



Education
Endowment
Foundation

Healthy Minds: Health Outcomes

Evaluation report and executive summary

March 2019

Independent evaluators:

Dr Grace Lordan and Professor Alistair McGuire





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

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

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

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus Trust (now part of Impetus - Private Equity Foundation) and received a founding £125m grant from the Department for Education.

Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.



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

Jonathan Kay
Research and Publications Manager

Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP
p: 020 7802 1653
e: jonathan.kay@eefoundation.org.uk
w: www.educationendowmentfoundation.org.uk

About the evaluator

The evaluation of the programme was led by the London School of Economics and Political Economy through the Centre for Economic Performance and LSE Health. Professor Alistair McGuire, Professor in Health Economics at the LSE, led the evaluation and along with Dr Grace Lordan undertook the statistical analysis. Professor McGuire has an international reputation in health economics, acted as an advisor to numerous UK government offices and research councils and been involved in a number of major clinical trials. Dr Grace Lordan is an Associate Professor in Behavioural Science at the LSE, with extensive econometric and programme evaluation knowledge.

Bounce Forward, a charity dedicated to teaching resilience skills managed the project. Lucy Bailey, at Bounce Forward, oversaw the project as a whole, including course development and the training of teachers, as well as all practical aspects of data collection and processing, and monitoring, guidance and support for the schools, but was not part of the analytical team. Healthcare Solutions, in cooperation with Bounce Forward, provided the logistics of gaining access to the participating schools, issuing of questionnaires and liaising with principal teachers.

The implementation and evaluation phases of the project were overseen by an advisory group, Chaired by Lord (Professor) Richard Layard, LSE.

Contents

Executive summary.....	4
Introduction	6
Methods	10
Implementation and process evaluation	33
Conclusion.....	34
References	37
Appendix A: EEF cost rating.....	40
Appendix B: Security classification of trial findings.....	41
Appendix C: Effect size estimation for Additional Secondary Outcomes.....	42

Executive summary

The project

The Healthy Minds (HM) course aims to improve health related outcomes for teenagers.). The course was made up of 14 modules (totaling 113 hours), based on existing evidence or guidance on health education, covering a range of topics including: social and emotional learning, relationships and healthy living content suitable for students in UK secondary schools. It was delivered to classes over the first 4 years of secondary school (when pupils are aged 11 to 15 years old). Lessons either replaced the one hour-a-week of PSHE timetabled lessons, or were built in to the school week at other times, and were taught by school staff (teachers or learning support assistants, who received full training in each element).

The trial started in schools in September 2013 and ended in July 2018. Thirty-four schools were recruited over two phases, 13 in 2013 and a second group of 21 in 2014. Five of the control schools from phase one of the trial were assigned to receive the intervention in the second phase, meaning that there was a total of 39 cohorts.

This evaluation focused on health outcomes as measured by the Child Health Questionnaire-CF87 (CHQ-CF87) (Schmidt, Garrett and Fitzpatrick, 2002). The primary outcome measure was the single scale of self-assessed general health drawn from the CHQ-CF87. This instrument also contains twelve other scales which are evaluated as part of this study as secondary outcomes, alongside other validated scales (the Short Mood and Feelings Questionnaire, the life satisfaction 0-10 ladder and the Child Anxiety Related Disorders (SCARED)). Outcomes were measured after both two years and four years of the programme. A second evaluation conducted by a team at NIESR includes both an impact evaluation looking at academic outcomes and an Implementation and Process Evaluation. This will be published in 2020.

HM was developed and delivered by the charity Bounce Forwards. This study was funded by the LSE, Rosetree (a charitable organisation), Hertfordshire Public Health, and the Education Endowment Foundation (EEF).

Key conclusions

1. Pupils in schools that received the Healthy Minds programme had higher average self-assessed general health (0.25 standard deviations) compared to similar pupils in other schools after four years. This finding has moderate to high security.
2. The evaluation also measured pupils' self-assessed general health after two years of the programme and found a similar impact (0.23 standard deviation difference).
3. Secondary outcomes associated with physical health, behaviour and external relations were generally positive.
4. Measures of internalised emotions (Emotional Difficulties, Self-Esteem, and Mental Health) were mainly close to zero and positive after four years of the programme, but in some cases were negative when collected after two years of the programme.

EEF security rating

The primary outcome has a moderate to high security rating. The trial was an efficacy trial, which tested whether the intervention worked under developer-led conditions in a number of schools. The trial was a well-designed randomised controlled trial, however 23% of the pupils who started the trial were not included in the final analysis because outcome data for them was not provided.

Additional findings


The study showed the HM course had a positive impact on the primary HRQoL outcome. Self-assessed general health score, was raised by approximately 0.25 standard deviations in the treatment group compared to the control group and this result was statistically significant. The interim result, after two years of the programme indicated that almost all the gain had already been achieved. These results could suggest that only two years of the programme are required. However, an alternative interpretation is that four years are needed to maintain the impacts. To fully understand this we would need to randomise some children to receive the programme for two years and some for four years.

Positive outcomes were also seen on the majority of secondary outcomes. The exceptions to this were the variables that capture internalising behaviour (Emotional Difficulties, Self-Esteem, and Mental Health), with negative impacts seen at the two year measurement point, although these revert to close to zero by the end of the programme.

Cost

Costs of providing training and resources to deliver HM within the existing PHSE timetabled slots are £23.50 per pupil per year.

Table 1: Summary of impact on primary outcome

Outcome/ Group	Effect size (95% confidence Interval)	P value	No. of pupils	EEF security rating	EEF cost rating
Global Health score	0.25 (0.019, 0.471)	0.035	7,362		£££££

Introduction

Background evidence

It is now well accepted that health related quality of life (HRQoL) in childhood, the multi-dimensional concept that includes domains related to physical, mental, emotional, and social functioning, has a long arm into adulthood. For example, there are many influential papers which highlight that childhood health significantly predicts adult labour market outcomes (Case *et al.*, 2005; Black *et al.*, 2007; Smith, 2009; Currie *et al.*, 2009; Currie, 2009, Case and Paxson, 2011; and Case and Paxson, 2010). There is also evidence that poor mental health correlates with long-term negative impacts into adulthood. For example, there is evidence that poor mental health impacts on the ability to work and earn as adults (Goodman *et al.*, 2011), educational attainment Gibb *et al.* (2012) and long run psychological disturbance (Collishaw *et al.*, 2004; Thapar *et al.*, 2012). Dimensions of HRQoL, such as behaviour and self-esteem, overlap with non-cognitive skills. We define non-cognitive skills as in Kautz, Heckman, Diris *et al.* (2014), as attributes which are not measured by IQ or achievement tests. There is a growing literature which underlines the importance of non-cognitive skills as measured in childhood on later life outcomes. For example, Heckman, Pinto, and Savelyev (2013), Heckman *et al.* (2011) and Lleras (2008) show that proxies for non-cognitive skills in childhood are strong predictors of a variety of adult outcomes, including educational attainment, labor market outcomes, and health. Overall there is ample evidence of a long arm for childhood and adolescent HRQoL (Heckman, Humphries and Veramendi, 2014).

Schools provide a major opportunity for a public health intervention aimed at improving aspects of HRQoL. While such interventions could crowd out traditional academic achievement, we know of no study that puts forth evidence that this is the case. Given the importance of HRQoL in determining later labor market outcomes, there are independent reasons as to why time should be set aside in schools to hone non-cognitive skills. Pushing this debate to one side, we note that there are growing examples of programs being rolled out in early childhood at the school level whose aim is to enhance certain aspects of HRQoL. Encouragingly, evidence that programs like this change later life outcomes for the better has also emerged. See for example evidence in favor of the Perry Preschool program in Heckman, Pinto, and Savelyev (2013) and The Abecedarian Program in Campbell, Conti, Heckman, Moon, and Pinto (2013). Both programs highlight differential effects by gender and are targeted at young children. For programs rolled out once children hit primary school the evidence suggests that programs succeed in their goals (see Durlak, Weissberg, Dymnicki, Taylor, and Schellinger (2011) for a meta-analysis of 213 school-based social and emotional learning programs), with studies with a longer follow up having a mean impact that is positive and statistically significant. Two things are worth noting about the received evidence. First, it is mainly US-based. Second, there are sparse examples of studies which roll out a course during core teaching hours for adolescent children in secondary school. We address this gap. That is, we provide evidence that a four-year program, constructed to augment HRQoL in adolescents, rolled out in secondary schools in the UK fulfilled its objectives. Encouragingly, Heckman and Kautz (2013) provide a compelling argument backed by empirical evidence that aspects of HRQoL are skills which can be improved throughout the life course. That is, it is possible to change the HRQoL of the treated regardless of their starting point.

Recently within the UK there has been an acknowledgement that personal, social, health and economic education (PSHE) at school may be a means to provide young people with the skills to become more self-aware, resilient to negative peer-pressures and to make more informed life-choices (House of Commons, 2015). In theory, if effective, these lessons may address an imbalance attributable to a poor home or family background. Heckman and Kutz (2014) and Kautz, Heckman, Diris *et al.* (2014) provide a general literature review of this area. They outline mechanisms and some empirical support to suggest that improvement in non-cognitive skills can improve educational achievement. However, they also highlight that there is a dearth of evidence relating to the impact of intervention programmes on adolescents. Moreover, they highlight that adolescent-based programmes tend to measure few

outcome dimensions and also focus on traditional educational outcomes (attainment) and employment success. For the American school environment, they do highlight, however, that the most successful adolescent programmes promoting non-cognitive skills integrate the course into traditional education.

It still remains unclear what schools should do to provide effective PHSE. The UK government has recently highlighted that the quality of PHSE is sub-optimal and that teaching in this area requires improvement in 40% of schools. There is a current investigation into how to improve the curriculum in this area (House of Commons Education Committee. Life lessons: PHSE and SRE in schools. Fifth Report of Session 2014-15. HC 145. London: HMSO). One of the recommendations from the House of Commons recent report was for PHSE to become statutory. Yet there was little evidence provided on what any statutory content should be provided within the curriculum.

Some programmes that develop PHSE type skills and knowledge have been scientifically evaluated and have been found to augment emotional wellbeing, behaviour and academic performance. In the previously mentioned meta-analysis of primary school-based programmes, Durlak et al (2011) found that the typical programme raised outcomes on social and emotional learning by around 11 percentile points. Two variables moderated positive outcomes: how well conducted the programmes were (absence of implementation problems); and how well-designed and integrated they were. This latter point is especially important as short-term, non-integrated PHSE type teaching is prone to fading effects (Bond and Hauf, 2004; Challen et al, 2011; Brunwasser et al, 2009).

It remains unclear what an integrated course, able to cover a range of the dimensions required by PHSE teaching, is capable of delivering in terms of achievable outcomes. A literature review of evaluated PHSE type programmes was undertaken to identify individual teaching modules within this area that had proven efficacy, would be able to be combined into an integrated PHSE course feasible to use in UK state schools (Coleman et al, 2011). This review identified 14 individual modules of proven effectiveness that combined into an integrated teaching package covering fundamental topics that could be offered through an integrated, statutory PHSE course. The aim of this study was to empirically evaluate this school-based intervention.

