46 (100%)	41 (100%)

Table 9. Summary of baseline and outcome sample sizes across five measures

Outcome measure	Intervention Baseline/outcome (% lost to attrition)	Control Baseline/outcome (% lost to attrition)	Total Baseline/outcome (% lost to attrition)
M&MF BD (primary outcome)	910/314 (66%)	752/330 (56%)	1,662/644 (61%)
M&MF ED (secondary outcome)	908/307 (66%)	736/314 (57%)	1,644/621 (62%)
BES cognitive empathy (secondary outcome)	871/286 (67%)	729/305 (58%)	1,600/591 (63%)
BES affective empathy (secondary outcome)	868/277 (68%)	713/295 (59%)	1,581/572 (64%)
Teacher SDQ total difficulties (secondary outcome)	711/207 (71%)	513/249 (51%)	1,224/456 (63%)
All four pupil-level measures	782/239 (69%)	635/242 (62%)	1,417/481 (66%)
All four pupil-level measures plus teacher SDQ	513/99 (81%)	367/100 (73%)	880/199 (77%)

Note: M&MF=Me and My Feelings, BD=behavioural difficulties, ED=Emotional difficulties, BES=Basic Empathy Scale, SDQ=Strengths and Difficulties Questionnaire

The complete missing data analysis is restricted to the primary outcome – but balance relating to the baseline versions of all outcomes is summarised in the Table below. Histograms of the primary outcome baseline data demonstrated that it was positively skewed, demonstrating that the majority of participants had low scores for BD (appendix G)

Table 10. Baseline means and effect sizes for the Nurturing Empathy Before Transition and control groups

	Complete aseline intention to treat sample			Subsample with baseline and outcome scores		
Outcome measure	Intervention mean (95% CI)	Control mean (95% CI)	Mean difference as effect size	Intervention mean (95% CI)	Control mean (95% CI)	Mean difference as effect size
M&MF BD	3.04 (2.88;3.2)	3.13 (2.96;3.3)	−0.02	2.62 (2.38;2.86)	2.92 (2.67;3.17)	−0.13
M&MF ED	7.19 (6.94;7.44)	7.4 (7.15;7.65)	−0.02	7.16 (6.76;7.56)	7.58 (7.20;7.96)	−0.12
BES cognitive empathy	33.4 (33.0;33.7)	33.5 (33.1;33.9)	−0.04	34.1 (33.5;34.7)	33.8 (33.2;34.4)	0.05
BES affective empathy	34.8 (34.4;35.3)	34.8 (34.3;35.3)	0.03	34.9 (34.1;35.7)	35.0 (34.3;35.8)	−0.02
Teacher-SDQ TD	7.9 (7.35;8.37)	8.1 (7.54;8.72)	−0.04	6.89 (5.97;7.81)	7.65 (6.84;8.46)	−0.11
Teacher-SDQ EP	1.9 (1.73;2.07)	2.1 (1.91;2.35)	−0.12	1.75 (1.45;2.05)	1.98 (1.67;2.29)	−0.10
Teacher-SDQ CP	1.1 (0.98;1.26)	1.3 (1.14;1.48)	−0.07	0.88 (0.64;1.12)	1.27 (1.04;1.5)	−0.21
Teacher-SDQ Hyp	3.6 (3.37;3.83)	3.3 (3.06;3.58)	0.14	2.97 (2.55;3.39)	3.23 (2.84;3.62)	−0.08
Teacher-SDQ PP	1.5 (1.35;1.61)	1.4 (1.23;1.53)	0.01	1.27 (1.03;1.51)	1.22 (1.01;1.43)	0.03
Teacher-SDQ PS	7.0 (6.82;7.18)	7.4 (7.17;7.59)	−0.10	7.41 (7.07;7.75)	7.69 (7.4;7.98)	−0.12

CI=confidence interval, M&MF=Me and My Feelings, BD=behavioural difficulties, ED=Emotional Difficulties, BES=Basic Empathy Scale, SDQ=Strengths and Difficulties Questionnaire, TD=total difficulties, EP=Emotional problems, CP=Conduct Problems, Hyp=Hyperactivity, PP=Peer Problems, PS=Prosocial behaviour

Outcomes and analysis

Primary analysis

Table 13 reports the findings for the ITT analysis of the primary outcome measure, the BD subscale of the M&MF self-report scale. When compared with the control group, a small negative effect size (ES) was observed (ES=−0.06 standard deviations (sds)), suggesting that pupils who received the NEBT programme less frequently self-reported behavioural problems. However, caution is needed because of the sizable attrition this evaluation experienced. Additionally, even if attrition had not been such a problem, the observed effect size was smaller than the impact evaluation was powered for, and the 95% confidence intervals span from −0.22 sds (reduced BD) through zero (no difference) up to +0.10 sds (increased BD). Therefore, the evidence is insufficient to conclude that NEBT led to a reduction in self-reported BD (as measured by M&MF).

Table 11. Intention to treat primary analysis of the behavioural difficulties subscale of the Me and My Feelings (M&MF) questionnaire

Combined	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
M&MF behavioural difficulties	314 (596)	2.86 (2.62;3.10)	330 (422)	3.15 (2.87;3.43)	644 (314;330)	−0.06 (−0.22;0.10)	0.50

Table 14 reports the findings for the ITT analysis of the primary outcome measure, the BD subscale of the M&MF self-report scale, split by cohort. When compared with the control group, a small positive effect size was observed (ES=0.09 sds) for cohort 1, suggesting that pupils who received the NEBT programme self-reported more behavioural problems. For cohort 2, a small negative effect size was observed (ES=−0.19 sds), suggesting that pupils who received the NEBT programme self-reported less behavioural problems. However, caution is needed because of the sizable attrition that this evaluation experienced. Additionally, even if attrition had not been such a problem, the observed effect size for cohort 1 was smaller than the impact evaluation was powered for, and the 95% confidence intervals span from −0.14 sds (reduced BD) through zero (no difference) up to 0.33 sds (increased BD). For cohort 2, the 95% confidence intervals span from −0.37 sds to −0.01 sds, both suggesting reduced BD but of different magnitudes. Therefore, there is only sufficient evidence to conclude that NEBT led to less self-reported BD (as measured by M&MF) for cohort 2, and that should still be interpreted with caution due to attrition.

Table 12. Intention to treat primary analysis of the behavioural difficulties subscale of the Me and My Feelings (M&MF) questionnaire, split by cohort (sensitivity analysis)

Separate cohorts	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
M&MF behavioural difficulties Cohort 1	124 (207)	2.89 (2.52;3.26)	151 (173)	2.89 (2.51;3.27)	275 (124;151)	+0.09 (−0.14;0.33)	0.44
M&MF behavioural difficulties Cohort 2	190 (389)	2.84 (2.52;3.16)	179 (249)	3.38 (2.98;3.78)	369 (190;179)	−0.19 (−0.37;−0.01)	0.04

Secondary analysis

Table 15 reports the findings for the ITT analysis of the secondary outcome measures: M&MF emotional difficulties subscale, BES cognitive empathy and BES affective empathy. When compared with the control group, no effect was found on emotional difficulties (ES=−0.00 sds), suggesting that there was no difference between groups in terms of reported emotional difficulties. The confidence intervals span from −0.17 sds (reduced emotional difficulties) through zero (no difference) to 0.17 (increased emotional difficulties). Therefore, alongside issues of attrition, evidence is insufficient to conclude that NEBT led to a reduction in self-reported emotional difficulties. Furthermore, when compared with the control group, a small negative effect was observed for cognitive empathy (ES=−0.01 sds), suggesting that those who received the NEBT showed next to no difference to controls. The confidence intervals span from −0.16 sds (less cognitive empathy) through zero (no difference) to 0.15 (increased cognitive empathy). We would therefore conclude that there is insufficient evidence to conclude that NEBT led to increased levels of cognitive empathy as measured by the BES. When compared to the control group, a moderate positive effect was observed for affective empathy (ES=0.19 sds), suggesting that pupils who received NEBT had increased affective empathy. The confidence intervals span from −0.03 sds (reduced affective empathy) through zero (no difference between groups) to 0.35 sds (increased affective empathy). These findings suggest that NEBT had a positive impact on affective empathy, but the slight overlapping of the confidence intervals alongside the attrition levels means that these need to be interpreted with caution.

Table 13. Intention to treat secondary analysis of the emotional difficulties subscale of the Me and My Feelings (M&MF) questionnaire and the Basic Empathy Scale (BES)

	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
M&MF emotional difficulties	307 (601)	7.13 (6.75;7.51)	314 (422)	7.22 (6.80;7.64)	621 (307;314)	0.00 (−0.17;0.17)	0.98
BES cognitive empathy	286 (585)	34.6 (33.0;35.3)	305 (424)	34.4 (33.8;35.0)	591 (286;305)	−0.01 (−0.16;0.15)	0.95
BES affective empathy	277 (591)	35.6 (34.8;36.4)	295 (418)	34.3 (33.5;35.1)	572 (277;295)	0.19 (0.03;0.35)	0.02

Strengths and difficulties

Table 16 reports the findings for the ITT analysis of the secondary outcome measure, the SDQ. When compared with the control group, a medium negative effect was found for the NEBT programme on BD (ES=−0.36 sds). The confidence intervals were −0.68 to −0.04, demonstrating fewer BD but differing magnitudes. We would, therefore, conclude that there is sufficient evidence to suggest that the NEBT

programme led to fewer BD. Similarly, when compared with the control group, a large positive effect was found for the NEBT programme on prosocial behaviour (ES=0.63 sds), with confidence intervals demonstrating medium to large effect sizes (0.36 to 0.90). We would, therefore, conclude that there is sufficient evidence to conclude that the NEBT programme led to increased prosocial behaviour. However, as with previous findings, due to notable attrition, these findings need to be treated with caution. In addition, given that teachers were not blinded to their pupils' allocation, further caution is warranted.

Exploratory analysis of all the subscales of the SDQ showed that when compared to a control group, a small negative effect was found on conduct problems (ES=−0.20 sds) but with confidence intervals that span from −0.42 (fewer conduct problems) through zero (no difference) to 0.10 (increased conduct issues). For peer problems, when compared to a control, a small negative effect was observed (ES=−0.40 sds), with confidence intervals from −0.76 sds (fewer peer problems) to −0.03 sds, suggesting a reduction in peer problems of varying magnitudes. We would, therefore, conclude that for conduct problems, there was evidence of an impact but the overlapping of confidence intervals means the evidence was insufficient to state that NEBT led to decreased conduct problems. There is sufficient evidence to conclude that the NEBT programme led to a decrease in peer problems, but, as with previous findings, due to notable attrition, these need to be treated with caution. In addition, given teachers were not blinded to their pupils' allocation, further caution is warranted (as previously discussed on pages 10 and 28). When compared to a control, a small negative effect was observed for hyperactivity (ES=−0.35 sds), and the confidence intervals suggest a negative effect of varying magnitudes (−0.63 to −0.07 sds).

Lastly, for emotional problems, when compared to a control, a small negative effect was observed (ES=−0.21 sds). However, confidence intervals spanned −0.52 sds (fewer emotional problems) through zero (no difference) to 0.11 (more emotional problems), suggesting that there is insufficient evidence to suggest that the NEBT programme led to reduced emotional problems as assessed by teachers.

Table 14.Intention to treat secondary analysis of the strengths and difficulties questionnaire (SDQ)

Combined cohorts	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
SDQ total difficulties	207 (504)	6.48 (5.58;7.38)	249 (264)	9.17 (8.22;10.12)	456 (207;249)	−0.36 (−0.68;−0.04)	0.03
SDQ pro-social	208 (534)	8.80 (8.53;9.07)	251 (267)	7.60 (7.30;7.90)	459 (208;251)	0.63 (0.36;0.90)	<0.01

SDQ emotional problems	213 (532)	2.09 (1.76;2.42)	252 (267)	2.42 (2.10;2.74)	465 (213;252)	−0.21 (−0.52;0.11)	0.21
SDQ conduct problems	213 (510)	0.84 (0.62;1.06)	253 (266)	1.53 (1.27;1.79)	466 (213;253)	−0.20 (−0.42;0.10)	0.09
SDQ peer problems	211 (535)	2.56 (2.16;2.96)	252 (265)	3.65 (3.24;4.06)	463 (211;252)	−0.40 (−0.76;−0.03)	0.03
SDQ hyperactivity	213 (531)	1.00 (0.79;1.21)	251 (267)	1.67 (1.44;1.90)	464 (213;253)	−0.35 (−0.63;−0.07)	0.02

Analysis in the presence of noncompliance

Compliance with the NEBT programme was specified using three conditions:

- **Instructor level:** Whether the NEBT instructor attended all four NEBT training sessions (=1) or not (=0)
- **School level:** Whether schools delivered at least eight of the nine NEBT themes (=1) or not (=0)
- **Pupil Level:** Whether pupils in NEBT schools attended at least eight of the nine NEBT themes (=1) or not (=0)

Overall compliance is achieved when a '1' is scored on all three conditions. Given the large issues with attrition, similar caution is needed in interpreting analyses in the presence of noncompliance. We summarise the analyses here, but more details can be found in Appendix E.

The CACE estimate was obtained using the STATA 'ivregress' command with adjustment to standard errors to acknowledge clustering at the school level. Table 17 presents the effect size estimates from the CACE analyses.

Table 15. CACE analysis

	SAP specified	Sensitivity
Samples: pupils (schools)		
NEBT	159 (10)	314 (23)
Control	330 (25)	330 (25)
	489 (35)	644 (48)
Effect Sizes (95% CI)		
ITT analysis	0.04 (−0.13;0.20)	−0.06 (−0.22;0.10)
CACE analysis using ivgreress	0.05 (−0.13;0.23)	−0.06 (−0.24;0.12)
F-test for exogeneity	F(1,35)=3.27; p=0.08	F(1,47)=2.92; p=0.09
CACE analysis using the formula	0.05 (−0.17;0.26)	−0.06 (−0.24;0.12)

Note: SAP=Statistical analysis, NEBT=Nurturing empathy before transition, ITT=intention to treat, CACE=Compiler Average Causal Effect

For the SAP-specified CACE analysis, compliance is estimated at 79%, and the correlation between group membership and compliance was high, at 0.84. The SAP-specified CACE estimate was 0.05 sds (CI: −0.13;0.23). This compares with the ITT analysis estimate of −0.06 (CI: −0.22;0.10). However, the ITT analysis was based on the complete case sample (n=644), and when this analysis is re-run with the restricted SAP-specified compliance sample (n=489), the ITT estimate had the same sign and was closer to the CACE estimate at 0.04 (CI: −0.13;0.20). The F-test for the first stage of the instrumental variable model was small and not statistically significant – F(1,35)=3.27, p=0.08 – suggesting that compliance can be considered to be exogenous. The alternative CACE estimate using the formula specified in the SAP provided a second estimate for the SAP-specified CACE and was close to what was obtained from the instrumental variable models, at 0.05 sds (CI: −0.17;0.26).

For the sensitivity CACE analysis, compliance is estimated at 89%, and the correlation between group membership and compliance was very high, at 0.90. The sensitivity CACE estimate was −0.06 sds (CI: −0.24;0.12), which compares closely to the ITT analysis estimate of −0.06 (CI: −0.22;0.10). The F-test for the first stage of the instrumental variable model was small and not statistically significant – F(1,47)=2.92, p=0.09 – suggesting that compliance can be considered to be exogenous. The alternative CACE estimate using the formula provided a second estimate for the sensitivity CACE and was close to what was obtained from the instrumental variable models, at −0.06 sds (CI: −0.24;0.12).

In summary, the ITT analyses found no evidence of impact for pupils being offered the NEBT programme, and the CACE analyses found no evidence of impact for pupils who engaged in the NEBT programme as intended. We therefore conclude that we found no evidence that NEBT had a statistically significant impact on pupil behaviour as measured by the self-reported M&MF BD scale. The notable caveat for these and all other impact analyses in this evaluation is the very high rate of attrition (61.3% lost to attrition, only 38.7% of pupils randomised to the NEBT programme were included in the impact analyses). This serves to undermine the validity of the causal conclusions drawn from these analyses, but from the data we analysed, we found no evidence of impact. The implications of the level of missing data are discussed later.

Missing data analysis

The NEBT evaluation suffered greatly from attrition. At the pupil level, only 644 out of the 1,662 pupils (39%) randomised to the NEBT or control groups were included in the ITT analysis of the primary outcome. This attrition was higher in the NEBT group (only 34.5% were included in the ITT analysis) than in the control group (43.9% included) and was observed in both cohorts 1 and 2 of the evaluation. Attrition was even higher for secondary outcomes, particularly the teacher-completed SDQ measures (only 37% were included in the ITT analysis for SDQ total difficulties).

Minimising attrition and reasons for missingness

Several things were undertaken to try to minimise the risk of attrition at various points across both trials (cohort 1 and cohort 2). Partway through the first trial year and prior to recruitment for year 2, we held a lessons learnt workshop with ROE, the YEF and the evaluation team at SHU. This provided all parties the opportunity to reflect on how recruitment and data collection had gone so far and what might need to change for the second trial (cohort 2).

The MoU provided schools with a detailed account of the evaluation and the intervention and acted as a method to ensure that schools were fully aware of what they were signing up for when they elected to be part of the NEBT trial. However, several schools dropped out of the trial when they found out that they were to be part of the NEBT programme rather than part of the control group. Whilst the evaluation activities could be burdensome in terms of time, these schools dropped out after providing their data for the evaluation. This suggests that for these schools, it wasn't the evaluation activities that were problematic. However, it is likely that a mix of evaluation burden, features of the programme and school-level factors (staff shortage, for example) may have contributed to attrition. A lessons learnt section on page 75 explores some of these reasons for missingness and ways to mitigate these issues in future trials. The level of commitment for the NEBT programme was significant, which was likely to be an influential factor in some school-level attrition.

Missing data analysis

As noted in the SAP, to avoid further school burden, it was agreed[14] that no additional pupil-level data should be collected from schools. This decision precludes the possibility of using multiple imputation to estimate missing data for sensitivity analyses. Additionally, the examination of the patterns in missing data is predominantly limited to school-level variables.[15] Further, most school-level data is restricted to schools in the English education system (see Table 10-11). Following recruitment difficulties, the NEBT evaluation was run over two cohorts: the first in 2022/23 and the second in 2023/24. In response to further recruitment difficulties for the second cohort, the sample of eligible schools was expanded to include primary schools in the Welsh education system. Whilst data on all state primary schools in England can be easily accessed for use in the missing data analysis,[16] a comparable data source is not currently available for primary schools in Wales. In Wales, data is only available on a school-by-school basis.[17] This data access issue is compounded by inconsistencies in the measurement of variables included for English and Welsh primary schools.[18] These differences reflect the divergence of the English and Welsh education systems since 1999, when schools in Wales came under the control of the Welsh National Assembly. One clear difference is how the Welsh system has moved away from the publication of school comparison 'league tables', which may explain the current lack of a Welsh primary school database. This issue carries into the National Pupil Database (NPD), where no data is held for pupils in Welsh schools.[19] These data issues (and the limitations they bring for longitudinal analyses) were part of discussions between the YEF and SHU before deciding to extend the sample into Wales for cohort 2. The evaluation data file includes the Welsh schools' unique reference

14 Through discussions between YEF, ROE and SHU.

15 The one pupil-level variable used is the baseline M&MF Behavioural Difficulties score.

16 See here: https://www.compare-school-performance.service.gov.uk/download-data

17 See here: https://mylocalschool.gov.wales/

18 For example, in England, concentrations of pupil deprivation is approximated using the %FSM in last 6 years variable whilst in Wales, a three-year average %FSM is used.

19 Data - SAIL Databank provides access to health and population data across Wales. Education data was collected until 2021 and the HAPPEN network also collects health and wellbeing data from primary schools across Wales. Schools participate in this voluntarily, rather than it being systematically collected from all Welsh primary schools. These issues prevent it being used for the current research.

numbers and, when provided, UPNs. If school- and/or pupil-level data for Welsh schools does become available, these codes may be useful for linking.

Patterns of missing data can be examined for all schools across two variables: at the pupil level, the baseline M&MF BD score and at the school level, the regional stratification variable. Analyses that include the remaining nine school-level variables[20] are limited to the subsample of English schools recruited to the NEBT evaluation.

Table 18 presents the mean baseline M&MF BD score for the whole ITT sample (n=1,662), the subsample included in the ITT analysis of the M&MF BD outcome (n=644) and the subsample excluded from the ITT analysis because of missing data (n=1,018). Table 18 also shows these three samples for the NEBT intervention and control groups.

Table 16. Comparing the mean baseline Me and My Feelings (M&MF) behavioural difficulties (BD) for pupil subsamples included and excluded from the intention to treat (ITT) analysis of the M&MF BD outcome

Included in ITT analysis of M&MF BD outcome	NEBT intervention schools		Control schools		All schools		Effect size
	n	Mean (95% CI)	n	Mean (95% CI)	n	Mean (95% CI)	(NEBT-Control)/sd
At baseline	910	3.04 (2.88;3.2)	752	3.13 (2.96;3.3)	1,662	3.08 (2.96;3.2)	−0.04
At outcome	314	2.62 (2.38;2.86)	330	2.92 (2.67;3.17)	644	2.77 (2.6;2.94)	−0.30
Missing	596	3.26 (3.06;3.46)	422	3.3 (3.06;3.54)	1,018	3.28 (3.13;3.43)	−0.02

Note: NEBT=Nurturing Empathy Before Transition, sd=standard deviation

The total sample of 1,662 pupils randomised had a higher baseline mean M&MF BD score (3.08) than the subsample of 644 pupils included in the ITT analysis of the M&MF BD outcome (2.77). The subsample of 1,018 pupils excluded from the ITT analysis had the highest baseline mean M&MF BD score (3.28). In other words, (self-reported) BD were higher at baseline for the subsample missing from the ITT analysis than for the subsample that was included.

The pattern observed for all schools is also seen within the NEBT intervention and control school samples. In intervention schools, pupils included in the ITT analysis reported fewer BD (2.62) than pupils with missing outcome data (3.26). In control schools, pupils included in the ITT analysis reported fewer BD (2.92) than pupils with missing outcome data (3.30). The difference between mean baseline M&MF BD scores for the included and missing subsamples was greater for the NEBT intervention group than for the control group.

20 School type, Ofsted rating, School size, Mean pupil KS2 reading score, Mean pupil KS2 grammar, punctuation and spelling score, Mean pupil KS2 maths score, %SEND, %EAL & %FSM in last six years

This results in a wider difference in the mean baseline M&MF BD for the analysed sample (effect size=−0.30 sds) compared with the sample at baseline (effect size=−0.04 sds).

The M&MF BD has a scale ranging from 0 to 12, with higher scores indicating greater self-reported BD. In a longitudinal analysis with M&MF BD used at baseline and outcome, as the baseline M&MF BD score approached zero, the likelihood of observing a positive impact diminished because of floor effects. These missing data analyses reveal that pupils with complete baseline and outcome M&MF BD scores had a lower mean baseline score than pupils with complete baseline but missing outcome M&MF BD scores. Therefore, one impact of attrition will be to limit the extent of the positive impact that could be possible. This is particularly the case for the NEBT sample (baseline mean for the complete analysis sample was 2.62 compared with 3.04 at baseline) but also has relevance for the control sample (baseline mean for the complete analysis sample=2.92 compared with 3.08 at baseline).

In summary, the NEBT evaluation suffered greatly from attrition in both cohorts. The extent of this attrition will undermine the validity of causal conclusions drawn from any of the impact analyses presented. This missing-data analysis provides some detail on whether and how this attrition resulted in an increased imbalance between the two groups. Further information on attrition using data measured at the school level can be found in Appendix (H).

IPE results

The evaluators worked with the ROE team to develop an initial evidence-based logic model and ToC (see Appendix D). This is grounded in existing research on empathy and pro-social interventions, outlining the inputs, mechanisms and intended short-, medium- and long-term outcomes of the NEBT programme. The evidence-based logic model was used to structure the approach to IPE data collection, analysis methods and synthesis of the findings. Using a theory-based evaluation approach also provided the opportunity to deepen knowledge of how and in what contexts mediating variables, individually and together, interact with inputs, outputs and emerging outcomes and are, in turn, impacted and changed by the trial.

