Adventure Learning
Statistical Analysis Plan

Evaluator (institution): Sheffield Hallam University
Principal investigator: Professor Tim Jay

PROJECT TITLE	Adventure Learning
DEVELOPER (INSTITUTION)	Outward Bound Trust, Commando Joe's
EVALUATOR (INSTITUTION)	Sheffield Hallam University
PRINCIPAL INVESTIGATOR	Prof. Tim Jay
SAP AUTHORS	Sean Demack, Sarah Reaney-Wood, Prof. Tim Jay
TRIAL DESIGN	Three-arm clustered randomised controlled trial with random allocation at the school-level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Year 9 (13-14 years), KS3
NUMBER OF SCHOOLS	97
NUMBER OF PUPILS	2328
PRIMARY OUTCOME MEASURE AND SOURCE	Pupil self-regulation after 1 year, using the Self-Regulation of Learning Self-Report Scale
SECONDARY OUTCOME MEASURE AND SOURCE	Pupil self-regulation immediate post-test (Self-Regulation of Learning Self-Report Scale) Student engagement immediate post-test and after 1 year (School Engagement Instrument) Pupil behaviour in school after 1 year (Strengths and Difficulties Questionnaire, teacher version) Mathematics attainment after 1 year (Progress Test in Maths; GL Assessment) Attainment after 2 years (GCSE Attainment 8)

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0	22 November 2019	*N/A*

Table of contents

Introduction

Adventure learning (the programme) is evaluating whether a 1-week adventure learning experience leads to increases in pupil self-regulation, student engagement, improved behaviour in school and mathematics attainment, either directly or indirectly. Adventure learning programmes have been shown to be effective in raising attainment (Cason & Gills, 1994; Hattie et al., 1997) and improving non-cognitive outcomes (Bobilya, Klalisch, & Daniel, 2011; Ewert & Yoshino, 2011; Bowen & Neill, 2013), with longer, sustained experiences deemed more effective (Rickinson et al., 2004). However, the reasons why are not fully clear (Hattie, Marsh, Neill, & Richards, 1997; Martin & Leberman, 2005). The purpose of this statistical analysis plan is to provide detail on the Adventure Learning randomised controlled trial (RCT) impact analysis. To date, this is the first RCT to look at whether Adventure Learning improves non-cognitive outcomes and attainment, and whether a residential element provides the catalyst for this change over a school-based programme. The trial tests two approaches to adventure learning: one provided by Outward Bound includes a residential experience, while one provided by Commando Joe's takes place on school grounds.

Design overview

Trial design, including number of arms		Three-arm clustered randomised controlled trial
Unit of randomisation		School
Stratification variables (if applicable)		N/A (simple randomisation used)
Primary outcome	variable	Self-regulation after 1 year
	measure (instrument, scale, source)	Self-regulation of Learning Self-Report Scale (SRL-SRS) (Toering, 2012)
Secondary outcome(s)	variable(s)	Self-regulation immediate post-test; Mathematics attainment after 1 year; General attainment after 2 years; Student engagement (immediate post-test and after 1 year); Pupil behaviour in school after 1 year
	measure(s) (instrument, scale, source)	SRL-SRS; Progress Test in Maths (PTM, GL Assessment); Attainment 8 (NPD); Student Engagement Instrument (SEI); Strengths and Difficulties Questionnaire (SDQ)
Baseline for primary outcome	**variable**	Self-regulation of Learning prior to randomisation
	measure (instrument, scale, source)	Self-regulation of Learning Self-Report Scale (SRL-SRS) (Toering, 2012)

Baseline for secondary outcome	**variable**	Mathematics attainment; School engagement; Pupil behaviour in school
	measure (instrument, scale, source)	KS2_MATSCORE, NPD; SEI; SDQ

Data will be collected at four times during the project, as in TABLE 1 (on a rolling basis, depending on the timing of the intervention in each school).