Intervention

This study is evaluating whether an evidence-based life skills course, Healthy Minds (HM), within PSHE curriculum over 4 years in secondary schools, can improve teenagers' well-being and non-cognitive skills, and improve their resilience.

The primary aim of the evaluation is to establish whether HM can improve teenagers' health-related quality of life (HRQoL). For this study HRQoL captures elements of a child's health and soft skills, as well as specific aspects relating to the child's family life. This a unique non-assessed (in terms of in school testing) study. The study draws together 14 modules into a cohesive program with respect to enhancing specific aspects of a child's health related quality of life. The items in this 14-module package, have been separately evaluated through various controlled trials and studies to be successful in similar audiences. These items are:

- Penn Resilience Programme
- Breathe (Mindfulness)
- Media Navigator
- From School to Life
- Unplugged (Part 1 and 2)
- Media Influences

- Resilience Revisited
- Sex Ed Sorted (Part 1 and 2)
- Relationship Smarts Plus
- School Health Alcohol Harm Reduction Programme (SHAHRP)
- Resilient Decisions
- Mental Illness Investigated
- Parents Under Construction
- Resilient Learners

Full details of each of the individual modules, and their evaluations, are described in Coleman *et al.* (2011). On the basis of the Coleman *et al.* (2011) review these modules were integrated to form a comprehensive course taught to pupils as a trial intervention in UK schools. This course was taught as a 113-hour universal programme delivered over the first 4 years of secondary school using one hour-a-week of timetabled lessons (replacing whatever non-standardized PSHE that had been historically timetabled for the same cohort) and taught by school staff who received full training in each module. The training covered 7 days of teacher training for Year 7 teachers; 6 days of training in Year 8; 2 day in Year 9; and 4 days in Year 10. Training covered all aspects of the 14 elements, with a core theme of resilience throughout. It specifically focused on the Penn Resilience Programme, Media Navigation and Breath (a mindfulness programme) in Year 7; From School to Life (a life skills programme), Unplugged (a substance abuse/misuse programme), Media Influences, Sex Education (part 1), Relationships, and a resilience reflection (“Review and Connect”) programme in Year 8; Relationships, Alcohol misuse, Sex Education (Part 2), and Resilience (“Resilient Decisions”) in Year 9; Mental Illness, Substance Misuse, Relationships (“Parents under Construction”) and Resilience in exams in Year 10. This training was provided by Bounce Forward and covered both the material, and appropriate teaching methods for the HM course.

Evaluation objectives

The study will evaluate whether HM, provided within the PSHE timeslot over a 4-year period in secondary schools, improved the HRQoL of those who received it. The primary purpose of the trial is to assess the 14-module course and training package as a whole. That is, we aimed to quantify the effect on teenagers’ HRQoL exposed to HM, as compared to those that continued with non-standardised PSHE offerings. The primary research question addressed is: Whether the programme improves pupils’ HRQoL, measured by the Child Health Questionnaire (CHQ-87). However, the study also examined whether the programme improved moods and feelings as measured through the Short Mood and Feelings Questionnaire (Angold and Costello, 1987), life satisfaction and on mental health as related to anxiety-type disorders as measured through the Child Anxiety Related Disorders (SCARED) questionnaire (Birmaher et al, 1999). Improvement is assessed statistically after correcting the standard errors for considering multiple outcomes.

Ethics and trial registration

Ethical review was undertaken through the LSE Ethics Committee. As data were anonymised at collection there was general support for the study. As well as the information sheet, the recruited schools were then provided with letters to be sent to individual parents asking for permission to be recruited into the study in the form of a decision to not participate. This letter detailed the objectives of the study and the anonymised nature of the data.

Data protection

Data were collected and coded by an independent data collection team, (an independent firm, Healthcare Solutions), with the coding using a unique (anonymised) pupil identifier ensuring pupil anonymity but retaining linkage within a longitudinal data set. Pupil names were not retained, and an anonymised data set was released to the statistical analysts at the end of June 2018.

The General Data Protection Regulations and the Data Protection Bill came into operation within the timescale of the study (25 May 2018). As individual parents had already been informed of the purposes of the study and given the possibility to withdraw, and as the data was held in anonymised form there was no further action required to become compliant with these regulatory requirements provided that we did not use the data for analysis other than agreed (i.e. assessing Healthy Minds on its recipients HRQoL). No teachers, parents or pupils had any data released. The LSE employed an intermediate company, Healthcare Solutions, to code and anonymise the data and all names were subsequently omitted from the data given to the LSE analysts.

Project team

The evaluation of the programme was led by Professor Alistair McGuire, Professor in Health Economics at the LSE. Professor McGuire has an international reputation in health economics, acted as an advisor to numerous UK government offices and research councils and been involved in a number of major clinical trials. Dr Grace Lordan, an Associate Professor in Behavioural Science at the LSE, with extensive econometric and programme evaluation knowledge led the statistical analysis.

Lucy Bailey at Bounce Forward was project manager, but not part of the analytical team, overseeing the project as a whole, including course development and the training of teachers, as well as all practical aspects of data collection and processing, and monitoring, guidance and support for the schools. She operated closely with Healthcare Solutions in the logistics of gaining access to the participating schools, issuing of questionnaires and liaising with principal teachers.

The implementation and evaluation phases of the project were overseen by an advisory group, Chaired by Lord (Professor) Richard Layard, LSE.

Methods

Trial design

Table 2: Trial information

Trial type and number of arms		Two-arm cluster randomised trial
Unit of randomisation		School
Minimisations variable(s) (if applicable)		Percentage of FSM pupils, percentage of pupils with GCSE grades A*-C, and single sex or mixed school
Primary outcome	Variable	Improving Health Related Quality of Life (HRQoL)
	measure (instrument, scale)	General health dimension score from the CHQ-CF87.
Secondary outcome(s)	variable(s)	Improving Health Related Quality of Life (HRQoL)
	measure(s) (instrument, scale)	Twelve sub-scales from the CHQ-CF87 scale. These sub-scales capture dimensions of HRQoL that represent aspects of physical health, emotional wellbeing and behaviour. We also assess the impact on other instruments which capture other aspects of HRQoL. These are the Short Mood and Feelings Questionnaire, the life satisfaction ladder and the Child Anxiety Related Disorders questionnaire (SCARED).

The study is based on a cluster randomised trial, with school level randomisation. Randomisation was conducted using minimisation and schools were identified according to whether the percentage of pupils eligible for Free School Meals (FSM) is less than 13 per cent, between 13 and 25 per cent or greater than 25%; whether the percentage of pupils with 5 GCSEs with grades A*-C is below 59 per cent or not; and whether the school is single sex or mixed. These criteria were used to aid identification of schools which matched our original intention of recruiting schools with poor attainment in above-average areas of deprivation. In the end recruitment encompassed a pragmatic element as school opt-in proved difficult.

The 4-year trial in schools began in 2013-2014. In an effort to minimise drop-out in school recruitment the control schools were initially based on wait-list control, where they would then be offered the course/treatment to subsequent pupil cohorts to offer an incentive for engagement. Given the timing of initiation of the study (in the middle of a school year) and the length of time of engagement with schools, the actual recruitment took place in two phases with a first wave initiating the intervention in September 2013, including a smaller than intended number of wait list control schools and a second wave initiating the intervention or providing a (straightforward) control year group from September 2014.

Table 3: School cohorts by study stage

Number of School Cohorts in Each Study Stage			
Study Time	Time =0 (Baseline)	Time=2 (Interim)	Time =4 (Endline)
Phase 1 (2013)			
Treatment	7	5	7
Wait List Control	6	4	6
Phase 2 (2014)			
Wait List Treatment	5	4	4
Treatment	11	6	11
Control	10	6	7

Table 3 details the number of school cohorts who were part of the project and their classification in each phase. In 2013 (Phase 1), 13 schools were recruited with 6 allocated to the (wait list) control arm and 7 to the treatment arm. Schools allocated to the (wait list) control arm were due to start treatment in 2014. In 2014 (Phase 2), 21 schools were recruited, with 10 allocated to the control arm and 11 schools to the treatment arm. This gave a total of 34 schools, but 39 school-cohorts. The data collection questionnaire was administered at baseline (t=0; either 2013 or 2014 depending on when schools entered the trial), at an interim point (t=2; two years after baseline), and at endline (t=4, two years after the interim administration; either 2017 or 2018 depending on timing of entry for schools).

The 39 school-cohorts reflect the design of the study, which in the first year of recruitment included 6 wait-list control schools in the first cohort (2013). These wait-list control schools were meant to progress to treatment schools, using the following year's (2014) entrance cohort of pupils. So, 34 schools and 39 school-cohorts formed the basis of the analysis. After recruitment and retention problems over the course of the study, 35 school-cohorts were included in the final analysis and 25 school-cohorts in the interim analysis. The larger number of school-cohorts in the final analysis reflecting a claw-back mechanism that attempted to collect data for all the 39 school-cohorts in the final administration of the questionnaire, with only 35 responding. More details are given in the Participant Flow Diagram (Figure 3 below).

Participant selection

To initiate recruitment for the study a list of all state maintained secondary schools in 42 local authorities in the South Eastern region of England was compiled from national records (EduBase – the database of all educational establishments in England and Wales, <http://www.education.gov.uk/edubase/about.xhtml>). The aim was to recruit schools with poor attainment serving pupils with above-average levels of deprivation. All 751 English schools were therefore assigned a score of 1-10 based on the decile in which they fell for each of: percentage of pupils making expected progress in English; percentage of pupils making expected progress in mathematics; percentage of pupils gaining at least 5 GCSEs at C or better including English and mathematics; and the percentage of pupils eligible for free school meals, based on 2012 GCSE and school census data from the Department for Education. A school scoring 40 was thus in the lowest (worst) decile for

progress and attainment at GCSE, and in the highest decile for the percentage of pupils eligible for free school meals. Excluding schools with missing data and those which were already involved in other interventions, this left 174 schools scoring 22 and above, who were invited to participate by letter. Schools expressing interest were sent a project information sheet, stating the requirements of the project and evaluation. Schools expressing interest amounted to 42, and after drop-out, the final number of schools willing to participate was 37, with a high representation from the South East of England and the Midlands (4 of the schools were from the Wolverhampton area).

The intention was to recruit all 37 schools. However, as noted above, there was school drop-out which began and continued throughout the recruitment phase. Recruitment generally proved difficult as study recruitment began late in the annual school planning cycle and it proved difficult to retain schools for a complex 4-year study involving a regular slot in their timetable for 'soft skills', from the beginning of the study. As 3 schools dropped out during the recruitment phase the study eventually recruited 13 participating schools in Phase 1 (2013), with 6 allocated to the (wait list) control arm and 7 to the treatment arm, and 21 participating schools in Phase 2 (2014), with 10 allocated to the control arm and 11 schools to the treatment arm. This gave a total of 34 schools, and 39 school-cohorts forming the baseline participants.

Retention problems led to further drop-out over the course of the study. Over time some schools were unable to maintain the teaching commitment or were unable to provide support for questionnaire administration. Interim data collection was therefore completed for 25 school cohorts only. A claw-back (re-engagement) mechanism was initiated for final data collection, where schools which had dropped out over the study period were contacted and asked if they were willing to participate in final data collection, and 35 school-cohorts were subsequently included in the final analysis.¹ We note that those classified as forming the treatment group subsequently did not necessarily administer the course in its entirety. Analysis is therefore an intention to treat design (however, our robustness does consider compliance). See Table 3 for details of the school cohorts recovered in each stage of the study.

