INTERVENTION	**INCLUSIVE / Learning Together**
DEVELOPER	UCL
EVALUATOR	University of Manchester
TRIAL REGISTRATION NUMBER	ISRCTN10751359 (UCL main trial)
TRIAL STATISTICIAN	Patricio Tronsco & Michael Wigelsworth
TRIAL CHIEF INVESTIGATOR	Michael Wigelsworth
SAP AUTHOR	Michael Wigelsworth & Patricio Tronsco
SAP VERSION	V4
SAP VERSION DATE	15/11/17
EEF DATE OF APPROVAL	
DEVELOPER DATE OF APPROVAL	

Table of Contents

Protocol changes

Due to concerns around the integrity and completeness of the secondary outcome data (specifically Teacher estimated Key Stage 4), Stop/Go criteria have been introduced (see 'analysis').

Introduction

INCLUSIVE is a secondary school-led intervention which combines changes to the school environment with the promotion of social and emotional skills and restorative practices through: the formation of a school action group involving students and staff supported by an external facilitator to review local data on needs, determine priorities, and develop and implement an action plan for revising relevant school policies/rules and other actions to improve relationships at school and reduce aggression; staff training in restorative practices; and a new social and emotional skills curriculum. The intervention combines strong fidelity of inputs, processes and core components with the capacity for tailoring non-core components to local needs. INCLUSIVE can be described as a multi-component universal SEL intervention. Using the classification system adopted in recent major reviews in this field (e.g. Blank et al., 2010), it combines *curricular* and *environmental* components. From the perspective of Humphrey's (2013) SEL taxonomy, it may be described as a hybrid programme in terms of its prescriptiveness, offering both 'manualised' content of core components and flexible, needs-led delivery of non-core components.

Study aims

An NHIR funded major cluster-randomised control trial of INCLUSIVE is currently underway (2013-2019) and is being led by University College London (UCL). This trial is designed to examine the effectiveness and cost-effectiveness of INCLUSIVE over three school years. Outcome measures in the UCL trial focus on changes in behaviour, specifically bullying and aggression.

The role of the University of Manchester as independent evaluator is to utilise the existing trial infrastructure to determine the impact of INCLUSIVE on the academic attainment of pupils in participating schools. Specifically:

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

Study design

The independent evaluation of INCLUSIVE benefits from the very clearly defined trial protocol that is already in place for the main UCL trial. In short, 40 secondary schools in England have been recruited and randomly assigned to implement the INCLUSIVE intervention over 3 years or continue usual practice (i.e. 'business as usual') during the equivalent period of time. Schools in the intervention arm will receive technical support and assistance from external facilitators for the first 24 months. The target cohort comprises pupils in Year 8 (i.e. aged 12/13) at the outset of the intervention. Outcome measures of aggression and bullying will be taken at baseline and 36 month follow-up (for the main UCL trial), with attainment-based outcome measures at school and student levels taken at 24 (provisional) and 36 months. Attainment measures are taken at time points that align with academic assessment and/or testing periods (specifically teacher estimated Key Stage 4 results and 'attainment 8' at Key Stage (assessed during national examinations). A process evaluation examining different aspects of implementation (e.g. fidelity/adherence) will take place throughout the main trial period

In light of the discussion with the EEF and INCLUSIVE project team, Manchester will be responsible for collecting the attianment measures (attainment 8, Maths GCSE results and English GCSE results), providing analysis and authoring the EEF report. As part of the NHIR main trial, IOE will be collecting process data, which will be shared with Manchester to augment the ITT analysis (see 'on-treatment analysis').

As such, specific trial details (e.g. number of pupils) are not available until UCL release the data to Manchester. Also, as the trial is blinded to condition, Manchester will not know the details regarding allocation to condition (analysis will be run as 'group1' and 'group 2').

Randomisation

As per UCL's protocol, allocation was at school level, allocated randomly in a ratio of 1:1 to intervention and control arms. Stratification was by:

i) single sex vs. mixed sex school (dichotomous categorical)
ii) ii) school level deprivation as measured by percentage of students eligible for free school meals (low/moderate 0 to 23%; high >23%, with 23% being the median for England)
iii) iii) school contextual value-added attainment (CVA) in GCSE exams (above and below median for England of 1,000). Value added (VA) score is a school-level measure of students' attainment in public exams adjusting for their attainment on entry to the school. VA rather than Ofsted ratings for schools was used as there is better evidence for VA being associated with violence rates (Tobler, Komro, Dabroski, Aveyard, & Markham (2011).