Table 19 presents IPE activities conducted across case study schools during the two-year fieldwork period and includes the geographical area, types of sessions observed and data collection methods used. Interviews with staff and focus groups with pupils were audio recorded and later transcribed, while session observations were summarised from field notes. Analysis was undertaken using the NVivo 12 software using a thematic approach (Braun and Clarke, 2006). The coding process combined deductive coding against IPE research questions and the logic model, with inductive coding to capture unexpected themes and insights. The following section syntheses findings from the three forms of IPE data sources, organised by research question. Additionally, supplementary material in the form of extended pupil focus group notes and observation notes, structured by case rather than by theme, are provided in Appendix F.

Table 17. Implementation and process evaluation (IPE) activity across the two years of IPE fieldwork

Year	School ID	Geographical area	Visit type			Interview/focus group				Observation
			Pre family	Family	Post family	Pupil	Instructor	Teacher	Head teacher	
1	Case study 1	Yorkshire	x			x	x	x	x	X
1	Case study 2	Yorkshire				x	x	x	x	
1	Case study 3	London	Limited to an online interview due to COVID				x			
2	Case study 4	London	x			x	x	x		X
2	Case study 5	Wales		x		x	x	x	x	X
2	Case study 6	Wales			x	x	x	x	x	X
2	Case study 7	Merseyside	x			x	x	x		X
2	Case study 8	Yorkshire		x		x	x			x

Note: These eight case studies for qualitative fieldwork (covering both England and Wales) represent more than 30% of the schools for which post-test data was collected and matched (total=23; see Figure 5).

RQ11. What are the key factors which influence the successful delivery of the NEBT programme in years 1 and 2?

Key points

- Several enablers to successful programme delivery were identified, including having a positive relationship between teachers and instructors, having a suitable physical space for sessions and receiving support and flexibility from both mothers and school settings.
- Conversely, barriers were also noted by interviewees. These included difficulties in the recruitment of mothers and babies for some schools. Sufficient time for preparation for sessions was raised as an issue, alongside difficulties in table tabling sessions. For some instructors the training felt overwhelming, particularly if they had limited experience of a similar role or programme.

Enablers

- Institutional support and sufficient resourcing

Schools that successfully implemented the programme demonstrated flexibility in integrating NEBT within their existing operations. This included accommodating extended family visits, allocating preparation time for instructors and prioritising sessions in the weekly schedule. The selection of instructors who were not classroom-based staff proved advantageous in some schools, facilitating their release for programme activities, including multi-day training. In one school, having a dedicated team of four staff members for wellbeing further supported implementation. Overall, strong institutional support and adequate resourcing emerged as important factors.

> *"*[The school has] *been great. They've just let me get on with it. I've said I need a good hour to plan my sessions, and it's like 'Yeah sure, is Monday morning okay, do the sessions on a Monday afternoon?' I'm like 'Yeah, fine.' So it's an hour to an hour and a half. It's fine. They just let me get on with it, and I just sort of do it, and they go, 'You alright? Do you need anything?' and just crack on".* Instructor, CS7

- Dynamics between instructors and teachers

In connection with the above, the quality of classroom dynamics between teachers and instructors significantly influenced programme delivery. Most case study schools reported productive relationships, with instructors feeling empowered to lead sessions and a sense of support for the programme. For example, the instructor in CS6 commented: "The class teacher I've got, he was amazing – he was like, you tell me, and I will work the class around it". However, in one school (CS8), the instructor felt that a lack of clear top-down endorsement undermined the instructor–teacher relationship, as the value and purpose of the programme were not made clear.

- Maternal engagement

Engagement from parents recruited by schools for the sessions emerged as another key success factor. In four schools (CS2, CS4, CS5, CS6), teachers and instructors highlighted mothers' positive contributions through their flexibility in attending sessions, their relaxed approach and their willingness to share their experiences with the children. "I mean the crying was quite poignant to them at the time because mum was really open about how she managed and how she felt when [baby] was crying". Instructor, CS5.

- Selection of the space and class

Physical space considerations and appropriate class selection also contributed to successful delivery. Two schools found art spaces particularly suitable, offering adequate room for circle-based activities while providing an environment distinct from regular classrooms. Additionally, some interviewees reported careful consideration of a class that was the right fit for the programme as a factor in success.

> *"It's just worked really well, and it's worked well in that class. Again, there is a mixture of pupils in there. They were quite diverse, the little group you had."* Instructor, CS5

Barriers

- Timetabling

Time constraints presented as a barrier in several ways across the schools. Four schools struggled to accommodate sessions within crowded timetables, with competing activities, such as school trips, forcing rescheduling. As two head teachers noted, time management remained an ongoing challenge within the school environment: "It's a big investment in time for other things that are going on in the school as well, so you've got to be aware of what's happening". Head teacher, CS6.

Programme scheduling created additional pressures. Two instructors reported difficulties delivering the required number of sessions within truncated timeframes. In CS3, a month-long delay between training and implementation led to condensed delivery, causing the instructor to be concerned that the training would not be fresh in her mind. Similarly, CS6 reported "trying to squish everything in" due to starting in November, shortly before the Christmas break. Some instructors specifically avoided scheduling family visits before holidays because they were concerned about the impact on post-session effectiveness due to the children's memories.

- Staff availability and planning time

Staffing limitations also affected programme delivery. The presence of supply teachers in two schools (CS2, CS8) complicated the initial implementation of sessions, as instructors felt they did not wish to cause further disruption. Preparation time posed a particular challenge for TAs who, unlike teachers, lacked designated planning, preparation and assessment time. For example, one instructor reported that it took her anywhere up to "a couple of hours" to prepare the sessions.

> *"You're timetabled up to every minute basically. Same with the lunch times – you don't have a lunch time or a break time because you're on [duty] outside. And in the morning, when you get here, you're preparing for the day. When the kids are going, you're finishing off and preparing for tomorrow"*. Instructor, CS2

Similarly, another instructor reported doing this work at home or after school. When asked whether they were compensated for prep time, they reported that although they were supposed to ask for preparation time, short staffing at the school made this unrealistic.

- Instructor suitability and prior experience

Instructors' previous experience may have presented a barrier to successful delivery. Instructors were selected from a mixture of roles, including some TAs and others in support roles, including a school receptionist, a family engagement lead and a member of a wellbeing team. Only one of the instructors had previous experience as a teacher in a school setting. Most instructors had limited or no prior experience delivering a programme like NEBT (one had delivered a bereavement intervention, Rainbows, and one had completed training in the Flourish programme, aimed at 11–14-year-olds).

Although feedback on the training, mentor support and teaching materials was generally positive, some challenges emerged. A subset of instructors found the three-day training intensive, describing the content

volume and complexity as “daunting” (instructor, CS8) and “overwhelming” (instructor, CS5). In some cases, instructors linked these feelings with their lack of formal training or experience in a classroom setting. As one instructor from CS6 noted, other trainees who were TAs already had a grounding in child development and classroom management skills, which she lacked: “I feel like they were kind of only learning the ROE [programme], and I was learning everything”.

- Suitability of training materials

The digital format of teaching materials presented additional hurdles at times. Half of the instructors reported difficulties navigating the iPad-based manual and printing resources, expressing a preference for paper materials. However, some linked this with the fact that it was their first time running the programme and expected to become more adept with the materials over time.

- Parent and baby recruitment

Initial recruitment challenges in finding a suitable parent and baby affected three schools (CS1, CS2, CS8), who found publicity channels, such as school blogs, social media channels and local posters (in a GP’s office), unfruitful.

> *“I know that some of them [instructors] struggled really [badly] trying to find a baby, but I was really lucky. I'm now panicking about next year”. Instructor, CS7*

While all schools eventually secured a parent and baby (one through a staff member, one through a local church via a school parent), one instructor expressed safeguarding concerns about their recruited family’s suitability, but they felt constrained by limited options.

> *“I did have some doubts at first because, being part of safeguarding, I know the family [has] had some issues involved with social care and things in the past, and I didn’t know if they were going to be the best choice, and I wasn’t 100% sure about* [mother’s] *commitment to the programme ... I think she has needed a lot of encouragement [during] the family visits. She sort of gives quite short answers, and she doesn’t really say a great deal”.* Instructor, CS8

An unusual situation was reported by the instructor in CS3, where the NEBT team directly recruited the participating family. The instructor was not able to explain why this had been the case, reporting that she “did not get a chance to look locally” for a family. This resulted in logistical complications due to the family’s distance from the school, though the mother's commitment to the programme helped overcome these challenges. Most schools demonstrated flexibility in scheduling family visits to ensure that these took place, adapting to circumstances, such as illness, holidays and train strikes; however, CS2 and CS8 reported missed sessions.

RQ12. What are the perceptions of pupils, teachers, deliverers and instructors about the effectiveness of the programme in years 1 and 2?

Key points

- Despite a few initial worries from some schools, the programme was well received overall, with improved levels of empathy, patience, tolerance and prosocial behaviours reported by many staff members. Additionally, pupils were able to speak about the programme and how they felt about it.
- Lower levels of impact were perceived where behaviour and wellbeing were already seen as good and where in-school support for the programme was seen as lacking.
- There was some evidence of different engagement patterns between girls and boys. Where necessary, adaptations were made for pupils with SEND, particularly in the pre- and post-visit sessions.
- Pupils were particularly well engaged with family visits and enjoyed seeing the baby. Focus groups indicated that some session topics had been better received and remembered than others. There were mixed perceptions of effectiveness for children from diverse socio-economic and family backgrounds.
- In addition, there were some significant unintended consequences for some schools, where some pupils became upset, in particular looked after children, those with caring responsibilities and those who felt they did not have the mother–baby bond central to the programme content.

This section begins by summarising comments provided about enthusiasm for the programme among school staff, parents and participating pupils before addressing interviewees' perspectives on the effectiveness of the programme in relation to the logic model.

- Reception among the school community

Interview responses indicated that the programme has been very well received within schools. A positive reception was reported in all but one case study school, with active student engagement being a common theme. Several schools explicitly mentioned strong leadership support for the programme and/or an intent to continue or expand the programme (CS2, CS3, CS5, CS6).

Some instructors and teachers had been concerned that certain elements of the programme, such as the take-home health messaging leaflets or the session which addressed shaking a baby, could prompt negative reactions from children's families. However, these fears proved unfounded, and none of the interviews reported parental complaints about the NEBT programme.

> *"[With the shaken baby session] we were thinking, 'Oh god, we're going to get a backlash; we're going to have parents saying, "What are you teaching our kids, what are you doing showing them something like that?"' but they didn't".* Instructor, CS7

"[T]here was another [leaflet] about smoking; I was just like, 'Oh my god, are we giving this out', and I was a bit worried because I thought that some of the children might turn around and say, 'But my mum', and you don't know what the children are going to say … but I had no response. No response from anyone. Not the children, or the parents. But I was worried about it". Teacher, CS5

- Pupil engagement

Reflecting the ToC (M1 and CMI3), staff typically perceived elements of the programme that facilitated experiential learning to be more effective for pupil engagement. Interviews specifically commented on high engagement with family visits (CS1, CS2, CS4, CS5) and with hands-on/practical activities, such as the demonstration, which involved shaking an egg in a jar to convey the dangers of shaking a baby (CS1, CS2, CS7). The opportunity to observe and interact with a real baby was perceived by many of the interviewees as a core aspect of the programme's effectiveness.

"They see the real baby, and they see it happening, the thing, in front of them. So nobody told them that the baby did this or does that. They can see the baby is doing, and they can relate [to] the baby, with the feelings that the baby has". Instructor, CS3

In the pupil focus groups, children were very positive about the family visits, stating that it had been fun and exciting to see the baby grow, develop and start interacting with the world around them. One child stated that seeing the baby made them happy; another noted that the highlight was being the first to hold the baby. They were able to recall things that the baby had done, for example, playing with toys and crawling. Pupils enjoyed singing to the baby, particularly when they were able to sing a song that they had created for the baby.

Session observations support these findings, with engagement observed to be generally good, particularly when the baby was present**.** Pupils were keen to ask and answer questions posed by the instructors and to participate in discussions. While some children were less keen to engage directly, such as the invitation to tickle the baby's toes, they did appear to be watching the baby.

Focus groups indicated that some topics had particularly stuck with the children. Pupils clearly remembered the dangers associated with smoking and with shaking babies.

"*[I] remember theme 2, with the dangers of shaking the baby …. and also being sick on the alcohol and second-hand smoke. I remember that".* Pupil, CS6

"Like that you should not shake babies because the fluid in their head will mix and it could cause a sign of death, blindness and cause your baby to not hear". Pupil CS4

In terms of aspects that were not considered to be as engaging, the lesson on nappies was specifically identified by staff in two schools (CS4, CS6). In the pupil focus groups, the session on nappies also elicited more negative responses, as did times when the baby cried, was sick or dribbled. At the same time, focus group comments suggested that the children better understood crying; they spoke about how you could tell a baby was tired by their body language, that as babies do not have language, they communicate by crying and that these cries may be differentiated in terms of tiredness or hunger.

"That you have two tones of crying …There is, like, a different one for sleeping, when you're sleepy. And then another one is when he's, like, hungry, and his mum understands that". Pupil, CS5

- Perceived impact of the programme in relation on the logic model

In general, teachers and instructors felt that participating in the programme had positively impacted children's empathy and prosocial behaviours. Several schools (CS1, CS2, CS7) reported clear improvements in empathy, with the teacher in CS2 estimating that "about 70% of them have got greater empathy and tolerance to other people now." Specific examples of this included more patient and caring interactions between children, a better understanding of each other's perspectives and increased tolerance of differences. Some comments linked the NEBT sessions with positive changes to in-classroom behaviour among participating children and, less commonly, in the playground or at home. For example, one teacher (CS1) spoke about two different pupils whose parents had commented on improvements in how they interacted with younger siblings.

A theme across most case study schools (CS1, CS2, CS4, CS6, CS7, CS8) was a perception of improved emotional awareness and expression. Examples given here included perceptions that the programme had enhanced pupils' emotional literacy, increased their willingness to discuss feelings (particularly among boys), improved their understanding of the validity of emotions and increased their confidence in sharing personal experiences. For example, one instructor (CS8) commented that the children had broadened their vocabulary for expressing emotion, using terms like 'frustrated' where they would previously rely on the more limited happy/sad.

Several of the schools commented on behavioural improvements among the participants, including calmer classroom behaviour, fewer 'fallings-out', more independent conflict resolution, less cliquey behaviour and better mixed-gender cooperation. In one school (CS5), an instructor perceived that the programme had "made a definite impact on [bullying] levels". The instructor reported that the school had been experiencing "massive issues" with bullying, primarily outside of school in the form of cyberbullying, and felt the programme had "come hand-in-hand at the right time" to help tackle this.

One teacher (CS1) perceived that the "calm environment" of the baby visit sessions had a prolonged effect, helping the children to be better attuned to each other's emotional needs:

> *"[Male pupil] was [sitting] on a chair; [he had] been picking his scab, so it was bleeding quite badly. And straight away, some of the kids were, like, 'No problem, we'll help you', and were calm with him, went and got him things that he needed. Whereas if it [were] a normal lesson, they'd ignore him."*
> Teacher, CS1

This was reflected in the perceptions of an SLT interviewee at the school, who reported that "since the children have been accessing it, particularly, that classroom behaviours have calmed down a lot" (SLT, CS1). Several focus group comments suggested that at least some of the pupils had learned to label the baby's emotions, reflect on their own emotions and bridge these to understanding the emotions of others (CMI3).

In terms of thinking about learning about emotions and how a baby might express itself, pupils spoke about how the baby's emotions changed quickly and how they used their body language, cries and gaze to communicate.

> *"So, if he needs to get something, he looks at us in the eye directly to communicate to give it to us".*
> Pupil, CS6

Pupils were also able to speak about what they had learned about their own emotions and how they treat others. They were able to name some of the specific emotions they had learned about, including sympathy and empathy, and to differentiate between them.

Some pupils stated that they might be more able to deal with their own emotions now and that this had given them confidence to share what they think or feel in class and about moving up the school.

> *"Because the first time I shared something with the class when we were learning about baby* [name], *I felt a bit embarrassed, but now I do it quite a lot, and I feel fine".* Pupil, CS4

> *"Since* [instructor] *has teached* [sic] *us all the lessons, I've felt more confident, and I used to be scared of going in [Year 6], but now I'm not that scared*". Pupil, CS4

Going beyond the baby and themselves, pupils spoke about how people might express their emotions and how, for example, anger might evoke different reactions from different people, that people have differing temperaments and how you might be able to tell how someone is feeling by their body language.

> *"Some people react with violence, some people react in language and then some people react in emotions".* Pupil, CS8

In addition, a couple of pupils linked what they had seen in the family sessions with their own lives, realising that their parents would have looked after them in the way that they saw the mother looking after her baby, and reported that this had changed the way that they behaved at home.

> *"How* [mother] *takes care of baby* [name] *and how she really loves him makes me feel like I remember to not be really rude to my mum and dad because I go back and remember when I was a baby, that is what my mum and dad would have done as well".* Pupil, CS4

Conversely, interviewees in two schools reported limitations in the programme's impact. In CS6, the teacher did not believe that the programme had improved mental wellbeing levels or reduced challenging behaviours because these were already at a good level: "I wouldn't say it was relevant because they are always quite well behaved, and we have [a] positive mindset and things like that anyway" (Teacher, CS6). In CS8, the instructor attributed limited impact to inadequate top-down support and unsuitable class selection, stating, "I would have picked a different one and certainly not one that's had such an up and down year" (Instructor, CS8).

Engagement and impact among diverse groups

Although the evaluation did not set out to collect the quantitative data needed to perform subgroup analyses, interviews asked teachers and instructors to comment on any observed differences in engagement or effectiveness between different groups, e.g. in relation to gender, SEND or socio-economic status, and to comment on any adaptations to delivery with the intention of making the sessions more inclusive.

Sex

Most schools commented on different engagement patterns among boys and girls, reporting immediate enthusiasm among girls ("I think the girls initially, at the start, were very much engaged and cooing over the baby. 'Oh, we've got a baby!'" Teacher, CS2) and initial resistance among boys ("At the beginning, the boys weren't engaged with the baby. They were a bit silly about it" Instructor, CS4). However, schools generally described an evolution in boys' engagement levels over the course of the programme, with most boys'

attitudes shifting positively through direct interaction with the baby. Two schools reported persistent self-consciousness or reluctance to take part in activities among small groups of older boys (CS2, CS8).

> *"The girls are straight on. Whereas the boys – sometimes they're a bit like, 'Oh no, I'm too cool for this' ... But if baby [name] comes towards them, you can see their faces light up. So, they are more reluctant to want to enjoy it, but they do. Whereas the girls straight away love this session".* Teacher, CS1

Some interviewees commented on differences in programme effectiveness between boys and girls. Boys were generally perceived as having "more distance to travel" in terms of expressing their emotions and initially being more reluctant to engage with the programme, although this was not universal ("I thought boys wouldn't, but they are probably more engaged than the girls sometimes" [Instructor, CS2]). One teacher (CS4) felt that the programme had helped quieter girls find their voices while conversely helping already confident boys become better listeners. Some specific positive changes included a boy developing his "gentle side" through interactions with the baby (CS1) and boys in a class gradually opening up:

> *"I think it was after – I think it was by about theme three or four, the boys really start opening up, 'Oh yeah, I cried because this happened/I kicked someone at football because I was angry, and I know I shouldn't have, but I said sorry', and it was like – wow".* Instructor, CS4

- SEND

Interviewees also discussed the engagement levels of pupils with SEND. While SEND pupils could find pre-/post-visit sessions a challenge due to greater focus and writing requirements (CS1, CS2, CS4, CS8), they were perceived as enjoying and engaging more when the baby was present. Some barriers noted in relation to inclusive adaptations were the difficulties faced by some SEND pupils in relation to change and new people (CS4) and that topics exceeded comprehension levels for some students:

> *"Some of them, like the lower ability ones, some of the stronger topics then have not even really registered with them. It's gone over their head because their understanding isn't there".* Instructor, CS5

Schools implemented various adaptations to facilitate the involvement of SEND pupils, including shorter stretches of participation (which could be extended if the pupil was comfortable to stay), sign language interpretation to include deaf children in songs, simplified materials prepared in advance of sessions (it is not clear whether this was agreed with NEBT) and flexibility about children's behaviour during the sessions (choosing to sit further away, doodling).

> *"There is a pupil with ASD … he really enjoys the family visits, but he struggles to engage with the pre and post, particularly with writing tasks and things like that, so I had a chat with the TA about how we could adapt it for him and sort of simplify things a bit more so that he could join in".* Teacher, CS8

Reflecting this feedback on engagement levels, some comments suggest that the programme may have been less effective for children with SEND, either because they had been less "actively involved" in the sessions (Instructor, CS4) or because some of the "stronger topics" were too advanced for some children to fully grasp (Instructor, CS5).

- Socio-economic status

Teachers offered mixed perspectives on the programme's effectiveness for children from diverse socio-economic and family backgrounds. In CS2, the teacher described the school as serving "a lot of single-parent families, a lot of broken families" in a deprived area. This teacher observed that children from complex family circumstances "really thrived" in the programme, discussing the NEBT lessons "constantly" at home. In contrast, parents of children with a more "secure home life" remained unaware of the programme. However, a teacher in CS5 presented a contrasting view. They believed that children from families with fewer resources to provide "exposure" to "life experiences" outside of school struggled to relate to some of the programme content. This teacher perceived that the disconnect between the content of the NEBT programme and the children's experiences hindered their engagement.

Unintended consequences

The most significant unintended consequence described by interviewees related to children from complex home situations. In four schools (CS2, CS5, CS7, CS8), staff reported that content had, in some cases, caused distress to children whose family circumstances were not reflective of the NEBT content – particularly looked after children, young carers, or those who otherwise lacked the close mother–baby bond portrayed in the programme. The CS5 instructor noted that sessions had "brought up some emotions" and led to tears for some children, especially for one looked after child. A similar situation arose in CS8, where a looked after child who had experienced the breakdown of their family placement during the programme shifted from being happy to participate to "almost uncomfortable to see the relationship with mum and baby" (Instructor, CS8). Some instructors talked about approaches they took to mitigate this discomfort: not insisting children actively participate and amending the language used to describe families to make this more inclusive ("I never say just 'mum and dad'. I say, 'or the adults who look after you'" (Instructor, CS2). Additional concerns emerged in relation to pupils who had experienced recent losses, including one child who had lost a baby sibling and another (in the school, not in the NEBT class) whose mother had passed away. In the case of the baby loss, the child was asked in advance of a session which touched upon infant death (in relation to shaking a baby), if they would prefer not to attend. Additionally, the school opted not to send home NEBT leaflets about behaviours that can cause harm to babies.

> *"The experiences that these children have, school is often their safe place, and I think they sometimes see that: look at baby* [name] *with his mum. It's upset some of them. You've had like, 'I haven't got a mum and dad', or 'I don't live with my mum and dad".* Instructor, CS5

RQ13. What fidelity issues are observed during years 1 and 2 of the trial?

Key points

- Data collected shows that fidelity to the programme was at a good level, although delivery took place flexibility in some settings to overcome challenges, for example timetabling.
- Schools reported making changes to language and also altering some content to fit in with the school population, in particular for pupils from Gypsy or Roma backgrounds, and where a school family had experienced the loss of a baby.
- Some instructors developed, extended or purchased additional resources to be used in sessions.
- Apart from a minority of cases of poor behaviour in sessions, where teachers intervened, Instructors ran sessions independently.

Were the sessions delivered as timetabled?

As outlined in Appendix E, across the wider trial, the baseline sample of schools (N=46) achieved 61% compliance in delivering eight or more topics from the NEBT curriculum, with the ITT complete case sample for the primary outcome (N=23), achieving 96% compliance. Interview data from the case study schools indicated that compliance was generally high, though instructors sometimes had to be flexible to maintain this level of delivery.

Across the eight case study schools, there was a diversity of approaches reported in terms of timetabling the sessions. There were a number of challenges to overcome in delivering all the content; however, all schools reported that they had covered, or expected to cover, all the themes. Two schools reported that sessions were not scheduled on a set day or time and instead shifted week to week to fit around other lessons; a teacher at one school commented that this made it difficult to ensure consistent attendance by the same children in every session.

In a few cases, school staff made the decision to pause NEBT sessions the week before a school holiday if a family visit was scheduled, shifting them to after the holiday to retain an unbroken pre-visit/visit/post-visit session format. This was considered preferable for the children's engagement and retention of the topic.

