Statistical Analysis Plan
for Developing Healthy Minds in Teenagers

London School of Economics and Political Science
Dr. Alistair McGuire, Dr. Grace Lordan and Prof. Richard Layard

Template last updated: March 2018

PROJECT TITLE	**Healthy Minds**
DEVELOPER (INSTITUTION)	How to Thrive
EVALUATOR (INSTITUTION)	London School of Economics and Political Science
PRINCIPAL INVESTIGATOR(S)	Professor Alistair McGuire, Dr Grace Lordan and Professor Richard Layard (all LSE)
TRIAL (CHIEF) STATISTICIAN	Dr Grace Lordan, LSE
SAP AUTHOR(S)	Professor Alistair McGuire, LSE
TRIAL REGISTRATION NUMBER	n.a.
EVALUATION PROTOCOL URL OR HYPERLINK	https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/developing-healthy-minds-in-teenagers/

Table of Contents

SAP version history

VERSION	DATE	REASON FOR REVISION
1.1 [*latest*]	18 July 2018	EEF request for a SAP prior to analysis
1.0 [*original*]		The EEF have agreed to fund the final 2 years of this 6 year study. The initial 4 years of the study have been funded from various sources including internal LSE research funds; the Templeton Education and Charity Trust; and the Rosetrees Trust. As such there were internal documents prepared for the original randomisation process, sample size calculations and future statistical analysis plan for this study, but none prepared for, or explicitly designed for the EEF at the time of randomisation as they were not funding body for this study at the time of initiation. They are, therefore, one of a number of funding organisations who have contributed to the study. Niether was a protocol for this specific study submitted to the EEF to secure the final stage funding (An application for funding reference number 3793 was approved based on the information detailed therein November 2016).

Introduction

This study is evaluating whether an evidence-based life skills curriculum (Developing Healthy Minds in Teenagers), within the Personal, Social, and Health Education (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 this curriculum can improve teenagers' emotional well-being. A secondary aim is to establish whether the curriculum also improves soft skills, including mental health, compared to the usual taught PHSE curriculum. We therefore term the study to be "character and well-being" study. It is a unique non-attainment (in terms of educational attainment) study based at assessing improvements in individual character, well-being and soft skills.

The purpose of the trial is to assess the curriculum and training package of an 14-module taught package as a complete whole.

The 14-module package, consists of individual elements which have been separately evaluated through various controlled trials and studies to be successful and are defined as:

- Penn Resilience Programme
- .breathe (Mindfulness)
- Media Navigator
- From School to Life
- Unplugged (Part 1 and 2)
- Media Influences
- Resilience Revisted
- Sex Ed Sorted (Part 1 and 2)

- Relationship Smarts Plus
- School Health Alcohol Harm Reduction Progarmme (SHAHRP)
- Resilient Decisions
- Mental Illness Investigated
- Parents Under Construction
- Resilient Learners

The programme has been taught to pupils in the intervention schools during a 140-hour universal programme delivered over the first 4 years of secondary school using one hour-a-week of the timetabled lessons (PSHE slot where timetabled) and taught by school staff, (teachers, learning support assistants, who have received full training in each module), covering social and emotional learning, relationships and healthy living amongst pupils in mainstream secondary schools.

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 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 maths; percentage of pupils gaining at least 5 GCSEs at C or better including English and maths; 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 similar projects, this left 174 schools scoring 22 and above, which 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 there was substantial drop-out and the final number of 36 schools willing to participate, included a matching set of 4 schools from the Wolverhampton area. In addition a small special school requested and was approved by EEF to participate.

The study is a cluster randomised trial, with school level randomisation. Randomisation was conducted using minimisation and schools were stratified 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 deprevation. It was largely pragmatic as school opt-in determined the final recruited sample. Consequently no balance analysis was undertaken at initiation.

The intention was to recruit all 30 – 34 schools. Half of the schools randomised with a treatment year group starting in school in September 2013 and half with the same year acting as the control group. The control schools would then provide the treatment to the follow year group starting in September 2014, a wait-list control. However school recruitment proved difficult as it started in January 2013 proving too late in the school planning cycle. So school recruitment has taken place in two phases with the first wave initiating the intervention in September 2013, including the wait list control year groups and the second wave initiating the intervention or providng a control year group from September 2014.

