Supporting Teachers and ChildRen in Schools (STARS): Incredible Years® Teacher Classroom Management Statistical Analysis Plan

Evaluator: National Foundation for Educational Research
Principal investigator: Ben Styles

Education Endowment Foundation

PROJECT TITLE	Supporting Teachers And childRen in Schools (STARS): Incredible Years® Teacher Classroom Management Evaluation
DEVELOPER (INSTITUTION)	University of Exeter
EVALUATOR (INSTITUTION)	National Foundation for Educational Research
PRINCIPAL INVESTIGATOR(S)	Dr. Ben Styles
SAP AUTHOR(S)	Connie Rennie
TRIAL DESIGN	Two arm randomised controlled trial. Random allocation at school level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Ages 5-6, Key Stage 1
NUMBER OF SCHOOLS	140
NUMBER OF PUPILS	5800 (based on a mean cohort size of 42 pupils per school)
PRIMARY OUTCOME MEASURE AND SOURCE	N/A – no longer in use
SECONDARY OUTCOME MEASURE AND SOURCE	Pupil classroom behaviour: Pupil Behaviour Questionnaire (PBQ) Pupil emotional and social well-being: Total Difficulties Score (SDQ) Pupil concentration: Hyperactivity Scale (SDQ) Pupil prosocial behaviour: Prosocial Scale (SDQ) Student-teacher relationship: Student Teacher Relationship Scale (revised version) (STRS)

SAP version history

VERSION	DATE	REASON FOR REVISION
2.0	14th May 2021	Primary outcome removed due to disruptions caused by Covid-19. Pupil Behaviour now the main secondary outcome Midpoint testing removed and three level models changed to two level models. Change due to Covid-19 disruptions.

Table of Contents

Introduction

The Incredible Years Teacher Classroom Management (IY-TCM) programme is a teacher training programme designed to support teachers in improving classroom behaviour through the use of specific classroom management techniques and behaviours. The programme focuses on improving teacher-student relationships by reinforcing positive behaviour, and creating a positive learning environment in the classroom.

Research suggests that between three and six percent of children in each school class will have significant social, emotional; or behavioural difficulties that impair their learning, as well as that of their peers, and further negatively affect their future life prospects; children from disadvantaged backgrounds are even more likely to be affected by these issues (Ford *et al.*, 2018). The IY-TCM programme aims to support teachers to deal with these issues by introducing techniques such as behaviour plans, incentive systems and relationship-building tools.

The purpose of this evaluation is to investigate the outcomes of pupils aged five to seven who are taught by IY-TCM trained teachers. The trial aimed to explore whether the IY-TCM programme gives teachers the tools to reduce emotional, social, and concentration problems, as well as anti-social problems within the classroom, and if working with these problems in the classroom ultimately leads to improved attainment. However, due to the cancellation of Key Stage 1 tests caused by Covid-19, we are no longer able to measure attainment outcomes, and therefore this trial will focus on the aforementioned behavioural outcomes.

The training for teachers is delivered over six whole days across a period of six months. For this evaluation, Year 1 and Year 2 teachers will receive six training days at monthly intervals during the academic year 2019/20. The trial pupils will be in Year 1 in the Year 2019/20, and will be taught by Year 1 teachers who will be receiving training throughout the year, and will be taught in Year 2 in the year 2020/21 by Year 2 teachers who will have received the training in the previous year. The trial cohort will be exposed to two years of teaching: in the first year by teachers whilst they are being trained, and in the second, by teachers fully trained on the programme. The outcomes of the pupils will be compared to students who have been taught business as usual (BAU) for the same period.

The primary outcome of maths attainment at Key Stage 1 will no longer be measured due to the cancellation of Key Stage 1 tests caused by Covid-19, and therefore pupil behaviour as measured by the Pupil Behaviour Questionnaire will be treated as the main secondary outcome. The other secondary outcome measures include pupil emotional and social well-being, concentration, prosocial behaviour, classroom behaviour, and the student-teacher relationship. Outcomes are further detailed below in the analysis section. The main secondary outcome will further be analysed separately for disadvantaged children, as measured by eligibility for free school meals, as subgroup analysis.

Eligible schools are mainstream primary schools i.e. all schools other than independent, selective, special, alternative provision and schools in special measures, within the target areas (University of Exeter delivery hubs), or within reasonable travelling distance of the areas of Reading, Cornwall, Bristol, Southampton, Dorset and Liverpool. Eligible teachers are all Year 1 and Year 2 teachers who have at least four days of classroom responsibility per week, or both partners of a job share. The intervention is delivered to whole classes of Year 1 and Year 2 in intervention schools, hence no pupil selection is necessary.