Data collection was carried out through questionnaires issued to individuals and conducted on school-sites at baseline (September 2013 or 2014), 21 months (June 2015 or 2016) and 42 months (June 2017 with the final questionnaires delivered during 2018). Individual questionnaires were completed under standard exam conditions within the individual schools, with participants informed at the start of the session that the survey data would be collated anonymously, and that parents, teachers or other pupils would not have any access to the data.

Sample size

The average English school has approximately 150 students per year, however in order to allow for absentees and students leaving the school over the course of the trial we based our calculations on 100 per year group. We apply conventional statistical significance of 0.05 and power of 0.80, and given that this is a HRQoL study, we assume intra-class correlation (ICCs) to be 0.06, as ICCs were reported to lie between 0.03 and 0.06 for a range of earlier comparable studies (Challen *et al*, 2011, UK Resilience programme evaluation: final report.). Based on these figures, and equal numbers of treatment and control schools, a sample size of 25 schools is required to detect an effect size of 0.3 standard deviations. This effect size is consistent with estimated standardised mean difference found in a number of studies of school interventions supported by mindfulness programmes, which are similar to a sub-set of the interventions proposed by HM, as well as cognitive behavioural interventions, which we might expect to have at least as great an impact as the HM programme, assessed here. Hattie, in a number of studies undertaking meta-analyses of proven effect sizes of various interventions in schools documents these effect size findings and sets them within a wider context of school interventions (see e.g. Hattie, 2011; 2015; 2018). To allow for drop-out of schools over the four-year period of follow-up,

¹ There was an additional special-needs school which participated in the study, but as it did not meet the inclusion criteria it was excluded from the analysis for this report.

pupil attrition and parental consent withdrawal we based sample size calculations on the recruitment of 30 schools, which would allow detection of an effect size of 0.28 change in standard deviation.

As noted above, 42 schools expressed interest and 37 schools initially agreed to participate, but the study faced school recruitment difficulties from initiation. It proved difficult to recruit schools for a complex 4-year study involving a regular slot in their timetable for HRQoL. Within the two recruitment phases and including the wait-list schools from the first phase, a total of 40 school-cohorts from 34 schools agreed to participation in the study. These are the participating schools represented in the data for this study and could be considered “as randomised”. Based on this number of schools, (schools rather than school cohorts to allow conservative estimation), the Minimum Detectable Effect Size (MDES) was calculated to be 0.28.

EEF Statistical Analysis Guidance suggests conducting a sub-group analysis for FSM pupils and including MDES calculations for this sub-group. As FSM identifiers were not available for the evaluation team these sub-group analyses and the accompanying MDES calculation are not reported.

Randomisation

Randomisation was conducted using minimisation given the predicted small sample size, such that the incremental allocation of individual schools was based on specific characteristics of schools, in an attempt to pursue our objective of recruiting schools with poor attainment and where pupils were drawn from above-average levels of deprivation.

Minimisation was based according to whether the percentage of pupils eligible for Free School Meals (FSM) was less than 13%, between 13 and 25% or greater than 25%; whether the percentage of pupils with 5 GCSEs with grades A*-C was below 59% or not; and whether the school is single sex or mixed. Randomisation was prepared and undertaken by Amy Challen, a Research Officer at LSE (CEP), who left LSE prior to data collection and analysis, which in the event allowed randomisation to be undertaken independent from the research analyst team for the Phase 1 schools. Phase 2 schools were randomised using the same basis by a member of the analytical team (Dr Grace Lordan). In all cases schools were allocated a unique identifier and the actual process of randomisation undertaken through the use of a random number generator routine in Excel, with schools randomly allocated to 0 (control) or 1 (treatment), so that the randomisation process mimicked the flipping of a coin.

There was an objective of minimising imbalance across the treatment and control schools using these characteristics. Given the difficulties with and the phasing of recruitment, explained in detail above, no balance analysis was undertaken at initiation.

Outcome measures

PHSE, as described by Ofsted, is aimed at delivering a planned programme of study to allow young people to acquire “the knowledge, understanding and skills they need to manage their lives”. Our study attempts to influence the PHSE curriculum through providing an integrated course to improve adolescent’s HRQoL. The primary outcome used was the change in the General Health single item scale embedded in the Child Health Questionnaire (CHQ-CF87). This item is a single measure of self-reported health which forms part of the overall CHQ-CF87 questionnaire.

The CHQ-CF87 is specifically designed for young people aged 10 to 18 (CHQ, 2013). The questionnaire is based on 87 items that measure physical and psychosocial health, divided across 14 multi-item scales on physical functioning, social-emotional role, social-behavioural role, social-physical role, pain, general behaviour, mental health, self-esteem, general health perceptions and family activities. That is, the questionnaire provides data on a child’s health and soft skills, as well as specific aspects relating to the child’s family life.

The CHQ-CF87 has been found to be reliable and sensitive within 10-18 year olds (Schmidt, Garratt, and Fitzpatrick, 2002). The questionnaire is suitable for and has been validated within a school context and takes a maximum of 20 minutes to complete. The questionnaire has been validated for use in the UK (see Schmidt, Garratt, and Fitzpatrick, 2002). Unfortunately, the one question that captures the scale relating to past health was omitted from the baseline questionnaire (it was dropped by the company commissioned to print the questionnaire in error). So, the study has thirteen scales on which to assess impact, with the General Health scale being the primary outcome, and the remaining twelve scales providing secondary outcomes.

It is well recognised that HRQoL, capturing soft skills, health and general well-being, cannot be assessed within a single measured outcome (Conti and Heckman, 2012; Decancq and Neuman, 2014; Khanemann and Krueger, 2006). This dictated the use of CHQ-CF87 in this study, which has multiple scales which is the major focus in this work. However, the study also gathered data on the Short Mood and Feelings Questionnaire (Angold and Costello, 1987) and the Child Anxiety Related Disorders (SCARED) questionnaire (Birmaher et al, 1999), as well self reported life satisfaction on the 0-10 ladder. All are validated instruments. The Short Mood and Feelings Questionnaire is administered in short-form with 13-questions addressing mainly issues of depression in children. It is not a diagnostic tool, but an indicator of the possible presence of symptoms of depression. The SCARED questionnaire is a tool to highlight issues of childhood anxiety using four domains: panic, separation anxiety, generalised anxiety and school phobia. The life satisfaction ladder returns a visual scaling of subjective life satisfaction assessment. Impacts on these additional measures are also presented with correction to the estimated standard errors to allow for multiple outcomes (discussed more completely below).

The CHQ-CF87, the Short Mood and Feelings Questionnaire and the Child Anxiety Related Disorders questionnaire were all administered through the same paper-based questionnaire given to pupils by the coding team, who are a commercial firm (Healthcare Solutions) and are distinct from the analysis team, during a class setting. Pupils were asked to answer the questionnaire under exam-type conditions, although it was explained that the questionnaire was not an exam, that there were no right or wrong answers and that no questionnaires would be returned to teachers or parents, and that all data would be anonymised. The full questionnaire takes approximately 40 minutes to complete, and although no counter-balancing was undertaken to assess fatigue effects, early piloting of the questionnaire did not reveal any such problems. The questionnaire was issued across all schools in two phases, one group beginning in 2013 and the other in 2014, with the questionnaire issued three times (2013; 2015; 2017 & 2014; 2016; 2018 respectively). These time periods correspond to baseline, inter, and ex post data collection and are two years apart. The coding team administered the questionnaires, collected and collated the data from the questionnaires, removed names and allocated a unique identifier (ID) to each of the questionnaires to allow panel construction and recorded data within Excel spreadsheets, which was released to the analysts at the end of the final data collection period (summer 2018).

Statistical analysis

Primary Outcome Analysis

The primary outcome was based on a change in the general health scale of the CHQ-CF87.

The primary empirical analysis was based on the following basic difference-in-difference specification. This is preferred over standard approaches to RCT analysis given that the course was administered over a 4 year period over which, alongside HM, the recipients would have experienced changes which could impact on their HRQoL outcomes. Using differences in differences allows us to assume that such changes are either common to the treatment or control school (e.g. common changes) (we also allow for specific schools to have different changes by adding school fixed effects are used in our robustness analysis, see below):

$$y_{ist} = \beta_0 + \beta_1 treatment_{ts} + \beta_2 year_t + \beta_3 treatment * year_{ts} + \epsilon_{ist} \quad (1)$$

where:

y_{ist} = the outcome variable

$treatment = 1$ if a school was chosen for treatment, regardless of whether they adhered to the treatment

$year_t$ is a set of yearly fixed effects based on the year the data was collected.

The coefficient β_3 captures the effect of being assigned the treatment, under an assumption of common trends (i.e. the treatment group would have continued on the same trajectory as the control group in the absence of the treatment). We interpret the primary analysis as an intention to treat effect. This is arguably the effect policy makers care about the most, as if the program is adopted there will be heterogeneity in how the program is rolled out at the school level. The expectation here is that the HM course will give a cumulative effect from building up over time. We expect β_3 to be most substantive when we compare across the baseline and end line data so this is the major focus in this work. This coefficient, β_3 , represents the average treatment effect of Healthy Minds (HM) overall. We also present results which analyse β_3 for the interim data in a robustness check. This reflects the average treatment effect of HM when half of the course was complete.

Estimation was undertaken through use of Stata (version 15). Standard errors are adjusted to allow for clustering at the school level and unknown heterogeneity (double HAC standard errors in Stata).

This basic analysis forms the basis of a common element of analysis running through into the secondary analysis, which analyses the multiple scales of the CHQ-CF87. A number of further model specifications were undertaken within the analysis, to include robustness checks and control variables, as detailed below.

Secondary Outcome Analysis

As already described, this is a multiple outcome study, with the major focus on the thirteen scales in the CHQ-CF87. In the initial SAP agreed with the EEF we proposed using exploratory factor analysis on these scales to extract underlying orthogonal factors that represent the independent dimensions of HRQoL that are captured within this instrument. We note that we have attempted this, but it was not successful. So, we have proceeded with more traditional corrections when analysing multiple outcomes testing. That is, alongside traditional t-testing we also document significance after applying the correction for multiple testing proposed by Benjamini, Y. and Hochberg, Y. (1995) to the p-values calculated in our routine analyses.

Additional analysis

A set of additional specifications will be used for robustness. Specifically, these are:

1. A robustness check will consider a more saturated version of equation 1 and add school level control variables. Through the addition of school fixed effects with suitable adjustment to equation (1), as we note that the treatment indicator in equation 1 drops out, however the interaction term which is the main point of interest remains.
2. We will consider additional robustness through the specification of pupil fixed effects (we note that the treatment indicator in equation 1 drops out, however the interaction term which is the main point of interest remains). This allows us to control for unobserved fixed pupil effects.
3. Robustness test of the impact of peer-group effects. To test whether there are significant peer group spillover effects associated with the treatment programme, captured by a “leave-me-out”

mean effect of other responders (based on the mean of programme effects witnessed in other class responders). So, an additional variable is included in equation 1 measuring the aggregate mean treatment effect (that is being used to define the specific y_{ist}) associated with all other responders for each i , based on leaving the specified individual out of the calculated mean effect, for each i .