Calculation of sample size

The current trial is adequately powered to detect effects on the primary outcome measures for which it was originally designed. However, measures of academic attainment present a quandary. First, the ICC for attainment in secondary schools (approximately 0.21[1]) is much larger than for aggression or bullying (which has been specified at 0.04). Second, the expected effect size (ES) for attainment is presumably much smaller than for aggression or bullying (which has been specified at approximately 0.23), as the former is presumably an indirect, distal outcome of the main intervention processes and effects. So we might reasonably expect an ES for attainment as low as 0.1.

Given this, the current trial is powered thus:

*Assuming N=190 per cluster, 40 clusters, ICC=0.21, Pre-Post Test Correlation=0.5, Power=0.8 and Alpha=0.05, an ES of 0.41 or larger would be detectable (using G*Power (Faul, Erdfelder, Buchner, & Lang, 2007)).*

This ES is useful benchmark as it corresponds directly to Hattie's (2009) 'hinge point' of ES = 0.4, at which, "the effects of innovation enhance achievement in such a way that we can notice real-world differences" (p.17). The trial is not powered to reliably detect effect sizes smaller than this.

[1] This figure was calculated using GCSE scores for English and Maths in the National Pupil Database (NPD).

Follow-up

Data is passed from UCL to Manchester post-hoc.

Outcome measures

Primary outcome

It is agreed that Manchester will report on attainment as the primary outcome, specifically Attainment 8 raw score for July 2018 examinations, as provided by the National Pupil Database. Coverage is based on matched pupil lists provided by UCL (see data-sharing protocol).

Secondary outcomes

- MATHS GCSE (KS4_APMAT) results for July 2018 examinations (as provided by the NPD)
- ENGLISH GCSE (KS4_APENG) results for July 2018 examinations (as provided by the NPD)

- Teacher estimated Key Stage 4 predicted scores (dependent on stop-go criteria as this is the only variable with likely large missingness) (July 2017) Maths and English (combined score) will be collected from schools directly.

Analysis

Analysis will be conducted as blinded to allocation condition.

For any of the above outcome measures, the standard procedure to be applied is as follows:

- Stop/Go criteria applied:

Analysis will only proceed once data returns match a pre-specified 'minimum acceptable quality limit', i.e. if there is 'too much' missing data at the pupil level, the analysis cannot proceed. This relates specifically to the teacher estimated Key Stage 4 predicted scores as the timing of this measure corresponds to additional data burden due to primary data collection from UCL. Objective thresholds for missing data are difficult to determine, and are dependent (in part) on the nature of the missingness (e.g. MCAR, MAR, NMAR) and the use of treatment procedures (e.g. imputation). Scheffer (2002) indicates that for Missing at Random (MAR), using imputation procedures, estimations diverge (in comparison to 'full' sets) at approximately 25% 'missingness'. Accordingly, the additional steps apply to outcome variables reaching at least 75% in returnable data are:

- Data cleaning and screening ahead in preparation for analysis
- Basic descriptive analysis
    - Production of descriptive statistics by allocation group (e.g. means, standard deviations) and visual data displays (e.g. error bar charts) to identify key trends vis-à-vis trial Hypotheses.
- Demonstration of equivalence at baseline (on the basis of primary and secondary outcome), specifically no significant differences at baseline.

If returns of Key Stage 4 predicted scores do not meet minimal criteria, i.e 25% of pupil level missing data, then no inferential analysis will be used for this outcome variable. Instead, descriptive trends will be reported.

Missing data

If passing the stop/go criteria, the extent of any missing data will be established. As the primary outcome is drawn from the NPD, there will be minimal missing data. However, for the secondary data there may be missing cases dependent on the rigour of school data-records and overall compliance to requests for information. Once data are available, differences between complete and missing cases will be examined to establish any pattern to the missingness. Logistic regression will be used to predict missingness, whereby each child will be coded as providing complete (0) or incomplete (1) outcome data, with treatment allocation, outcome data and demographic variables as explanatory variables (Pampaka, Hutcheson, & Williams, 2017). This will be done to test whether the missingness found in the data is of a random nature and subsequently inform the selection of auxiliary variables for the multilevel multiple imputation procedure.