Every school reported at least one incident of needing to push back a session, citing a variety of reasons: a mother/baby being unavailable (illness, holiday), train strikes, a COVID outbreak and a conflict in the school timetable (e.g. a trip). Similarly, two schools reported delayed starts. To cover all the content in fewer weeks, interviewees across schools described doubling up sessions in some weeks.

> *"Because the terms after Christmas were really short, they are only like five weeks long, so some weeks we had to do two lessons in one week because otherwise, we're not going to fit it all in before we break up in July. I think next year, if you've done the training and you could find a baby, you could start in September, and then you've kind of got a bit more time to fit everything in*". Instructor, CS6

Were the sessions delivered as intended?

- Modifications to language, content or delivery

A minor adaptation mentioned by several interviewees was modifying the Canadian English language used in the materials to terms that would be more familiar to children in the UK. Relatedly, in session observations, Welsh schools were observed to do the countdowns for songs and deep breathing in Welsh.

> *"Some of the research – all their data will say, 'In Canada, this this this …'. So I've changed it to find out UK-based. When it says diaper, I change it to nappy. Anything that's wording that's not how we speak. The sign language that was Canadian, I've changed to British. It felt more appropriate for them to know".* Instructor, CS2

An instructor in one school (CS5) described modifying the ROE session delivery approach to better fit with school practices. ROE's approach promotes an equal weighting between children's and adults' thoughts, feelings and opinions; acknowledges contributions but avoids praise; and aims to create a more relaxed, risk-free learning environment.[21] They linked this to feeling that some boundaries to behaviour were necessary for sessions to run smoothly and to avoid confusing the children when they had to adhere to the school's typical behaviour expectations in other classes.

A contextual factor for one school was the consideration of pupils from Gypsy Roma and Traveller (GRT) families. This included omitting topics, such as breastfeeding, to be inclusive of these children: "We've got a lot of traveller children, and you can't do anything about reproduction and how babies eat because the parents will take them out of the class" (Instructor, CS6). It was judged that if these adaptations were not made, the GRT children would have been removed by their parents on subsequent NEBT session days and would therefore have missed out on a full day of schooling.

Several interviewees raised concerns about the content of NEBT, in particular, the centrality of a close mother–baby bond, being potentially alienating or upsetting for some children. One instructor described making small adjustments to language with the aim of making the content more inclusive of diverse family arrangements:

> *"Maybe, for the improvements, they could say that mother doesn't mean that you had to give birth. Maybe they could be just more in-depth about ... [it could be] the person that is in charge of looking after you. Maybe some [adaptations] like that might be ... I do always say, when I say stuff, I never say just 'mum and dad', I say, 'or the adults who look after you'".* Instructor, CS2

In at least one school, the decision was made not to send home some of the health messaging materials (in this case, about the harms to babies caused by smoking) out of sensitivity towards a recent baby loss in the school community.

- Supplementary materials or instructions

Some instructors and teachers described going above and beyond the NEBT materials, creating supplementary resources or expanding on the directed approach to delivering the sessions. These were discussed in terms of knowing the needs of the specific children in the class and ensuring that they would be able to engage fully with the content. For example, a teacher at CS2 talked about some of the content not being fully comprehensible for SEND or children with lower attainment and taking time to break down

21 More information about the Roots of Empathy learning environment ethos can be found here: ROE_Freeing_up_Children_to_Learn_May_2014.pdf

the material into more detail: “Sometimes, I’ve taken them out and just had a bit more in-depth one-to-one conversation about it if I can see that they’re not understanding it”.

A teacher at CS4 mentioned the instructor bringing in extra resources, such as her own child’s dummies and blankets, and felt this made it more engaging for the children. Similarly, the instructor at CS5 talked about buying some board books for the children to read to the baby after the prescribed content for the session had been delivered. Another example was an instructor making additional prompt cards:

> *“For the last one, the safety one, it was about getting them in groups to discuss things, and I thought, there’s nothing to prompt them – do you know what I mean? So I made little prompt cards about which room I [wanted] them to focus on when it’s looking at safety. I gave them a room each and said write it on there. So just thinking and knowing the children really and thinking”.* Instructor, CS1

- Classroom roles

In all case study schools, interviewees reported that the instructors had taken the lead on delivering the sessions as intended, while teachers had adopted a supportive role. This support sometimes included classroom management if behaviour issues arose during the session (CS2). Instructors described seeking teacher assistance when needed but otherwise handling delivery of the content themselves. Session observations confirmed this dynamic, with teachers present in all but one observed session and assuming a supportive role, particularly around behaviour management. School CS8 was an exception, where pupil behaviour proved challenging during the observed session. Unlike the other sessions observed, this one appeared poorly planned, with no defined learning outcomes. In the other four sessions we observed, instructors – sometimes with teacher input – maintained a good level of discipline and were able to bring pupils back to the topic or task at hand. During these sessions, instructors effectively linked learning points and topics back to the pupils’ own lives and experiences, for example, around tradition and communication. In one example, the CS6 instructor was able to address contentious issues while linking them to the topic under discussion.[22]

[22] A longer description of this observation can be found in appendix F, page 119.

RQ14. What does the trial indicate about scalability?

Key points

- Recruitment of a mother and baby was challenging for several schools and was experienced as stressful for some.
- In addition, with schools increasingly busy, many instructors found that the time for in-school session preparation was limited and therefore this took place in their own time. Given the low levels of TA pay this feels problematic.
- Most schools have an established Social Emotional Mental Health (SEMH) programme. There may no be the appetite to introduce a new programme, particularly in England where curriculum time in this area is very limited.

The IPE fieldwork has surfaced a few issues around the potential for further scaling the NEBT programme in England and Wales. The first of these centres around the recruitment of the parent and baby, which some schools found challenging, and it did not always work as well as had been hoped. This also appeared to present a degree of stress for some instructors.

Schools are very busy places and staff members – whether TAs, teachers or those in other roles – are under increasing pressure. The interviews with NEBT instructors found that many were doing unpaid preparation work in their own time, and some felt unsupported by the school. As TAs are amongst the lowest-paid and least-trained staff in schools, this may feel particularly unfair. Some schools may already have established SEMH programmes, making the introduction of an additional one feel impractical, especially in English schools, where such areas are not a primary focus within the curriculum.

Usual practice

Most schools in the intervention group already had an established PSHE curriculum, such as Jigsaw[23] (CS2, CS5, CS6) or SEAL[24] (CS2). Interviewees at four of the schools described the approach as trauma-informed, while others were not sure or did not comment on this. Existing wellbeing approaches mentioned included a cross-class daily wellbeing check-in and community building initiative referred to as Crew (CS1), a Thrive[25] room and practitioners (CS4), and zones of regulation (CS8). Some commented on the no-praise/-reward approach used in NEBT as a particular difference from usual practice. Displacement of existing provision varied; where interviewees commented on this, some mentioned separate PSHE running as usual (CS6) and others mentioned a partial replacement: “We haven't fit in as much of [Jigsaw] as we could have because this kind of took that place” (Instructor, CS5). This may be problematic if pupils are missing out on key PSHE content, and it is worth considering for future evaluations and is worth raising with schools when planning the delivery of the programme.

[23] Jigsaw provides online schemes of learning for PSHE and RE. See: https://jigsaweducationgroup.com/

[24] SEAL stands for Social and Emotional Learning, an approach to promoting social and emotional skills that is widely used in English and Welsh schools. See: https://www.sealcommunity.org/node/1735. See:

[25] Thrive is a whole school, trauma- informed, mental health and wellbeing initiative. See: https://www.thriveapproach.com/.

There was some evidence of spillover to non-NEBT classes. In one school (CS2), the instructor reported using learning from NEBT with a Year 3 class, and in another (CS4), the teacher reported that Year 5 children from other classes were sometimes present in NEBT sessions.[26] In another school (CS1), the participating class shared learnings during 'Crew', a pre-existing daily school community-building initiative during which children from different classes within a year group spent time in mixed groups:

"What we've started doing is, the crew that's from my class will discuss it with the other Year 5 class because sometimes they get a bit jealous that they're not allowed to take part … So we're trying to pass on the learning to them as well". Teacher, CS1

There was no evidence to suggest that control schools received the ROE programme, but no information was collected that detailed whether they were running similar interventions.

Cost information

Due to issues with attrition and the prioritisation of the collection of outcome data, no further burden was placed on schools to collect further cost data. As such, the information given below has been supplied by ROE and through the evaluator's knowledge of programme delivery.

The table below provides a summary of the ROE NEBT intervention estimated costs. For the current trial, the NEBT programme instructor was located within the participating school.

All ROE programmes in the UK and the Republic of Ireland use a combination of true community volunteers and non-instructional staff from schools. That is, ROE recruits volunteer instructors from both the community and participating schools, whereas in normal circumstances, schools volunteer staff about 50% of the time or less. A 50/50 blend of staff and volunteers allows the training of local mentors who can support the development of new instructors and ongoing instructor growth. This also circumvents the school-based staff's lack of capacity and/or any administrative barriers to their serving as mentors. However, for purposes of this evaluation, this research design was restricted to recruiting only school-based instructors. The costings below are based on what was delivered for the evaluation, but we also provide costings for the typical/usual model delivered by ROE.

ROE covers the cost of training volunteer instructors through philanthropy, so there are no training costs for the community instructor model detailed below.

However, ROE offers both school and community instructor models, with the community instructor model costing less. As such, costs for both delivery models are included below.

Total costs for the school-based instructor model are £1,631.15 in 2024 prices for delivery to 30 pupils in a school. This equates to £54.36 per pupil. For a community-based instructor model, the cost is £800 equating to £26.66 per pupil.

[26] It is not typical practice to have children who are not regular participants present in NEBT sessions, as the intended implementation model seeks to build a risk-free learning environment.

Table 18. Summary costs

	Total costs	**Cost per participant**
School-based instructor model without instructor training costs*		
Setup	£398.40	£13.28
Recurring	£1,235.75	£41.19
Total	£1,634.15	£54.47
School-based instructor model with instructor training costs		
Setup	£3,098.40	£103.28
Recurring	£1,235.75	£41.19
Total	£4,334.15	£144.47
Community instructor model **		
Setup	£800	£26.66
Recurring	-	-
Total	£800	£26.66

*ROE usually covers instructor training through philanthropy

**No further costs, other than programme set-up materials, have been provided for the community instructor model. As the instructor is recruited from the community or non-instructional staff in the school, this model costs the school considerably less than the school instructor model.

Pre-requisite costs

For the school-based instructor model, it is assumed that prior to the intervention, schools have an appropriate member of staff who can be trained to deliver the NEBT programme. For the community instructor model, it is assumed that a member of staff in the school will work with the ROE team to identify a suitable community instructor. As ROE typically works with regions and schools that have actively requested the ROE programme, the support both within the school and in the community needed to ensure successful programme delivery is expected to be in place.

In addition, instructors will have access to a classroom to deliver the intervention for each session for all nine topics, will have suitable facilities to welcome a parent and baby into school and will be able to contact parents to opt their child out of the intervention if requested. It is also assumed that members of staff in the school will be able to give time to liaise with ROE when setting up the programme.

Set-up costs

The set-up costs detailed in this section are those which occur at the beginning of the NEBT intervention, regardless of delivery model. ROE provide four days of training to trainee instructors; instructors have to locate a suitable family and schedule when family visits take place. For community instructors, there is an additional introductory session for the community instructors to meet the school staff. As detailed above, ROE covers the cost of training instructors via philanthropy.

Recurring costs

Recurring costs need to happen year-on-year and include programme delivery and the instructors' time required to prepare for the sessions. In addition, there is yearly training that instructors have to attend and mentoring support for instructors. Volunteer families are not compensated for their time in the ROE programme.

Items not included in the cost analysis

Data on the time required for instructors to identify a suitable family to take part in the intervention was not collected. From the IPE analysis, it is clear that the time taken varied between schools, with some reporting struggling to identify a parent and baby. In the school-based instructor model, this time would be at a cost to the school in the form of instructor time, but it is not accounted for in these costs.

In addition, the cost for schools to contact parents for opt-out/opt-in consent has also not been factored into this model. This may not be relevant for future programmes where evaluations are not taking place, but schools may differ in their policies.

Lesson planning time for instructors has not been included in the model separately. The calculation assumes that the hours for programme delivery (35) cover planning time. However, it is worth noting that planning time was highlighted as an issue for instructors in the current evaluation, with one instructor stating they didn't receive planning time and that it took her anywhere up to "a couple of hours" to prepare the sessions. Lack of available planning time may be due to individual school pressures and the level of support from the leadership team. Schools may nonetheless wish to consider costing in additional planning time to ensure sessions can be delivered as expected per the programme manual.

Instructors reported needing to print lesson materials from the iPad provided when planning sessions. These costs have not been included as a record of printing required per session, per school, is not available.

Contextual factors impacting cost estimates

ROE reports that for many programmes, the full cost of training volunteer instructors for schools can often be covered by philanthropic funding. As such, the cost estimates reported are a guide only, and future schools may explore ways to cover costs by alternative means.

The instructors in the ROE NEBT trial were from a range of support positions in the school (e.g. TAs or SENCO). We have based the hourly wage for the cost calculations on the average of the hourly wage of Higher Level Teaching Assistants (HLTAs) without a SEN specialty (£12.26), HLTAs with a SEN specialty (£13.03), TAs (£12.26) and educational support assistants (£12.26); the estimates were taken from gov.uk. The average hourly wage used was £12.45.

Table 19. List of items considered for cost estimates

Category	Description	Phase
Staff		
Instructor time training	Four days of training at a location suitable for the location of the school (where possible): 32 hours of training estimated.	Set-up
Family identification and introductions	Instructors should identify a suitable family, and an introductory session should take place	Set-up
Instructor time for delivery and mentoring	Programme delivery of nine topics: three lessons a month for a school year. Delivery can be flexible in some cases to meet the needs of the school.	Recurring yearly
Professional development activities	Direct mentoring, professional development and access to the Virtual Training Institute 35 hours in total	
Programme		
Start-up materials	Startup materials (three t-shirts, an infant book for volunteer baby and parent, Roots of Empathy [ROE] book by Mary Gordon, pamphlets for the junior and senior curriculum, ROE information booklet and infant safety pamphlets for volunteer parents, program description pamphlets for the ROE host classroom teacher and the head teacher, ROE 'Lives Here' posters for the school and a bulletin board topper)	Set-up
Curriculum	Availability via the dashboard and ongoing updates, plus access to a wide variety of professional development videos	Recurring
Programme infrastructure	Annual programme evaluation and ROE's operational expenses and administrative costs	
Buildings and facilities		

Classroom space	Classroom space to deliver the sessions, including a space large enough to allow children to sit in a circle/together rather than at separate tables.	Recurring

Table 20. Cost of implementing the programme for the school in the school-based model

Price year: 2024 **Category**	**Set-up or recurring**	**Schools**	**Total**
Staff Instructor training–trainee instructor hours	Set-up	£398.40	£398.40
Instructor training delivery costs*	Set-up	£2,700	£2,700
Programme delivery, preparation and professional development	Recurring – yearly	£435.75	£435.75
Programme Start-up materials and curriculum, direct mentoring, professional development and access to the Virtual Training Institute for instructors	Recurring	£800.00	£800.00
Buildings and facilities Classroom space	Recurring	£0.00	£0.00
Incentives None	N/A	£0.00	£0.00
Family compensation for time	N/A	£0.00.	£0.00

Cost per participant			
Average number of pupils per school class	-	-	30
Set-up costs per participant	-	£103.28	£103.28
Recurring cost per pupil	-	£41.19	£41.19
Total cost per pupil	-	£144.47	£144.47

* Instructor training costs £2,700, usually covered by philanthropy

**For a community model, the cost would be £800, £26.66 per pupil.

Conclusion

Key conclusions
NEBT had a **small impact** on reducing children's self-reported behavioural difficulties. After the programme, children in NEBT schools reported slightly lower levels of behavioural difficulty compared to their counterparts in schools that did not receive NEBT. This result has an **extremely low** security rating.
NEBT had no impact on children's self-reported emotional difficulties or their cognitive empathy (understanding others' thoughts). It had a moderate impact on their self-reported affective empathy (empathy with others' emotions). NEBT showed a large impact on reducing teacher-reported behavioural difficulties, and this impact was driven by large impacts on reducing peer problems and hyperactivity and moderate impacts on reducing conduct and emotional problems. NEBT also showed a large impact on improving teacher-reported pro-social behaviour. These are the secondary outcomes, which should be interpreted with even more caution.
A very high level of attrition from the evaluation significantly weakens our confidence in the findings. 61% of children who started the trial were not included in the final analysis. 23/46 NEBT schools dropped out shortly after randomisation. School concerns regarding the time taken to deliver NEBT and measurement burden may have contributed to attrition.
61% of intervention schools delivered 8 out of 9 themes from the NEBT curriculum. TAs often made amendments to session scheduling to ensure that the content could be covered in time.
Positive relationships between teachers and the teaching assistants delivering NEBT, physical space for the sessions, and flexibility from mothers and school settings supported delivery. Barriers to delivery included challenges in recruiting mothers for some schools and insufficient time for TAs to prepare for sessions.

Evidence to support the logic model

The findings from the IPE and impact evaluation from both cohort 1 and cohort 2 provide some support for the original logic model (page 15); however, due to issues with attrition and the likely lack of randomness in this, it has to be interpreted with caution. In addition, across the course of the trial, from the intervention schools and the ROE team, we learnt of several program inputs that were not captured in the original ToC and logical model.

The original logic model and ToC failed to capture the following inputs and causal mechanisms for instructors:

- **Causal mechanism for instructors:** Feeling part of a collective group undertaking this journey together was identified as a causal mechanism for instructors. For example, experiences of attending the original training sessions or accessing the community portal resources. Being part of the ROE community is thought to contribute positively to the instructors' experience of the programme.
- **Input for pupils and families:** Health promotional materials, such as 'Don't shake the baby' and 'Smoking in pregnancy' were not included in the original logic model, but are a key part of the programme.

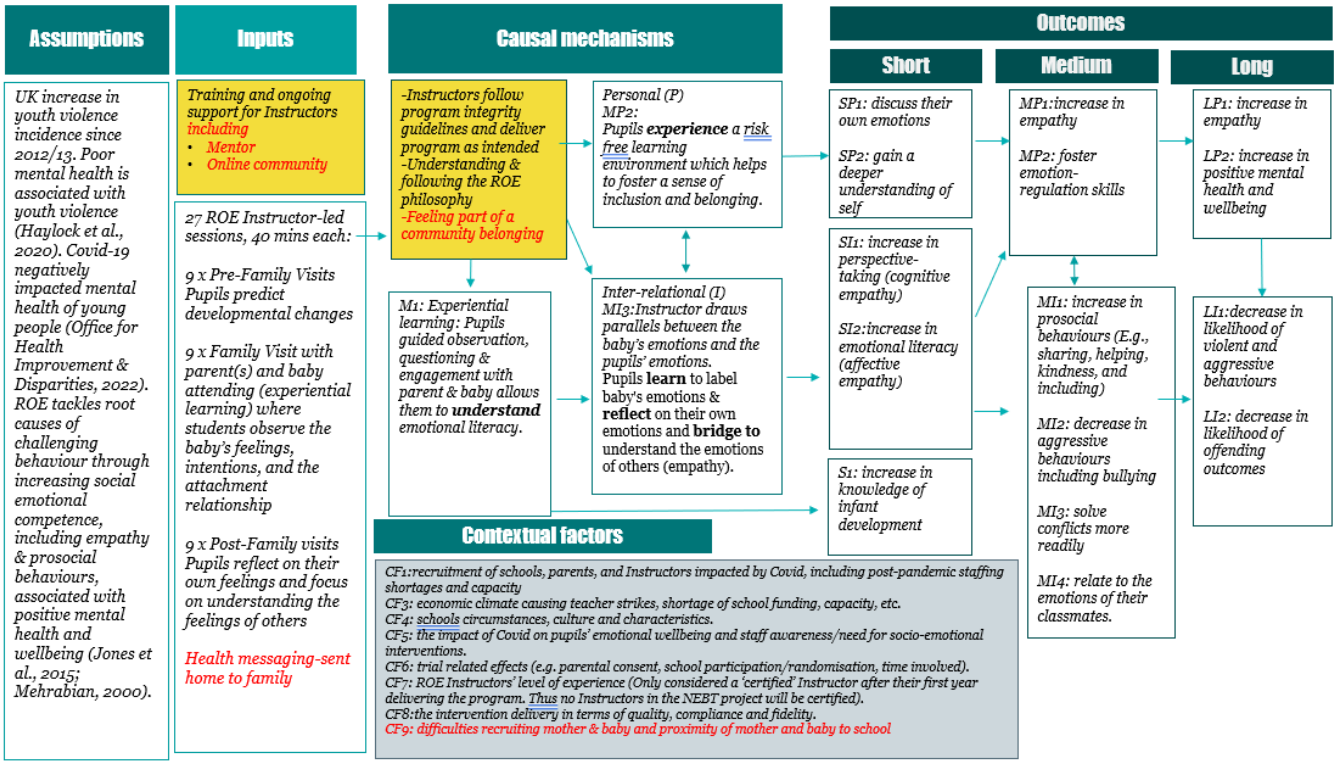


Figure 6. Updated logic model post evaluation

Interpretation

The primary outcome for the trial was BD measured using the self-report M&MF scale (Deighton et al., 2013). The impact analysis found a very small negative impact (reduction in BD) in both cohorts combined when compared to a control. However, as the CIs overlapped, the findings actually represent a reduction in some pupils, no change in others and an increase in some pupils. Given overlapping CIs, levels of attrition across the trial and the fact that attrition is not likely to be random, there is insufficient evidence to support the claim that the NEBT programme led to a decrease in BD. These findings are, however, supported by the rich IPE data that details teachers' perspectives on changes in their pupils' behaviour, which they attributed to the programme. These changes could be evidenced by improved classroom behaviour and fewer occurrences of bullying. The findings on the SDQ, as part of the impact analysis, also demonstrated a reduction in BD (ES=−0.36) and an increase in prosocial behaviour (ES=0.63). This supports the findings from previous research on the ROE programme by Santos et al. (2011), Wrigley, Makara and Elliot (2015) and Connolly et al. (2018), where teacher perceptions of prosocial behaviour were observed, but unfortunately,

given that these were teacher perceptions, they suffer the same limitations. These findings need to be interpreted with caution, as the teachers were aware that their pupils had undertaken the ROE programme. A lack of blinding is associated with overestimation of intervention effects, particularly when the outcomes of interest are subjective (Wood et al., 2008; Hrobjartsson et al., 2014). It is important to acknowledge that, to be valid and reliable, the teacher SDQ requires that the teacher knows the pupil well. In the context of a year-long programme, it would be impossible to meet these requirements of the SDQ and the methodological feature of blinding. One alternative to the teacher SDQ would have been to use the pupil self-report SDQ. However, due to the age of the pupils in this trial, we were not able to change the teacher SDQ to a pupil self-report SDQ.

The findings from the impact analysis and the IPE on affective empathy converge. The impact analysis reported a small positive impact of the intervention on affective empathy with an effect size of 0.19. These are in line with the findings of Wrigley, Makara and Elliot (2015) of increases in affective empathy only. As with the other findings, attrition and a lack of randomness in attrition impact the security of these findings. Instructors and teachers commented on the change in their pupils' emotional literacy and willingness to discuss feelings. Further supporting this, the pupil focus groups highlighted that pupils were able to talk about their emotions, understood how others (the baby) may have felt and why, and appreciated the importance of their emotions. They demonstrated empathic concern, for example, when discussing that babies may feel frustrated at their inability to communicate what they need. These findings are in support of the ToC and also the broader literature on empathy development.

Whilst the findings from the pupil focus groups and teacher interviews are suggestive of pupils behaving in ways that align with a good level of cognitive empathy, the impact analysis does not report a significant impact on cognitive empathy for the intervention group when compared to the control group. Cognitive empathy is an important dimension of the construct of empathy. The reasons for this disparity are unclear, but it is possible that the pupils' levels of cognitive empathy were higher than their levels of affective empathy from the start or that the programme developed affective empathy levels more. Given that the findings of Wrigley, Makara and Elliot (2015) also show increases in affective empathy only and teacher perceptions of improvements in both affective and cognitive empathy, the latter may be more plausible.