TABLE 1. TIMETABLE FOR DATA COLLECTION

Time	Code	Data Collection Activity
January – June 2019	T1	Baseline SRL-SRS, SDQ, SEI
October 2019 – January 2020	T2	Immediate post-test SRL-SRS, SEI
October 2020	T3	Post-test after 1 year SRL-SRS **(primary outcome)**, SEI, SDQ, PTM
October 2022	T4	General attainment GCSE Attainment 8

Sample size calculations overview

TABLE 2. MDES CALCULATIONS

		Protocol		Randomisation	
		OVERALL	**FSM**	**OVERALL**	**FSM**
Minimum Detectable Effect Size (MDES)		0.21	0.25	0.22	0.25
Pre-test/ post-test correlations	level 1 (pupil)	0.5	0.5	0.5	0.5
	level 2 (school)	0.25	0.25	0.25	0.25
Alpha		0.025	0.025	0.025	0.025
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two	Two	Two	Two
Average cluster size					
Number of schools	OB	33	33	33	33
	CJ's	33	33	32	32
	Control	33	33	32	32
	total	99	99	97	97
Number of pupils	OB	792 (24 per school)	396 (12 per school)	792 (24 per school)	396 (12 per school)
	CJ's	792 (24 per school)	396 (12 per school)	768 (24 per school)	384 (12 per school)
	control	792 (24 per school)	396 (12 per school)	768 (24 per school)	384 (12 per school)

	Protocol		Randomisation	
total	2376	1188	2328	1164

Please see Appendix 1 for details on how these MDES estimates were calculated. The Alpha value is specified as 0.025 as the trial involves multiple comparisons. A Bonferroni comparison has been used (appendix 1). The estimates in the table correct for two comparisons (OB v control & CJ's v control). Appendix 1 gives estimates for three comparisons (with OB v. CJ's as an additional comparison, which will be reported as an exploratory analysis). Level 1 pre-/post-test correlations are estimated based on test-retest reliability of the primary outcome measure (between .69 and .84 over 4 to 6 weeks; Toering et al., 2012), We have reduced to a more conservative .5 due to the longer duration between pre- and post-test in this trial. Level 2 correlations are estimated as .25 following recommendations from EEF. School-level ICC has been estimated as 0.05. There is little information in the literature regarding likely ICC for non-cognitive measures, however this is anticipated to be lower than that for measures of attainment (typically 0.1-0.2). A power analysis for the 2-level RCT, with 66 schools per comparison and 24 pupils per school has been carried out for the protocol. Similarly, a power analysis was undertaken for the FSM subsample (approximately 12 pupils per school, given sampling requirements in the protocol). The target number of schools for recruitment was 99 (33 per arm), but 97 schools entered randomisation; 33 schools to the OBT arm, 32 schools to the CJ arm, and 32 schools to the control. The results of this analysis estimate that, for the main ITT analyses the design will be able to detect an effect size of 0.22 or higher as statistically significant (α<0.025) with a statistical power of 0.80. For the FSM subsample the design will be able to detect an effect size of 0.25 or higher as statistically significant (α<0.025) with a statistical power of 0.80.

The risk of attrition for this evaluation is relatively high. The pupils recruited for this study are 'hard to reach' pupils or those achieving lower than expected. As such, it is presumed that there will be a higher level of attrition due to school moves, exclusions and lack of engagement. There are two key issues that could result in attrition; reduced sample size and randomisation. Assuming that missing data has no impact on randomisation' with 10% attrition[1] the MDES estimates are 0.23 for the whole sample and 0.28 for the FSM sample; with 20% attrition[2] the MDES estimates are 0.26 for the whole sample and 0.32 for the FSM sample. However, it is rather naive to assume that the missing data resulting from pupil

[1] This is 10% attrition at school level (i.e. number of schools reduced to 30 per comparison group) and at the pupil level (20 pupils assumed per school; 10 FSM pupils assumed per school).
[2] This is 20% attrition at school level (i.e. number of schools reduced to 27 per comparison group) and at the pupil level (16 pupils assumed per school; 8 FSM pupils assumed per school).

attrition will have no impact on randomisation. Therefore, the impact of missing data will be explored as set out below; this includes missing data analyses and possible imputation of missing data.