A balance table is included to show the balance of characteristics across the treatment and control populations at baseline.

Missing Data analysis

The ex ante inclusion criteria in this work specified that a student must have responded to the global health one item questionnaire on the CHQ-CF87 (i.e. the primary outcome) to be included in this study. Missing data within the other 12 sub scales were imputed following the validated algorithm provided within the CHQ-CF87 coding book (HealthActCHQ, 2013). Once this algorithm was complete we had no substantive missing data within the returned questionnaires.

Non-compliance analysis

As not all treatment schools completed the delivery of the amended (treatment) PHSE curriculum over the 4 years of delivery, an analysis of compliance will be based on the difference-in-difference equation identifying the “intensity” of treatment effect given above in the first robustness check.

We consider this by estimating the following equation:

$$y_{ist} = \beta_0 + \beta_1 \text{delivery}_s + \beta_2 \text{year}_t + \beta_3 \text{delivery} * \text{year}_{ts} + \epsilon_{ist} \quad (2)$$

In equation (2) $\text{delivery}=1$ if a school delivered the program to the satisfaction of the Bounce Forward team. That is, the Bounce Forward team have a record of the school completing the HM course in full over the four years. Delivery is then equal to 0 if a school was a control. In total 13 treated school-cohorts completed the program in full.

We note that schools who did not deliver the program to a satisfactory standard are excluded from the robustness. Given that the schools who selected out of the study are likely to be systematically different from those that remain, there is also a likelihood that those that remain differ from the controls. Thus, we also include school fixed effects in a separate estimation of equation (2). We are aware that those that selected out are lower SES so we expect that including these effects will attenuate the estimate of β_3 , which captures the effect of the intervention.

Intra-cluster correlation analysis

A decomposition of the overall, within and between school-level mean effects will be calculated for all primary and secondary outcome variables. The main regressions, given above, will include school fixed effects and cluster robust estimation of the variance matrix. This will be based on calculating the following decomposition of variance:

$$1/C \sum_{s=1}^S \sum_{t=1}^T (y_{ts} - \bar{y})^2 = \frac{1}{C} \sum_{s=1}^S \sum_{t=1}^T (y_{ts} - \bar{y}_s)^2 + \frac{1}{C} \sum_{s=1}^S (y_s - \bar{y})^2$$

where y is the outcome variable of interest, with s being the school and t being the number of returned questionnaires administered in any given school in any year, and C is the total number of returned questionnaires across all schools in total.

Effect size calculations

The effect size will be returned as β_3 from equation (1). We also present the effect divided by the unconditional variance of the outcome measure as per EEF analysis guidance. This allows for better comparison across other EEF projects.

Implementation and process evaluation

This study did not include a formal implementation and process evaluation. Nevertheless, estimates of the costs are presented below.

Costs

Costs were collected by Bounce Forward based on the training days given to teachers, the replacement teacher cost and the printing of material. The major costs involved related to the trial itself and the training undertaken by teachers to deliver the HM course. The trial costs are not relevant (i.e. they are sunk) to the evaluation and therefore are not reported. The training costs were associated with 19 days of teacher training at £190 per day, plus the teacher replacement costs of those on training estimated at £160 per day. These costs total £6,640 over the 4 years of the trial or £1,662.50 per annum. Assuming one teacher can teach a course of three classes and that this totals 90 students per school year, and that the printing of associated materials totals £5 per student per school year the total cost of the HM course is £23.50 per student per school year.

Timeline

The Table 4 and Figure 1 outline the timeline of the analysis and the structure of the participant flows. The timeline illustrates the role of Phase 1 treatment and wait list control schools, as well as Phase 2 treatment and control schools. It documents the intended administration of the data questionnaires. Given the large degree of drop-outs, as noted above, a number of schools were clawed-back for final data collection to retain sample size across baseline and endline data collection and this is reflected in the participant flow chart.

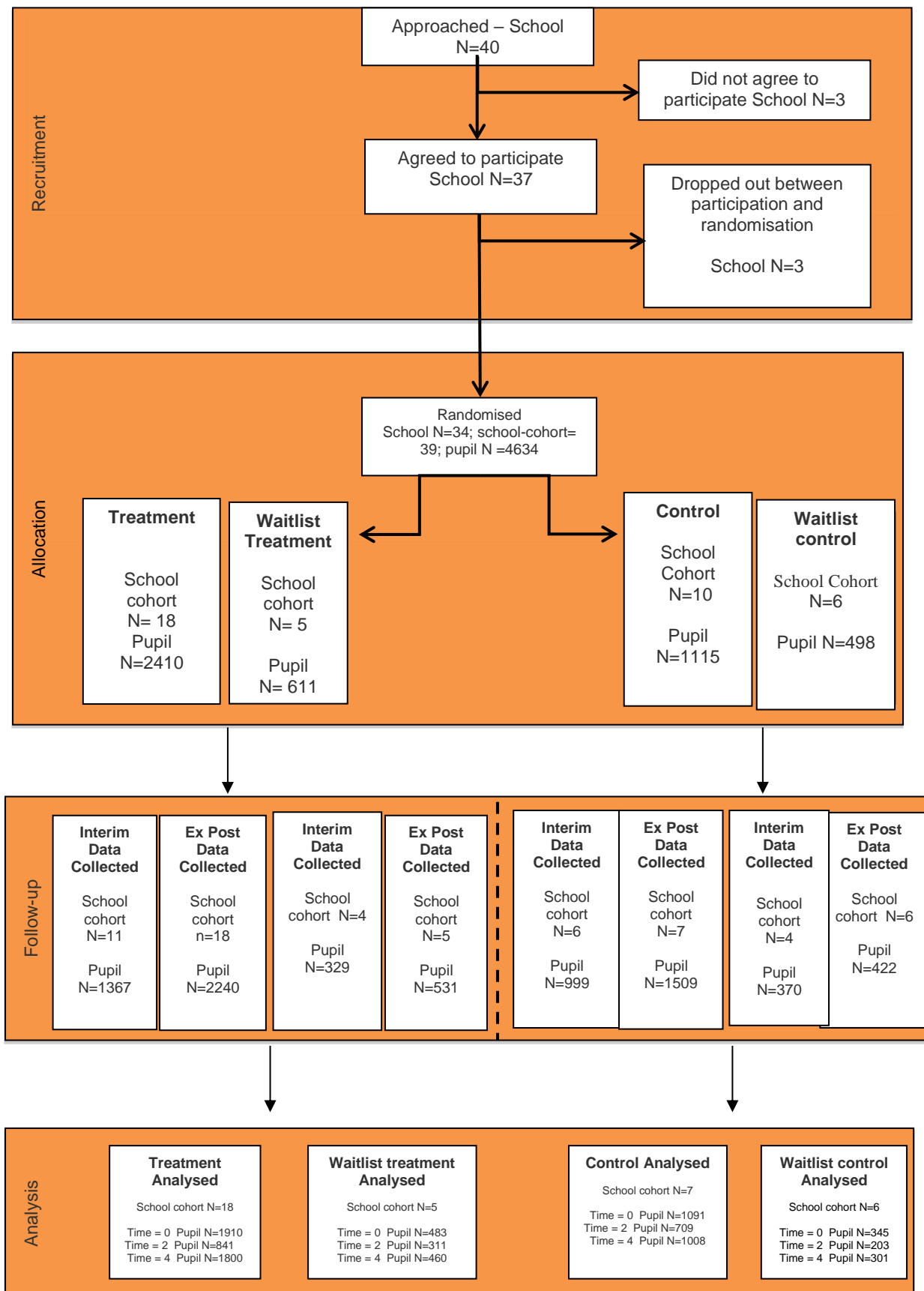
Figure 1 provides the actual data collection stages and participant flow at the school, school-cohort and pupil level respectively. There were 23 treatment school-cohorts (including 5 wait list treated cohorts) and 16 control school-cohorts (including 6 wait list control cohorts) available for an analysis which includes baseline and endline data. To allow for selection effects, the robustness analysis described above includes analyses that add school and pupil fixed effects. For a pupil to have been included in our analysis they need to have been assigned by the coders a valid (anonymised) identifier and provided a valid response for the primary outcome question. Details of the number of pupils included at T=0 (baseline in 2013 or 2014), T=2 (interim in 2015 or 2016) and T=4 (endline in 2017 or 2018) is provided in the analysis box of the flow diagram.

Table 4: Study Timeline

Date	Activity
January 2013	Approach potential participating schools
January 2013 through September 2013	Recruitment of schools
April – September 2013	Allocation to Phase 1 schools (pupils enter study September 2013) and Phase 2 schools (pupils enter study September 2014)
July 2013	Year 1 Teacher Training for Phase 1 schools
September 2013	HM teaching begins for Phase 1 schools
September 2013	Data (baseline) questionnaires administered to Phase 1 schools (Treatment and wait-list control schools)
July 2014	Year 1 Teacher Training for Phase 2 schools and Year 2 Teacher Training for Phase 1 schools
September 2014	HM teaching begins for Phase 2 schools
September 2014	Data (baseline) questionnaires administered to Phase 2 schools (Treatment, wait-list treatment schools and control schools)
July 2015	Year 2 Teacher Training for Phase 2 schools and Year 3 Teacher Training for Phase 1 schools
May/June 2015	Data (interim) questionnaires administered to Phase 1 schools (Treatment and wait-list control schools)
July 2016	Year 3 Teacher Training for Phase 2 schools and Year 4 Teacher Training for Phase 1 schools
May/June 2016	Data (interim) questionnaires administered to Phase 2 schools (Treatment, wait-list treatment and control schools)
July 2017	Year 4 Teacher Training for Phase 2 schools
May/June 2017	Data (endline) questionnaires administered to Phase 1 schools (Treatment and wait-list control schools)
May/June 2018	Data (endline) questionnaires administered to Phase 2 schools (Treatment, wait-list treatment and control schools)
August/September 2018	Data released and analysis undertaken

Participant flow

Figure 1 Participant and school cohort flow



The minimum detectable effect size was based on a benefit defined through a movement in standard deviations across a difference in means as measured by β_3 estimated from equation (1). Table 5 gives information based on the designed study of the treatment intervention (protocol), as compared to the actual randomisation of the recruited schools, after withdrawal at baseline. As can be seen from Table 5 the protocol envisaged balance between the control and treatment schools; 15 schools each and 1,500 pupils each. However, the noted recruitment and retention issues meant that this was not achieved. In the end 34 schools were recruited in a wait-list control, control and treatment design that defined 39 school-cohorts, with 2589 pupils in the treatment cohort and 1711 pupils in the control cohort.

Attrition

Randomisation was undertaken at the school level. As well as recruitment difficulties, over the 4-year study period retention also proved to be problematic as schools faced an increasingly difficult environment, with increasing financial and performance pressures as well as course changes in mainstream subjects. Through a claw-back mechanism, efforts were made to match school-cohort participation at the beginning and the end of the study, to ensure baseline and endline comparisons. While largely successful there remained considerable attrition. At the pupil level attrition was a loss of 1,065 pupils (4,634 – 3,569) equivalent to 23% of the pupils. At the school-cohort level 4 school-cohorts were lost at the final data collection, with 35 school-cohorts forming overall analysis.