Several schools reported adapting the materials that ROE had provided before delivering them to their classes. The guidance from ROE is clear that the programme materials should not be adapted; however, schools felt that some of the language, the level of content and the topics were inappropriate for their settings. It is unclear whether these changes had a substantial impact on the meaning of the materials or whether this would have reduced the impact of the programme, but the finding does need to be considered if the programme continues in English and Welsh schools. In addition to this, four of the case study schools reported unintended negative consequences of the programme due to the disparity between the pupils' individual circumstances and the content of the NEBT sessions. This was especially the case for one looked after child who struggled to engage in the sessions following the breakdown of their family placement. Considering that the number of looked after children in England was 83,630 on 31 March 2024 (DfE,2025) it is logical to assume that this would be an issue across schools, rather than specific to one school in this trial. In addition, the content of some of the NEBT leaflets around baby loss was difficult for some children due to family experiences. Whilst these experiences were not universal, serious consideration needs to be given to these issues by ROE when delivering their programmes in England and Wales.

For this programme to be a success in schools in England and Wales using a school-based instructor model, buy-in is needed from the whole staff team, particularly the senior leadership team. Where the programme

worked well, instructors were given time to plan the sessions, had a good relationship with their classroom teacher and had support from their SLT. In the absence of such support, instructors were forced to use their own time to plan the sessions and gain confidence with the programme materials. It would be logical to assume that this would have had an impact on the instructors' experience of the NEBT programme and, potentially, the quality of the sessions delivered.

The commitment required from schools to successfully run the NEBT programme was high, and for some schools, this was a barrier to staying on the trial. As discussed in the attrition section on pages 41–44 of the report, attrition across both the cohort 1 and cohort 2 trials was high, with several intervention schools dropping out when they found out they were to be part of the intervention. The point at which schools drop out can provide insight into the reasons for attrition, when a clear reason is not provided. Given that schools were only informed of their randomisation allocation after they had completed all the baseline evaluation data collection, evaluation burden seems unlikely to be the cause. In those schools that continued to deliver the programme, there were several instances within the IPE data collection that highlighted the difficulties schools had in setting up the programme and finding the time to prepare for and run the sessions. Instructors were required to recruit their own parents and babies, which often took significant time and resources.

We are unable to explore how these findings generalise across different cohorts and ethnicities, as these demographic characteristics were not systematically captured as part of this trial. This is a limitation of the evaluation and is a direct outcome of this programme and of it's evaluation being part of the YEF's first grant round. Since its inception, the YEF has much clearer and more systematic processes for the collection and analysis of equity and diversity data that evaluators are expected to follow. However, as part of the IPE, teachers did talk about the differential approaches or impacts for their pupils based on gender. Studies have reported differences in the development of empathy by gender (e.g. Chen et al., 2014), but due to a lack of data on pupils' genders as part of this evaluation, we were unable to explore this systematically. Again, from the IPE data, we understand that some of the NEBT materials were considered inappropriate by the parents of pupils from some minoritised groups. Such differences in content acceptability are likely to lead to differential outcomes for some groups, and future work should aim to explore this.

Unfortunately, given the level of attrition across the trial, which was much higher than anticipated, we must treat any findings with caution. Levels of attrition and some of the reasons for this are discussed below, in Limitations and lessons learned.

In summary, the NEBT programme was well received by most schools that were included in the IPE. The schools reported positive changes in their pupils and spoke highly of the programme training and content. The impact analysis supported some of the changes teachers reported observing in their pupils, particularly a reduction in BD and increases in emotional empathy. Unfortunately, as noted, these findings need to be treated with caution due to the levels of attrition experienced across the trial.

Limitations and lessons learned

As discussed earlier in the report, this evaluation is one of the YEF's launch grant round projects commissioned in 2019, before the global COVID-19 pandemic. As the pandemic developed, many of the practical steps of project setup had to be put on hold or altered. As was the case with most other school-

based interventions taking place around that time, it is clear that the delivery of NEBT was affected by the pandemic, which created methodological challenges for this evaluation. The project increased in length (going from a two-year project with a one-year evaluation to a five-year project with two one-year evaluations), which meant that there were staff changes in all the organisations. Whilst every effort was made to minimise any impact this may have had on the trial, it is important to recognise this possible impact.

It is also well documented that the effects of the pandemic have lasted beyond the initial outbreak in schools, and increased prevalence of mental health issues has been reported (Clemens, Deschamps and Fegert, 2020; Bell et al., 2023), alongside challenges with pupil absence and staffing challenges, which all lead to learning losses in schools. It is important that the findings and conclusions of this evaluation be considered in this context.

Limitations of the programme

Levels of attrition across both cohort 1 and cohort 2 significantly limit the confidence we can have in the outcomes of the trial. We are aware that the ROE programme has two modes of delivery, one in which delivery is undertaken by a school-based instructor and one in which community volunteers are recruited to deliver the programme. The levels of attrition across this trial mainly indicate that the school-based instructor model may not be a sustainable way of delivering the programme in an English and Welsh context. Furthermore, levels of support within a school for programme activities may have differed across schools and may have impacted family recruitment and attrition levels. Moreover, the use of TAs as instructors may not be in line with their expected classroom roles.[27] Had community instructors delivered the programme, the commitment of schools and the resultant burden may have been lower. A split delivery model was suggested by ROE, which would include both community and school-based instructors. However, this suggestion came after the recruitment to the first cohort of the trial had begun; as such, the original evaluation design was maintained. Furthermore, for instructors in this trial, it was their first year of delivery. It is important to note that ROE doesn't usually carry out evaluations that include instructors within their first year, and usual organisational practice requires a year for full ROE certification.

The lack of adaptation from the Canadian context and expectations of family norms were highlighted in the IPE data as issues arising in the programme. Schools had to adapt the programme content to meet the needs of their learners (removing the Canadian context and differentiating for learners with specific needs), and aspects of the programme seemed to prescribe the 'nuclear family', which is potentially biased towards families from specific backgrounds. This led to unwanted consequences during programme delivery, where pupils from families which did not fit the norms described in the programme became upset and distressed. These findings suggest that a review of the programme materials at the start of future deliveries may be beneficial to ensure they are representative of a range of family norms (including non-heteronormative) from different social and demographic backgrounds and take into account the local school context. However, it is worth noting that ROE advises against adapting materials for delivery, and it is unclear whether the adaptations had implications on the findings, e.g. when compared with other evaluations.

Limitations of the evaluation

[27] https://neu.org.uk/advice/member-groups/support-staff/hltas-and-cover-supervisors

The lack of pupil data and subgroup analysis: A significant limitation of the evaluation is the lack of baseline and endpoint characteristic data (ethnicity and gender). This meant that we have been unable to comment on how these findings generalise to different racial and ethnic groups. As this trial was part of the first grant round for the YEF, several structures that required data on these characteristics to be collected had not been put in place. Given the amount of data being collected as part of the trial for the primary and secondary outcomes, the collective decision was made not to collect other data to avoid a significant burden on schools. However, this does need to be acknowledged as a significant limitation, given the findings from the IPE highlighting issues with engagement in the NEBT programme among some minoritised groups because of the topics being covered.

Long-term outcomes: The findings from this evaluation relate to short- and medium-term outcomes only, which means the longitudinal impact of the NEBT programme remains unclear. Intervention schools were advised to complete their endpoint data within the two weeks following completion of the programme. This was to avoid inflated outcomes that can be reported if endpoint data is collected at the end of the programme (for example, in the last session).

The quantity of data collected for the evaluation of this trial was large and is likely to have contributed to the level of missing data for the trial. In addition, ROE conducts an internal evaluation (an expected and integral part of implementation), which a small number of schools mistook for the independent evaluation. The teacher SDQ (Goodman, 2001) was included as a secondary outcome, fitting the funder's requirement for the inclusion of core measures across all evaluations. However, considering the findings from this trial, we feel that when interventions are at the class level, inclusion of the teacher SDQ is too burdensome. Furthermore, given the nature of these types of trials, blinding of outcome assessors (teachers, youth workers, etc.) is extremely challenging, and the opinions they may have formed about the intervention are likely to bias their responses. As such, we would advise that the SDQ is only included in trials of class-level interventions if it is a primary outcome. Furthermore, self-report SDQs should be referenced, as they are more robust than teacher or youth worker responses when blinding isn't possible. However, when deciding on using the self-report SDQ, researchers should be mindful of the limitations with the validity of the measure that have been highlighted in more recent research (Black, Mansfield and Panayiotou, 2020).

Conversations with schools during the recruitment and data collection periods suggested that schools were confused about what the programme and the evaluation would entail. Whilst the evaluation MoU was designed to minimise the risk of this, it was clear from these communications that schools were not engaging fully with these materials and lacked a full understanding of their responsibilities. ROE was responsible for the recruitment of schools, with the evaluation team managing the randomisation process to maintain independence between the two processes and minimise imbalance or bias in treatment assignment. This model of recruitment may have contributed to the confusion from schools, as some inaccuracies were communicated to schools during recruitment about the RCT. Furthermore, from the perspective of capacity, the team at ROE reported difficulties in managing the number of schools needed for a trial. The split nature of the trial design was intended to help with this, but it also brought with it some difficulties.

Future research and publications

We intend to publish a peer-reviewed paper from the NEBT trial, but at this point, this is just in the planning stage until this report is published.

Future research questions:

Several future research questions have emerged from the evaluation, in particular, working on the limitations and lessons learnt sections of this report. We would suggest that, should future large-scale evaluations of the NEBT programme be funded in schools in England, issues around attrition should be carefully considered.

- Do pupils from minority groups experience the NEBT programme differently from those in non-minority groups?
- Is there a differential impact of the NEBT programme for minority groups?
- How does instructor type impact programme sustainability?
- Learning from the unintended consequences of this trial, how can the NEBT programme be developed further to ensure the programme considers diverse pupils with traumatic life experiences?
- Is there an ethical issue with TAs doing the majority of NEBT delivery?

References

Austin, P. (2008). Assessing balance in measured baseline covariates when using many-to-one matching on the propensity score. *Pharmacoepidemiology and Drug Safety* 17: 1270-1225.

Bell, I. H., Nicholas, J., Broomhall, A., Bailey, E., Bendall, S., Boland, A., Robinson, J., Adams, S., McGorry, P., and Thompson, A. (2023). The impact of COVID-19 on youth mental health: A mixed methods survey. *Psychiatry research* 321: 115082. https://doi.org/10.1016/j.psychres.2023.115082

Black, L., Mansfield, R., and Panayiotou, M. (2020). Age appropriateness of the self-report Strengths and Difficulties Questionnaire. *Assessment* 28(6): 1556-1569. https://doi.org/10.1177/1073191120903382

Bloom, H. S., Richburg-Hayes, L., and Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis* 29(1): 30-59. https://doi.org/10.3102/0162373707299550

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2): 77-101, https://doi.org/10.1191/1478088706qp063oa

Browne G., Gafni A., Roberts J., Byrne C., and Majumdar B. (2004). Effective/efficient mental health programs for school-age children: A synthesis of reviews. *Social Science & Medicine* 58(7): 1367-1384. https://doi.org/10.1016/S0277-9536(03)00332-0

Clemens, V., Deschamps, P., Fegert, J.M. *et al.* (2020). Potential effects of "social" distancing measures and school lockdown on child and adolescent mental health. *European Child and Adolescent Psychiatry* 29, 739-742. https://doi.org/10.1007/s00787-020-01549-w

Connolly, P., Miller, S., Kee, F., Sloan, S., Gildea, A., McIntosh, E., Boyer, N., and Bland, M. (2018). A cluster randomised controlled trial and evaluation and cost-effectiveness analysis of the Roots of Empathy schools-based programme for improving social and emotional well-being outcomes among 8- to 9-year-olds in Northern Ireland. NIHR Journals Library.

Deighton, J., Tymms, P., Vostanis, P., Belsky, J., Fonagy, P., Brown, A., Martin, A., Patalay, P., and Wolpert, W. (2013). The development of a school-based measure of child mental health. *Journal of Psychoeducational Assessment* 31: 247. https://doi.org/10.1177/0734282912465570

Deighton, J., Humphrey, N., Belsky, J., Boehnke, J., Vostanis, P., and Patalay, P. (2018). Longitudinal pathways between mental health difficulties and academic performance during middle childhood and early adolescence. *British Journal of Developmental Psychology* 36(1): 110-126. https://doi.org/10.1111/bjdp.12218

Dong, N., & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. [Software]. http://www.causalevaluation.org/.

Durlak J. A., Weissberg R. P., Dymnicki A. B., Taylor R. D., and Schellinger K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Childhood Development* 82: 405-432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

Forbes, D. (2013). Blinding: An essential component in decreasing risk of bias in experimental designs. *Evidence-Based Nursing* 16: 70-71.

Gaffney, H., Farrington, D. P., and White, H. (2021). Anti-bullying programmes. *Anti-bullying Programmes*. Toolkit Technical Report. Youth Endowment Fund.

Gaffney, H., Farrington, D. P., and White, H. (2021). Social Skills Training. Toolkit technical report. Youth Endowment Fund.

Gerber, A. S., and Green, D. P. (2012). Field experiments: Design, analysis, and interpretation. W.W. Norton & Company.

Gibbons, S., McNally, S., and Montebruno, P. (2025). Pupil Absence and the COVID-19 Pandemic. Centre for Economic Performance.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry* 40(11): 1337-1345. https://doi.org/10.1097/00004583-200111000-00015

Hamoda, H. M., Chiumento, A., Alonge, O., Hamdani, S. U., Saeed, K., Wissow, L., and Rahman, A. (2021). Addressing the consequences of the COVID-19 lockdown for children's mental health: Investing in school mental health programs. *Psychiatric services* 72(6): 729-731. https://doi.org/10.1176/appi.ps.202000597

Haylock, S., Boshari, T., Alexander, E.C., Kumar, A., Manikam, L., and Pinder, R. (2020). Risk factors associated with knife-crime in United Kingdom among young people aged 10–24 years: A systematic review. *BMC Public Health* 20: 1451. https://doi.org/10.1186/s12889-020-09498-4

Hedges, L. V., and Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation Review*, 37(6): 445-489. https://doi.org/10.1177/0193841x14529126

Hewitt, C., and Togerson, D. (2006). Is restricted randomisation necessary? *British Medical Journal Research Methods* 332: 1506-1508. https://doi.org/10.1136/bmj.332.7556.1506

Hrõbjartsson, A., Emanuelsson, F., Skou Thomsen, A.S., Hilden, J., and Brorson, S. (2014). Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. *International Journal of Epidemiology* 43(4): 1272-83. https://doi.org/10.1093/ije/dyu115

Ho, D., Imai, K., King, G., and Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15: 199-236. https://doi.org/10.1093/pan/mpl013

Humphrey, N., and Panayiotou, M. (2022). Bounce Back: Randomised trial of a brief, school-based group intervention for children with emergent mental health difficulties. *European Child & Adolescent Psychiatry* 31(1): 205-210. https://doi.org/10.1007/s00787-020-01612-6

Jo, B., Asparouhov, T., Muthen, B., and Ialongo, N. (2008). Cluster randomized trials with treatment noncompliance. *Psychological Methods* 13(1): 1-18. https://doi.org/10.1037/1082-989X.13.1.1

Jolliffe, D., and Farrington, D. P. (2006). Development and validation of the Basic Empathy Scale. *Journal of Adolescence* 29(4): 589-611. https://doi.org/10.1016/j.adolescence.2005.08.010

Lätsch, D., Stauffer, M., and Bollinger, M. (2017). Evaluation of the Roots of Empathy program in Switzerland, years 2015 to 2017. Full Report. Bern University of Applied Science.

Lee, J. (2020). Mental health effects of school closures during COVID-19. *The Lancet Child & Adolescent Health* 4(6): 421. https://doi.org/10.1016/s2352-4642(20)30109-7

Office for Health Improvement and Disparities. (2022). Covid-19 mental health and wellbeing surveillance: Report. GOV.UK. Retrieved May 6, 2022, from https://www.gov.uk/government/publications/covid-19-mental-health-and-wellbeing-surveillance-report

Panayiotou, M., Humphrey, N., and Wigelsworth, M. (2019). An empirical basis for linking social and emotional learning to academic performance. *Contemporary Educational Psychology* 56: 193-204. https://doi.org/10.1016/j.cedpsych.2019.01.009

Patalay, P., Deighton, J., Fonagy, P., Vostanis, P., and Wolpert, M. (2014). Clinical validity of the Me and My School questionnaire: A self-report mental health measure for children and adolescents. *Child and Adolescent Psychiatry and Mental Health* 8(17). https://doi.org/10.1186/1753-2000-8-17.

Santos, R. G., Chartier, M. J., Whalen, J. C., Chateau, D., and Boyd, L. (2011). Effectiveness of school-based violence prevention for children and youth: Cluster randomized field trial of the Roots of Empathy program with replication and three-year follow-up. *Healthcare Quarterly* 14: 80-91. http://www.longwoods.com/articles/images/HQ_vol14_ChildHealth_Issue2_Santos.pdf

Schochet, P.Z., and Chiang, H.S. (2011). Estimation and identification of the complier average causal effect parameter in education RCTs. *Journal of Educational and Behavioral Statistics* 36(3): 307-345. https://psycnet.apa.org/doi/10.3102/1076998610375837

Schonert-Reichl, K.A., Smith, V., Zaidman-Zait, A., and Hertzman, C. (2012). Promoting children's prosocial behaviors in school: Impact of the "Roots of Empathy" program on the social and emotional competence of school-aged children. *School Mental Health* 4: 1-21. https://doi.org/10.1007/s12310-011-9064-7

Sussman J. B., and Hayward, R. A. (2010). An IV for the RCT: Using instrumental variables to adjust for treatment contamination in randomised controlled trials. *BMJ* 340:c2073. https://doi.org/10.1136/bmj.c2073

Tilbrook H. E., Hewitt C. E., Aplin J. D., Semlyen A., Trewhela A., Watt I., and Torgerson D. J. (2014). Compliance effects in a randomised controlled trial of yoga for chronic low back pain: a methodological study. *Physiotherapy* 100(3): 256-62. https://doi.org/10.1016/j.physio.2013.10.001

Tomé, G., Almeida, A., Ramiro, L., and Gaspar, T. (2021). Intervention in schools promoting mental health and well-being: A systematic review. *Global Journal of Community Psychology Practice* 12(1). https://doi.org/10.17161/gjcpp.v12i1.20659

van Noorden T. H., Haselager G. J., Cillessen A. H., and Bukowski W. M. (2015). Empathy and involvement in bullying in children and adolescents: A systematic review. *Journal of Youth & Adolescence* 44(3): 637-657. https://doi.org/10.1007/s10964-014-0135-6

Wood L., Egger M., Gluud L. L., et al. (2008). Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. *BMJ* 336: 601-605.

Wrigley, J., Makara, K., and Elliot, D. (2015). Evaluation of Roots of Empathy in Scotland 2014-2015. Final Report for Action for Children (unpublished).

Appendix A: Effect size estimation

As specified in the SAP, effect sizes were calculated using Hedges' g, as specified in the following equation, where T is the treatment mean, C is the control mean, δ^2_{sch} is the school level variance and δ^2_{pup} is the residual (within-school, between-pupil) variance for the null/empty model:

$$ES = \frac{(T - C)_{adjusted}}{\sqrt{\delta^2_{sch} + \delta^2_{pup}}}$$

The headline effect size will be calculated from the group allocation (intervention/control) coefficient in the full analysis model specified in equation P1.

i.e. $\boldsymbol{\beta_1 = (T - C)_{adjusted}}$

The (square root of) the unconditional total variance (from the empty/null multilevel regression model) is used as the denominator.

Total (unconditional) variance = $\boldsymbol{\delta^2_{sch} + \delta^2_{pup}}$

The numerator and denominator used to calculate the effect size (and 95% confidence intervals) for the M&MF Behavioural Difficulties primary outcome are shown in the Tables below. These estimates are then used to illustrate how the effect size was calculated.

Appendix table 1: Effect size estimation

	Model Coefficients / Estimates	
	Coefficient: β_1 (Standard error)	95% Cis
M&MF Behavioral difficulties	-0.13 (0.196)	(-0.52; +0.25)

Appendix table 2: Effect size estimation

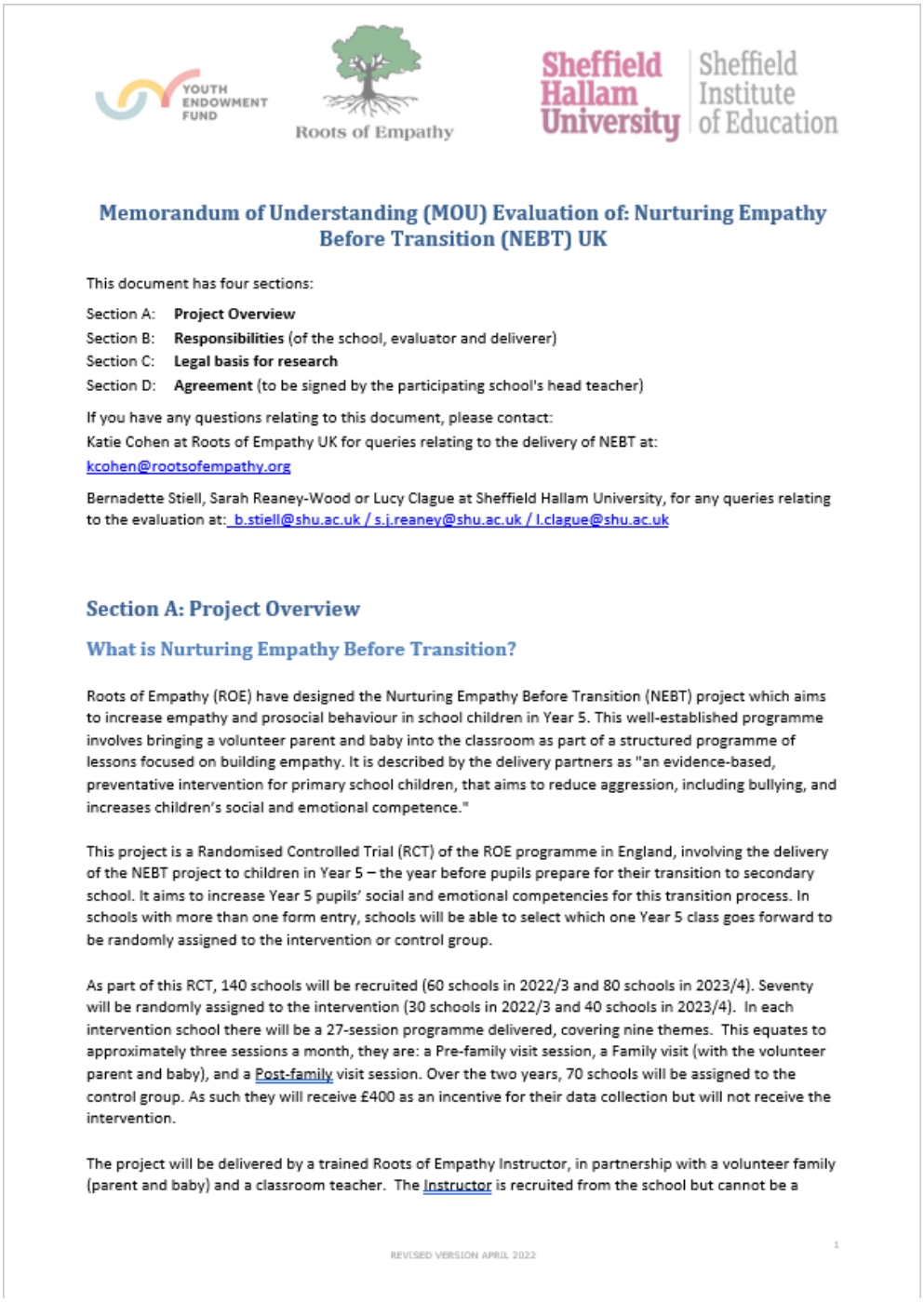
Outcome	Null (empty) model Variance Decomposition			Denominator
	School δ^2_{sch}	Residual (pupil) δ^2_{pup}	Total $\delta^2_{sch} + \delta^2_{pup}$	$\sqrt{\delta^2_{sch} + \delta^2_{pup}}$
M&MF Behavioral difficulties	0.26	5.54	5.80	2.408

The ITT effect size (-0.06 sds) and confidence intervals (-0.22 to +0.10 sds) were estimates by dividing the coefficient by the variance denominator.

i.e.	Effect size	= -0.13 / 2.408 = -0.0554 ~ -0.06 sds
95% Cis:	Lower	= -0.52 / 2.408 = -0.2150 ~ -0.22 sds
	Upper:	= +0.25 / 2.408 = 0.1043 ~ +0.10 sds

Appendix B: Recruitment documents

Please include the recruitment documents sent to settings and participants/ parents (MoU, information sheets, privacy notices, withdrawal forms etc., as applicable).