Analysis

For all the analyses in this trial the raw score for each of the non-cognitive outcomes; the self-regulation of learning self-report measure (SRL-SRS), the Student Engagement Instrument (SEI) and the total difficulties score of the Strengths and Difficulties questionnaire (SDQ) will be used. The analysis aims to follow an intention to treat (ITT) approach. As such, pupils will be included in the analysis even they have not full provided data or complied with the intervention. However, due to the nature of the sample in this study (hard to reach pupils) a higher level of attrition may be expected at the point of testing 1 year post-intervention.

The primary outcome measure is the Self-Regulation of Learning Self-Report Scale (SRL-SRS) taken at T3 – when pupils will be at the start of Year 10. The total score for the SRL-SRS will be computed following the scoring criteria reported in Toering et al (2012).

School engagement will be measured using the Student Engagement Instrument (SEI; Appleton et al. 2006) and the scores will be computed following the guidelines available at: http://checkandconnect.umn.edu/docs/SEI%20Administration,%20Scoring,%20and%20Results%20Use%20ICI.pdf.

The Strengths and Difficulties questionnaire (SDQ) will be scored following the guidelines available at: https://www.sdqinfo.com/py/sdqinfo/c0.py and in Appendix 2.

The mathematics assessment at T2 will be the digital Progress Test in Maths, from GL Assessment. The raw score will be used as the outcome measure, as this removes risks associated with incomplete data for participants' dates of birth. Sample size is large enough that distribution of age across groups will be close to equal, and so using standardised score would not confer additional validity This data will be collected and scored via GL Assessment's online platform. General attainment at T4 will be measured as an Attainment 8 score (DfE, 2016). This is calculated by adding up the point scores for a pupil's eight qualifying subjects and dividing by 10. This will be obtained from NPD and matched to the other data collected prior to randomisation using pupil UPN/ PMR.

Identifiers within the models are fixed effects based on the guidance provided in the EEF statistical analysis guide. The Adventure Learning is an efficacy trial and therefore conditional interference only will be made, we will not be generalising beyond the sample of schools included in the study. All multi-level models will be conducted in R, version 3.6.1 (or

later if appropriate). Statistical uncertainty will be expressed as standard errors of multilevel model coefficients and use of 95% confidence intervals.

Impact analysis

Primary outcome analysis

RESEARCH QUESTION 1A

Does participation in the Outward Bound intervention lead to changes in self-regulation of learning compared to a passive control?

RESEARCH QUESTION 1B

Does participation in the Commando Joe's intervention lead to changes in self-regulation of learning compared to a passive control?

Exploratory analysis

RESEARCH QUESTION 2

Does participation in the Outward Bound intervention lead to different changes in self-regulation of learning compared to participation in the Commando Joe's intervention?

Secondary outcome analysis

RESEARCH QUESTION 3[3]

Is there an immediate effect of an adventure learning intervention on self-regulation?

RESEARCH QUESTION 4

Does an adventure learning intervention lead to changes in pupil behaviours in schools after approximately 1 year?

RESEARCH QUESTION 5

Does an adventure learning intervention lead to changes in mathematics attainment after approximately 1 year?

RESEARCH QUESTION 6

Does an adventure learning intervention lead to changes after two years in general attainment at GCSE (attainment 8)?

RESEARCH QUESTION 7

Does an adventure learning intervention lead to changes in student engagement?

[3] Research questions 3 to 7 each involve 3 comparisons – Outward Bound v. Control; Commando Joe's v. Control; Outward Bound v. Commando Joe's

Analyses to address these research questions are described below.