Table 5: Minimum detectable effect size at different stages

		Protocol	Randomisation	
		OVERALL	OVERALL	
MDES		0.28		0.28
Pre-test/ post-test correlations	level 1 (pupil)	0.00		0.00
	level 2 (class)	0.00		0.00
	level 3 (school)	0.00		0.00
Intracluster correlations (ICCs)	level 2 (class)	0.00		0.00
	level 3 (school)	0.06		0.06
Alpha		0.05		0.05
Power		0.8		0.8
One-sided or two-sided?		2		2
Average cluster size		100		121
Number of schools	Intervention	15		18*
	Control	15		16*
	Total	30		34
Number of pupils	Intervention	1500		2589
	Control	1500		1711
	Total	3000		4,634

Note: * Given the wait list design some schools enter both the control group and treatment group. In Phase 1 (2013) 7 schools were recruited to treatment and 6 as (wait-list) control. In Phase 2 (2014) 11 schools were recruited to treatment and a further 10 as control schools. Phase 2 also included 5 wait-list treatment schools. So there are 23 school-cohorts in the treatment group, as 5 were wait list schools, who entered the 1st Phase as wait list control schools. There had been 6 wait-list control schools in Phase 1 but one of these schools dropped-out by Phase 2. So the number of school-cohorts is 39, based on 34 recruited schools. See Table 3 and Figure 2 for school cohorts by stage and pupil participation respectively.

Pupil and school characteristics

Given the difficulties noted above in the recruitment of the schools, and that schools were subsequently phased in over time there was no initial (pre-tests) undertaken. Instead baseline data was gathered after a school had been assigned as a treatment or control school. The basic school characteristics and initial pupil numbers at the time of recruitment into the study are reported below in Table 6.

As can be seen most of the schools were receiving “Good” OFSTED performance scores around the time of entry into the study. Only 3 schools had an “Outstanding” achievement OFSTED score; 2 in the treatment group and 1 in the control group. Overall 9 received “Requires improvement” OFSTED ratings at the time of entry; 5 in the treatment and 4 in the control. Most were medium to large comprehensive schools with a mix of Academy, Academy convertor and Foundation status, although 1 school in the

treatment group was a Community school and 1 school in the control group was a voluntary grant aided at the time of randomisation.

Table 6: Baseline comparison

School-level (categorical)	Intervention group		Control group	
	n/N (missing)	Count (%)	n/N (missing)	Count (%)
Academy	892/3012 (0)	4 (22%)	585/1613 (0)	5 (31%)
Academy converter	904/3021 (0)	6 (33%)	902/1613 (0)	9 (56%)
Foundation	72/3021 (0)	1 (5.5%)	54/1613 (0)	1 (6.25%)
Community	830/3012 (0)	6 (33%)	0	0
Free school	0	0	72/1603	1 (6.25%)
Voluntary aided	182/3021 (0)	1 (5.5%)	0	0
Ofsted rating				
Outstanding		2		1
Good		11		10
Requires improvement		5		4

Table 7 reports the baseline raw difference in mean scores for both the treatment and the control school-cohorts. As can be seen for the raw scores the primary outcome scores and a number of the secondary outcomes the raw scores in the treatment group fall below those of the control group. However, as reported in Table 8, by the end of the study the treatment group had caught-up to some degree with the control group in terms of raw scores. As these are raw means only the general trend is noted.

Table 7: Baseline Raw differences in scores

Unadjusted differences in means							
Primary outcome	Global Health score from the CHQ-CF87						
	T= 0.005 C = 0.014 g= -0.009						
Secondary outcomes:	Physic. Function	Emotional difficulty	Behav. difficulty	Self-Esteem	Phy. Difficulty	Pain & Discomfort	General Behaviour
	T= -0.106 C= -0.052 g=-0.047	T= -0.094 C= -0.055 g=-0.036	T= -0.083 C= -0.047 g=-0.033	T= -0.027 C= 0.053 g=-0.082	T= -0.062 C= -0.039 g=-0.022	T= 0.028 C= 0.062 g=-0.035	T= -0.098 C= 0.059 g=-0.155
Secondary outcomes:	Global behaviour	Mental Health	General health	Family Activities	Family Cohesion		
	T= -0.085 C= 0.035 g=-0.119	T= -0.052 C= 0.024 g=-0.075	T= -0.039 C= 0.025 g=-0.064	T= -0.071 C= -0.002 g=-0.069	T= 0.029 C= -0.056 g=0.085		

Note: g is effect size reported as Hedge's g

Table 8: Endline Raw differences in scores

Unadjusted differences in means							
Primary outcome	Global Health score from the CHQ-CF87						
	T= 0.065 C = -0.063 g=0.126						
Secondary outcomes:	Physic. Function	Emotional difficulty	Behav. difficulty	Self-Esteem	Phy. Difficulty	Pain & Discomfort	General Behaviour
	T= -0.025 C= 0.023 g=-0.048	T= -0.042 C= -0.050 g=-0.009	T= -0.033 C= 0.042 g=-0.075	T= -0.078 C= 0.082 g=-0.160	T= -0.004 C= -0.001 g=-0.003	T= 0.053 C= -0.054 g=0.107	T= -0.049 C= 0.058 g=-0.108
Secondary outcomes:	Global behaviour	Mental Health	General health	Family Activities	Family Cohesion		
	T= -0.037 C= 0.028 g=-0.065	T= -0.069 C= 0.085 g=-0.154	T= 0.040 C= 0.045 g=-0.086	T= -0.028 C= 0.027 g=-0.055	T= 0.055 C= -0.069 g=0.124		

Note: g is effect size reported as Hedge's g

Outcomes and analysis

Table 9 below gives the *unadjusted* differences in the mean scale levels for the items in the CHQ-CF87 questionnaire between the treatment and control schools at baseline. Scales have been standardised to have a mean of zero and standard deviation. A negative sign denotes that the average treated child had a worse outcome than at baseline, conversely a positive sign denotes that they were better off. Table 9 also documents standard errors for these differences in brackets. For seven of the thirteen outcomes, there are no significant differences between the average treated and control child. This includes the primary outcome (Global health) scale, where the unadjusted mean difference is not statistically significant and negative (-0.009). However, significant differences are observed in six of the remaining outcomes, with the treated child doing notably worse at initiation in the behaviour, global behaviour, self-esteem and general health scales. The treated children are slightly better off in the family activities and family cohesion scales (0.034 and 0.084 respectively). The same picture is painted if we consider the Hedges g effects reported for the baseline data in table 7. However, these are all *unadjusted* mean differences. Our modelling strategy in equation (1) assumes that without the HM course these

differences would remain fixed. The robustness analysis described above outlines alternative approaches which relaxes this assumption.

Table 9: T-tests for difference in unadjusted mean scores

Unadjusted differences in means							
Primary outcome	Global Health score from the CHQ-CF87						
	-0.009 (0.036) N=2,976						
Secondary outcomes:	Physic. Function	Emotional difficulty	Behav. difficulty	Self-Esteem	Phy. Difficulty	Pain & Discomfort	General Behaviour
	-0.053 (0.041)	-0.039 (0.040)	0.064 (0.039)	-0.081* (0.037)	-0.023 (0.039)	-0.026 (0.036)	-0.157 (0.038)
Secondary outcomes:	Global behaviour	Mental Health	General health	Family Activities	Family Cohesion		
	-0.120*** (0.037)	-0.077 (0.042)	-0.064* (0.37)	0.034* (0.037)	0.084** (0.037)		

Table 10 documents the main results from our standard difference in difference models (see equation 1). All of the reported estimates are given with the statistical significance at conventional levels unadjusted for multiple comparisons. Given that the primary outcome is part of the 13-item scales returned from the CHQ-CF87 questionnaire adjustment was also made to allow for multiple comparisons. Those reported in the Tables for both the primary and secondary analysis in **bold** indicate that these results remain significant at the 5% level after adjustment for multiple comparisons and those reported in **bold** and *italics* remain significant at the 10% level after adjustment for multiple comparisons. We follow the methods proposed by Benjamini and Hochberg, 1995 when making these corrections and also report the associated p values. We note that when commenting on the results we focus only on those variables that maintain significance under Benjamini and Hochberg, 1995 correction.

Table 10 also documents a number of robustness check. These are, in order of rows, a model which adds a school effect to equation (1), a model that adds pupil effects to equation (1), a model that adds a number of school and pupil level control variables to equation (1), a model that incorporates compliance by estimating equation (2), a model that adds school effects to equation (2) and a model that adds the “leave me out” mean effect.

As can be seen for the primary outcome reported in Table 10, the effect is of the expected positive sign and significant in the baseline result. Students exposed to HM have global health attainment that is 0.245 standard deviations (s.d.) higher than children in the control group. In all of the robustness analysis conducted in this work the primary effect remains of the expected positive sign and statistically significant. This result is consistent with a general finding that more than 60% of individuals in the treatment group return a self-assessed general health improvement which is above that of the control arm individuals as a result of the intervention. This effect size remains robust across a number of different specifications. In particular, in comparing the analysis based on compliance and the main (intent-to-treat) analysis the effect size is very stable. Introducing pupil fixed effects reduces the effect size somewhat, but it remains positive and significant.

The interim results report an improvement obtained after 2-years of teaching HM. The improvement in the primary outcome for the interim analysis is of similar magnitude to that at the end of the program, although slightly smaller (a gain of 0.234 s.d. as opposed to 0.245 s.d. for the final analysis). In aggregate this could be interpreted as the programme benefits being gained within the first two years. While this is true, the findings are also consistent with the maintenance and persistence of these positive programme effects over time, with a slight improvement in benefit over the last two years. However, this conclusion is reached without comparison to any counterfactual of switching from treatment to control teaching during the last phase of study. It nonetheless is the case that the interim finding and its relationship to the final outcome reflects a persistence of the structured programme teaching on the primary outcome.

Table 10: Effect size estimation for Primary Outcome: General Health score from the CHQ-CF87

Outcome	Adjusted differences in means	Population (n)	Missing (n)	CI 95% Level	Adjusted P
Primary outcome: Global Health score from the CHQ-CF87					
Baseline difference-in-difference estimates	0.245** (0.110)	7,326	0	0.018 0.471	0.042
Including School effects	0.232* (0.121)	7,326	0	-0.017 0.481	0.067
Including Individual pupil effects	0.150* (0.075)	6,173	1153	-0.002 0.294	0.089
Including school and pupil covariates	0.212** (0.099)	7,326	0	0.014 0.410	0.046
Compliance Analysis	0.237* (0.128)	5,533	1793	-0.019 0.493	0.068
Compliance Analysis with School Fixed Effects	0.145 (0.109)	5,533	1793	-0.073 0.371	0.124
Add “leave me out” mean effect	0.153*** (0.047)	7,326	0	0.056 0.250	0.007
Difference-in-difference estimates (interim results)	0.234** (0.094)	5,821	1505	0.046 0.422	0.030

Notes: The valid population for analysis is equal to all students who filled in the single item primary outcome question in phase 1 (2013/2014) or phase 3 (2017/2018). In total there are 7326 cases. Please see figure 1 for more information on how many students did not answer this item and are excluded from the analyses here. Higher values imply better health. All Outcomes are standardised to have a mean of 0 and standard deviation of 1, so effect Standard Errors are clustered at the school level. * * * * * denotes significance using standard t testing. **Bold** font indicates the treatment effect is significant at the 5% level of significance after the Benjamini, Y. and Hochberg, Y. (1995) multiple comparison correction. **Bold and Italic** indicates the treatment effect is significant at the 10% level using the same correction. N is always the actual observations in the panel. For the compliance analysis with school fixed effects we only include schools that fully comply and the control schools.