YOUTH ENDOWMENT FUND

Roots of Empathy

Sheffield Hallam University | Sheffield Institute of Education

Memorandum of Understanding (MOU) Evaluation of: Nurturing Empathy Before Transition (NEBT) UK

This document has four sections:

Section A: **Project Overview**
Section B: **Responsibilities** (of the school, evaluator and deliverer)
Section C: **Legal basis for research**
Section D: **Agreement** (to be signed by the participating school's head teacher)

If you have any questions relating to this document, please contact:
Katie Cohen at Roots of Empathy UK for queries relating to the delivery of NEBT at:
kcohen@rootsofempathy.org

Bernadette Stiell, Sarah Reaney-Wood or Lucy Clague at Sheffield Hallam University, for any queries relating to the evaluation at: b.stiell@shu.ac.uk / s.j.reaney@shu.ac.uk / l.clague@shu.ac.uk

Section A: Project Overview

What is Nurturing Empathy Before Transition?

Roots of Empathy (ROE) have designed the Nurturing Empathy Before Transition (NEBT) project which aims to increase empathy and prosocial behaviour in school children in Year 5. This well-established programme involves bringing a volunteer parent and baby into the classroom as part of a structured programme of lessons focused on building empathy. It is described by the delivery partners as "an evidence-based, preventative intervention for primary school children, that aims to reduce aggression, including bullying, and increases children's social and emotional competence."

This project is a Randomised Controlled Trial (RCT) of the ROE programme in England, involving the delivery of the NEBT project to children in Year 5 – the year before pupils prepare for their transition to secondary school. It aims to increase Year 5 pupils' social and emotional competencies for this transition process. In schools with more than one form entry, schools will be able to select which one Year 5 class goes forward to be randomly assigned to the intervention or control group.

As part of this RCT, 140 schools will be recruited (60 schools in 2022/3 and 80 schools in 2023/4). Seventy will be randomly assigned to the intervention (30 schools in 2022/3 and 40 schools in 2023/4). In each intervention school there will be a 27-session programme delivered, covering nine themes. This equates to approximately three sessions a month, they are: a Pre-family visit session, a Family visit (with the volunteer parent and baby), and a Post-family visit session. Over the two years, 70 schools will be assigned to the control group. As such they will receive £400 as an incentive for their data collection but will not receive the intervention.

The project will be delivered by a trained Roots of Empathy Instructor, in partnership with a volunteer family (parent and baby) and a classroom teacher. The Instructor is recruited from the school but cannot be a

REVISED VERSION APRIL 2022

1

classroom teacher. In the past Teaching Assistants and or Parent Liaison workers have taken on the role of Instructor. The Instructor receives training, is assigned an ROE mentor and given ongoing professional development. For more information visit the Roots of Empathy website.

Evaluation Outline

The Youth Endowment Fund (YEF) has funded Roots of Empathy (ROE) to deliver the Nurturing Empathy Before Transition programme in schools for two years starting in Autumn 2022. The YEF has commissioned a team at Sheffield Hallam University's Institute of Education – SIOE (the evaluation team), led by Bernadette Stiell and Sarah Reaney-Wood (Co-principal investigators), and Lucy Clague (Project manager) to independently evaluate the NEBT project.

The intervention and evaluation will in brief, be as follows:
Year 1 – Cohort A: September 2022-July 2023. First year of the Randomized Control Trial (RCT). Sixty schools are recruited to the RCT. Schools are randomized with 30 schools becoming intervention schools and 30 schools assigned as control schools. All schools complete online baseline and follow up measures administered by SHU. SHU will visit six case study schools for qualitative fieldwork.
Year 2 – Cohort B: September 2023-July 2024. Second year of RCT. A further 80 schools are recruited to the RCT. All activities are repeated as year one above.

The evaluation conducted by SHU will consist of:

- **A Randomized Control Trial (RCT)** with all 140 schools involved. An RCT is a way of comparing outcomes of pupils who have received an intervention with those who have not, through completion of a pre-intervention (baseline) and post-intervention (follow up) measure (set of questions). For this intervention, a combination of online measures will be used, each consisting of a small number of questions relating to pupils' perceptions of their own behaviour and characteristics. The measures used will be:
    - Strengths and Difficulties Questionnaire (SDQ) – completed by the class teacher for each pupil (25 items)
    - Me and My Feelings measure – completed by pupils (16 items)
      Basic Empathy Scale (BES) – completed by pupils (20 items)
    - Some number of additional SHU questions – completed by pupils and their teacher

Me and My Feelings and BES measures will be completed by Y5 pupils at two time points. In addition, the class teacher will be asked to complete a teacher report version of the SDQ for each child in their Y5 class at two time points, with additional questions on any other socio-emotional interventions or activities taking place in the school.

Additional data will be collected alongside the outcome measures, such as the pupils name, gender, date of birth (DOB) and unique pupil number (UPN). This data would be used for the trial to match pupil data and their responses.

For the purpose of research, at the end of the study pupil data supplied to the SHU evaluation team by schools/ROE will be shared with the Department of Education (DfE). The DfE will pseudonymise the data, so it is no longer possible to identify any individual young person from the study data. The DfE will then transfer the data to the YEF Data Archive, which is stored in the Office for National Statistics (ONS) Secure Research Service. The YEF is the 'controller' of the information in the Archive. It will be possible for information in the Archive to be linked with information about the pupils from the **National Pupil Database**

(NPD) and the Police National Computer (PNC). This will help approved researchers find out the long-term impact of the projects funded by YEF. More information can be provided on this and how pupil data security will be ensured.

It is YEF's intention to retain data in the YEF archive indefinitely. This is consistent with the Information Commissioner's Office (ICO)'s view that data can be held indefinitely for research and archiving purposes. However, YEF will commit to reviewing every 5 years, following submission to the YEF archive, to see whether there is a continued benefit to storing the data and its potential use in future research. For transparency, the precise terms of data sharing will be stated in a fair processing notice that will be provided to all schools, specifying the personal data (pupil names, pupil ID numbers, FSM status, KS2 attainment) to be processed, in line with General Data Protection Regulation (GDPR) guidelines that came into force from 25 May 2018.

- **Qualitative data collection** will take place in 12 intervention schools (6 in year one and 6 in year two) where staff are willing to take part in this aspect. This would consist of a researcher visiting the school on a day that the programme was being delivered, to observe a session and talk to pupils and staff involved. The plan for the half day school visit would include:
    - Observing an ~~instructor~~ training session
    - Observing the ROE session (where feasible)
    - Undertaking a focus group with a sample of pupils after the session
    - Interviewing the ROE Instructor
    - Interviewing the teacher present during the ROE lesson
    - Interviewing a member of SLT (where possible)

The case study visits are undertaken to understand how the NEBT project is delivered in schools and any early impacts perceived by those involved. SHU will also collect:

- Monitoring information (MI) data from intervention schools to monitor recruitment and retention on the programme and will analyse data collected by ROE as part of the project.
- A register of the pupils attending each NEBT session, collated by the YS class teacher, to measure pupil 'dosage'. This will enable us to compare pupil attendance with their measure scores/outcomes to evaluate the impact of the intervention.
- Instructors will be asked to complete a simple checklist of session activities to evaluate the extent or 'fidelity' of the programme delivery in each session.

The evaluation is not an assessment of individual pupils, staff, or schools, but is about understanding how the ROE programme works overall.

Timeline of evaluation activities Year one:

Dates	Activity
By the end of June 2022	60 schools recruited to year one (Cohort A). Schools Sign and return MOU (this document) to ROE team (also signed by SHU and ROE)
Sept 2022	Y5 pupils and class teachers in all 60 schools complete baseline measures over a 2-week period. Schools are then randomised – randomly allocated to either intervention or control
Oct 2022	Nurturing Empathy delivery begins in 30 cohort A intervention schools 30 control schools continue with 'business as usual'
Oct 2022- July 2023	SHU team visit six intervention schools (case studies) to observe NEBT programme and interview teachers and pupils
June-July 2023	ROE programme finishes in Cohort A intervention schools Follow up (teacher and pupil) measures completed by all schools.
July 2023	ROE to share MI data with SHU

Evaluation FAQs

Is my school eligible?

This project is currently open to state-funded, primary schools in disadvantaged areas of England including the following locations: Yorkshire, London (Greater), East and West Midlands and Merseyside. To take part, schools will need to read through this Memorandum of Understanding (MoU) and agree to the trial requirements outlined within it.

Does my school have to take part in the project?

No, you can choose whether your school takes part.

Does my school have to take part in a case study visit?

No, if you participate in the project and are allocated to the intervention group, you may be contacted by the SHU evaluation team and invited to take part in a case study visit, but you are under no obligation to agree to this.

How does my school benefit?

If you participate, your school will be randomly allocated to an intervention or control group. Those in the intervention group will receive the ROE programme free of charge for an academic year. At the end of the year the certified ROE Instructor can deliver further ROE programmes in school. Control schools will be offered an incentive of £400, that will be paid in two parts (£200 upon receipt of baseline data and £200 upon receipt of endpoint data). Participating schools will also have the opportunity to be directly involved with a high-quality evaluation that can inform future provision of the ROE programme. In addition, schools taking part in a case study visit will allow staff and pupils an opportunity to talk through and reflect on the ROE programme in their school.


Roots of Empathy


What does sign up mean for my school?

By signing this MOU, you are agreeing for your school to take part in the trial, once the required 60 schools have been recruited by the end of June 2022. You are committing to undertaking the baseline and follow up measures with your selected Y5 class of pupils. You are agreeing for your school to be randomly allocated to either the intervention or control group. **Therefore, your school may or may not receive the ROE programme depending on this allocation.** You are also agreeing for your contact details and some key monitoring information (MI) data to be passed on to the evaluation team at SHU. The team at SHU will contact you to arrange for the measures to be completed. The trial will not proceed if insufficient schools are recruited to the project.

How will school and pupil data be protected?

For SHU to contact participating schools to arrange the above data collection, SHU will need for ROE to **securely** share with the evaluation team: the names of participating schools, Y5 teacher and instructor contact names, email addresses and telephone numbers, and a scan of the signed MoU. **This data will be stored securely on password protected computers which are only accessible to the evaluation team. Details of data to be shared as part of the RCT are outlined below in the responsibilities section.**

The researchers will also have access to the data ROE, school staff and pupils are asked to complete and submit at different points in the programme. This data will be used as part of the evaluation. Sheffield Hallam Universities privacy policy can be found here.

Section B - outlines full responsibilities for all parties connected to evaluation

Responsibilities of ROE and SHU

Allocation to intervention and control groups

By the end of September 2022, 60 schools in Cohort A will be allocated to the intervention or control group. This will be done by SHU through a process of randomisation. We recognise that schools allocated to the control group are likely to be disappointed not to be taking part in the ROE programme. However, we hope that the incentive of £400 will recognise the essential role played by control schools.

Conducting the trial

The team at SHU will contact all schools in the trial with details of how to complete the measures outlined above with pupils and teachers, including a link to the combined measures and instructions on how to facilitate pupil completion. The responses will be collected and stored securely on password protected computers by the research team at SHU. The team at SHU will be available to answer questions from school staff or parents about the evaluation throughout the trial process.

Data protection

SHU will not use pupil names or school names in any report arising from the research.

SHU and ROE will strictly comply with current data protection legislation, including the GDPR. Under GDPR Article 6, Paragraph 1e, the legal basis for this project is it being a 'public task', as the research is being conducted to evaluate the impact of an approach to building social emotional skills and empathy that has potential benefits for pupils participating in the trial and beyond. Therefore, parental consent will not be sought for participation, but parents/carers are free to withdraw their child from data collection and analysis at any time until the data is archived at the end of the project. Information on how to withdraw will be

Responsibilities of Sheffield Hallam University across the evaluation

Consent and ethics

- SHU will strictly comply with current legislation in relation to data processing and storage. Following the signing of this MOU, a data sharing agreement will be sent to all schools, outlining all aspects of data collection and sharing process.
- Under General Data Protection Regulation (GDPR) Article 6, Paragraph 1e, the legal basis for this project is it being a 'public task'. Given ethical considerations, separate permission will also be sought from school staff, pupils, parents/carers for additional data collection as detailed in this document.
- For any qualitative data, verbal consent would also be taken before proceeding with any interview or focus group.
- SHU will provide information sheets that make pupils, staff, and the ROE parent being observed aware of the expectations underpinning their involvement. It will make clear that participating pupils/school staff are free to withdraw from the evaluation. For case study visits where the ROE parent may be observed during a lesson, the information sheet will make it clear that they are not part of the evaluation and that the school and ROE are responsible for their wellbeing.
- SHU will ensure that this evaluation has been assessed and approved by an independent ethics committee at Sheffield Hallam University.

Data

- At all points, SHU are responsible for retrieving and processing data, including MI information, as part of this evaluation using secure transfer methods such as SHU Zend To which SHU will send links to.
- No pupil/staff names or school names will be used in any report arising from the research.
- Pupil names, D.O.B and UPN will be used for matching of data only. Once data is matched ready for analysis, these identifiers will be removed.
- SHU are responsible for the qualitative and quantitative analysis of all the data collected.

Communications

- SHU will be the point of liaison for schools on anything related to the evaluation throughout the course of the evaluation.
- SHU will liaise with ROE and YEF throughout the course of the evaluation.

Fieldwork

- All researchers visiting schools will hold a current enhanced DBS (formerly CRB) certificate.
- SHU researchers will undertake case-study visits at times convenient for the school.

Responsibilities of Roots of Empathy

Communication

- Acting as a point of contact for sending signed MoUs between schools and SHU
- Dealing with any queries from school staff related to the delivery of the ROE programme.
- Sending scanned copies of signed MoUs and consent to SHU in PDF format for records

Data

- Securely sharing school, instructor and Y5 teacher contact details and ROE monitoring data (as previously outlined securely) with SHU.

Delivery
Delivering the ROE nine-theme programme to intervention schools and supporting schools as necessary.

Responsibilities of all schools

- Allocate a school-based lead (likely to be the Instructor or Y5 class teacher) to the trial that can be the main contact for all queries from SHU.
- Complete and sign this Memorandum of Understanding (MOU) if in agreement with the terms outlined.
- Complete a spreadsheet template at the start of the trial including the names, D.O.B and UPN of each pupil in the Y5 class taking part in the trial. This is to allow SHU to be able to match individual pupils baseline data to their endpoint data.
- Enable the Y5 class teacher to complete the online pre (Sept 2022) and post (July 2023) intervention measure on behalf of all pupils in their class, with additional questions on any other socio-emotional activities or interventions taking place in the school.
- Facilitate all participating Y5 pupils to complete the online pre (Sept 2022) and post (July 2023) intervention measure.
- Keep SHU up to date with any pupils that have left the school or withdrawn from the trial, so that they can ensure accurate records at all time points.
- Ensure the wellbeing of the parent and baby during all NEBT sessions, in keeping with the philosophy of the programme and outlined in the ROE MOU.

Additional responsibilities of intervention schools

This is in addition to ROEs standard MOU and agreement with intervention school expectations in relation to the running of the NEBT programme.

- Ensure that the ROE sessions are timetabled and delivered as planned and let the ROE team know if this is not possible.
- Take an attendance register of pupils attending each ROE session to be sent securely with ZendTo to the team at SHU.
- Allow ROE to share their school, Instructor and Y5 class teacher contact data and the ROE Instructor checklist of session activities.
- Agree to be *contacted* by SHU to be invited to take part in a case study visit. Engagement in this activity is voluntary and separate permission will be sought from teachers and parents/carers for additional research activities as detailed in this document. Data generated from research activities will be handled in accordance with the GDPR requirements and the fair processing notice.
- By the end of September 2022, issue information to parents/carers of Year 5 pupils about the project and provide details to ROE of any pupils whose parents/carers do not wish for them to be part of the data collection and analysis. As research is being conducted on the legal basis of a public task, schools are not obliged to seek parental consent unless they prefer to do so for their own purposes. But parents/carers should know that they can ask for their pupils' data to be excluded at any stage of the trial. SHU will provide suggested relevant phrasing for the parent/carer communication to be sent out in September 2022. Given ethical considerations, separate permission will be sought from parents/carers for additional research activities as detailed in this document.
- To ensure the parent and baby involved in any lessons observed as part of the evaluation are provided with an information sheet (provided by SHU), explaining that parents and their babies are not part of

provided for schools, parents and carers. If a parent/carer decides to withdraw, this would mean that no data on their child would be included in the evaluation and the child would not be required to take the measures (surveys) but can still participate in the NEBT sessions. If you would like further information, please contact the delivery team at ROE using the contact details at the end of this document.

SHU and ROE will act jointly as data controllers during the evaluation period and the YEF will be the data controller once the data is archived at the end of the project. A data sharing agreement will detail the personal data to be shared, and a fair processing notice will be sent to all participating schools as per GDPR requirements.

Type of personal data	Variables/types	Data sharing flow
Pupil characteristics	Pupil names, DOB, UPN	SHU/DfE*/PNC/YEF
MI data	ROE attendance data, school names	ROE/SHU/ONS/YEF
Qualitative data (observation/focus groups/ interviews)	Audio recordings, anonymised transcripts	SHU
Measure (survey) data	Pupil/teacher attitude data	SHU/ONS/YEF

*DfE will pseudonymise the data - so no personally identifying data will be transferred to PNC (Police National Computer) or YEF

SHU has undergone a full review and approval processes through the university ethics committees (Ref: ER19810112. All researchers visiting schools will hold a current enhanced DBS (formerly CRB) certificate and have completed additional NSPCC safeguarding training.

If a sub-sample of pupils or teachers are involved in additional qualitative data collection activities (observations, focus groups and interviews) these will also be covered by GDPR legal basis as a public task. In keeping with good research ethical practices, participant consent will be sought for participation in the additional qualitative research activities.

the evaluation and that responsibility for their wellbeing and that of the baby lie with the school and ROE and not the research team.

Additional responsibilities of control schools

- Continue with business as usual for the academic year of the trial (and completing the measures as outlined above).

Section C: Legal Basis for Research

SHU undertakes research as part of its function for the community under its legal status. Data protection allows us to use personal data for research with appropriate safeguards in place under the legal basis of public tasks that are in the public interest. A full statement of your rights can be found by clicking our privacy policy for research participants.

All University research is reviewed to ensure that participants are treated appropriately, and their rights respected. This study was approved by the University Research Ethics Committee. Further information can be found on our research ethics and integrity page.

Below we outline key contacts should you or any participants have any concerns (this information will also be included in parental/participant information sheets).

You should contact the Data Protection Officer if:

- you have a query about how your data is used by the University
- you would like to report a data security breach (e.g. if you think your personal data has been lost or disclosed inappropriately)
- you would like to complain about how the University has used your personal data DPO@shu.ac.uk

You should contact the Head of Research Ethics (Dr Mayur Ranchordas) if:

- you have concerns with how the research was undertaken or how you were treated ethicssupport@shu.ac.uk

Postal address: Sheffield Hallam University, Howard Street, Sheffield S1 1WBT
Telephone: 0114 225 5555

Section D: Agreement

If the above terms are acceptable, please sign (electronic signature or scanned signature) and date both copies, keeping one copy for your records and returning the other to the Roots of Empathy team before XXX/the end of June 2022. If unsure of any aspects of the delivery or the evaluation, please don't hesitate to contact ROE or SHU (details at top of form)

Name of School _____

Headteacher	Full name: Work email: Signature:	Date
School based lead (SBL)	Full name: Work email: Signature:	Date
Sheffield Hallam University (Principal Investigator)	Full name: Bernadette Stiell Signature:	Date
Roots of Empathy UK Manager	Name: Katie Cohen Signature:	Date

YOUTH ENDOWMENT FUND

Roots of Empathy
Racines de l'empathie

Sheffield Hallam University | Sheffield Institute of Education

Invite Roots of Empathy to your school!

Want to help reduce bullying, aggression and violence in your school?

Want to surround your students with a positive and lasting experience that will help them to bring empathy to their future lives as citizens, parents and leaders of tomorrow?



We are seeking schools to participate in an exciting opportunity!

Roots of Empathy is an evidence-based programme that has shown, in two decades of independent research on three continents, to have a significant effect in reducing levels of aggression, including bullying, among primary school children while raising social and emotional competence and increasing empathy.

With support from the Youth Endowment Fund and evaluation from Sheffield Hallam University, Roots of Empathy is delivering The Nurturing Empathy Before Transition (NEBT) project, in order to find out whether the Roots of Empathy programme is effective in raising levels of pro-social behavior (helping, caring, sharing, including and kindness) and empathy and reducing aggression, including bullying and engagement in risky behaviour, and supporting children as they prepare to make the important and often scary transition to secondary school.

Your participation in the Nurturing Empathy Before Transition project:

- We are recruiting schools from **areas of social disadvantage** to participate. Once recruited schools will be randomly assigned to an intervention or control group.
- **The control group will:**
  - Collect teacher and student questionnaires (at the beginning and end of the school year)
  - Receive a £400 incentive payment on completion of questionnaires
- **The intervention group will:**
  - Receive the Roots of Empathy programme in their school in October 2022 for their **Year 5** students
  - Receive training and mentoring for a school based Roots of Empathy Instructor, all programme equipment and gold standard curriculum which can be used in subsequent years
  - Collect teacher and student questionnaires (at the beginning and end of the school year)

To find out more about what's involved, please go to our website:
www.rootsofempathy.org and **click on Participate**

For more information please contact:
Katie Cohen
kcohen@rootsofempathy.org
Tel: 0771-408-3042
Expression of Interest required from schools ASAP

YOUTH ENDOWMENT FUND

Roots of Empathy
Racines de l'empathie

Sheffield Hallam University | Sheffield Institute of Education

Nurturing Empathy Before Transition Project Sequencing

Key Milestones	Timescales for 2022/23
School Recruitment Interested schools make a commitment to participation by Signing MOU	April - June 2022
Schools accepted on to the Trial Schools will be notified they have been accepted on to the programme	April –June 2022
Roots of Empathy – will work with schools to nominate their Instructor and find a volunteer family	May – July 2022
Baseline data collection for Year 5 class Intervention and Control Groups	September 2022 - Two weeks into term
Randomisation and assignment of control or Intervention status	End of September 2022
Intervention schools: Confirm Instructor and Volunteer Family **Control schools:** Business as usual	Early October 2022
3 Day Instructor Training	End-October 2022
Programme Implementation in Intervention schools: Instructor works with classroom teacher and volunteer family to deliver Roots of Empathy to Year 5 students	Beginning of November 2022 – June 2023
Post intervention data collection: From Intervention and control schools	June 2023

"Roots of Empathy is about changing the world, child by child."

Mary Gordon, Founder/President

Roots of Empathy UK
c/o Oxford House
Derbyshire Street, Bethnal Green, London E2 6HG
internationallirootsofempathy.org | rootsofempathy.org | @rootsofempathy
Charity No.1170754

Roots of Empathy
Racines de l'empathie

YOUTH ENDOWMENT FUND

Sheffield Hallam University | Sheffield Institute of Education

Roots Of Empathy (ROE) is pleased to be working in partnership with the Youth Endowment Fund (YEF) and Sheffield Hallam University (SHU), to conduct the first independent evaluation of ROE's Award Winning programme in England. ROE has been delivering transformational programmes in primary schools in England since 2012.

ROE is an international charitable organisation that offers empathy-based programming for children. Created 25 years ago by social entrepreneur, educator, auther, parenting expert and child advocate Mary Gordon – ROE is considered a model of social innovation. Over two decades of independent academic research across three continents has consistently shown that the programme dramatically reduces aggression (including bullying) and increases pro-social behaviours (including caring, sharing, helping and inclusion) among children who receive it.

In the ROE programme, a parent and baby (who is two to four months old at the start of the programme) from the community visit a classroom nine times over the course of a school year. A trained ROE Instructor visits with the family to guide children as they observe the relationship between the baby and its parent(s). The Instructor also visits before and after each family visit to reinforce teachings. There are 27 classroom visits in total in a ROE programme.