Afterwards, we will also perform an analysis using complete cases and a sensitivity analysis using multiple imputation (via the REALCOM-Impute extension to MLWin). Accordingly, multiple imputation procedures will be carried out in REALCOM-Impute, using the missing at random assumption (Carpenter, Goldstein, & Kenward, 2011). This will enable us to include both partially and completely observed cases of all schools and pupils in the analysis, thereby reducing the bias associated with attrition. The imputation model will be built using the logistic regression procedure described above and bearing in mind the main model of interest (primary ITT analysis). Therefore, the variables that will be included in the imputation models will comprise: treatment allocation, demographic variables (specifically, gender and FSM eligibility), prior attainment (KS2 scores) and the outcome variable (e.g. KS4 scores) will be entered as auxiliary variables and used to impute missing values. Following general guidelines about multilevel multiple imputation (Carpenter, Goldstein, & Kenward, 2011), REALCOM-Impute will be set to run for 5,000 iterations, with a burn-in period of 500 iterations. We will store 10 imputed datasets, allowing for 500 iterations to run between them, to ensure that they are independent. The pooling of results to obtain the final model coefficients will be done afterwards in MLwiN. Results using complete-case analysis will be then compared to the results using the multiply-imputed datasets.

Primary intention-to-treat (ITT) analysis (Key Stage 4)

An ITT analysis (according to intention-to-treat principles, e.g. ignoring noncompliance, protocol deviations and other events that take place after randomisation (Gupta, 2011)) will be conducted for the primary outcome variable (attainment 8). This will be conducted through the construction of 2-level (school, pupil) hierarchical models (random effects at the school level, utilising robust standard errors) to account for the nested nature of data using MLWin Version 2.36.

We will employ a model with Key Stage 2 scores (as a prior attainment control) and group allocation (e.g. group 1 / group 2, as allocation to condition is blinded) included at school level.

This model has the following algebraic form:

$$y_{ij} = \beta_{0ij} + \beta_1 group2_{0j} + \beta_2 KS2_{ij} + \quad \textbf{Eq. 1}$$
$$\beta_{0ij} = \beta_0 + u_{0j} + e_{ij}$$
$$where:$$
$$u_{0j} \sim N(0, \sigma_u^2)$$
$$e_{ij} \sim N(0, \sigma_e^2)$$

Where y_{ij} is the standardised outcome variable at KS4 (attainment 8) of the i-th pupil in the j-th school; β_0 is the intercept or overall average; β_1 is the effect of the allocation of the j-th school to group 2 of the intervention; β_2 is the effect of the standardised outcome $KS2_{ij}$ of the i-th pupil in the j-th school; The second line of equation corresponds to the random part of the multilevel model, where u_{0j} is the unique effect of the j-th school, which follows a normal distribution with mean 0 and variance σ_u^2; and finally e_{ij} represents pupils' heterogeneity, which is also assumed to be normally distributed with a mean of 0 and a variance σ_e^2.

Considering equation 1, the expected value of the KS4 outcome of a pupil in a school allocated to group 2 of the intervention ($\widehat{g_2} = [\hat{y}|group2 = 1]$) would be as follows:

$$E(\widehat{g_2}) = \beta_0 + \beta_1 + \beta_2 \quad \textbf{Eq. 2}$$

While in comparison, the expected value of the outcome for a pupil in a school allocated to group 1 of the intervention ($\widehat{g_1} = [\hat{y}|group2 = 0]$) would be:

$$E(\widehat{g_1}) = \beta_0 + \beta_2 \quad \textbf{Eq. 3}$$

As per specific requests from the EEF we will also employ a multilevel model which includes other co-variates that were used in the design (the randomisation factors shown above), specifically: -

i) single sex vs. mixed sex school (dichotomous categorical)
ii) ii) school level deprivation as measured by percentage of students eligible for free school meals (low/moderate 0 to 23%; high >23%, with 23% being the median for England)
iii) iii) school contextual value-added attainment (CVA) in GCSE exams (above and below median for England of 1,000).

Results obtained by both models will be compared, in terms of change of the magnitude or direction of coefficients but the second model will be considered the headline figure for the report.

This model will have the following algebraic form:

$$y_{ij} = \beta_{0ij} + \beta_1 group2_{0j} + \beta_2 KS2_{ij} \quad \textbf{Eq. 4}$$
$$+ +\beta_3 mixed_{0j} + \beta_4 highdepriv_{0j} + \beta_5 belowCVA_{0j}$$

The random part of equation 4 remains unchanged from equation 1 and hence it was suppressed for simplicity. Given that model 2 (equation 4) is built up from model 1 (equation1),

both models are nested and can be in a straightforward way, by using goodness of fit measures, such as the likelihood ratio test (deviance test), as well as the AIC (Akaike Information Criterion).

Interim analyses

IF secondary data meets stop/go criteria THEN:

The protocol for the ITT analysis will be followed prior to analysis of Key Stage 4 data.

ELSE

Preliminary analysis (descriptive statistics) are provided alongside analysis of Key Stage 4 data.