Table 3 summarises the planned ITT analysis for the primary outcome of the Adventure Learning trial (dual primary outcome, RQ 1a and 1b) and further exploratory analysis (RQ 2) that includes the primary outcome measure. RQ1 will be answered by constructing two multi-level linear models (pupils nested in schools) for the primary outcome SRL-SRS score at time point three (T3). The first (RQ 1a) model will measure the impact of the OBT programme on SRL-SRS compared with the business as usual control group (OBT v control). The second (RQ 1b) model will measure the impact of the Commando Joe programme on SRL-SRS compared with the business as usual control group (CJs v control). The CRT design for evaluating the impact of Adventure Learning trial was powered for these two comparisons. A third comparison (RQ 2) will be undertaken as exploratory analysis where the impact of the OBT and CJ programmes is directly compared (OBT v CJ). The pupil-level baseline SRL-SRS covariate will be included in all models. As no other variables were used as part of randomisation (i.e. to stratify), no other variables will be included in the analysis, as per EEF statistical guidance.

TABLE 3. SUMMARY OF PLANNED ANALYSES OF SELF-REGULATION FOR RQ 1A, 1B, 2 AND 3

Analysis and Sample	*Level 1 (pupil) Variables*	*Level 2 (school) Variables*	*Outcome variable*
RQ 1a ITT sample	Baseline SRL-SRS collected Jan-June 2019	Group (1=OBT; 0=control);	SRL-SRS at T3
RQ 1b ITT sample	Baseline SRL-SRS collected Jan-June 2019	Group (1=CJ; 0=control);	SRL-SRS at T3
RQ 2 ITT sample	Baseline SRL-SRS collected Jan-June 2019	Group (1=OBT; 0=CJ);	SRL-SRS at T3
RQ 3 ITT sample	Baseline SRL-SRS	Group (1=OBT; 0=Control);	SRL-SRS T2
RQ 3a ITT sample	Baseline SRL-SRS	Group (1=CJ; 0=Control);	SRL-SRS T2
RQ 3b ITT sample	Baseline SRL-SRS	Group (1=OBT; 0=CJ);	SRL_SRS T2

Secondary outcome analysis

RESEARCH QUESTION 4

Does an adventure learning intervention lead to changes in pupil behaviours in schools after approximately 1 year?

Table 4 below summarises the planned ITT analysis for the secondary outcome of the Adventure Learning trial, pupil behaviour. Total strength and difficulties questionnaire (SDQ) score will be calculated following the scoring instructions detailed in appendix 2. As with RQ1, three comparisons will be undertaken[4] (OBT v control; CJ v control and OBT v CJ) by constructing three separate multi-level linear models. Baseline SDQ will be included as a covariate in each of the models.

TABLE 4. SUMMARY OF THE PLANNED SECONDARY OUTCOME ITT ANALYSIS FOR RESEARCH QUESTION 4

Analysis and Sample	*Level 1 (pupil) Variables*	*Level 2 (school) Variables*	*Outcome variable*
RQ 4 Final Analysis ITT sample	Baseline SDQ collected Jan-June 2019	Group (1=OBT; 0=control);	SDQ T3
RQ 4a Final Analysis ITT sample	Baseline SDQ collected Jan-June 2019	Group (1= CJ; 0=control)	SDQ T3
RQ 4b	Baseline SDQ collected Jan-June 2019	Group (1=OBT; 0=CJ);	SDQ T3

RESEARCH QUESTION 5

Does an adventure learning intervention lead to changes in mathematics attainment after approximately 1 year?