The YEF was established in March 2019 with a £200m endowment and ten-year mandate from the Home Office, to research, support and evidence the best ways to protect children from becoming involved in violence in the long term. SHU is one of the leading new universities in the UK for research, with an international reputation.

With support from the YEF and evaluation from SHU, ROE is delivering The Nurturing Empathy Before Transition (NEBT) programme, in order to find out whether the programme is effective in raising levels of empathy and pro-social behavior, reducing aggression and engagement in risky behavior, and supporting children as they prepare to make the important and often scary transition to secondary school.

The Randomised Control Trial* will run in 140 primary school classrooms in England over a two-year period, starting in 2022. Half of these school classrooms (70) will be randomly selected to receive the ROE programme, as part of their PSHE (Personal, Social and Health Education) lessons, and the remaining half will form a 'control group' who will continue with their normal lessons, without the programme.

In a world where our differences increasingly tend to alienate us from one another, there is a need to build empathy so we can find our shared humanity. We all ache to belong. When children are empathic, they include others because they understand what it feels like to be left out. We need to nurture empathy in our children so that they will build a caring, peaceful and civil society that is inclusive of all.

Roots of Empathy UK
c/o Rich Mix Cultural Foundation, 35-47 Bethnal Green Road, London, England E1 6LA
international@rootsofempathy.org | rootsofempathy.org | @rootsofempathy

Charity No.1170754

*In order to understand the impact of the Roots of Empathy programme on children, the YEF has funded a randomised controlled trial (RCT) to evaluate the Roots of Empathy programme.

An RCT is a way of comparing the outcomes of pupils who have received the programme with those who have not, to see if the programme makes a difference. Schools will be randomly allocated to either an intervention group (group that receives the Roots of Empathy programme) or a control group (group that does not receive the Roots of Empathy programme but will receive an incentive for their participation). Randomization occurs in an effort to reduce potential bias in the study.

Appendix C: SHU materials

Nurturing Empathy before Transition project
Sheffield Hallam University
Fair Processing Notice

Introduction

This document accompanies the information sheet and consent form and outlines the responsibilities of Sheffield Hallam University (SHU) in handling personal data collected from participants as part of the Nurturing Empathy before Transition project evaluation. The trial is being funded by the Youth Endowment Fund (YEF). Participants include: teachers and pupils.

From 25 May 2018 the General Data Protection Regulation (GDPR) replaces the Data Protection Act and governs the way that organisations use personal data. Personal data is information relating to an identifiable living individual.

Transparency is a key element of the GDPR and this Data Protection Statement is designed to inform participants about:

- how and why, SHU will use personal data collected in this evaluation
- how personal data will be pseudonymised and by whom
- why YEF archive pseudonymised data for use in future research
- data protection processes for archived data
- what participants' rights are under GDPR, and
- how to contact us to exercise those rights.

This Fair Processing Notice is compliant with the Data Protection Act 2018

Participants' Rights

One of the aims of the General Data Protection Regulation (GDPR) is to empower individuals and give them control over their personal data. The GDPR gives participants the following rights:

- the right to be **informed**
- the right of **access**
- the right to **rectification**
- the right to **erase**
- the right to **restrict** processing
- the right to **data portability**
- the right to **object**
- rights in relation to **automated decision making and profiling**

For more information about these rights please see: https://ico.org.uk/for-organisations/guide-to-data-protection/principle-6-rights/ and: https://www.shu.ac.uk/about-this-website/privacy-policy/data-subject-rights/subject-access-request

Participants can contact SHU at any time to:

- request copies of any personal data held by SHU (**a subject access request**)
- exercise **other rights** (e.g. to have inaccurate data rectified, to restrict or object to processing)
- **query** how data is used by SHU
- report a **data security breach** (e.g. if there are concerns that personal data has been lost or disclosed inappropriately)
- **complain** about how SHU have used personal data.

Details of who to contact are provided at the end of this statement.

Why are we processing participants' personal data?

It is necessary for SHU to process some personal data, to evaluate the impact of the Nurturing Empathy before Transition project. This will add to the research literature on empathy development, prosocial behaviour and crime.

Retention

After the evaluation with YEF is complete, SHU will retain participants' anonymised data for research and knowledge-exchange purposes, including presentations at professional or academic conferences, or publications in professional or academic journals, for a period of ten years after the last publication arising from the evaluation. After this period, SHU will review the longer-term archival value of the data.

Respecting confidentiality

In the production of professional or academic publications or presentations, all data will be fully anonymised and no individual or school will be identified or identifiable.

What is the legal basis for processing activities?

SHU are the Data Controllers for the data collected as part of the Nurturing Empathy before Transition project evaluation, up until the evaluation has finished. After the evaluation is finished (in 2024), the pupil data collected will be sent to the Department for Education (DfE) (at which point SHU cease to be responsible for the data), where it will be pseudonymised and transferred to the secure archive, which is being held by the Office for National Statistics in their Secure Research Service (SRS). Once the data has been transferred to the SRS, the Youth Endowment Fund become responsible for the data.

No pupils will be individually identifiable in the data archived and archived data will be kept indefinitely. Further information on YEF's data archive can be found below.

The processing of personal data through the Nurturing Empathy before Transition project evaluation is defined under GDPR as a specific task in the public interest. The legal basis for processing your personal data is 'Public Task' (Article 6 (1) (e)). https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/public-task/

Which Personal Data will we collect and use?

In order to provide our services as evaluators we need to collect and use some personal data. Below is a list of what this **may** include for the trial:

Type of personal data	Pupil	Teacher/Form tutor	School based lead	NEBT Instructors
Names	X	X	X	X
Contact details		X	X	X

Personal characteristics data including: Pupil Name, UPN, FSM status and gender	X			
Me & My Feelings responses	X			
Empathy questionnaire responses	X			
Pupils behaviour survey (filled in by tutor)	X			
Data on participation in the Nurturing Empathy before Transition project and school practices	X	X	X	X ('instructor checklist' in the MoU)
Observation the Nurturing Empathy before Transition programme	X	X		X

For SHU Nurturing Empathy before Transition case study schools:

Type of personal data	**Pupil**	**Teacher**	**Pastoral team**	**SLT**	**School based lead**	**NEBT Instructors**
Interview responses		X	X	X	X	X
Observation in school	X					X
Focus group data	X	X				

Using the information we receive from schools; we may also obtain data from sources such as the DfE Schools Comparison Service.

Who will we share personal data with?

The privacy of personal data is paramount and will not be disclosed unless there is a justified purpose for doing so. Data may be shared between SHU and:

- **YEF** for the purposes of research and evaluation**.** This includes submitting project data to the archive via the Department for Education (DfE) at the end of the project. At this point, YEF becomes a data controller, and DfE becomes a data processor (see data archiving section below for further information and the YEF privacy notice YEF_Data_Guidance_Participants_Nov2020.pdf (youthendowmentfund.org.uk)).
- **Transcribers**, who we may ask to produce transcripts of audio recordings of interviews and focus groups. If this is the case SHU will ensure that appropriate contracts and/or data-sharing agreements are in place and that the transcribers process personal data in accordance with the GDPR and other applicable legislation.

Archived data process

The Youth Endowment Fund (YEF) want to understand whether taking part in any of the programmes they fund prevents young people from being involved in criminal activity and/or violence in later life. In order to do this they archive the data evaluators collect, for future analysis.

The archive process is as follows:

Step one: Sheffield Hallam University will send the data collected during the Nurturing Empathy before Transition project to the Department for Education (DfE). DfE will then match each young person's data to their pupil matching reference number held within the National Pupil Database (NPD). Once this is complete, DfE will then delete the personal information sent, making it now impossible to identify any individual from the information.
Step two: Once the data has been pseudonymised (described in step-one), the DfE will then transfer the data to YEF's secure archive, held by the Office for National statistics (ONS) in the Secure Research Service (SRS).

Step three: The DfE link information they hold in the NPD with criminal justice information stored by the Ministry of Justice (MoJ) in the Police National Computer (PNC). The DfE will transfer this lined data into the SRS, so it can be linked with the pseudonymised data held in the YEF archive. This will allow YEF to fulfil their aim of researching the long-term impact participation in their projects has on young people and their involvement in criminal activity.

Step four: The archive will only be able to be accessed and the data within it used for research in accordance with the ONS's 'Five Safes' framework.

Safe people: Researchers wanting to use the data in the archive must apply for access, have relevant qualifications/experience and be trained by the ONS in how to use the data.

Safe projects: Restrictions are placed on how the data can be used and YEF must approve and support all research proposals intending to use the data. Each proposal must provide evidence to confirm that the research they have planned will serve the 'public good', has been approved by an ethics panel and has been assessed and passed by the ONS's research accreditation panel.

Safe settings: Data cannot be moved outside of the SRS, all analysis conducted by researchers must take place within the SRS. The ONS use secure technology, physical security and procedures and protocols to ensure the protection of the data when being used. Internet access whilst in the SRS is prohibited and monitoring software is used to record each user's activity.

Safe data: Researchers will only have access to de-identified data, to ensure that no individuals can be identified from the data.

Safe Output: Once a project is complete, researchers must get their outputs approved by two members of the ONS who will independently review the research outputs to ensure privacy and confidentiality is safeguarded. Any research undertaken in the SRS must be published.

In addition to the security and protection the ONS provides in the SRS, the YEF archive is also protected by laws which:

- prevent information being used in a way that may cause substantial damage or distress to an individual
- Make it unlawful for anyone to use the data to make decisions or take actions against an individual

SHU NEVER sell personal data to third parties

Security

SHU takes a robust approach to protecting the information they hold. This includes the installation and use of technical measures including encryption of data, firewalls and intrusion detection and prevention tools on networks and

segregation of different types of device; the use of tools on University computers to detect and remove malicious software and regular assessment of the technical security of SHU systems. SHU staff monitor systems and respond to suspicious activity. SHU also has Cyber Essentials certification.

Alongside these technical measures there are comprehensive and effective policies and processes in place to ensure that SHU users and administrators of information are aware of their obligations and responsibilities for the data they have access to. Access to project data is restricted to the research and evaluation teams and administrators associated with the project. Any sharing of the data with other researchers would require approval by the SHU Faculty of Social Sciences and Humanities ethics committee who will ensure that all data protection requirements are met. Training is provided to new staff joining SHU, and existing staff have training and expert advice available if needed.

Further Information and Support

For further information about how SHU use personal data see:
https://www.shu.ac.uk/about-this-website/privacy-policy/privacy-notices/privacy-notice-for-research
https://www.shu.ac.uk/about-this-website/privacy-policy/information-governance-policy

The Information Commissioner is the regulator for GDPR. The Information Commissioner's Office (ICO) has a website with information and guidance for members of the public:
https://ico.org.uk/for-the-public/

If there are any concerns about the way this project processes personal data, please raise these with the project teams.

Contact details
SHU
Dr Sarah Reaney-Wood
Sheffield Hallam University S1 1WB
s.j.reaney-wood@shu.ac.uk

OR

Governance Services
City Campus, Howard Street
Sheffield S1 1WB
foi@shu.ac.uk
0114 225 5555

If you have an ongoing concern, you can contact the Information Commissioner's Office, the body responsible for enforcing data protection legislation in the UK, at https://ico.org.uk/concerns/

Sheffield Hallam University | Sheffield Institute of Education

YOUTH ENDOWMENT FUND

Roots of Empathy

Intervention School Information Sheet: Case Study evaluation of Nurturing Empathy Before Transition

Background

Dear School,

Your school is taking part in the **Nurturing Empathy Before Transition (NEBT)** programme and is one of the intervention schools. The programme is run by Roots of Empathy UK (ROE), and funded by the Youth Endowment Fund (YEF). This letter briefly explains what the programme is and what is involved for your school in terms of being an evaluation case study.

What is 'Nurturing Empathy Before Transition'?

- The project aims to increase empathy and pro-social behaviour (such as helping, sharing and co-operating with others) in Year 5 pupils by developing their emotional and social skills before transition to secondary school.
- This project uses the established NEBT programme, which involves bringing a parent and baby into the classroom as part of a structured programme of lessons focused on building empathy. It is described as "an evidence-based, preventative intervention for primary school children, that aims to reduce aggression, including bullying, and increases children's social and emotional competence".
- The programme consists of 27 lessons delivered over the school year, covering nine topics. Three sessions a month will take place: a pre session; parent and baby session; and a post-family visit session.

For further information on the delivery of the NEBT programme, please contact: Kcohen@rootsofempathy.org

The Case Study Evaluation

The NEBT Evaluation: In order to find out whether the programme raises levels of empathy and pro-social behaviour in Y5 pupils, the YEF has funded an evaluation of NEBT which will be undertaken by Sheffield Hallam University. As part of this evaluation the evaluators will be carrying out a case study at your school. As part of this the following evaluation activities will be taking place over a half day school visit by the evaluation team:

- Observation of the ROE session (where feasible)
- Undertaking a focus group with a sample of pupils after the session
- Interview with the ROE Instructor
- Interview with the teacher present during the ROE lesson
- Interview with a member of SLT (where possible)

For further information on the evaluation of NEBT, please contact Josephine Booth Josephine.booth@shu.ac.uk

Sheffield Hallam University | Sheffield Institute of Education

YOUTH ENDOWMENT FUND

Roots of Empathy

The evaluation's ethical approach

This evaluation has been checked and approved by the University Research Ethics Committee (UREC) - (www.shu.ac.uk/research/quality/ethics-and-integrity) [Ethics Reference: ER19810112]. Below we outline the key things we will do to ensure this is fair and respectful:

- All responses in pupil focus groups and staff interviews will be anonymous and confidential. This means no individual pupil, staff member or school will be identifiable in any way in our reports or publications.
- Any data collected will be stored confidentially for up to ten years, as is standard university practice. All data will be stored securely on password protected and encrypted computer servers and will be held in compliance with the General Data Protection Regulation (GDPR) before secure transfer to DfE at the end of the project (see below).
- Parents/carers have the right to withdraw their child from the evaluation process and data collections at any point, as do members of staff taking part in the case study interviews.

Right to change your mind

You are free to withdraw your school from the evaluation at any time. You can also ask the team not to use your school's focus case study data up to two weeks after the session has taken place, this is called a Subject Access Request. If you do decide to withdraw your data, it will not be included in the evaluation. Please contact josephine.booth@SHU.ac.uk.

Data protection: Data security and right to withdraw

Personal data will be kept securely on a password protected University folder accessible only to members of the evaluation team. The SHU evaluation team will comply with General Data Protection Regulation (GDPR legal basis: public task Article 6 (1e)) and the SHU Data Protection Policy Statement. Please refer to the link for more information: https://www.shu.ac.uk/about-this-website/privacy-policy/privacy-notices/privacy-notice-for-research

After the evaluation is finished (in 2024), the pupil data collected (name, UPN, DOB) will be sent to the Department for Education (DfE) (at which point SIOE cease to be responsible for the data), where it will be pseudonymised and transferred to the secure archive, which us being held by the Office for National Statistics in their Secure Research Service (SRS). This data will be matched to national pupil data (NPD) (for example, attainment) and criminal justice information in the Police National Computer (PNC). This will allow future research to take place into whether the NEBT project has an impact on educational attainment and criminal records. Once the data has been transferred to the SRS, the Youth Endowment Fund become responsible for the data. No pupils will be individually identifiable in the data archived and archived data will be kept indefinitely. For more detail please see: https://youthendowmentfund.org.uk/wp-content/uploads/2020/10/YEF_Data_Guidance_Participants_Nov2020.pdf

You should contact the Data Protection Officer if:

- you have a query about how your data is used by the University
- you would like to report a data security breach (e.g. if you think your personal data has been lost or disclosed inappropriately)
- you would like to complain about how the University has used your personal data DPO@shu.ac.uk

You should contact the Head of Research Ethics (Dr Mayur Ranchordas) if:

- you have concerns with how the research was undertaken or how you were treated ethicssupport@shu.ac.uk

Postal address: Sheffield Hallam University, Howard Street, Sheffield S1 1WBT
0114 225 5555

Sheffield Hallam University | Sheffield Institute of Education

YOUTH ENDOWMENT FUND

Roots of Empathy

Outputs and what happens next: The evaluation report will be published on the YEF website and the findings may also be disseminated at educational research conferences and in academic or professional journals. No individual pupils, staff or schools will be named in reporting.

THANK YOU FOR TAKING THE TIME TO READ THIS INFORMATION SHEET

Please answer the following questions by indicating the response that applies

	YES	NO
1. I have read the Information Sheet for this research and have had details of the study explained to me.	☐	☐
2. Any questions have been answered to my satisfaction and I understand that I may ask further questions at any point.	☐	☐
3. I understand that I am free to withdraw my data without giving a reason for my withdrawal, or to decline to answer any particular questions without any consequences for the programme or to me.	☐	☐
4. I wish to participate under the conditions set out in the Information Sheet.	☐	☐
5. I agree for the interview to be audio recorded for note taking purposes	☐	☐

By signing below, you indicate that you have read and understood this information sheet and voluntarily decided to take part in the activities outlined in it.

Participant's Name: **Date:**
Participant's Signature:

Researcher's name: **Date:**
Researcher's signature:

3

Pupil Information Sheet: Evaluation of Nurturing Empathy Before Transition

We want to tell you about some new activities that will be happening in your school. A charity called Roots of Empathy have some interesting new lessons and we want to find out if they will help you in your life. The activities that will be happening are part of a project called **Nurturing Empathy Before Transition (NEBT)**

What is a charity?

It is a group of people who do things to mainly help others

What is empathy?

Empathy is being able to understand how someone else feels and being able to feel with them. Like understanding that bullying might make a person feel sad or feel scared.

Nurturing Empathy Before Transition (NEBT) project

- The NEBT project aims to help you to understand how others feel, manage your own feelings and how you act towards other people, before you move to secondary school.
- NEBT project uses the Roots of Empathy programme that lots of other children around the world have also had in their school.
- While you are in Year 5, you will participate in the Roots of Empathy programme which has 27 lessons in total. In some of these lessons, a Mum and/or Dad and their baby will come into the classroom. You will learn about many things, such as why the baby cries and how you can tell if baby is happy.You will also be given opportunities to discuss your feelings and listen to the feelings of others.

To find out whether the Roots of Empathy programme helps children to understand how people feel and helps the way they treat others, an organisation called the Youth Endowment Fund (YEF) has given some money for the Roots of Empathy lessons to be tested to see if they work (this is called an evaluation). Your school is one of 70 schools in England who will be having these lessons.

What is an evaluation? It is a way of finding out if something is working in the way it should.

What will I have to do?

Researchers at Sheffield Hallam University (SHU) are carrying out the evaluation of the Roots of Empathy programme to find out whether these lessons do help children.

Year 5 pupils having these Roots of Empathy lessons will be asked to answer some questions online in September and again in July. Researchers will also visit some of the schools to ask pupils and teachers what they thought about the lessons. This is called a ***'focus group'*** and will also include other children in your class.

Do I have to take part in the focus group?

No, you do not have to take part in the focus group as part of the evaluation.

What if I change my mind after I've said yes?

If you decide to take part and then change your mind, that's OK. Just talk to your parent/carer, teacher or the researcher. They will be able help you and arrange for you not to take part anymore. This is called having a **'right to withdraw'** and can happen at any point in Year 5. You will not be in trouble if you don't want to take part.

Do my parents know that I will be taking part?

Your parents have been sent a letter to tell them that you and your school are taking part in the evaluation. They have been given information about the lessons and the evaluation. They have been asked if it OK for you to take part.

What will happen to the answers I give?

This evaluation has been checked and approved by the University Research Ethics Committee (UREC) - (www.shu.ac.uk/research/quality/ethics-and-integrity).

- The researchers will not tell anyone else what you say. They will not name you or the school in any reports they write afterwards. This means your answers will be anonymous and confidential. No one **will be able to tell that those answers have come from you.**
- It is not a test and **there are no right and wrong answers**. So, as an example, if you are asked if you have enjoyed having a baby in the class and you haven't, it is OK to say no!
- The information that we collect from you, such as your name and date of birth will be kept very safe. All the information will be stored and protected using a password and will follow the General Data Protection Regulation (GDPR) rules.

Data protection: Confidentiality and right to withdraw

The statement below is included on all information sheets as this is important for research. Please ask a trusted adult to explain the information below if you need help to understand this!

Personal data will be kept securely on a password protected University folder accessible only to members of the evaluation team. The SHU evaluation team will comply with General Data Protection Regulation (GDPR legal basis: public task Article 6 (1e)) and the SHU Data Protection Policy Statement. Please refer to the link for more information: https://www.shu.ac.uk/about-this-website/privacy-policy/privacy-notices/privacy-notice-for-research

After the evaluation is finished (in 2024), the pupil data collected (name, UPN, DOB) will be sent to the Department for Education (DfE) (at which point SIOE cease to be responsible for the data), where it will be pseudonymised and transferred to the secure archive, which us being held by the Office for National Statistics in their Secure Research Service (SRS). This data will be matched to national pupil data (NPD) (for example, attainment) and criminal justice information in the Police National Computer (PNC). This will allow future research to take place into whether the NEBT project has an impact on educational attainment and criminal records. Once the data has been transferred to the SRS, the Youth Endowment Fund become responsible for the data.
No pupils will be individually identifiable in the data archived and archived data will be kept indefinitely.