Table 11: Effect size estimation for Additional Secondary Outcomes

Adjusted differences in means												
Category	H	I	E	I	H	H	E	E	I	H	H, E	N/A
Secondary outcomes:	Physic. Function	Emotional Diff	Behav. Di	Self-Esteem	Phy. Difficul	Pain & Discomfort	General Behav	Global behaviour	Mental Health	General health	Family Activities	Family Cohesion
Baseline	0.162 (0.128)	0.052 (0.086)	0.123* (0.067)	-0.040 (0.078)	0.297*** (0.068)	0.206* (0.116)	0.145** (0.061)	0.139 (0.143)	0.067 (0.052)	0.151** (0.070)	0.118* (0.062)	0.121 (0.101)
Confidence Interval 95% level	-0.102 0.426	-0.125 0.230	-0.015 0.260	-0.200 0.120	0.157 0.437	-0.033 0.446	0.146 0.061	-0.139 0.418	-0.039 0.173	0.006 0.294	-0.010 0.246	-0.087 0.328
Adjusted P	0.569	1.000	0.128	1.000	0.000	0.165	0.027	1.000	0.442	0.052	0.101	0.790
N (Missing)	7326 (0)	7300 (26)	7278 (48)	7272 (54)	7209 (117)	7212 (114)	7030 (296)	7051 (275)	6429 (897)	7277 (49)	7177 (149)	7189 (137)

Notes: Category H=Health, I=Internalising Behaviour and E=Externalising Behaviour. Higher values indicate better HRQoL. See also notes to Table 5 and 10

Turning to Table 11, given the text of the questions, the secondary outcomes may be viewed as each capturing health, internalising behaviour and externalising behaviour. The exceptions are family activities, which captures the effects of *both* health and externalising behaviour on family activities, and family cohesion which does not fall under any of these categories. The top row of Table 6 assigns a category to each of the remaining outcomes. Here H = health, I=internalising behaviour and E = externalising behaviour. Using these categories it is clear that the HM gains go to augmenting health and externalising behaviour. Thus, we may conclude that HM promotes these outcomes. In contrast, the three variables that capture internalising behaviour have coefficients that are centred around zero (i.e not economically meaningful), and not significant.

The robustness analysis (see Table 1 Appendix) follows this pattern. We note the exception is the interim analysis, which demonstrates negative effects on the variables that capture internalising behaviour, which revert to zero by the end of the program. We also note the self esteem measure is always negative, and although centered on zero for our baseline specification, does become more substantive and statistically significant in alternate specifications. In contrast, the effects on externalising behaviour and health are positive, even if often not significant. There are a number of reasons why this may have occurred. First, it is possible that HM augments externalising behaviour through a path that causes internalising behaviour to deteriorate in the interim. This may occur if students who are treated engage in more self-introspection and become more critical of themselves. Second, it is possible that there is not enough content in HM that relates to positively affecting internalising behaviour. Third, it is possible that the students who are treated, end up being more in touch with their feelings, and answer the questions on internalising behaviour differently. In contrast, the questions that they are asked for the externalising and health categories are concerned with actions, behaviours and symptoms that are easily observed and arguably more objective. Finally, it is possible that the teachers are better at teaching the elements of the program on health and externalising behaviour, over and above internalising behaviour. We cannot not disentangle any of these explanations and these are areas which should be explored as HM goes forward to tighten its delivery and outcomes. However, we can exploit the data on the elements of the course covered to say a bit more about certain groups of modules and their effects on each of the three categories. This analysis lies outside the statistical analysis plan given to the EEF.

Table 12: Effect size estimation for Additional Secondary Outcomes

Adjusted differences in means								
Secondary outcomes:	Anxiety Disorder	Pain Disorder	Generalised Anxiety Disorder	Separation Anxiety	Social Anxiety Disorder	Significant School Avoidance	Mood and Feelings	Life Satisfaction
Baseline D&D	-0.029	-0.091**	0.031	-0.062*	-0.011	-0.168***	0.112	0.182**
	(0.043)	(0.036)	(0.024)	(0.031)	(0.058)	(0.033)	(0.320)	(0.040)
Confidence	-0.118	-0.166	-0.119	-0.126	-0.130	-0.236	-0.547	-0.001
Interval 95% level	0.060	-0.016	0.081	0.002	0.107	-0.100	0.773	-0.735
Adjusted P	1.000	0.022	0.418	0.091	1.000	0.000	1.000	0.068
N	6364	6364	6364	6364	6364	6364	7014	6976
(missing)	(962)	(962)	(962)	(962)	(962)	(962)	(312)	(350)

Notes: Lower values in anxiety disorder, pain disorder, generalised anxiety disorder, separation anxiety, social anxiety disorder and significant school avoidance imply worse health. These are binary outcomes. Higher values of Mood and Feelings and Life Satisfaction indicate better health. These outcomes are standardised to have a mean of 0 and standard deviation of 1. Standard errors are clustered at the school level. The *, **, *** denotes significance using standard t testing. **Bold** font indicates the treatment effect is significant at the 5% level of significance after the Benjamini, Y. and Hochberg, Y. (1995) multiple comparison correction. **Bold and Italic** indicates the treatment effect is significant at the 10% level using the same correction.

Table 12 documents some additional estimates which consider additional outcomes based on two additional questionnaires other than the CHQ-CF87. The first five columns of Table 12 report the outcomes relating to the SCARED questionnaire. The outcomes are binary and are assigned “equal to 1” if the screen for child anxiety related disorders (SCARED) instrument suggests that the respondent has i) anxiety disorder, ii) pain disorder, iii) generalised anxiety disorder, iv) separation anxiety, v) social anxiety disorder and vi) significant school avoidance. Higher scores signify worse mental health. Three of the scores (pain disorder, separation anxiety and significant school avoidance) were negative and significant which suggests better mental health.

The seventh column considers the standardised score from the Short Form Mood and Feelings questionnaire. The score was positive but not significant for the general mood and feelings question

Finally, the eighth column documents estimates for life satisfaction. Here, the children were asked to indicate on a scale of one to ten ‘Overall, how satisfied are you with your life nowadays?’ In our regressions we standardise the score to have a mean of zero and standard deviation of one. This general life satisfaction score is positive and significant. From Table 9, it is clear that HM had significant and augmenting effects in four out of eight of these ancillary outcomes.

Cost

The training costs were associated with 19 days of teacher training at £190 per day, plus the teacher replacement costs of those on training estimated at £160 per day. These costs total £6,640 over the 4 years of the trial or £1,662.50 per annum. Assuming one teacher can teach a course of 3 classes of 30 pupils to a total of 90 students per school year, and that the printing of associated materials totals £5 per student per school year the total cost of the HM course is £23.50 per student per school year. On these assumptions the total cost to a school over the 4-year period of study would be £7,240. These are the estimated, actual costs incurred as associated with implementation by individual schools given in Table 13 below. As shown in Table 14, which details the cumulative costs, once teacher training is undertaken the incremental costs are low.

Table 13: Cost of delivering Healthy Minds

(1)	(2)	(3)	(4)	(5)
Item	Type of cost	Cost	Total cost over 4 years	Total cost per pupil per year
One-off teacher training	Teacher training cost per school	£3,600 per school	£3,600 per school	£10 (assuming 1 teacher teaches 3 classes of 30 students each)
	Replacement teacher cost during training	£3,040 per school	£3,040 per school	£8.50 (assuming 1 teacher teaches 3 classes of 30 students each)
Total set-up cost			£6640 per school	£18.50 (assuming 1 teacher teaches 3 classes of 30 students each)
Material printing costs per student	Material per student	£5 per student school year	£600 (assuming 30 students per year)	£5
Total			£7,240	£23.50

Notes: we assume that each trained teacher teaches 3 classes of 30 students each. To get a per pupil per year cost, for schools of various sizes, £6640 in column (4) should be divided by $4 \times i \times j$. Here, i is the number of pupils a teacher will teach in each class and j is the number of classes. It is necessary to divide by 4 to get a per year cost as the costs in column (4) relate to the total cost of HM in 4 years. Do note, that schools undertaking HM will have a large proportion of costs front loaded as they relate to training (see Table 14).

Table 14: Cumulative costs of Healthy Minds

	Year 1	Year 2	Year 3	Year 4
Healthy Minds	£6,790	£6,940	£7,090	£7,240

Implementation and process evaluation

The implementation and process evaluation will be published as part of the full evaluation report in 2019.²

² There is a planned implementation and process evaluation associated with the NIESR study which will report on the educational attainment aspects of the programme in the forthcoming year.

Conclusion

This working paper reports on a large trial in which schools were randomised to a 4-year programme of standardised PHSE teaching, where teachers had been trained to deliver 113-hours of a structured course. A total of 34 secondary schools (39 school-cohorts) were followed-up over the 4-year period, with 3,021 students in the treatment arm and 1,613 students in the control arm. The trial assessed, on an intent-to-treat basis, whether this programme had an impact on students' HRQoL. This assessment was undertaken using a validated questionnaire, the CHQ-CF87, supplemented with three others, given that it is well-recognised that multiple skills and attributes are involved in the measurement of HRQoL, behaviour and emotional well-being and that it is important to capture all relevant dimensions.

This analysis is important as the standard of PHSE teaching within secondary schools has been heavily criticised recently by the House of Commons Education Committee (2015) and there is currently an on-going debate over the optimal content of such a curriculum. Moreover, it is increasingly recognised that positive interventions made during adolescence can have long-term positive consequences over an individual's lifespan (Coleman et al, 2011; Colishaw et al, 2004; Durlak et al, 2010, 2011; Kautz et al, 2016).

The study is also important, as it is one of the few non-US trial based assessments of the effect of a course change on HRQoL, behaviour and emotional well-being in UK adolescents. Even in the US, where there is a longer history of such studies, interventions aimed at adolescents tend not to have long follow-up and therefore they lack the depth to properly evaluate the intervention or prove the persistence of the effect. Even when considering non-cognitive skills, most adolescent trials are aimed at schooling performance and employment outcomes (Kautz, T. Heckman, J., Diris, R., et al., 2014). Durlak et al (2010) in a meta-analysis of after-school programmes (APS) aimed at improving personal and social skills, noted the small number of studies reviewing impacts on adolescents. Across all age groups they noted that many studies were not based on randomisation, but rather through identifying controls as pupils not in the ASP. They also noted, in line with our findings that the most successful programmes were those associated with the incorporation of formal teacher training.

While recruitment and retention for such a large study was difficult, a total of 31 schools (35 school-cohorts) were retained for baseline-endline comparisons, using a clawback mechanism. This clawback was based on the administration of the final study questionnaire to pupils not only in schools retained throughout the study, but in schools which had withdrawn from treatment over the course of the study. As the analysis was conducted on an intent-to-treat basis this meant we retained the appropriate sample size to conduct our analysis with statistical confidence.