Design overview

Trial design, including number of arms		Two arm cluster randomised controlled trial
Unit of randomisation		School
Stratification variables (if applicable)		Geographic area; School size proxy (number of reception teachers)
Primary outcome	variable	N/A
	measure (instrument, scale, source)	N/A
Secondary outcome(s)	variable(s)	Pupil classroom behaviour Pupil emotional and social well-being Pupil concentration Pupil prosocial behaviour Student Teacher relationship
	measure(s) (instrument, scale, source)	Pupil Behaviour Questionnaire (PBQ) Total Difficulties Score (SDQ) Hyperactivity Scale (SDQ) Prosocial Scale (SDQ) Student Teacher Relationship Scale (revised version) (STRS)
Baseline for primary outcome	**variable**	N/A
	measure (instrument, scale, source)	N/A
Baseline for secondary outcome	**variable**	Pupil classroom behaviour Pupil emotional and social well-being Pupil concentration Pupil prosocial behaviour
	measure (instrument, scale, source)	Pupil Behaviour Questionnaire (PBQ) Total Difficulties Score (SDQ) Hyperactivity Scale (SDQ) Prosocial Scale (SDQ)

Sample size calculations overview

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.17	0.20	0.17	0.20
Pre-test/ post-test correlations		-	-	-	-
	level 1 (student)	0.51	0.51	0.51	0.51
	level 2 (school)	-	-	-	-
	level 1 (student)	-	-	-	-

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Intracluster correlations (ICCs)	level 2 (school)	0.15	0.15	0.15	0.15
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two	Two	Two	Two
Average cluster size		42	10.5	40.6	10.2
Number of schools	intervention	70	70	70	70
	control	70	70	69	69
	total	140	140	139	139
Number of pupils	intervention	2940	735	2718	680
	control	2940	735	2924	731
	total	5880	1470	5642	1411

The ICC and pre and post correlation assumptions were drawn from a previous NFER trial using KS1 maths scores. The recruitment target of 140 schools was only just missed (139), and the assumption of an average cluster size was close to the actual average cluster size at randomisation (40.6). Therefore the trial was still powered to detect an effect of 0.17 in the primary outcome. Due to the cancellation of the KS1 tests, the PBQ is being used as the main secondary outcome. It is theorised that attainment will be causally impacted by a change in pupil behaviour, therefore we are confident that the trial is also powered to detect an effect as measured by the PBQ, a reliable measure of pupil behaviour (Allwood *et al*., 2018).

Analysis

Outcome Measurements

1. Pupil Behaviour Questionnaire (PBQ)

The PBQ is a tool developed by Exeter University, and has been previously used as an outcome measure in an IY-TCM evaluation; it was validated in 2018 (Allwood *et al.*, 2018). Whereas the Strengths and Difficulties Questionnaire focuses upon indicators of pupil mental health and individual behaviours, this scale captures behaviours that have a negative impact on the classroom. As such, it relates to an important aspect of the programme; managing behaviour within the classroom. It is a six-item scale, each rated on a three point scale; (never happens, occasionally happens, frequently happens). The scores range from 0-12 with a higher score reflecting worse behaviour. It has a Cronbach's alpha value of 0.85 (Allwood, et al., 2018).

2. Pupil emotional and social well-being

Emotional and social wellbeing will be measured using the Total Difficulties score of the Strengths and Difficulties Questionnaire (SDQ, Goodman, 2001). The SDQ was chosen as it is a reliable measure of pupils' emotional and social wellbeing (Goodman, 1997). It consists of 25 items, split into five subscales with five items each (emotional symptoms, conduct problems, hyperactivity/inattention, peer relationship problems, prosocial behaviour, each scored 0-2). The Total Difficulties score (TDS) is the sum of emotional symptoms, conduct problems, hyperactivity/inattention and peer relationship problems subscale scores. It has a possible range of between 0 and 40, with a higher score representing a higher level of difficulty. The score will be coded using the syntax published by Youth in Mind[1].The TDS will be analysed as a binary outcome, (in line with categorization used in the previous trial, Ford, *et al.*, 2018), as well as a continuous outcome. A score of 12 or above represents the 80th percentile of the British school-age population, and those above the score of 12 have been classified as 'strugglers' in previous analysis of the IY-TCM programme (Ford *et al.*, 2018). This cut-off point will be used for the purpose of comparing to previous research.