Archived information will only be accessed and used for research in accordance with the ONS's 'Five Safes' framework. For more detail please see: https://youthendowmentfund.org.uk/wp-content/uploads/2020/10/YEF_Data_Guidance_Participants_Nov2020.pdf

You should contact the Data Protection Officer if:

- you have a query about how the data is used by the University
- you would like to report a data security breach (e.g. if you think your personal data has been lost or disclosed inappropriately)
- you would like to complain about how the University has used your personal data DPO@shu.ac.uk

You should contact the Head of Research Ethics (Dr Mayur Ranchordas) if:

- you have concerns with how the research was undertaken or how you were treated ethicssupport@shu.ac.uk

Postal address: Sheffield Hallam University, Howard Street, Sheffield S1 1WBT
0114 225 5555

Sheffield Hallam University | Sheffield Institute of Education

YOUTH ENDOWMENT FUND

Roots of Empathy
Racines de l'empathie

Pupil Consent form

	YES	NO
1. I have read this Information Sheet and have had things explained to me if I did not understand them.	☐	☐
2. I could ask questions, and someone has answered them. I understand that I can ask more questions if I want to at any time.	☐	☐
3. I understand that I do not have to answer any questions I don't want to.	☐	☐
4. I understand that I can change my mind about taking part and can ask for my information to be removed later if I want to. I do not have to say why.	☐	☐
5. If I choose not to take part, it's OK.	☐	☐
6. I understand that no one will be able to tell that those answers have come from me.	☐	☐
7. I would like to take part in this research, as set out in the Information Sheet.	☐	☐
8. I am OK with the focus group being recorded.	☐	☐

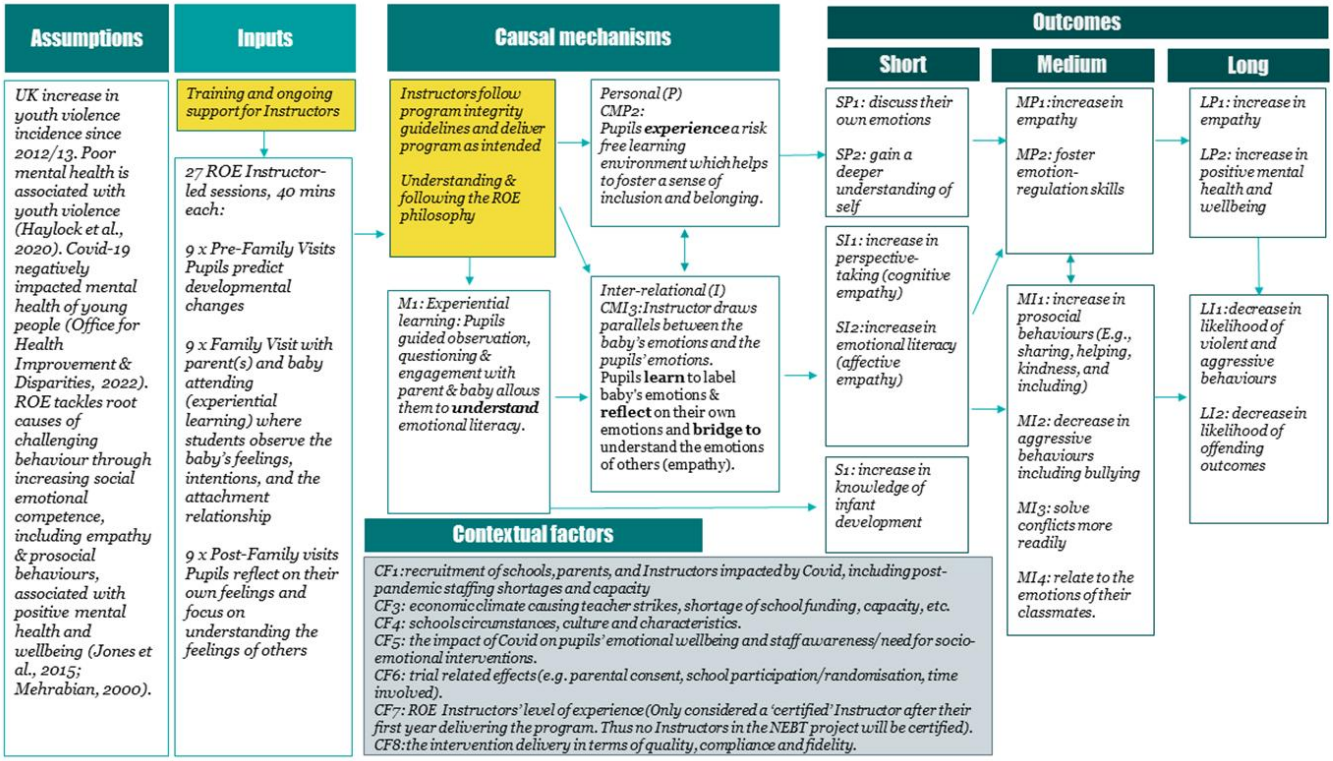
☐

My name: ..... Date:

Researcher's name:..... Date:

4

Appendix D: Finalised NEBT logic model


Assumptions
Inputs
Causal mechanisms
Outcomes
Short
Medium
Long
UK increase in youth violence incidence since 2012/13. Poor mental health is associated with youth violence (Haylock et al., 2020). Covid-19 negatively impacted mental health of young people (Office for Health Improvement & Disparities, 2022). ROE tackles root causes of challenging behaviour through increasing social emotional competence, including empathy & prosocial behaviours, associated with positive mental health and wellbeing (Jones et al., 2015; Mehrabian, 2000).
Training and ongoing support for Instructors
27 ROE Instructor-led sessions, 40 mins each:
9 x Pre-Family Visits Pupils predict developmental changes
9 x Family Visit with parent(s) and baby attending (experiential learning) where students observe the baby's feelings, intentions, and the attachment relationship
9 x Post-Family visits Pupils reflect on their own feelings and focus on understanding the feelings of others
Instructors follow program integrity guidelines and deliver program as intended
Understanding & following the ROE philosophy
Personal (P) CMP2: Pupils experience a risk free learning environment which helps to foster a sense of inclusion and belonging.
M1: Experiential learning: Pupils guided observation, questioning & engagement with parent & baby allows them to understand emotional literacy.
Inter-relational (I) CMI3: Instructor draws parallels between the baby's emotions and the pupils' emotions. Pupils learn to label baby's emotions & reflect on their own emotions and bridge to understand the emotions of others (empathy).
SP1: discuss their own emotions
SP2: gain a deeper understanding of self
SI1: increase in perspective-taking (cognitive empathy)
SI2: increase in emotional literacy (affective empathy)
S1: increase in knowledge of infant development
MP1: increase in empathy
MP2: foster emotion-regulation skills
MI1: increase in prosocial behaviours (E.g., sharing, helping, kindness, and including)
MI2: decrease in aggressive behaviours including bullying
MI3: solve conflicts more readily
MI4: relate to the emotions of their classmates.
LP1: increase in empathy
LP2: increase in positive mental health and wellbeing
LI1: decrease in likelihood of violent and aggressive behaviours
LI2: decrease in likelihood of offending outcomes
Contextual factors
CF1: recruitment of schools, parents, and Instructors impacted by Covid, including post-pandemic staffing shortages and capacity
CF3: economic climate causing teacher strikes, shortage of school funding, capacity, etc.
CF4: schools circumstances, culture and characteristics.
CF5: the impact of Covid on pupils' emotional wellbeing and staff awareness/need for socio-emotional interventions.
CF6: trial related effects (e.g. parental consent, school participation/randomisation, time involved).
CF7: ROE Instructors' level of experience (Only considered a 'certified' Instructor after their first year delivering the program. Thus no Instructors in the NEBT project will be certified).
CF8: the intervention delivery in terms of quality, compliance and fidelity.

Appendix E: Analysis in the presence of non-compliance

Compliance to the NEBT programme was specified using three conditions:

- **Instructor Level:** Whether the NEBT instructor attended all four NEBT training sessions (=1) or not (=0).
- **School Level:** Whether schools delivered at least eight of the nine NEBT themes (=1) or not (=0).
- **Pupil Level:** Whether pupils in NEBT schools attended at least eight of the nine NEBT themes (=1) or not (=0).

Overall compliance is achieved when a '1' is scored on all three conditions.

ROE provided data for the Instructor and School level conditions whilst pupil attendance was collected directly from schools. In both Cohorts 1 and 2 of the evaluation, there was a very high drop-out rate for the NEBT intervention schools. In Cohort 1, eight of the 16 schools randomised to NEBT dropped out and in Cohort 2, 15 of the 30 schools randomised to NEBT dropped out. These schools dropped out shortly after being randomised and so none will have met the Instructor or School level compliance conditions. Unfortunately, the NEBT schools dropped out of participating in the NEBT programme and evaluation data collections. This means that, whilst we know that 50% of schools that were randomly allocated to receive the NEBT programme did not comply (because they dropped out), we obtained no outcome data for these schools. Attrition at the school[28] and pupil[29] levels is very unlikely to be random and so very likely to have undermined randomisation and (therefore) the validity of drawing causal conclusions from any of the impact analyses in this evaluation. This methodological weakness carried into these CACE analyses and increased further because of the particularly poor[30] response from schools for the pupil attendance Compliance condition. In both cohorts 1 and 2, the pupil attendance data represented the sixth and final data collection for the impact evaluation of the NEBT programme. In addition to these six evaluation data collections, ROE collected data for their own, internal, evaluation. The extent of data collection activities seems to have been a notable burden on schools; reflected in the very high levels of missing data.

We begin the CACE analysis by presenting the compliance conditions at the instructor and school level in Table 3.

Appendix Table 3 21:Instructor and school level compliance

	Baseline Sample (N=46 schools)	ITT complete case sample for primary outcome (N=23 schools)
Instructors attended all 4 training sessions	34 (74%)	22 (96%)

28 50% (23 out of 46) of NEBT schools and 39% (16 out of 41) control schools dropped out of the evaluation.

29 At the pupil level, attrition was 65.5% in NEBT schools and 56.1% in control schools.

30 Across both cohorts, we obtained pupil attendance data from 10 of the 23 intervention schools that did not withdraw from the NEBT programme.

8+ topics delivered from NEBT curriculum recorded by instructors	28 (61%)	22 (96%)

Across the whole sample of 46 schools randomised to the NEBT programme in cohorts 1 or 2, compliance was higher at the instructor level (34 instructors, 74%) compared with the school level (28 schools, 61%). However, 23 (50%) of these schools dropped out of the NEBT programme and evaluation and we obtained no outcome data for pupils in these schools. Within the subsample of 23 schools that did not drop out of the NEBT programme, Compliance was nearly universal with the same 22 schools (96%) meeting both instructor and school level compliance conditions.

The sample of 22 schools identified as meeting the instructor and school level compliance conditions included 300 of the 314 pupils included in the ITT analysis of the M&MF Behavioural Difficulties primary outcome, with 14 pupils in the one school that did not meet the conditions. Table 4 presents descriptive statistics for the M&MF Behavioural Difficulties primary outcome for these two pupil subsamples.

Appendix Table 4 22: M&MF Behavioural Difficulties summary for pupils in schools that met the instructor and school level compliant conditions compared with pupils in the school that did not meet the compliant conditions.

	Compliant (N=300)		Noncompliant (N=14)	
	N=	Mean (95% CIs)	N=	Mean (95% CIs)
Mean M&MF BD outcome scores (95% CIs)	300	2.81 (2.57; 3.05)	14	3.79 (1.88; 5.70)

The subsample of pupils in schools which met both instructor and school level compliant conditions had a lower mean M&MF BD score at outcome compared with pupils in the school that did not meet these conditions. In other words, on average, pupils in compliant schools had lower self-reported behavioural difficulties at compared with pupils in the non-compliant school. However, the 95% confidence intervals for both subsamples overlap, indicating that the difference may be positive, null or negative (or that the difference between the two groups is not statistically significant). Given this, we conclude that, from these descriptive analyses, we found no evidence that compliance at the instructor or school level had an impact on the primary outcome. These descriptive analyses are limited in two ways. First, the very high attrition at both school and pupil levels undermines the validity of drawing firm conclusions. Second, these are bivariate descriptive analyses that do not take account of the baseline M&MF BD score, the clustering of data at the school level or the (regional) stratification of the randomisation. Whilst the CACE analyses presented below address the second limitation, the problem of attrition will remain and so caution is required when interpreting these findings.

Pupil attendance of NEBT lessons was collected from just 10 of the 23 schools that did not drop out of the NEBT programme. Table 5 presents the pupil level compliance condition alongside the mean M&MF BD score and 95% confidence intervals.

Appendix Table 5 23:Pupil level Compliance

	N=	M&MF BD Outcome Mean (95% CIs)
Pupils recorded as attending 8+ NEBT themes	125	3.03 (2.64; 3.42)
Pupils recorded as not attending 8+ NEBT themes	20	3.75 (2.97; 4.53)
Pupils with missing attendance details	169	2.62 (2.29; 2.95)
Total number of pupils in NEBT intervention group with baseline & outcome M&MF BD scores	314	2.86 (2.62; 3.10)

The subsample of pupils known to have attended at least eight of the NEBT themes had a lower mean M&MF BD score compared with pupils known not to have attended at least eight themes. At the same time, the mean M&MF BD score for the subsample of pupils with missing attendance details was lower than the mean score for pupils known to have attended at least eight themes. In other words, on average, pupils who attended at least eight NEBT themes had lower self-reported behavioural difficulties compared with pupils who did not attend at least eight themes but lower self-reported behavioural difficulties compared with pupils with no attendance details. However, the 95% confidence intervals for all subsamples overlap. Given this, we conclude that, from these descriptive analyses, we found no evidence that compliance at the pupil level had an impact on the primary outcome. As with the instructor and school level compliance conditions analyses, these descriptive analyses are limited in two ways. First, the very high attrition at both school and pupil levels undermines the validity of drawing firm conclusions. Second, these are bivariate descriptive analyses that do not take account of the baseline M&MF BD score, the clustering of data at the school level or the (regional) stratification of the randomisation. The CACE analyses presented below addresses the second limitation but the problem of attrition will remain and so caution is required when interpreting these findings.

The SAP noted that compliance was not expected to be close to 100%. The high withdrawal rate of NEBT schools reflects this expectation; 23 of the 46 schools withdrew (50% in each cohort). Unfortunately, the schools that withdrew from delivering the NEBT programme also withdrew from all or most evaluation activities. Within the subsample of 23 NEBT schools that did participate in the programme and evaluation data collections, compliance was very high at the school- and instructor-levels. If all schools that withdrew from the NEBT programme did not deliver at least eight NEBT themes, across the whole sample we estimate compliance at the school level to be 48% (22 out of the 46 NEBT schools randomised). However, our analyses that explore the relationship between compliance and the impact of the NEBT programme is limited to the subsample of 23 schools that provided baseline and outcome primary outcome data. In this 'complete case' subsample, school-level compliance was much higher at 96% (22 out of 23 schools). When the pupil attendance threshold is included, the sample diminishes further to just 10 schools.

We have adhered to the specification of compliance in the SAP; a pupil is identified as compliant if they are known to have attended at least eight NEBT themes in a school that delivered at least eight NEBT themes with an instructor who attended all four NEBT training sessions. We follow this specified definition with one that includes the full ITT sample by assuming 100% attendance of pupils in the 12 NEBT schools. This second CACE analysis was not specified in the SAP and is in response to the notable issue of missing data.

Table 6 below illustrates the specified and sensitivity versions of the compliance variable used in the CACE analyses.

Appendix Table 6 24: Compliance Measures

	SAP specified	Sensitivity
Measures of compliance to NBT:		
Compliance:	125 (79%)	280 (89%)
Non-Compliance	34 (21%)	34 (11%)
Total	159	314

The CACE estimate was obtained using the STATA 'ivregress' command with adjustment to standard errors to acknowledge clustering at the school level. Table 18 presents the effect size estimates from the CACE analyses.

Appendix Table 7 25: CACE analysis

	SAP specified	Sensitivity
Samples: pupils (schools)		
NBT	159 (10)	314 (23)
Control	330 (25)	330 (25)
	489 (35)	644 (48)
Effect Sizes (95% CIs)		
ITT analysis	+0.04 (-0.13; +0.20)	-0.06 (-0.22; +0.10)
CACE analysis using IV (ivgreress)	+0.05 (-0.13; +0.23)	-0.06 (-0.24; +0.12)
F-test for exogeneity	F(1,35)=3.27; p=0.08	F(1,47)=2.92; p=0.09
CACE analysis using the formula	+0.05 (-0.17; +0.26)	-0.06 (-0.24; +0.12)

For the SAP specified CACE analysis, compliance is estimated at 79% and the correlation between group membership and compliance was high at 0.84. The SAP specified CACE estimate was +0.05 sds (CIs: -0.13; +0.23). This compares with the ITT analysis estimate of -0.06 (CIs: -0.22; +0.10). However, the ITT analysis was based on the complete case sample (n=644) and when this analysis is re-run with the restricted SAP specified compliance sample (n=489), the ITT estimate has the same sign and was closer to the CACE estimate at +0.04 (CIs: -0.13; +0.20). The F-test for the first stage of the instrumental variable model was small and not statistically significant (F(1,35)=3.27, p=0.08) suggesting that compliance can be considered to be exogenous. The alternative CACE estimate using the formula specified in the SAP provides a second estimate for the SAP specified CACE and was close to what was obtained from the instrumental variable models at +0.05 sds (CIs: -0.17; +0.26).

For the sensitivity CACE analysis, compliance is estimated at 89% and the correlation between group membership and compliance was very high at 0.90. The sensitivity CACE estimate was -0.06 sds (CIs: -0.24; +0.12) which compares closely to the ITT analysis estimate of -0.06 (CIs: -0.22; +0.10). The F-test for the first stage of the instrumental variable model was small and not statistically significant (F(1,47)=2.92, p=0.09) suggesting that compliance can be considered to be exogenous. The alternative CACE estimate using the formula provides a second estimate for the sensitivity CACE and was close to what was obtained from the instrumental variable models at -0.06 sds (CIs: -0.24; +0.12).

In summary, the ITT analyses found no evidence of impact for pupils being offered the NEBT programme and the CACE analyses found no evidence of impact for pupils who engaged in the NEBT programme as intended. We therefore conclude that we found no evidence that NEBT had a statistically significant impact on pupil behaviour as measured by the self-reported M&MF BD scale. The notable caveat for these and all other impact analyses in this evaluation is the very high rate of attrition (61.3% lost to attrition, only 38.7% of pupils randomised to the NEBT programme included in the impact analyses). This serves to undermine the

validity of drawing causal conclusions from these analyses; but from the data we analysed we found no evidence of impact.

Appendix F: IPE Focus Group and Observations Supplementary Material

Pupil focus groups

As noted in Table 7 focus groups were held with pupils who had taken part in the programme in the five, year 2 case study schools (CS4-8). The same schedule of questions was used at every school, with participating pupils selected by school staff. Overall, pupils were willing to share their experiences; however, responses from pupils at CS7 school were noticeably shorter compared to other settings. It is important to note that the instructor was present throughout this focus group (CS7), and at one point, another teacher entered the room with a pupil to discuss behaviour, which may have influenced the pupils' responses. The focus groups were analysed inductively using NVivo. The following section summarises the findings.

Children were very positive about the family visits, stating that it had been fun and exciting to see the baby grow, develop and start interacting with the world around them. One child stated that seeing the baby made them happy, another noted that the highlight was being the first to hold the baby. They were able to recall things that the baby had done, for example playing with toys and crawling. Slightly more negative responses were given around the baby crying, being sick, dribbling and the session around nappies. Pupils enjoyed singing to the baby, particularly when they were able to sing a song that they had made up for the baby.

Some pupils were able to recall the themes and speak about the topics and activities, for example temperament and transitional objects, as well as potential dangers to the baby (sharp objects, being left alone), changing nappies, routines, the cost of raising a baby, whether disposable or washable nappies were best for the environment and how to hold and feed a baby.

> *We always have a pre and post lesson after our visit and for our pre lesson we talk about our theme, and we have nine themes that we have done, and this theme was about everything that we've learned about in the other eight themes.* Pupil, CS8

They also spoke about how you could tell a baby was tired by their body language, and that as babies do not have language they communicate by crying, and that these cries may be differentiated in terms of tiredness or hunger.

> *That you have two tones of crying…There is like a different one for sleeping, when you're sleepy. And then another one is when he's like hungry and his mum understands that.* Pupil, CS5

> *I learnt that when babies cry they're trying to say to you that they're tired, hungry, lonely, or want to play with you – and of course babies cannot communicate like us so they cry instead.* Pupil, CS4

> *How different babies communicate.* Pupil, CS7

One pupil was able to articulate that this inability to express their needs must be upsetting and frustrating for a baby.

> *they can't tell us what they're frustrated. They've just got to cry and then some parents might not know what they mean so they might say 'Oh, can you stop crying' and they will be like why are you telling me to stop crying* Pupil, CS5

Several pupils noted that they had learned how to look after a baby from the programme and this would help when they had babies, or to look after their younger siblings.

In terms of thinking about learning about emotions, and how a baby might express itself, pupils spoke about how the baby's emotions changed quickly, how they used their body language, cries and gaze to communicate.

> *So if he needs to get something he looks at us in the eye directly to communicate to give it to us.* Pupil, CS6

Pupils spoke about how people might express their emotions and how, for example, anger might evoke different reactions from different people, people have differing temperaments and how you might be able to tell how someone is feeling by their body language.

> *some people react with violence, some people react in language, and then some people react in emotions.* Pupil, CS8

They were also able to articulate that people have differing emotions

> *It's taught us that it is okay to have lots of different emotions and how to deal with some of them and stuff.* Pupil, CS6

> *everyone has got a different personality and everyone has got different feelings.* Pupil, CS8

Pupils were also able to speak about what they had learned about their own emotions and how they treat others and were also able to name some of the emotions they had learned about including sympathy and empathy, and to differentiate between these.

> *Like you can understand other people's feelings*. Pupil, CS6

> *It's taught us to care about others more than we care about ourselves.* Pupil, CS4

Some pupils stated that they might be more able to deal with their own emotions now, as well as share their feelings with peers and that this had given them confidence in class, in sharing what they think or feel and in moving up the school.

> *because the first time I shared something with the class when we were learning about baby* [name] *I felt a bit embarrassed, but now I do it quite a lot and I feel fine.* Pupil, CS4

> *since* [instructor] *has teached* [sic] *us all the lessons I've felt more confident and I used to be scared of going in Y6, but now I'm not that scared.* Pupil, CS4

Pupils were also able to talk about emotions in terms of temperament and link this to the baby's behaviour and what they had learned about neurology.

> *.. about like their temperament when they cry. So like they can have a low temperament or like a high one* Pupil, CS4

> *I even learned that when babies learn new things then the neurons join together and it makes like a memorable mind for them.* Pupil, CS4

A couple of pupils linked what they had seen in the family sessions with their own lives, realising that their parents would have looked after them the way that they saw the mother looking after her baby, and that this had changed the way that they behaved at home.

> *How* [mother] *takes care of baby* [name] *and how she really loves him makes me feel like I remember to not be really rude to my mum and dad, because I go back and remember when I was a baby, that is what my mum and dad would have done as well* Pupil, CS4

> *So basically I used to be really rude to my parents but now I've seen Baby* [name]*, it's made me a bit more polite to my parents and my sister.* Pupil, CS4

Pupils clearly remembered the topics and themes around smoking and the associated chemicals in cigarette smoke and the dangers around shaking babies.

> *remember theme 2 with the dangers of shaking the baby…. and also being sick on the alcohol and second-hand smoke. I remember that.* Pupil, CS6

> *like that you should not shake babies because the fluid in their head will mix and it could cause a sign of death, blindness, and cause your baby to not hear*. Pupil, CS4

Overview of pupil focus groups

Pupils were generally happy to speak to interviewers and were responsive to questions. They were able to recall themes, activities and topic areas, and in some cases link them to their wider lives, and families. They also willingly shared what they had learned, using terminology from the intervention such as communication, temperament, and neurons. Additionally, they were able to discuss emotions and how individuals may differ in their feelings and reactions. A few pupils indicated that they had gained confidence in speaking up and sharing their thoughts and experiences in class through the intervention.

The family visits were clearly the highlight of the programme, with pupils clearly attached to 'their' baby and speaking about interacting with the baby and the development that they had seen with excitement.

Session Observations

As noted in Table 27, one observation took place in year 1, and five session observations took place in schools in year 2 (CS4-8). The observer recorded notes in a proforma, which were later typed up. The observations are presented here in a descriptive format by type of session observed (e.g. pre, baby, and post), and by school.

Pre family sessions

Three pre family visit sessions were observed; one on Theme 6 – safety in year 1. The two undertaken in year 2 were on the temperament topic.

<u>Case study 1</u>

During the session the teacher sat on the floor with the pupils, while the instructor led the session using her notes. The topic was not clearly introduced at the beginning of the session. Following the 'hello song' the pupils took turns to have the class tablet to take pictures during the session for the notice board/displays.

The instructor asked pupils how they felt about seeing the baby, to which they replied that they were happy and excited. the session then moved onto thinking about the baby's feelings and how could they tell how she was feeling. There was discussion around observing the baby's behaviour and mood (the baby cried and was comforted by her mum).

There followed some recall work on not smoking and drinking during pregnancy, followed by a recap from the pre family visit session. This was not explored in much detail, although some effort was made to link this topic to home contexts. The instructor then asked mum some questions about the baby's progress and milestones. The session was slow paced. The session ended with the 'goodbye' song.

Case study 4

The instructor asked pupils to consider '*what else makes us who we are?*'. The pupils offered various suggestions including beliefs, organs, names, languages, friends, church/religion, and where you've lived. The instructor often asked pupils to expand on their suggestions and also asked questions for example '*Who has ever lived in a different country*?'. Prior to the introducing the book that forms the focus of the lesson, the instructor asks what makes us 'the same but different'? Following a discussion, she reveals that we all share 'emotions', before introducing 'Welcoming Babies' by Margy Burns Knight. Several extracts were read out, giving diverse insights into the different traditions used throughout the world to introduce babies. Pupils are asked about their family traditions.

The pupils appeared less engaged when the book sections were being read out. However, when they were asked questions around traditions they appeared to reengage and shared their experiences. Pupils remained focussed during the final activity which asked them to think about what they might ask baby Ryan if he could talk. This was supported by the instructor and teacher with the prompt of '*Think about baby Ryan's point of view and how he would answer it'.* Questions were kept for Ryan's mum to answer in the following session.

Case study 7

The second 'temperament' session observed followed the same format, with pupils expressing that factors such as their personality, attitude, humour, interests, language, religion, and family contribute to their identity. As in the other school, the instructor then introduced the book 'Welcoming Babies', and the pupils had a discussion around family traditions, which led on to a conversation around baby Libby. This ranged from how babies should be looked after and changes that they had seen in Libby and her development – smiling, crawling, teething and rolling for example.

A task then centred around what pupils would like to know from the baby's perspective - if he could talk what questions would you ask him? Pupils wrote down their questions and these were passed around the class for others to try to answer. The focus was around trying to get the pupils to think from the perspective of the baby and what they know about Libby from her mum and from watching her grow and develop. For example, what is Libby's favourite food, what does she enjoy. Again, the questions were saved to ask Libby's mum.

The session ended with the instructor summarising learning – that everyone is all the same and all different and that these differences should be celebrated. The observer noted that the pupils seemed disappointed

when the lesson was over but were excited to see Libby at the next session. Pupils remained seated at the desks in this session.