Assessments have been carried out, through questionnaires, at baseline (September 2013 or 2014), 9 months (June 2014 or 2015), 21 months (June 2015 or 2016), 33 months (June 2016 or 2017), and 42 months (June 2017 with the final questionnaires delivered during 2018). Data are held by the data collection team, (an independent, from the LSE, firm (HcareSolutions)) coded through the use of a unique (anonymised) pupil identifier and will be released to the LSE statistical analysts by the end of June 2018, conditional on all schools having had assessments undertaken by that date. The data collection team are an independent company (HCare Solutions) contracted by LSE to issue the questionnaires, collect and code the data annually. Ensuring pupil anonyminity, but retaining linkage within the longitudinal data set.

A parallel, but distinct study, using the teaching intervention and assessing academic achievement was funded separately by the EEF and will be analysed separately. The academic study is being conducted by NIESER. NIESER have no involvement in the chararacter and wellbeing study. LSE have offered, under Professor Alistair Mcguire, to comment on the the academic's study methodological approach. This has been taken up in the past, but is opt in for NIESER going forward. Both of these studies will be produce separate reports for the EEF.

Design overview

Trial type and number of arms		**Two-arm, cluster randomised**
Unit of randomisation		School
Minimisation variables (if applicable)		Proportion FSM, GCSE Grades, and Single vs Mixed Sex School.
Primary outcome	Variable	Improving individual well-being, as measured by change in individual health
	measure (instrument, scale)	General health dimension on the CHQ-CF87 scale

Secondary outcome(s)	variable(s)	Improving individual “character and non-cognitive skills”
	measure(s) (instrument, scale)	CHQ-CF87 scale, The Short Mood and Feelings Questionnaire, the Child Anxiety Related Disorders questionnaire, and a general health scaling

Sample size calculations overview

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
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		15	
	control	15		15	
	total	30		30	
Number of pupils	intervention	1500		2589	
	control	1500		1711	
	total	3000		4300	

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 “character and well-being” 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. DFE). Repeated measures were not used to adjusted for correlations across the measures as primary results were initially to be based on baseline and final questionnaire returned outcome measures. 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. To allow for drop-out of schools over the four-

year period of follow-up, pupil attrition and parental consent withdrawal we aimed to recruit 30 schools, which would allow detection of an effect size of 0.28.

The study faced recruitment difficulties from the beginning. In particular, it proved difficult to recruit schools for a complex 4 year study involving a regular slot in their timetable for 'soft skills'. Initially the plan was to recruit 30 – 34 schools and we recruited 36 plus the small special school. Although the sample calculations had been undertaken on a total number of 30. On randomisation we lost 2 schools (who were randomised as control), and a further 1 school early on so the total dropped to 33. Over time we lost a further 6 schools (who were unable to maintain the teaching commitment) to interim data collection but have retained all schools for the final administration of the questionnaire. As of the beginning of 2018 there are 17 treatment and 17 control schools who have agreed collection, based on these schools, the MDES is 0.29. However a number of treatment schools, 4 in total, stopped the treatment part way through although they have administered the final questionnaire, allowing baseline and final data collection in 34 schools plus the small special school that will be looked at seperately. Primary analysis will remain based on an intent-to-treat design. Post-hoc MDES calculations with the analytical sample will be included in the report.

As a result of recruitment difficulties and differences between schools, there was imbalance in the number of pupils in the trial, with a smaller than anticipated number of pupils in the control arm.

EEF Statistical Analysis Guidance requires 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.

Analysis

Primary outcome analysis

The parallel study assessing academic achievement, which was fully funded by the EEF, as associated with the intervention is detailed separately, subject to it's own SAP and will be analysed independently from the "character and well-being" study. The academic achievement analysis is *NOT* dealt with here.

For the "character and well-being" study at the time of randomisation the primary outcome was defined as change in the General Health single item score embedded in the Child Health Questionnaire (CHQ-CF87), a self-report health measure designed for young people aged 10 to 18 (CHQ, 2013). This questionnaire has been validated for use in the UK (see Schmidt LJ, Garratt AM, Fitzpatrick R. 2001). 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) in each school with the first group completing in 2017 and the second in 2018.

The primary outcome was based on a change in general health as this provides an overall aggregate measure of well-being. Sample size was estimated on the basis of this.

The primary analysis will be based on the following basic difference-in-difference using an ITT specification:

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

where:

y_{it} = the outcome variable (CHQ-CF87 General Health score for each year)

Equation (1) will be estimated with all baseline, interim and endline data; and then separately for baseline and each year of data collection during treatment. The intuition here comes from the curriculm building up over time, so we expect β_3 to be the most substantive when we consider the baseline and endline data. This β_3 also represents the total effect of the program.