Causal Effects in the presence of Non-Compliance

Complier Average Causal Effect (CACE) analysis will be used. Compliance (fidelity) will be examined through data gathered by UCL, using a bespoke measurement tool (see appendix 1) that provides a binary score for key intervention components. Once this measure is provided, the SAP and Protocol will be amended accordingly.

Additional analyses

Manchester will also consider temporal relations between outcome variables included in both UCL's and Manchester's datasets that might help to empirically validate the intervention logic model. Specifically, do proximal changes in behaviour, specifically bullying and aggression, explain later, distal changes in academic attainment?

We note that the logic model included in the pilot trial protocol (http://www.nets.nihr.ac.uk/__data/assets/pdf_file/0017/53135/PRO-09-05-05.pdf) specifies student health rather than academic outcomes as the end point in the causal chain. Although a generic model could be applied from extant theory (see for example, CASEL, 2007) and empirical evidence (e.g. Durlak et al, 2011; Sklad et al, 2012) pertaining to the influence of SEL interventions on attainment, we feel that that this project merits the explicit, a priori development (and subsequent empirical validation) of a reworked logic model for INCLUSIVE that takes into consideration attainment as a distal outcome variable. This will allow an assessment of the likelihood of 'theory failure' in the event of null results.

Subgroup analyses

In addition, further exploratory models will be run to examine specific subgroup effects. A further model for each hypothesis subgroup will be constructed that will include the specific variable as a cross-level interaction term (e.g. FSM*Allocation group). An intervention effect at the subgroup level will be noted if the co-efficient associated with the interaction terms noted above are statistically significant. These will subsequently be converted to Hedge's *g* as per EEF reporting standards.

As per EEF's request, free school meal eligibility (everFSM; Yes/no) will be considered. Other subgroups are dependent on the results for UCLs analysis of the behavioural data. If UCL data indicates differential effects for specific subgroups (see below), Manchester will then explore the possibility for any subsequent effect on attainment, however this is included as a tentative anaylsis, given the likely low power. Possible subgroups include:

i) Gender (male / female) ii) victims of bullying (yes/no) iv) perpetrators of bullying (yes/no) v) adolescents presenting in the clinical, subclinical and normal ranges for aggression at baseline

('normal' is omitted, and interaction effects are included for 'sub-clinical' and clinical as compartors (interaction effects)).

Effect size calculation

In all cases, effect sizes will be reported using Hedge's g (Cohen's d bias corrected) and accompanied by 95% confidence intervals as per EEF specifications.

Report tables

The EEF trial report template[2] contains several tables whose structure is pre-specified. Evaluators should paste these into the SAP and populate them with their chosen variables. Templates for any tables and charts additional to those in the report template should also be specified in the SAP.

Table 1: Summary of impact on primary outcome

Group	Effect size (95% confidence interval)	Estimated months' progress	EEF security rating	EEF cost rating
Group 1 vs. Group 2				

Table 2: Timeline

Date	Activity

Table 3: Minimum detectable effect size at different stages

Stage	N [schools/pupils] (n=intervention; n=control)	Correlation between pre-test (+other covariates) & post-test	ICC	Blocking/ stratification or pair matching	Power	Alpha	MDES
Protocol							
Key Stage 3							

[2] https://educationendowmentfoundation.org.uk/evaluation/resources-centre/writing-a-research-report/

Year 10							
Analysis (Key Stage 4 data)							

Table 4: Baseline comparison

Variable	Intervention group (N=1560)		Control group (N=1525)	
School-level (continuous)	**n (missing)**	**Mean (SD)**	**n (missing)**	**Mean (SD)**
Size – number of full-time equivalent (FTE) students on roll				
Attendance – overall absence (% half days)				
FSM – proportion of students eligible for free school meals				
EAL – proportion of students speaking English as an additional language				
SEND – proportion of students with SEND				
Attainment – proportion of pupil achieving level 4+ in English and maths				
Pupil-level (categorical)	**n (missing)**	**Percentage**	**n (missing)**	**Percentage**
Sex – proportion of male students				
FSM – proportion eligible for free school meals				
EAL - proportion speaking English as an additional language				
Of those with SEND provision				
Pupil-level (continuous)	**n (missing)**	**Mean (SD)**	**n (missing)**	**Mean (SD)**
Prior attainment – Key Stage 2 KS2 Maths and reading combined (equal weighting)				

Table 5: Primary analysis

Outcome	Raw means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n in model (intervention; control)	Hedges *g* (95% CI)	*p*
KS4							
Subgroup							
free school meal eligibility							
Gender							
Victim of bullying							
Perpetrator of bullying							
Aggression: **Clinical** **Subclinical**							

References

Blank, L., Baxter, S., Goyder, L., Guillaume, L., Wilkinson, A., Chillcot, J. (2010). Promoting wellbeing by changing behaviour: a systematic literature review and narrative synthesis of the effectiveness of whole secondary school behavioural interventions. *Mental Health Review Journal, 15*, 43-53

Carpetner, J., Goldstein, H., & Kenward, M. (2011). REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types. *Journal of Statistical Software, 45*, 1-14.