Family visit sessions

Two of these sessions were observed.

Case study 5

Overall, the session observation in CS5 was positive, with most pupils engaged in activities, answering questions and joining in with singing and breathing exercises in these sessions. Some pupils chose not to tickle the baby's toes or to interact with the baby more directly. Those who did not sing were still watching the baby and listening.

Pupils were keen to answer questions about how the baby (David) might be feeling e.g. 'I think he missed us', 'he loves us', 'he is happy'. Milestones were highlighted by the instructor and mum e.g. clapping and four new teeth, with the pupils observing the baby chewing on a toy to help him with this. The instructor talked about David learning to clap and how proud his mother was of this achievement, and linked this with how pupils feel when they progress at school, and that their parents are also proud when they make progress. When David clapped the pupils clapped too. Mum also noted that David was nearly crawling and was now rolling and bum shuffling – a pupil stated that he too had bum shuffled. The instructor asked David's mum about his interactions with new people, prompting a discussion on adaptability and how to determine if David is happy.

The class talked about how David communicates with his mum, with pupils answering that he looks to her, moves towards her, and cries. This was linked to the use of language in the playground – what is ok and not ok, and what to do if pupils are unkind. The importance of how we communicate was highlighted e.g. name calling is a way of communication. The discussion also explored broader ways of speaking to others and the potential positive or negative impacts. The instructor asked questions e.g. 'how do you communicate with your family?', 'can your mum tell when you aren't happy?', and spoke about how families become 'attuned' to one another. This was then linked to neuroscience in terms of verbal and nonverbal communication, and how Alex has not yet developed the capacity for speech so has to rely on other forms of communication to tell people how he feels.

Case Study 8

The second baby session observation was the final visit from the baby and mother to CS8. It was noted that the session lacked a clear focus and structure evident in other sessions observed, with the instructor struggling to hold the children's attention. This may have been a result of its nature as the final session, which appeared to be a celebration rather than a learning opportunity. It was observed that the instructor often spoke directly to the baby's mother rather than engaging with the pupils, while any questions asked to pupils about baby Lizzy were random and unrelated, which further detracted from any potential learning focus. Pupils were not as engaged in the session activity, with several children refusing to participate in singing for example. Overall, while some students enjoyed interacting with the baby, the lack of focus made the session feel disjointed and unproductive, with low levels of engagement and a lack of boundaries around pupil behaviour. During the celebration session, the children were each given a squash and a cupcake. They were thrilled and happily chatted with one another while enjoying their cake.

Post family visit sessions

Case Study 6

One post baby visit session was observed. Pupils were observed to be engaged throughout the session, which was primarily discussion based, with handouts to supplement this. In this school the instructor was the school receptionist, who had clear rapport with pupils.

The first topic discussed was broadly around safety. For example, '*Who is responsible for keeping you safe*?'. Participants talked through strategies to keep safe online and in real life e.g. 'not giving out too much information' and what to do when near a body of water. This was then connected to the theme of courage, with a discussion on how it can involve 'standing up in the face of adversity' and exploring the concept of 'injustice'. The instructor gave Viola Desmond, the Civil Rights Activist, as an example, and described her role in 1940/50s America. This led to a discussion about whether it was right that Viola wasn't allowed to go on mixed buses. When two pupils defended this policy of segregation, the instructor dealt with this effectively and shifted discussion to enable challenge to this opinion, with several pupils responding.
The instructor then introduced Greta Thunberg (none of the class appeared aware of who she was) and explained how she had coordinated strikes across schools to raise awareness about climate change. The instructor spoke about how she used to attend climate protests when she was younger and risked arrest because she felt it was the right thing to do. Following on from this the discussion turned to the work of Marcus Rashford – a more recognisable name to some - and his campaign in the pandemic to secure 3.5 million free meals for children who would have got them during a normal school week. This was also used to challenge the earlier expressed views on segregation – as Marcus Rashford is black, should he have only campaigned for free school meals for black children?

The discussion then turned to gender norms, with some of the boys conveying outdated views about 'men do all the work and women just stay at home and look after the baby'. Again, the instructor remained neutral, asking for wider group views and also challenging the boys as to what might happen if a woman earns more, or a man wants to look after the children. The instructor used the opportunity to link the earlier topics together and gives a historical reference to how previously women couldn't vote or have their own bank account. It should be noted that the comments made in this session led to whole school session around inclusivity and gender roles.

Overview of observations

Most of the sessions observed were felt to be quite informal in nature, with pupils sitting on the floor in front of the instructor in both pre and post visits sessions, returning to their desks to do the tasks set, before returning to the front of the classroom for the discussions. The exception was in CS7, where pupils stayed at their desks throughout the session.

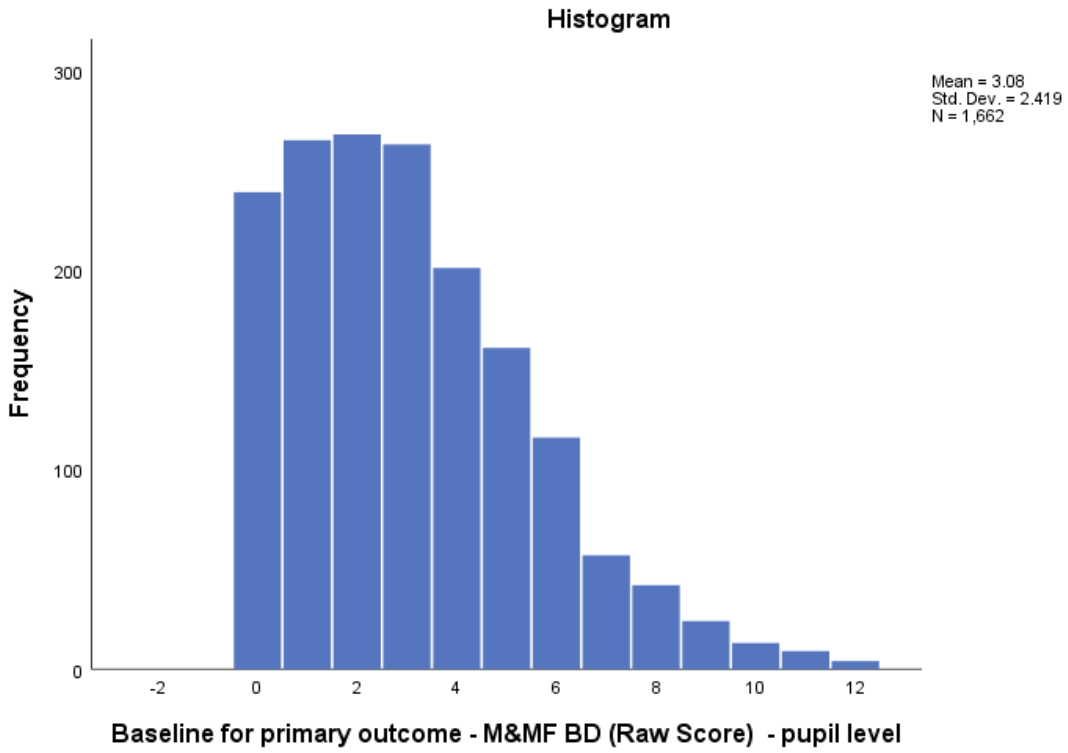
In terms of pupil engagement in sessions, this was observed to be generally good, particularly when the baby was present and while participating in discussions and tasks. Pupils were keen to ask and answer questions posed by the instructors, and to engage in discussions. While some children were less keen to engage directly with the baby – toe tickling etc. – they did appear to be watching the baby. Pupils appeared

less engaged during the reading of passages from the 'Welcoming Baby' books and when they were not actively participating in the sessions through questions or interaction with others.

Teachers were present in all bit one observed session. Where present this was in a supportive role, particularly around behaviour. The exception to this was in school RF, where, as noted above, pupil behaviour was an issue. Additionally, this session, in contrast to the others observed, felt poorly planned with no defined learning outcomes. In the other four sessions instructors, sometimes with teacher input where necessary, maintained a good level of discipline and were able to bring pupils back to the topic or task in hand. In these sessions instructors linked the learning points and topics back to the pupils' own lives and experiences, for example around tradition and communication. The instructor at CS6 was able to address contentious issues and link these to the topic under discussion.

In terms of adaptation, Welsh schools did the countdowns for songs and deep breathing in Welsh.

Appendix G. Histogram of primary outcome



Data for the primary outcome Me and My Feelings, behavioural difficulties was positively skewed.

Appendix H. Continuation of attrition-school level

The remaining variables used to examine the impact of attrition are measured at the school level. For these variables, we present distributions and summary statistics at both school and pupil levels. It is worth noting that that the measures shown in Tables Y2 to Y5 below relate to the school that the NEBT and control samples are attending and not directly to the NEBT and control pupil ITT samples. In other words, they provide an indirect perspective on the school context for the NEBT and control groups rather than a direct comparison of the randomised samples themselves. For this reason, the interpretation of these tables focuses on the school level.

Appendix table 8 presents subsamples across the five regions used to stratify the randomisation. Additionally, attrition rates across the five regions are shown.

At the school level, attrition ranged between 40% (Wales) and 67% (London) for the NEBT group and between 22% (Wales) and 60% (Yorkshire) for the control group.

Appendix Table 8: Geographic Regions for ITT samp 1Table 8: Geographic Regions for ITT samples randomised at baseline and analysed at outcome.

	ITT sample randomised		ITT sample analysed		% Attrition	
	Intervention n (%)	Control n (%)	Intervention n (%)	Control n (%)	Intervention	Control
School Level						
London	12 (26.1%)	11 (26.8%)	4 (17.4%)	6 (24%)	66.7%	45.5%
Merseyside	13 (28.3%)	12 (29.3%)	8 (32%)	8 (32%)	46.2%	33.3%
Midlands	4 (8.7%)	4 (9.8%)	2 (8.7%)	2 (8%)	50.0%	50.0%
Yorkshire	7 (15.2%)	5 (12.2%)	4 (17.4%)	2 (8%)	42.9%	60.0%
Wales	10 (21.7%)	9 (22%)	6 (26.1%)	7 (28%)	40.0%	22.2%
Total	46 (100%)	41 (100%)	23 (100%)	25 (100%)	50.0%	39.0%

At the school level, attrition resulted in an increased regional imbalance between the NEBT and control schools; most evidently in London (baseline 26.1% of NEBT schools & 26.8% of control schools, outcome; 17.4% NEBT & 24.0% control) and Wales (baseline; 21.7% NEBT & 22.0% control, outcome; 26.1% NEBT & 28.0% control).

The rest of the missing data analysis draws on data obtained from the Government school performance tables website[31]. As noted above, this data is not available for pupils in Wales and so the subsample of 299 pupils (168 in NEBT and 131 in control schools) in 19 Welsh schools (10 NEBT, 9 control) are excluded from the remaining missing data analyses. This restricts the sample to 1,363 pupils (742 in NEBT and 621 in control schools) in 68 English primary schools (36 NEBT, 32 control). For reference, the attrition rate for

31 https://www.gov.uk/school-performance-tables

the subsample of pupils in English schools was 62.6% (64.8% for pupils in NEBT schools and 59.9% for pupils in control schools).

Appendix table 9 presents school type for the NEBT and control samples randomised at baseline and analysed at outcome along with attrition rates at the pupil and school levels.

Appendix Table 926:Types of School for ITT samples randomised at baseline and analysed at outcome. Subsample of 68 English primary schools

	ITT sample randomised		ITT sample analysed		% Attrition	
	Intervention n (%)	Control n (%)	Intervention n (%)	Control n (%)	Intervention	Control
School Level						
State Maintained	12 (33.3%)	14 (43.8%)	7 (41.2%)	9 (50%)	41.7%	35.7%
Voluntary Aided/Controlled	6 (16.7%)	4 (12.5%)	3 (17.6%)	2 (11.1%)	50.0%	50.0%
Academy or Free School	18 (50%)	14 (43.8%)	7 (41.2%)	7 (38.9%)	61.1%	50.0%
Total	36 (100%)	32 (100%)	17 (100%)	18 (100%)	52.8%	43.8%

At the school level, across school types, attrition ranged between 42% (State Maintained) and 61% (Academy or Free School) for the NEBT group and between 36% (State Maintained) and 50% (Academy or Free School or Voluntary aided/controlled) for the control group

At the school level, 33% of the baseline NEBT schools were state maintained, 17% were Voluntary aided or controlled and 50% were Academies or Free Schools. This compares with the control group where 44% of schools were state maintained schools, 13% in Voluntary aided or controlled schools and 44% in Academies or Free Schools. In the analysed sample 41% of NEBT schools were state maintained, 18% were Voluntary aided or controlled and 41% in Academies or Free Schools. This compares with the control group where 50% of schools in the analysed sample were state maintained schools, 11% were Voluntary aided or controlled and 39% were Academies or Free Schools.

Appendix Table 10 presents Ofsted ratings for NEBT and control samples randomised at baseline and analysed at outcome along with attrition rates at the pupil and school levels.

Appendix Table 10 27:School Ofsted Ratings for ITT samples randomised at baseline and analysed at outcome. Subsample of 68 English primary schools

Pupils	ITT sample randomised		ITT sample analysed		% Attrition	
	Intervention n (%)	Control n (%)	Intervention n (%)	Control n (%)	Intervention	Control
School Level						
Outstanding	3 (8.3%)	3 (9.4%)	3 (17.6%)	2 (11.1%)	0.0	33.3
Good	28 (77.8%)	27 (84.4%)	12 (70.6%)	16 (88.9%)	57.1	40.7

Requires Improvement	5 (13.9%)	2 (6.3%)	2 (11.8%)	0 (0%)	60.0	100.0
Total	36 (100%)	32 (100%)	17 (100%)	18 (100%)	52.8	43.8

Attrition increased with declining school Ofsted ratings. Whilst five of the six schools rated as 'Outstanding' were maintained in the analysed sample, this was only the case for two of the seven schools rated as 'Requires Improvement'. In the Control group, all three schools rated as 'Requires Improvement' were lost to attrition. Three schools with 'Requires Improvement' ratings were also lost to attrition in the NEBT group but two were maintained in the analysed sample. Following attrition, the NEBT group includes schools that were predominantly classed as 'good' (71%) with some 'outstanding' (18%) and some 'requires improvement' (12%). This compares with the control group which includes schools that were predominantly classed as 'good' (89%) with some 'outstanding' (11%) but none classed as 'requires improvement'.

Appendix Table 11 presents the school-level means for KS2 attainment (in reading; grammar punctuation and spelling & maths) and for five school level pupil context variables (school size, % with SEND EHC plan, % with SEND support, % with English as an additional language and %FSM in last six years) for NEBT and control samples randomised at baseline and analysed at outcome.

To help examine the relationship between attrition and the balance between the NEBT and control groups, effect size statistics are shown which estimate the standardised difference between the two groups randomised at baseline and analysed at outcome.

Appendix Table 11 28: School level KS2 attainment & pupil contexts for ITT samples randomised at baseline and analysed at outcome.

	NEBT Intervention		Control		All		Effect Size
	n	Mean (95% CI)	n	Mean (95% CI)	n	Mean (95% CI)	(NEBT-Control)/sd
KS2 Reading (READ_AVERAGE)							
At Baseline	36	103.9 (103.1;104.7)	32	104.1 (103.3;104.9)	68	104 (103.4;104.6)	-0.10 sds
At Outcome	17	103.9 (102.7;105.1)	18	104.1 (103.0;105.1)	35	104 (103.2;104.8)	-0.07 sds
KS2 Grammar, Punctuation and Spelling (GPS_AVERAGE)							
At Baseline	36	103.3 (102.2;104.4)	32	104.4 (103.4;105.4)	68	103.8 (103.1;104.6)	-0.35 sds
At Outcome	17	103.2 (101.6;104.9)	18	104.2 (102.8;105.7)	35	103.7 (102.7;104.8)	-0.30 sds
KS2 Maths (MAT_AVERAGE)							
At Baseline	36	103.2 (102.3;104)	32	103.5 (102.5;104.5)	68	103.3 (102.7;104.0)	-0.11 sds
At Outcome	17	102.8 (101.5;104.2)	18	103.2 (102;104.4)	35	103.0 (102.1;103.9)	-0.15 sds
School Size (TOTPUPS)							
At Baseline	36	356.1 (295.6;416.7)	32	371.8 (296.0;447.7)	68	363.5 (315.9;411.2)	-0.08 sds
At Outcome	17	366.4 (276.0;456.7)	18	372.4 (290.8;454.1)	35	369.5 (309.6;429.3)	-0.03 sds
% pupils with a SEND EHC plan (PSENELSEN)							
At Baseline	36	3.9 (2.8;5.0)	32	3.9 (2.1;5.7)	68	3.9 (2.8;4.9)	-0.01 sds
At Outcome	17	2.3 (1.7;3.0)	18	2.9 (1.8;3.9)	35	2.6 (2.0;3.2)	-0.30 sds
% pupils with a SEND support (PSENELKN)							

At Baseline	36	19.4 (17.6;21.2)	32	18.2 (15.3;21.1)	68	18.8 (17.2;20.5)	0.17 sds
At Outcome	17	19.1 (16.5;21.7)	18	19.3 (15.2;23.5)	35	19.2 (16.8;21.7)	-0.03 sds
% pupils where English not first language (PNUMEAL)							
At Baseline	36	34.6 (26.2;43.0)	32	31.4 (23.6;39.2)	68	33.1 (27.4;38.8)	0.13 sds
At Outcome	17	29.9 (18.8;41.0)	18	29.4 (18.1;40.6)	35	29.6 (21.8;37.4)	0.02 sds
% pupils classed as FSM at any time during the past 6 years (PNUMFSMEVER)							
At Baseline	36	45.5 (41.0;50.0)	32	42.0 (37.2;46.9)	68	43.9 (40.5;47.2)	0.25 sds
At Outcome	17	44.7 (39.0;50.5)	18	43.6 (36.1;51.1)	35	44.2 (39.5;48.8)	0.08 sds

Attrition is observed to result in greater balance in school-level KS2 attainment between the NEBT and control groups. This is illustrated by the smaller effect sizes for the sample analysed at outcome compared with the baseline sample. The one exception was KS2 maths attainment where the effect size at baseline (-0.11 sds) increased slightly in the sample analysed (-0.15 sds).

In terms of school size, on average, NEBT schools were smaller than control schools at baseline and outcome. In terms of effect sizes, the difference between the two groups reduced from -0.08 sds at baseline to -0.03 sds at outcome.

The percentage of pupils that had SEND support was similar for the NEBT and control school samples. A greater difference between the two groups is observed at baseline +0.17) compared with the sample analysed at outcome (-0.03 sds). However, some imbalance is observed with the percentage of pupils with a SEND EHC plan. At baseline, the difference between the two groups was relatively small -0.01 sds) but the difference is seen to increase with the sample analysed at outcome -0.30 sds). This illustrates that, in the analysed sample, the schools that the control group attended had a higher percentage of pupils with a SEND EHC plan (2.9%) compared with the schools that the NEBT group attended (2.3%).

In terms of the percentage of pupils who do not have English as their first language, at baseline and outcome, the schools that the NEBT group attended had a higher concentration of EAL pupils compared with the schools that the control group attended; but this difference was reduced with attrition (effect size =+0.13 sds at baseline and +0.02 sds at outcome).

In terms of the percentage of pupils classed as eligible and claiming Free School Meals (FSM) in the last six years, a higher concentration is seen with schools that the NEBT group attended compared with the control group schools at baseline (+0.25 sds). However, this difference is seen to reduce with in the sample analysed at outcome (+0.08 sds).

Appendix Table 2912: Summary of baseline & outcome sample sizes across five measures

Outcome Measure	Intervention Baseline / Outcome (% lost to Attrition)	Control Baseline / Outcome (% lost to Attrition)	Total Baseline / Outcome (% lost to Attrition)
M&MF BD (Primary Outcome)	910 / 314 (66%)	752 / 330 (56%)	1662 / 644 (61%)
M&MF ED (Secondary Outcome)	908 / 307 (66%)	736 / 314 (57%)	1644 / 621 (62%)

BES Cog Empathy (Secondary Outcome)	871 / 286 (67%)	729 / 305 (58%)	1600 / 591 (63%)
BES Aff Empathy (Secondary Outcome)	868 / 277 (68%)	713 / 295 (59%)	1581 / 572 (64%)
Teacher-SDQ Total Difficulties (Secondary Outcome)	711 / 207 (71%)	513 / 249 (51%)	1224 / 456 (63%)
All four pupil-level measures	782 / 239 (69%)	635 / 242 (62%)	1417 / 481 (66%)
All four pupil-level measures plus teacher SDQ	513 / 99 (81%)	367 / 100 (73%)	880 / 199 (77%)

The complete missing data analysis is restricted to the primary outcome – but balance relating to the baseline versions of all outcomes is summarised in the Table below

Appendix table 13 30: Baseline Means & Effect sizes for NEBT & Control Groups

	Complete Baseline ITT Sample			Subsample with Baseline and Outcome Scores		
Outcome Measure	Intervention Mean (95% CIs)	Control Mean (95% CIs)	Mean Difference as Effect Size	Intervention Mean (95% CIs)	Control Mean (95% CIs)	Mean Difference as Effect Size
M&MF BD	3.04 (2.88;3.2)	3.13 (2.96;3.3)	-0.02	2.62 (2.38; 2.86)	2.92 (2.67;3.17)	-0.13
M&MF ED	7.19 (6.94;7.44)	7.4 (7.15;7.65)	-0.02	7.16 (6.76; 7.56)	7.58 (7.20; 7.96)	-0.12
BES Cog Empathy	33.4 (33.0; 33.7)	33.5 (33.1;33.9)	-0.04	34.1 (33.5; 34.7)	33.8 (33.2; 34.4)	+0.05
BES Aff Empathy	34.8 (34.4; 35.3)	34.8 (34.3; 35.3)	+0.03	34.9 (34.1; 35.7)	35.0 (34.3; 35.8)	-0.02
Teacher-SDQ TD	7.9 (7.35; 8.37)	8.1 (7.54; 8.72)	-0.04	6.89 (5.97;7.81)	7.65 (6.84;8.46)	-0.11
Teacher-SDQ EP	1.9 (1.73; 2.07)	2.1 (1.91; 2.35)	-0.12	1.75 (1.45;2.05)	1.98 (1.67;2.29)	-0.10
Teacher-SDQ CP	1.1 (0.98; 1.26)	1.3 (1.14; 1.48)	-0.07	0.88 (0.64;1.12)	1.27 (1.04;1.5)	-0.21
Teacher-SDQ Hyp	3.6 (3.37; 3.83)	3.3 (3.06; 3.58)	+0.14	2.97 (2.55;3.39)	3.23 (2.84;3.62)	-0.08
Teacher-SDQ PP	1.5 (1.35; 1.61)	1.4 (1.23; 1.53)	+0.01	1.27 (1.03;1.51)	1.22 (1.01;1.43)	+0.03
Teacher-SDQ PS	7.0 (6.82; 7.18)	7.4 (7.17; 7.59)	-0.10	7.41 (7.07;7.75)	7.69 (7.4;7.98)	-0.12

Appendix I. YEF Security Rating

Rating	Design	MDES Outcome: Threshold*	Attrition	Initial score		Adjustments		Final score
5	Randomised design	Offending: <=0.1 SDQ tot: <= 0.3 Other: <= 0.2	0-10%	**0**				
4	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	Offending: 0.11 – 0.19 SDQ tot: 0.31 – 0.39 Other: 0.21 – 0.29	11-20%					
3	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	Offending: 0.2 – 0.29 SDQ tot: 0.4 – 0.49 Other: 0.3 – 0.39	21-30%					
2	Design for comparison that considers selection only on some relevant confounders	Offending: 0.3 – 0.39 SDQ tot: 0.5 – 0.59 Other: 0.4 – 0.49	31-40%					

1	Design for comparison that does not consider selection on any relevant confounders	Offending: 0.4 – 0.49 SDQ tot: 0.6 – 0.69 Other: 0.5 – 0.59	41-50%					
0	No comparator	Offending: >= 0.5 SDQ tot: >= 0.7 Other: >= 0.6	>50%					**0**

*MDES requirements vary by outcome measurement. Offending: Offending data collected through self-report or admin data; SDQ tot = SDQ total difficulties score; Other: all other outcomes, incl. SDQ externalising and internalising