Table 5 below summarises the planned ITT analysis for the secondary outcome of the Adventure Learning trial, mathematics attainment using a GL mathematics test. As with RQ1 and RQ2, three comparisons will be undertaken (OBT v control; CJ v control and OBT v CJ) by constructing three separate multi-level linear models. Mathematics attainment at KS2 will be included in each of the models as a covariate. Whilst the multilevel design accounts for school-level clustering of the outcome data, the design does not account of

[4] This relates to comparisons of secondary outcomes that the trial design is not powered specifically to detect. Therefore, all three can be considered as exploratory analyses.

within-school clustering (e.g. at a class level). Demack (2019) highlighted the importance of accounting for class level clustering in trial design because the widespread practice of grouping students according to measured attainment results in very strong class-level clustering (particularly in England and particularly in secondary mathematics). Demack (2019) illustrated how ignoring clustering at a class level can result in trials being underpowered and/or with falsely inflated statistical sensitivity. However, both adventure learning interventions are targeted at individual pupils rather than classes. This means that pupils who participate in the adventure learning trial are not naturally clustered into mathematics classes (or sets), they are drawn from multiple classes across the year group. Whilst it is likely that each individual pupil will be taught maths in a 'setted' class, it does not follow that all selected pupils will be in one (or two) specific classes. This made the inclusion of a class level unfeasible and so the design does not attempt to take into account of class-level clustering of outcome data. The targeted nature of the adventure learning interventions is likely to mean that the estimated 24 pupils in each school will be relatively homogenous compared with untargeted, class-based interventions. Schools' approaches to selecting pupils for participation may vary and this may result in school-level differences in pupil compositions. The design and planned analyses acknowledge this through the inclusion of a school level. It is assumed that, within each school, a consistent approach was used to select pupils and that this was undertaken at the year group rather than class level. Data on this is to be collected in the IPE. Therefore, it is assumed that there will be no (zero) within-school clustering of outcome data (including maths attainment)[5].

TABLE 5. SUMMARY OF THE PLANNED SECONDARY OUTCOME ITT ANALYSIS FOR RESEARCH QUESTION 5

Analysis and Sample	*Level 1 (pupil) Variables*	*Level 2 (school) Variables*	*Outcome Variable*
RQ 5 ITT sample	KS2_MATSCORE	Group (1=OBT; 0=control);	PTM (GL)
RQ 5a ITT sample	KS2_MATSCORE	Group (1=CJ; 0=control);	PTM (GL)
RQ 5b	KS2_MATSCORE	Group (1=OBT; 0=CJ);	PTM (GL)

[5] This is an assumption rather than a fact. It would be technically feasible to collect class-level detail on the entire year group to examine the validity of this assumption (e.g. are pupils actually all drawn from one lower set maths class?) but this is beyond the scope (or resource) of this evaluation.

RESEARCH QUESTION 6

Does an adventure learning intervention lead to changes after two years in general attainment at GCSE (attainment 8)?

Table 6 below summarises the planned ITT analysis for the secondary outcome of the Adventure Learning trial, general attainment using the attainment 8 score. As with RQ1, RQ2 and RQ3, three comparisons will be undertaken (OBT v control; CJ v control and OBT v CJ) by constructing three separate multi-level linear models. KS2 maths score (KS2_MATSCORE) will be included in each model as a covariate. KS2_READSCORE, KS2_GPSSCORE will also be included in each model as covariates.

TABLE 6. SUMMARY OF THE PLANNED SECONDARY OUTCOME ITT ANALYSIS FOR RESEARCH QUESTION 6

Analysis and Sample	*Level 1 (pupil) Variables*	*Level 2 (school) Variables*	*Outcome Variable*
RQ 6 ITT sample	KS2_MATSCORE	Group (1=OBT; 0=control);	Attainment 8
RQ 6b ITT sample	KS2_MATSCORE	Group (1=CJ; 0=control);	Attainment 8
RQ 6c	KS2_MATSCORE	Group (1=OBT; 0=CJ);	Attainment 8

RESEARCH QUESTION 7

Does an adventure learning intervention lead to changes in student engagement?