3. Pupil concentration

Pupil concentration will be measured using the hyperactivity subscale of the SDQ. This is a five item scale that reliably measures levels of restlessness and distractibility. Each item is scored between 0-2 giving a range of scores between 0-10. Higher scores reflect lower levels of concentration. This subscale was chosen as it fits closely with the theory of change and improvements were found in previous investigations of the programme (Ford *et al.*, 2018).

4. Pupil prosocial behaviour

Pupil prosocial behaviour will be measured using the prosocial subscale of the SDQ. This is a five item scale that reliably measures levels of helpfulness and kindness. Each item is scored between 0-2 giving a range of scores between 0-10. A higher score reflects higher levels of prosocial behaviour. This subscale was chosen as it fits closely with the theory of change and improvements were found in previous investigations of the program (Ford *et al.*, 2018).

5. Student Teacher Relationship Scale (STRS)

The STRS is an instrument designed for teachers of children aged between 3 and 12 which measures a teacher's perception of conflict, closeness and dependency with a specific child (Pianta, 2001). The short version of the instrument was suggested for this trial as it allows measurement of pupil-teacher relationship, a focus of the intervention that is not captured in other outcome measures. Following internal discussion, NFER proposed that some questions, through emphasis and phrasing, were inappropriate for a UK context. Following discussion with EEF, a revised version was produced jointly by NFER and the developer team. The revised version is currently being trialled by Exeter University, and depending on the trial results, will be used at the end point of the trial only. Details on the finalised version will be included in the report appendices.

Originally we had planned to take a midpoint measurement of the PBQ and the SDQ (at the end of the first year of the intervention), however due to disruptions caused by Covid-19, this was not possible. As such, measurements of the SDQ, PBQ and STRS will be collected at one follow-up time-point only; at the end of the second year of the intervention. All measures

[1] http://www.sdqinfo.org/py/sdqinfo/c0.py

were measured at the baseline apart from the STRS which is being measured at end point only.

Primary outcome analysis

The primary analysis of Maths attainment using the KS1 Maths scores will no longer be analysed in this trial, due to the tests being cancelled. Therefore pupil behaviour, as measured by the Pupil Behaviour Questionnaire, will be elevated to the main secondary outcome measure, and will be analysed using the same methods previously planned for the primary outcome.

Secondary outcome analysis

The main secondary outcome analysis of pupil behaviour will be 'intention-to-treat' (ITT). The model will be a multilevel model with two levels (school and pupil). Pupils who have measurements at baseline and follow-up on the PBQ will be included in the model, regardless of whether their teacher attended the IY-TCM training.

The dependent variable for the model will be the PBQ total raw score at follow-up with the following covariates:

- Prior behaviour as measured by the baseline PBQ
- A series of dummy geographical variables (randomisation stratifiers)
- A series of dummy school size variables (proxy measure: number of reception teachers per school (randomisation stratifiers))
- Intervention allocation dummy variable

This model will determine whether the IY-TCM programme has an impact on pupils' behaviour. The analysis for all multi-level models in this investigation will be run in R using the lme4 package.

Two models will be run on the secondary outcome of pupil emotional and social well-being as measured by the Total Difficulties Score (TDS) from the SDQ. A logistic regression will be run using the cut-off score of 12 or above as a binary outcome, and a linear regression will be run on the TDS raw score as a continuous outcome. The models will be 'intention-to-treat' (ITT). They will be multilevel models with two levels (school and pupil). Pupils who have measurements at baseline and follow-up will be included in the models, regardless of whether their teacher attended the IY-TCM programme or not.

The dependent variable for the logistic regression model will be a dummy variable representing a score of 12 or above on the TDS at follow-up with the following covariates:

- Baseline score of the TDS
- A series of dummy geographical variables
- A series of dummy school size variables (proxy measure: number of reception teachers per school)
- Intervention allocation dummy variable

The dependent variable for the linear regression will be the raw score of the TDS, with the same covariates as the above model. These models ewill determine whether the IY-TCM programme has an overall impact on pupils' emotional and social wellbeing.