Durlak, J., Weissberg, R., Dymicki, A., Taylor, R., & Schellinger, K. (2011). The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions. *Child Development, 82*, 405-432.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191

Gupta, S. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research, 2*, 109-112.

Hattie, J. (2009). *Visible Learning*. Oxon: Routledge

Humphrey, N. (2013*). Social and Emotional Learning: A Critical Appraisal*. London: Sage Publications.

Scheffer, J. (2002). Dealing with missing data*. Research Matters in the Information and Mathematics Sciences, 3*, 153-160.

Sklad, M., Diekstra, R., Ritter, M. D., Ben, J. & Gravesteijn, C. (2012), Effectiveness of school-based universal social, emotional, and behavioral programs: Do they enhance students' development in the area of skill, behavior, and adjustment? *Psychology in the Schools*, 49(9), 892–909

Tobler, A., Komro, K., Dabroski, A., Aveyard, P., & Markham, W. (2011). Preventing the link between SES and high-risk behaviors: 'value-added' education, drug use and delinquency in high-risk, urban schools. *Prevention Science, 12*, 211-221. DOI: 10.1007/s11121-011-0206-9*.*

APPENDIX 1 – UCL FIDELITY CHECKLIST

Intervention component	Aspect of component	Good differentiator?	Might be important?	Measureable?	Cut-off points
Action group meetings	6 AGMS per year	**Yes** – 7/20 held six meetings in both years 1 and 2. Better indicator in year 3 (still waiting to confirm certain quantification of qualitative data)	Yes - central to theory of change	Facilitator diaries	1 point will be given if 6 or more meetings were held as defined in protocol
	Policies/rules reviewed	**Moderate** 17/20 schools reviewed rules or policies in years 1 or 2	Yes - central to theory of change	Facilitator diaries/interviews, meeting minutes	**Binary y/n** Based on Facilitator diaries, meeting minutes, and action plans
	Implementation of locally decided actions	**Moderate** 16/20 schools implemented locally decided actions in years 1 and 2	Yes - central to theory of change	Facilitator diaries / minutes / action plans/facilitator interviews/ routine monitoring forms	**Binary y/n**
	Perceived range of students and staff in AGMs	**No** 20/20 schools had a range of student and staff.	Yes - central to theory of change	AGM survey Qs 4-5	**Binary above/below 2, indicating perceived diversity** Data will be averaged over AGM survey participants at each school and over Y1-2 and Y1-3
	AGM well led	**No** 20/20 schools reported being well led by participants	Yes - evidence from pilot, important to the theory of change	AGM survey Q10	**Binary above/below 2, indicating good leadership**

					Data will be averaged over AGM survey participants at each school and over Y1-2 and Y1-3
Curriculum	Delivered (5+ hours?/units)	**Yes** 9/20 schools delivered the curriculum in years 1-2	Yes - central to the theory of change	Curriculum audit/Curriculum interviews	**Binary y/n** Based on whether they delivered 5 or more hours OR completed more than unit 1 0=delivered no LT curriculum, less than 5 hours, or only unit 1 1= delivered 5 or more hours, or more than unit 1
Training	Number of staff received in-depth training	**Moderate** 15/20 schools trained 5 or more staff members	Yes - process evaluation suggests is key for culture change	Training register	**Binary above/below 5 people** 5 people were specified for training in protocol
	Use of RP strategies to prevent or react to misbehaviour	**Yes** 12/20 schools had >85% of staff members report that they used restorative practices.	Yes-central to theory of change	Staff survey	**Binary above/below 85%** Cut-offs will be data-driven

A table will be created with schools on the X axis and measures on the Y axis. Each aspect of the intervention component (column 2) will be assessed on a scale from 0-1. The cut-offs within each quantitative aspect will be defined based on the study protocol, and if specific measures were not set, based on data-driven cut off points. Each school will have 2 final scores: one for years 1-2 with a range from 0-8, and one for year 3 with a range from 0-4. The scale does not need to be weighted as the intervention components gauged to be most important have more aspects scored within them. For example, according to preliminary analysis of the data so far, the Action Groups are more important than the curriculum. Therefore, they have more factors to be evaluated.