The primary outcome measure was the general health score embedded within the CHQ-CF87, a well-validated and assessed children's questionnaire, which has 14 distinct subscales relating to HRQoL. We use 13 of the 14 available sub-scales, as one sub-scale was not measured at baseline. The evaluation found that the primary outcome was improved by 0.245 standard deviations in the treated group, which remains significant even after adjusting the standard errors for multiple comparisons. Assuming normal distributions, this is consistent with 60% of the control group's measure lying below the average score in the treated population. In other words, students who were treated with HM improved their position by 10 percentiles (out of 100). This effect size remains robust across a number of different specifications.

If we compare these final baseline-endline results for the primary measure with the interim results obtained after 2-years of programme teaching the benefit is roughly of the same magnitude, although slightly smaller (0.234 standard deviations). In aggregate this might be taken to mean that the programme benefits are attributed to the first two years. However, caution must be exercised here as the measured primary outcome benefit is not assessed with respect to switching from treatment to control teaching. The finding of similar interim and final outcome benefits are consistent not only with

maintenance and slight improvement in course benefit, but could also be argued that this shows persistence of the taught structured programme on the primary outcome, or indeed that the positive effect needs continued maintenance to be preserved.

The secondary outcome measure results, based on the CHQ-CF87, are more mixed both across the specific measures and across the various specifications introduced to ensure robustness. Although if emphasis is placed on the baseline-endline comparisons outcomes, rather than on the interim analysis, these are largely positive and are significant across a number of dimensions (notably physical health and externalising behaviour outcomes are positive and significant, and outcomes related to internalising behaviour are more often centered around zero and not significant). For the main results five out of a total of 13 dimensions of the CHQ-CF87 questionnaire are of positive sign and statistically significant after adjustment is made for multiple comparison testing. The interim analyses does reveal negative and significant effects on self-esteem, and these type of effects are also documented for additional robustness which examines the baseline and end-line data on the same outcome. This is the only consistent anomaly (i.e finding of negative effects in a number of specifications), and we have taken care to write up competing hypotheses of why this may have occurred. However, we cannot know true cause of this anomaly given the data at hand and do hope it will be considered in future research as HM moves forward.

Of the other secondary outcome questionnaires, the SCARED questionnaire which screens for child anxiety related disorders suggests that children who received HM had improvements in three out of the six outcomes (pain disorder, separation anxiety and significant school avoidance). The analysis revealed no gains to the Short Form Mood and Feelings questionnaire, but positive and significant increases to life satisfaction. Overall, HM had significant and augmenting effects in four out of eight of these secondary outcomes.

The study has a number of limitations. From the beginning recruitment and retention were problematic. This meant that the original intention of a balanced cohort of similar schools randomised across the treatment and control groups was not attained. While generally the study schools were serving pupils with above-average deprivation as intended, the mix of schools (Academy, Academy convertor, Community, etc) was not balanced across the two groups. Recruitment also had to be rolled out over two phases. It was not possible to roll-out purely on a wait-list control basis, and Phase 2 schools were split into treatment and (pure) control schools. The high rates of drop-out affected retention. A claw-back mechanism was implemented to ensure that there was adequate power to ascertain whether or not the HM course was an improvement over non-standardised PHSE. Overall the problems of retention may reflect the fact that PHSE is not given adequate direction and, as noted by the recent House of Commons Education Committee (2015) report there is a sense of profound purposelessness in the teaching of this subject matter; it is often seen as an add-on rather than a central feature and is sometimes even dropped from a school's timetabling. Under such circumstances it is perhaps not surprising that retention for a trial was difficult. This does not necessarily mean that with appropriate authority that the HM course could not be scaled up and rolled-out successfully.

The evaluation of health related quality of life (HRQoL) in childhood, the multi-dimensional concept that includes domains related to physical, mental, emotional, and social functioning, is not straightforward. It is not possible to assess outcomes within a single measure. Consequently, we used a range of measures assess the impact of the intervention. We have avoided arbitrary aggregation of domains by relying on a validated instrument, the CHQ-CF87 as our preferred measure. This questionnaire provided the primary outcome, a general health score, and a number of secondary outcomes that we presented as representing categories capturing domains of health, internalising behaviour and externalising behaviour. We also considered the mood and feelings questionnaire, a life satisfaction ladder and an anxiety questionnaire (SCARED). It is not clear how all these secondary outcomes relate to each other, and given that they are based on three different questionnaires we have analysed them separately. Comparison across these instruments remains a challenge.

The overall conclusion, is that the HM course showed statistically significant improvements across a full range of HRQoL, behavioural and emotional well-being. The trend across a wider range of study measures, particularly when comparing baseline and end line data, is supportive of this conclusion. Moreover, these benefits are achieved at low cost, and would appear able to be scaled up as teaching is based on a small degree of additional training. This finding comes at a time when PHSE is being considered as a compulsory subject within the course.

References

- Benjamini, Y. and Hochberg, Y., (1995), Controlling the False Discovery Rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B (Methodological)*, 57, 289-300
- Black, S., Devereux, P., and Salvanes, K., (2007), From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes, *The Quarterly Journal of Economics*, 122(1): 409-439.
- Bond, L. A., & Hauf, A. M. C. (2004), Taking stock and putting stock in primary prevention: Characteristics of effective programs, *J. of Primary Prevention*, 24:199–221.
- Bonell, C., Fletcher, A. & McCambridge, J., (2007), Improving school ethos may reduce substance misuse and teenage pregnancy, *British Medical Journal*, 334:614–16.
- Brunwasser SM, Gillham JE, and Kim ES., (2009), A meta-analytic review of the Penn Resiliency Program's effect on depressive symptoms, *J Consult Clin Psychol.* 77(6):1042-54.
- Challen, A., Machin, S , Noden, P., & West, A., (2011), *UK Resilience Programme Evaluation: Final Report*. Department for Education, Research Report DFE-RR097 .
- Coleman, J., Hale, D. & Layard, R., (2011), *A Model for the Delivery of Evidence-Based PSHE (Personal Wellbeing) in Secondary Schools*, Centre for Economic Performance DP 1071. London, LSE.
- Collishaw, S., Maughan, B., Goodman, R. & Pickles, A., (2004), Time trends in adolescent mental health. *J. of Child Psychology and Psychiatry*, 45(8):1350-62.
- Conti, G. and Heckman, J., (2016), The Economics of Child Well-Being, IZA Discussion Paper, 6930.
- Conti, G. and Heckman, J., (2016), The effects of two influential early childhood interventions on health and behaviour, *Economic Journal*, 126, F28-F65
- Currie J., (2009), Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development. *Journal of Economic Literature*, 47(1): 87-122.
- Currie J., (2011), Inequality at Birth: Some Causes and Consequences. *American Economic Review*, 101(3): 1-22.
- Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, 82(1):405-32.
- Durlak, J. A., Weissberg, R. P., & Pachan, M., (2010), A meta-analysis of after-school programs that seek to promote personal and social skills in children and adolescents. *American J. of Community Psychology*, 45:294–309.
- Formby, E., Coldwell, M., Stiell, B., Demack, S., Stevens, A., Shipton, L., et al., (2011), *Personal, Social, Health and Economic (PSHE) Education: A mapping study of the prevalent models of delivery and their effectiveness*. Department for Education.
- Goodman, A., R. Joyce, & J. P. Smith, (2011), The long shadow cast by childhood physical and mental problems on adult life. *PNAS*, 108(15):6032-37.
- Hattie, J. A. C. (2011). Which strategies best enhance teaching and learning in higher education? In D. Mashek & E. Hammer (Eds.), *Empirical research in teaching and learning: Contributions from social psychology* (pp. 130–142).
- Hattie, J. (2015). The applicability of Visible Learning to higher education. *Scholarship of Teaching and Learning in Psychology*, 1(1), 79-91.
- Hattie, J., (2018), Visible-learning web site, <https://visible-learning.org/hattie-ranking-influences-effect-sizes-learning-achievement/> (accessed November 2018)
- HealthActCHQ, (2015), *The CHQ Scoring and Interpretation Manual*, Boston:MA
- Heckman, J. J. and Kautz, T., (2014a), Fostering and measuring skills: Interventions that improve character and cognition. In J. J. Heckman, J. E. Humphries, and T. Kautz (Eds.),

- The Myth of Achievement Tests: The GED and the Role of Character in American Life*, pp. 341–430. Chicago, IL: University of Chicago Press.
- Heckman, J. J. and Kautz, T., (2014b), *Fostering and measuring skills: Interventions that improve character and cognition*. Technical report, IZA Discussion Paper No. 7750.
- Heckman, J.E. Humphries, and T. Kautz (eds.), *The Myth of Achievement Tests: The GED and the Role of Character in American Life*, Chicago: University of Chicago Press,
- Heckman, J. J., R. Pinto, and P. A. Savelyev, (2013), Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review* 103 (6), 1–35.
- Heckman, J. J., J. E. Humphries, and G. Veramendi (2014). Education, health and wages. Unpublished manuscript, University of Chicago, Department of Economics
- House of Commons Education Committee, (2015), *Life lessons: PHSE and SRE in schools*. Fifth Report of Session 2014-15. HC 145. London: HMSO
- Humphrey, N., Lendrum, A., & Wigelsworth, M., (2010), *Social and emotional aspects of learning (SEAL) programme in secondary schools: National evaluation*. Department for Education.
- Kautz, T. Heckman, J., Diris, R., et al., (2014), *Fostering and Measuring Skill: Improving cognitive and non-cognitive skills to promote lifetime success*, NBER Working Paper, 20749, Cambridge, NBER (also published as an OECD Report under the same name).
- Klein JB, Jacobs RH, Reinecke MA., (2007), Cognitive-behavioral therapy for adolescent depression: a meta-analytic investigation of changes in effect-size estimates. *J Am Acad Child Adolesc Psychiatry*, 46, 1403–13
- Lleras, C., (2008), Do skills and behaviors in high school matter? The contribution of noncognitive factors in explaining differences in educational attainment and earnings. *Social Science Research* 37 (3), 888–902.
- Schmidt, L. J., Garratt, A.M., and Fitzpatrick, R. (2002) Child/parent-assessed population health outcome measures: a structured review, *Child: Care, Health and Development*, 28, 227-237
- Silk JS, Vanderbilt-Adriance E, Shaw DS, et al., (2007), Resilience among children and adolescents at risk for depression: mediation and moderation across social and neurobiological contexts. *Dev Psychopathol*, 19, 841–65.
- Thapar,A., Collishaw, S., Pine, D. and Thapar, A, (2012), Depression in adolescence, *Lancet*, 379, 1056-1067
- Weisz JR, McCarty CA, Valeri SM., (2006), Effects of psychotherapy for depression in children and adolescents: a meta-analysis. *Psychol Bull*, 132, 132–49:

Table 1 Appendix: Effect size estimation for Additional Secondary Outcomes

Adjusted differences in means												
Secondary outcomes:	Physic Funct	Emotional Diff	Behav. di	Self-Esteem	Phy. Difficul	Pain & Discomfort	General Behav	Global behaviour	Mental Health	General health	Family Activities	Family Cohesion
Baseline	0.162 (0.128)	0.052 (0.086)	0.123* (0.067)	-0.040 (0.078)	0.297*** (0.068)	0.206* (0.116)	0.145** (0.061)	0.139 (0.143)	0.067 (0.052)	0.151** (0.070)	0.118* (0.062)	0.121 (0.101)
N	7326	7300	7278	7272	7209	7212	7030	7051	6429	7277	7177	7189
Plus School Effects	0.125 (0.111)	0.049 (0.050)	0.058 (0.047)	- 0.168*** (0.059)	0.289*** (0.079)	0.238*** (0.053)	0.137** (0.056)	0.045 (0.056)	0.038 (0.075)	0.143** (0.067)	0.184* (0.096)	0.244*** (0.066)
N	7300	7278	7272	7290	7290	7209	7212	7030	7051	7277	7177	7189
Plus pupil Effects	-0.001 (0.056)	0.003 (0.071)	0.031 (0.075)	-0.141** (0.067)	0.115* (0.067)	0.208** (0.096)	-0.047 (0.078)	0.296*** (0.078)	-0.006 (0.081)	0.018 (0.072)	-0.054 (0.084)	0.095 (0.080)
N	6151	6135	6133	6145	6089	6094	5933	5950	5356	6137	6077	6080
Plus school & pupil Controls	0.131 (0.104)	0.076 (0.102)	0.075 (0.102)	-0.127 (0.098)	0.286*** (0.103)	0.217** (0.098)	0.150 (0.102)	0.034 (0.102)	0.099 (0.105)	0.169* (0.100)	0.196* (0.101)	0.226** (0.100)
N	7300	7278	7272	7290	7209	7212	7030	7051	6429	7277	7177	7189
Compliance	0.159 (0.135)	-0.022 (0.097)	0.056 (0.080)	-0.023 (0.097)	0.263*** (0.083)	0.250* (0.129)	0.176** (0.073)	0.195 (0.147)	0.031 (0.079)	0.150* (0.087)	0.079 (0.073)	0.102 (0.106)
N	5508	5488	5484	5505	5434	5450	5311	5327	4813	5490	5415	5424
Compliance with FE	0.204** (0.090)	0.043 (0.039)	0.075 (0.044)	-0.102 (0.068)	0.323*** (0.066)	0.212*** (0.060)	0.097* (0.052)	0.101* (0.055)	0.022 (0.074)	0.188** (0.068)	0.137 (0.082)	0.247*** (0.087)
N	5533	5508	5488	5484	5505	5434	5450	5311	5327	4813	5490	5415
Balance	0.124 (0.146)	0.037 (0.089)	0.109 (0.065)	0.027 (0.073)	0.307*** (0.065)	0.222 (0.131)	0.107** (0.051)	0.118 (0.137)	0.307*** (0.065)	0.211*** (0.066)	0.133** (0.057)	0.085 (0.106)
N	5943	5943	5943	5943	5943	5943	5943	5943	5943	5943	5943	5943
“leave me out” mean effect	0.056 (0.124)	0.004 (0.088)	0.077 (0.064)	-0.082 (0.077)	0.247*** (0.058)	0.254** (0.108)	0.144* (0.072)	0.140 (0.135)	0.063 (0.043)	0.161** (0.066)	0.125*** (0.039)	0.171 (0.106)
N	7324	7291	7313	7301	7224	7212	7051	7051	6429	7289	7221	7189
Mean imputation)	0.152 (0.129)	0.049 (0.086)	0.115 (0.067)	-0.038 (0.078)	0.289*** (0.066)	0.206* (0.116)	0.137** (0.062)	0.140 (0.135)	0.067 (0.052)	0.151** (0.071)	0.106 (0.064)	0.121 (0.101)
N	7324	7291	7313	7301	7224	7212	7051	7051	6429	7289	7221	7189
D&D (interim data)	0.179 (0.202)	-0.135* (0.078)	-0.093 (0.072)	-0.214* (0.115)	0.141 (0.096)	0.272*** (0.069)	-0.180* (0.090)	0.087 (0.131)	-0.192** (0.078)	-0.098 (0.120)	0.133 (0.086)	0.165 (0.116)
N	5803	5782	5775	5797	5722	5744	5564	5588	4465	5783	5690	5695




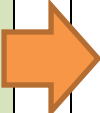



Notes: Higher values imply better health. All Outcomes are standardised to have a mean of 0 and standard deviation of 1, so effect Standard Errors are clustered at the school level. * * * * * denotes significance using standard t testing. **Bold** font indicates the treatment effect is significant at the 5% level of significance after the Benjamini, Y. and Hochberg, Y. (1995) multiple comparison correction. **Bold and Italic** indicates the treatment effect is significant at the 10% level using the same correction. N is always the actual observations in the panel.

Appendix A: EEF cost rating

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

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

Appendix B: Security classification of trial findings

Rating	Criteria for rating			Initial score		Adjust		Final score
	Design	Power	Attrition*					
5 	Well conducted experimental design with appropriate analysis	MDES < 0.2	0-10%			Adjustment for Balance [0]		
4 	Fair and clear quasi-experimental design for comparison (e.g. RDD) with appropriate analysis, or experimental design with minor concerns about validity	MDES < 0.3	11-20%					
3 	Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity	MDES < 0.4	21-30%	3				3
2 	Weakly matched comparison or experimental design with major flaws	MDES < 0.5	31-40%			Adjustment for threats to internal validity [0]		
1 	Comparison group with poor or no matching (E.g. volunteer versus others)	MDES < 0.6	41-50%					
0 	No comparator	MDES > 0.6	>50%					

- **Initial padlock score:** lowest of the three ratings for design, power and attrition. This study was a well conducted randomised trial designed to achieve a MDES of 0.28 and with a total pupil-attrition of 23%= [3] padlocks
- **Reason for adjustment for balance** (if made): Baseline variables presented good balance with a difference of ES=-0.009 in the primary outcome
- **Reason for adjustment for threats to validity** (if made): No threats to validity were reported
- **Final padlock score:** initial score adjusted for balance and internal validity = 3 padlocks

Appendix C: Effect size estimation for Additional Secondary Outcomes

Adjusted differences in means												
Secondary outcomes:	Physic Funct	Emotional Diff	Behav. di	Self-Esteem	Phy. Difficul	Pain & Discomfort	General Behav	Global behaviour	Mental Health	General health	Family Activities	Family Cohesion
Baseline	0.162 (0.128)	0.052 (0.086)	0.123* (0.067)	-0.040 (0.078)	0.297*** (0.068)	0.206* (0.116)	0.145** (0.061)	0.139 (0.143)	0.067 (0.052)	0.151** (0.070)	0.118* (0.062)	0.121 (0.101)
N	7326	7300	7278	7272	7209	7212	7030	7051	6429	7277	7177	7189
Plus School Effects	0.125 (0.111)	0.049 (0.050)	0.058 (0.047)	- 0.168*** (0.059)	0.289*** (0.079)	0.238*** (0.053)	0.137** (0.056)	0.045 (0.056)	0.038 (0.075)	0.143** (0.067)	0.184* (0.096)	0.244*** (0.066)
N	7300	7278	7272	7290	7290	7209	7212	7030	7051	7277	7177	7189
Plus pupil Effects	-0.001 (0.056)	0.003 (0.071)	0.031 (0.075)	-0.141** (0.067)	0.115* (0.067)	0.208** (0.096)	-0.047 (0.078)	0.296*** (0.078)	-0.006 (0.081)	0.018 (0.072)	-0.054 (0.084)	0.095 (0.080)
N	6151	6135	6133	6145	6089	6094	5933	5950	5356	6137	6077	6080
Plus school & pupil Controls	0.131 (0.104)	0.076 (0.102)	0.075 (0.102)	-0.127 (0.098)	0.286*** (0.103)	0.217** (0.098)	0.150 (0.102)	0.034 (0.102)	0.099 (0.105)	0.169* (0.100)	0.196* (0.101)	0.226** (0.100)
N	7300	7278	7272	7290	7209	7212	7030	7051	6429	7277	7177	7189
Compliance	0.159 (0.135)	-0.022 (0.097)	0.056 (0.080)	-0.023 (0.097)	0.263*** (0.083)	0.250* (0.129)	0.176** (0.073)	0.195 (0.147)	0.031 (0.079)	0.150* (0.087)	0.079 (0.073)	0.102 (0.106)
N	5508	5488	5484	5505	5434	5450	5311	5327	4813	5490	5415	5424
Compliance with FE	0.204** (0.090)	0.043 (0.039)	0.075 (0.044)	-0.102 (0.068)	0.323*** (0.066)	0.212*** (0.060)	0.097* (0.052)	0.101* (0.055)	0.022 (0.074)	0.188** (0.068)	0.137 (0.082)	0.247*** (0.087)
N	5533	5508	5488	5484	5505	5434	5450	5311	5327	4813	5490	5415
Balance	0.124 (0.146)	0.037 (0.089)	0.109 (0.065)	0.027 (0.073)	0.307*** (0.065)	0.222 (0.131)	0.107** (0.051)	0.118 (0.137)	0.307*** (0.065)	0.211*** (0.066)	0.133** (0.057)	0.085 (0.106)
N	5943	5943	5943	5943	5943	5943	5943	5943	5943	5943	5943	5943
“leave me out” mean effect	0.056 (0.124)	0.004 (0.088)	0.077 (0.064)	-0.082 (0.077)	0.247*** (0.058)	0.254** (0.108)	0.144* (0.072)	0.140 (0.135)	0.063 (0.043)	0.161** (0.066)	0.125*** (0.039)	0.171 (0.106)
N	7324	7291	7313	7301	7224	7212	7051	7051	6429	7289	7221	7189
Mean imputation)	0.152 (0.129)	0.049 (0.086)	0.115 (0.067)	-0.038 (0.078)	0.289*** (0.066)	0.206* (0.116)	0.137** (0.062)	0.140 (0.135)	0.067 (0.052)	0.151** (0.071)	0.106 (0.064)	0.121 (0.101)
N	7324	7291	7313	7301	7224	7212	7051	7051	6429	7289	7221	7189
D&D (interim data)	0.179 (0.202)	-0.135* (0.078)	-0.093 (0.072)	-0.214* (0.115)	0.141 (0.096)	0.272*** (0.069)	-0.180* (0.090)	0.087 (0.131)	-0.192** (0.078)	-0.098 (0.120)	0.133 (0.086)	0.165 (0.116)
N	5803	5782	5775	5797	5722	5744	5564	5588	4465	5783	5690	5695

Notes: Higher values imply better health. All Outcomes are standardised to have a mean of 0 and standard deviation of 1, so effect Standard Errors are clustered at the school level. * ** *** denotes significance using standard t testing. **Bold** font indicates the treatment effect is significant at the 5% level of significance after the Benjamini, Y. and Hochberg, Y. (1995) multiple comparison correction. **Bold and Italic** indicates the treatment effect is significant at the 10% level using the same correction. N is always the actual observations in the panel.

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

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

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

This document is available for download at www.educationendowmentfoundation.org.uk



Education
Endowment
Foundation

The Education Endowment Foundation

9th Floor, Millbank Tower

21–24 Millbank

London

SW1P 4QP

www.educationendowmentfoundation.org.uk