The same linear regression models will be run on the following secondary outcome measures: pupil concentration, prosocial behaviour and the student teacher relationship. Each model will be assessed using a two-level multilevel model (school, pupil) with the following covariates:

- Baseline measurements of the respective scales
- A series of dummy geographical variables
- A series of dummy school size variables (proxy measure: number of reception teachers per school)
- Intervention allocation dummy variable

The exception to this is the student teacher relationship model, which will use the baseline PBQ score as the baseline measure. These models will determine whether the IY-TCM programme has an overall impact on pupils' concentration, prosocial behaviour and student teacher relationship respectively.

Subgroup analyses

As FSM-eligible pupils represent a particularly important subgroup, a separate analysis of FSM-eligible pupils will be carried out as per standard EEF practice. Sample size calculations run prior to the Covid-19 disruptions indicated that a sample of 1470 FSM eligible pupils would ensure enough power to detect an MDES of 0.2 on the primary outcome. At analysis stage we will conduct ES calculations for the main secondary outcome instead, results of which will be reported. This will be conducted for the full sample of pupils, as well as the FSM subgroup. The subgroup model will mimic the model used to assess the main secondary outcome (PBQ), however with FSM pupils only.

A further subgroup analyses will be run on the whole sample, identifying subgroups using an interaction term. A two level multilevel model (school and pupil) will be run on the PBQ with the following covariates:

- Prior behaviour as measured by the PBQ
- A series of dummy geographical variables
- A series of dummy school size variables (proxy measure: number of reception teachers per school)
- Intervention allocation dummy variable
- A dummy variable indicating FSM eligibility
- An interaction variable FSM*intervention

This model will determine whether the IY-TCM programme has a differential impact on FSM as compared to non-FSM pupils' behaviour.

Imbalance at baseline

We will create a baseline comparison table comparing proportion of baseline characteristics between the intervention and control group, both for the randomised groups, and the analysed groups. The table will include the following variables:

School level

- School level KS2 maths attainmentSchool level FSM quintile
- Region
- School Type

- Ofsted rating
- School size

Pupil level

- PBQ mean and SDTDS mean and SD
- FSM eligibility

Due to randomisation, we would not expect any significant imbalance in baseline characteristics between the intervention and control pupils for randomised groups. Furthermore, pupil behaviour and the TDS score will be included as covariates in the respective models, and therefore any potential imbalance in analysed groups will be taken into consideration by the model, and will not influence the effect estimate. However, we will produce a baseline comparison table to show if balance was achieved, and to inspect whether attrition may have introduced bias.

Missing data

We will assess the level and pattern of missing data from the primary model. We will report the number and proportion of complete and missing cases included/not included in the primary model, including extent of missing covariates. In order to assess the missingness mechanism, we will run a multilevel logistic regression model on whether a case had follow-up data for the main secondary outcome, regressed on the covariates of the same model plus other school level and available pupil level variables. In conjunction with qualitative judgement on the reasons for missing, if a covariate is found to significantly predict missingness, we will run the main secondary model with the extra variables included. As per EEF guidance, if the substantive model estimates with and without the covariates are similar, we may conclude that the completers analysis is unbiased. If they differ, it is likely the outcome is missing not at random and sensitivity analysis will need to be carried out.

Should sensitivity analysis be required, this would be preceded by multilevel multiple imputation. Multiple imputation of chained equations is a flexible method that can handle different types of variables (Hughes *et al*., 2014). (The number of datasets is dependent on the amount of missing data but a minimum would be five datasets, with a minimum of ten iterations. The number of iterations will increase proportional to the level of missing data (White *et al.,* 2011). These iterations are necessary as with only one dataset, the parameter estimates have more sampling variability. Multiple iterations also help in generating the estimates of the standard errors to accurately reflect the uncertainty about the missing values (Allison, 2012). Once the substantive model is run on the multiply imputed data, sensitivity analysis will be carried out using different values of delta to reflect a ‘missing not at random’ scenario. Resulting intervention coefficients and their standard errors will be assessed and compared to the substantive model. All results will be reported.

Compliance

The University of Exeter will collect attendance data of teachers at each training session, through the use of attendance logs. We will produce descriptive statistics on this data. Compliance will be measured as a dichotomous variable indicating whether a teacher attended four out of the six sessions or above. Therefore compliance is a teacher level variable. In the case where the students are taught by the same teacher across the two years, the compliance will apply to the one teacher only. If the students are taught by

different teachers between Y1 and Y2, both teachers will have had to meet the four out of six criteria.