$treatments$ = 1 if a school was chosen for treatment, regardless of whether they took up the treatment

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

β_3 is then the effect of being assigned the treatment, with no account for compliance i.e. an individual is treated if they are compliers or never takers. We note that there are no always takers in this analysis making this interpretatble as an intention to treat effect. This is arguably the effect policy makers care about the most, as if the program is rolled out there will be heterogeneity in how the program is rolled out at the school level.

Estimation will be through use of Stata (version 15), as well as R. Standard errors will be adjusted to allow for clustering at the school level and unknown heterogeneity (double HAC standard errors in Stata).

This basic analysis will form the basis of a common element of analysis running through into the secondary analysis. A number of further model specifications will be undertaken within the analysis, to include robustness checks and control variables, as detailed below.

Secondary outcome analysis

It is well recognised that character and well-being cannot be assessed within a single measured outcome (Conti and Heckman, 2012; Decancq and Neuman, 2014; Khanemann and Krueger, 2006). This partly dictated why the primary outcome measure relied on a General Health score, which formed the calculation for the sample size. It is also the rationale behind the collection of a number of secondary outcomes: some based on the other measures embedded within the CHQ-CF87 (ten plus two single item questions), as well as the The Short Mood and Feelings Questionnaire (Angold and Costello, 1987) the Child Anxiety Related Disorders

(SCARED) questionnaire (Birmaher et al, 1999), and a general health scaling (based on the visual analogue associated with the EQ5D (EuroQol Group, 1990)). This gives fifteen additional measures which each have many associated sub questions. In order to allow for the various dimensions of character and wellbeing and circumvent the multiple comparisons problem in statistical analysis, we will utilise exploratory factor analysis on the sub questions of each item, and extract the underlying orthogonal factors that represent the independent dimensions of character that we capture with the included instruments. The labels applied to these factors will be intuitive to the items loading on them, and we expect them to follow the themes that these validated instruments set out to gather. The estimation applied to these latent factors will follow equation (1) above. We will also supply in an appendix an analysis following equation (1), which draws on the instruments outcomes as defined by their authors. We will clearly sign post the issue of the multiple comparison problem and its impact on standard errors.

We note specifically that the CHQ-CF87 87 items measure physical and psychosocial health, divided across 10 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. Two further single-item questions are also asked on change in general health and family cohesion. The CHQ-CF87 is designed for young people aged 10 to 18 and has been found to be reliable and sensitive to change for this age range. The questionnaire is suitable for, and has been validated within a school context and takes a maximum of 20 minutes to complete.

The CHQ-CF87, the Short Mood and Feelings Questionnaire, the Child Anxiety Related Disorders questionnaire, and the general health (EQ5D) scaling were all administered through the same paper-based questionnaire given to pupils by the coding team, who are distinct from the analysis team, during a class setting. 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 and recorded data within Excel spreadsheets, which will be released to the analysts in full at the end of the data collection period (summer 2018).

Additional analyses

A set of additional specifications will be used for robustness. These are:

1. A robustness check will consider a more saturated version of equation 1 and add a number of pupil and school level control variables For example size of class/respondents, %female in class/respondents, %free school meals). We will also consider robustness to the addition of i) school 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) and ii) pupil fixed effects
2. 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*.

Imbalance at baseline

Although originally estimated upon a 50-50 split of control and treatment schools, the under- (37%) and over- (63%) recruitment of control and treatment school respectively is not expected to affect the ability to detect the MES at given levels of significance and power. A balance table will be included to show the balance of characteristics across the treatment and control populations at baseline, and for each subsequent year of questionnaire administration.

Missing data

Missing data will be initially assumed to be missing at random (MAR) and altered through an inverse weight probability of not being missing in all specifications, apart from the robustness specification where peer effects are considered. Inverse probability weighting is a preferred method to apply to missing data when the incomplete cases provide little information, as is likely to be the case in our study (Seaman and White, 2011). Here a logistic regression based on whether the variable is missing or not will be regressed on a number of explanatory variables that are fully observed (to be determined once the data are “released” in the summer of 2018). From this fitted model the predicted probability for each pupil with similar characteristics (i.e. values on the explanatory variable) for the missing variable will be returned and used in a “fully” populated model. The results form this “fully” populated model will be compared to the individual-deletion model, where individuals with missing covariate data are deleted from the regression analysis.