Table 7 below summarises the planned ITT analysis for the secondary outcome of the Adventure Learning trial, student engagement (SEI). As with RQ1, RQ2 and RQ3, three comparisons will be undertaken (OBT v control; CJ v control and OBT v CJ) by constructing three separate multi-level linear models. Baseline SEI scores will be included in each model as a covariate.

TABLE 7. SUMMARY OF THE PLANNED SECONDARY OUTCOME ITT ANALYSIS FOR RESEARCH QUESTION 4.

Analysis and Sample	Level 1 (pupil) Variables	Level 2 (school) Variables	Outcome Variable
RQ7 ITT sample	Baseline SEI	Group (1=OBT; 0=control);	SEI T3
RQ 7a ITT sample	Baseline SEI	Group (1=CJ; 0=control);	SEI T3
RQ 7b	Baseline SEI	Group (1=OBT; 0=CJ);	SEI T3

Subgroup analyses

Sub group analysis will be conducted for FSM pupils using the same multi-level linear model approach as described above for RQ 1a and 1b. The FSM subgroup will be identified by the EverFSM_6_P indicator from the NPD data set.

Additional analysis

No additional analysis is planned. All primary, secondary and exploratory analyses are described above.

Longitudinal follow-up analysis

The above discussion of the models explains the longitudinal nature of the analysis. No additional follow-up is planned.

Imbalance at baseline

Imbalance at baseline will be assessed by creating a cross-tabulation of pupil characteristics (FSM status, Gender, School status (Ofsted) and baseline pupil self-regulation) by trial arm. Pupil and school-level categorical characteristics will be reported as counts and percentages in each category, while continuous variables will be summarised using mean and standard deviation. Differences in pupil-level pre-tests will be reported as effect sizes for each intervention arm.

Missing data

The baseline and ITT samples will be compared to help illustrate the impact of missing data for the primary outcome variable only. This will firstly be done descriptively by tabulating missing cases across the categories of variables included in the ITT analysis (SRL-SRS baseline, OBT Centre, CJ instructor). Reasons for any missingness will be summarised and

a multi-level logistic regression model (1=in ITT model; 0=not in ITT model) will examine whether missingness is associated with school/intervention and/or pupil-level covariates including; baseline SRL-SRS, KS2 attainment, FSM status.

If over 5% of cases in the baseline sample are missing from the headline ITT analysis, we will adopt the following approach for screening and addressing missing data.

Screening stage: We will examine whether data is missing completely at random (MCAP), missing at random (MAR) or missing not at random (MNAR). A series of binary variables will be generated for all variables in the final ITT analysis that measures whether a case is missing (=1) or not (=0). Logistic regression will be used to examine whether missing data can be statistically accounted for using the other variables in the ITT analysis with additional school-level variables included. When variables are found to account for a statistically significant proportion of variation in missing data, we will proceed to one of the next two stages.

- For instances where only data is missing in the SRL-SRS outcome measure, we will add any additional covariates that were found from the screening stage to the final ITT model and re-estimate the effect size.
- For instances where data is missing in the SRL-SRS baseline measure and where the screening stage identified variables that did account for variation in this missing data, we will construct a Multiple Imputation model using all variables listed for stage 1. The Multiple Imputation model will be estimated using 'STATA MI' to create 20 imputed data sets. These imputed data sets will be used to re-estimate the effect of RME and the standard error (Rubin, 2004).

Compliance

Compliance is defined at the pupil level only, with a pupil considered compliant if they attended three out of five of the scheduled intervention days. Compliance is not expected to be close to 100% and CACE analysis with Instrumental Variable (IV) approach using Two Stage Least Square (2SLS) analysis will be conducted on the primary outcome only (Sussman & Hayward. 2010; Tilbrook et al. 2014).

The purpose of this analysis is to estimate the impact of adventure learning for pupils deemed to have 'complied' (i.e. attended 3+ days of the intervention) with their intervention. Binary compliance codes will be constructed (1=complied, 0=not complied) for each intervention. These binary compliance variables will be used to estimate the IV for the OB and CJ interventions respectively for the first stage model.