As per EEF guidance a two-stage least squares multi-level model will be used to calculate the Complier Average Causal Effect (CACE) estimate (Angrist and Imbens, 1995). The first stage of the model will be compliance regressed on all covariates that are used in the main secondary outcome model and the group allocation variable. The second stage of the model will regress the main secondary outcome on the covariates used in the main model and will also include a covariate representing the teacher's estimated level of compliance from the first stage of the model. No interaction term between intervention and compliance will be included as the design of the investigation means contamination is impossible, i.e. pupils will either be in an intervention school, or a control school, and no control pupils can 'comply' and all intervention pupils will have a compliance value. The coefficient of the estimated compliance measure is the CACE estimate of the compliance effect. We will use the R package ivpack to perform the CACE analysis on the main secondary outcome only.

Intra-cluster correlations (ICCs)

A two-level multilevel model of the intervention regressed on pupil behaviour will be run without any covariates, to estimate the ICC, in this case, the proportion of the total variance accounted for by the school level.

Furthermore, the ICC for the main model will be calculated from the output of the main secondary model, (including the covariates). It will again, be the proportion of the total variance that is accounted for by school level variance.

Effect size calculation

Effect sizes and confidence intervals will be calculated for all outcome models.

For the two-level multilevel models, the numerator for the effect size calculation will be the coefficient of the intervention group from the model. The denominator will be the total variance from a multilevel model without covariates, i.e. equivalent to Hedges' g.

$$ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{\sqrt{\sigma_S^2 + \sigma_{error}^2}}$$

Confidence intervals for each effect size will be derived by multiplying the standard error of the intervention group model coefficient by 1.96. These will be converted to effect size confidence intervals using the same formula as the effect size itself. In the case where the outcome is a binary variable, the effect size will be presented as an odds ratio and differences in probabilities.

References

Allison, P. (2012). 'Why maximum likelihood is better than multiple imputation', *Statistical Horizons*, 9 July [online]. Available: https://statisticalhorizons.com/ml-better-than-mi [8 January 2020].

Allwood, M., Allen, K., Price, A., Hayes, R., Edwards, V., Ball, S., Ukoumunne, O. C. and Ford, T. (2018). 'The reliability and validity of the pupil behaviour questionnaire: a child classroom behaviour

assessment tool', *Emotional and Behavioural Difficulties*, **23**, 4, 361- 371 [online]. DOI 10.1080/13632752.2018.1478945.

Angrist, J. D. and Imbens, G. W. (1995). 'Two-stage least squares estimation of average causal effects in models with variable treatment intensity', *Journal of the American statistical Association*, **90**, 430, 431-442 [online]. DOI 10.1080/01621459.1995.10476535.

Ford, T., Hayes, R., Byford, S., Edwards, V., Fletcher, M., Logan, S., Norwich, B., Pritchard, W., Allen, K., Allwood, M., Ganguli, P., Grimes, K., Hansford, L., Longdon, B., Norman, S., Price, A. and Ukoumunne, O. C. (2018). 'The effectiveness and cost-effectiveness of the Incredible Years® Teacher Classroom Management programme in primary school children: results of the STARS cluster randomised controlled trial', *Psychological Medicine*, **49**, 5, 828-842 [online]. DOI 10.1017/S0033291718001484.

Goodman, R. (1997). 'The Strengths and Difficulties Questionnaire: A Research Note', *Journal of Child Psychology and Psychiatry*, **38**, 5, 581-586 [online]. DOI 10.1111/j.1469-7610.1997.tb01545.x.

Goodman, R. (2001). 'Psychometric properties of the strengths and difficulties questionnaire', *Journal of the American Academy of Child & Adolescent Psychiatry*, **40**, 11, 1337-1345 [online]. DOI 10.1097/00004583-200111000-00015.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J.R., Tilling, K. and Sterne, J. A. C. (2014) 'Joint modelling rationale for chained equations', *BMC Medical Research Methodology*, **14**, 28 [online]. DOI 10.1186/1471-2288-14-28.

Pianta, R. (2001). *Student–Teacher Relationship Scale–Short Form*. Lutz, FL: Psychological Assessment Resources, Inc. Pianta, R. C. and Steinberg, M. (1992). 'Teacher-child relationships and the process of adjusting to school', *New Directions for Child and Adolescent Development*, 1992, 57, 61-80 [online]. DOI 10.1002/cd.23219925706.

White, I. R., Royston, P. and Wood, A. M. (2011) 'Multiple imputation using chained equations: Issues and guidance for practice', *Statistics in Medicine*, **30**, 4, 377–399 [online]. DOI 10.1002/sim.4067.