Further robustness analysis of missing data will also be undertaken based on standard multiple imputation techniques (Rubin, 1978) where the missing values are imputed based on the predictive distribution of the missing values estimated from a regression of the variable(s) with missing values against the variables without missing values and the dependent variable. Additional explanatory variables that are fully observed (to be determined once the data are “released” in the summer of 2018) may also be used to define the predictive distribution.The coefficients from this regression are then used, in conjunction with the non-missing values to generate predicted values for the missing data. This distribution of predicted values is used to produce a complete data set which can then be analysed using statistical methods for complete data. Further, this approach will draw a number of values from the predictive distribution of the missing values and then the complete data analysis are repeated N times once with each imputation substituted. The final estimate of the missing data parameter is then the sum of the parameter values obtained for each imputation divided by N and the variance attributed to this variable is gained from the variation in the N parameter values. There is no authoritative recommendation in the

literature on the best way to acoount for clustered data when imputing missing values. We propose to include cluster indicators in our imputation model. This appears to be the most efficient solution when working with few clusters and many observations, as in our case (Graham, 2009). Neither is there authoritative guidance on the number of imputations to perform. Rubin (2008) suggests 10 imputations will address issues of point estimate efficiency, but clustering would suggest further imputation is required. Parameter replication also suggests further imputation. Again there is no agreement on the optimal number of imputations to be undertaken, the precise number in our analysis will be determined through a 2-stage procedure suggested by Hippel (2018).

Compliance

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:

$$y_{ist} = \beta_0 + \beta_1 delivery \ + \beta_2 year_t + \beta_3 delivery * year_{ts} + \epsilon_{ist} \qquad (2)$$

In equation 1 $delivery$=1 if a school delivered the program to the satisfaction of the How to Thrive team. That is, the How to Thrive team have a record of the school completing the curriculm in full up to the particular year of enquiry. Delivery is then equal to 0 if a school was a control. We note that this is a per-protocol analysis, and differs from (1) only in that we code as 0 the schools who did not deliver the program to a satisfactory standard. 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

Intra-cluster correlations (ICCs)

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:

$$\mathbf{1/C\Sigma_{S=1}^{S}\Sigma_{t=1}^{T}(y_{ts} - \overline{y})^2 = \frac{1}{C}\Sigma_{s=1}^{S}\Sigma_{t=1}^{T}(y_{ts} - \overline{y_s})^2 + \frac{1}{C}\Sigma_{s=1}^{S}(y_s - \overline{y})^2}$$

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 calculation

The effect size will be returned as $\boldsymbol{\beta_3}$ from equation (1). We will 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.

References

Adrian Angold & Elizabeth J. Costello, 1987; Developmental Epidemiology Program; Duke University

Birmaher, B., Brent, D.A., Chiappetta, L., Bridge, J., Monga, S., & Baugher, M. (1999). Psychometric properties of the Screen for Child Anxiety Related Emotional Disorders (SCARED): A replication study. Journal of the American Academy of Child and Adolescent Psychiatry, 38 (10),1230-6.

Challen, A., Machin, S , Noden, P., & West, A. (2011), UK Resilience Programme Evaluation: Final Report. Department for Education, Research Report DFE-RR097 .

CHQ, (2013), CHQ-CF87 Scoring and Interpretation Manual, CHQ, Boston, USA

Conti, G. and Heckman, J., 2012, The Economics of Child Well-Being, IZA Discussion Paper No. 6930, Insitute for Labor Studies, Germany

Decancq, K. and Neumann, D., (2014), Does the choice of well-being measure matter empirically? An illustration with German data, IZA Discussion Paper No. 8589, Insitute for Labor Studies, Germany

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.

EuroQol Group, (1999), EuroQol – a new facility for the measurement of health-related quality of life, Health Policy, 16, 199-208

Graham, J. W., (2009), Missing data analysis: Making it work in the real world. Annual Review of Psychology, 60, 549–576.

Hippel, P., (2018), How many imputations do you need? A two-stage calculation using a quadratic rule?, Sociological Methods and Research, forthcoming

Khaneman, D., and Krueger, A., (2006), Developments in the measurement of subjective well-being, Journal of Economic Perspectives, 20, 3-24

Landgraft, J, Abetz, L., Ware, J. ,(1996) The CHQ User's Manual. Boston: The Health Institute, New England Medical Centre

Rubin, D., (1978), Multiple Imputation for Nonresponse Surveys, John Wiley, New York.

Schmidt LJ, Garratt AM, Fitzpatrick R., (2001) *Instruments for Children and Adolescents: a Review*, Report from the Patient-reported Health Instruments Group Programme to the Department of Health.

Seaman, S., and White, I., 2011, A Review of Inverse Probabiltiy Weighting for dealing with missing data", Statistical Methods in Medical Research, 22, 278-295.