This first stage will model the pupil-level binary compliant variable(s) using the baseline SRL-SRS variable along with additional school level variables that are available via the school

census. This model will be used to generate the predicted compliance (1 or 0) IV for OB and CJ. The second stage model will use the predicted compliance in place of the group identifier in order to generate the CACE estimates for OB and CJ respectively. 'STATA IVRegress' will be used for this Instrumental Variable approach for estimating CACE.

Intra-cluster correlations (ICCs)

The pre-test for Adventure Learning will be self-regulation of learning (year 8) and the post-test will be self-regulation of learning (year 10). For both pre and post-test, ICCs at the school level will be estimated using a null (empty) 2-level variance components model. Within the analyses, a table will present the variance decomposition for the two levels (school and pupil) along with the ICC estimates.

Effect size calculation

The causal impact of Adventure Learning on the primary outcome of self-regulation at T3 will be determined using Hedges' g effect sizes. Hedges' g effect sizes will standardise the difference in self-regulation between control and intervention schools into units of standard deviations. As shown in equation 1.1 below**:**

Equation1.1 $$ES = \frac{(Intervention - Control)_{adjusted}}{\sqrt{\delta^2_{sch} + \delta^2_{pup}}}$$

A Hedges g effect size statistic will be calculated for each ITT comparisons (OBT v control; CJs v control) in the Adventure Learning trial. The upper and lower 95% confidence intervals for the estimated impact [$(Intervention - Control)_{adjusted}$] will be divided by the same standard deviation shown in equation 1.1 to estimate the 95% confidence intervals for Hedges g.

For example, for OBT:

Equation1.2 $$ES = \frac{(OBT - Control)_{adjusted}}{\sqrt{\delta^2_{sch} + \delta^2_{pup}}}$$

Where:

- δ^2_{sch} is the school level variance and δ^2_{pup} is the pupil level variance for the SRL_SRS outcome from the empty/null multilevel model.
- $(OBT - Control)_{adjusted}$ is the mean difference between the attainment of pupils in OBT schools and pupils in control schools in the original raw (SRL-SRS) units. This is obtained from the coefficient for the school-level 'group' variable from the ITT analyses.

References

Bobilya, A., Kalisch, K., & Daniel, B. (2011). An investigation of the Outward Bound final expedition. *Journal of Experiential Education, 33*(4), 356–359.

Bowen, D., & Neill, J. (2013). A meta-analysis of adventure therapy outcomes and moderators. *The Open Psychology Journal, 6*, 28-53.

Cason, D., & Gillis, H. L. (1994). A meta-analysis of outdoor adventure programming with adolescents. *Journal of Experiential Education, 17*(1), 40-47.

Demack, S. (2019) Does the classroom level matter in the design of educational trials? A theoretical and methodological review. *EEF Research Paper Series, No. 003*, May 2019 (https://educationendowmentfoundation.org.uk/public/files/Publications/Does_the_classroom_level_matter.pdf)

Department of Education. (2016). *Progress 8: How Progress 8 and Attainment 8 measures are calculated*. London: Department of Education.

Ewert, A., & Yoshino, A. (2011). The influence of short-term adventure-based experiences on levels of resilience. *Journal of Education and Outdoor Learning, 11*(1), 35-50.

Hattie, J., Marsh, H., Neill, J., & Richards, G. (1997). Adventure education and Outward Bound: Out-of-class experiences that make a lasting difference. *Review of Educational Research, 67*(1), 43-87.

Martin, A., & Leberman, S. (2005). Personal learning or prescribed educational outcomes: A case study of the Outward Bound experience. *Joumal of Experiential Education, 28*(1), 44-59.

Rickinson, M., Dillon, J., Teamey, K., Morris, M., Young Choi, M., Sanders, D., & Benefield, P. (2004). *A Review of Research on Outdoor Learning.* London: Field Studies Council.

Rubin, D. (2004). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons.

Sussman,J.B. & Hayward,R.A. (2010). An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials. *BMJ 340.* doi:0.1136/bmj.c2073

Tillbrook et al. (2014). Compliance effects in a randomised controlled trial of yoga for chronic lower back pain: a methodological study. *Physiotherapy 100*(3) pp252-262. https://doi.org/10.1016/j.physio.2013.10.001

Toering, T., Elferink-Gemser, M. T., Jonker, L., van Heuvelen, M. J., & Visscher, C. (2012). Measuring self-regulation in a learning context: Reliability and validity of the Self-Regulation of Learning Self Report Scale (SRL-SRS). *International Journal of Sport and Exercise Psychology, 10*(1), 24-38.

Appendix 1

From Bloom et al. (2007), the Minimum Detectable Effect Size (MDES) for a 2-level CRT is:

$MDES2LCRT=MK-L-2\sqrt{(ICCSch(1-RSch2)P(1-P)K)+((1-ICCSch)(1-Rpup2)P(1-P)Km)}$

Where:

P is the proportion of schools who receive the intervention (=0.50)

$Rpup2$ is the pupil-level covariate explanatory power (=0.50x0.50=0.25)

$RSch2$ is the school-level covariate explanatory power (=0.25x0.25=0.0625)

$ICCSch$ is the cluster (school) level Intra Cluster Correlation coefficient (=0.05)

K is the number of clusters (schools =66 or 70)

Km is the total number of individual (pupils=792x2=1,584; 840x2=1,680)

m is the number of pupils per school (=24)

L is the number of cluster level covariates (=2)

$MK-L-2$ is the t-distribution multiplier with K-L-2 degrees of freedom - which equates to between 62 (with 33 schools per arm) and 66 (35 schools per arm) degrees of freedom.

1:1 or multiple comparisons.

A 1:1 comparison involves (for example) comparing the OB group with the control group. This ignores the increased risk of type I errors associated with multiple comparisons (Wason et al., 2014). In this trial there are 3 groups; two intervention groups (OB, CJ) and a business-as-usual control. Therefore, there will be up to 3 comparisons:

- OB v Control
- CJ v Control
- OB v CJ

The first two of these relate to testing whether the two interventions have an impact relative to the business as usual control group. The third might be seen as an exploratory analysis that tests whether a significant difference exists between the two interventions. To control for the inflation of type I errors associated with multiple comparisons, a Bonferroni correction is applied by dividing the probability of making a type I error by the number of proposed comparisons.

The power analyses now proceeds in three stages, first assuming a 1:1 comparison, second assuming two comparisons (OB v control & CJ v control) and finally assuming three comparisons (OB v CJ; OB v control; CJ v control).

NOTE - for the headline ITT analyses, two comparisons are assumed - representing a comparison of each intervention with the control group.

Assuming 1:1 comparison and a 2-tailed test; α=0.05, α/2=0.025; statistical power of 0.80.
$M62$ = 2.8464; $M66$ = 2.8437.
Therefore, assuming 33 schools per arm (66 in each comparison);
$MDES2LCRT$=2.8464√0.00464 =0.1939 ~ 0.19
Assuming 35 schools per arm (70 in each comparison);
$MDES2LCRT$=2.8437√0.00438 =0.1881 ~ 0.19
So, if multiple comparisons are ignored, the MDES estimate is 0.19 standard deviations.

Assuming two comparisons (OB v control; CJ v control) and a 2-tailed test; Bonferroni correction used for α, per comparison = α /2 = 0.025, two-tailed = 0.0125; Statistical power of 0.80.
$M62$ = 3.1446. ; $M66$ = 3.1408.
Therefore, assuming 33 schools per arm (66 in each comparison);
$MDES2LCRT$=3.1446√0.00464 =0.214 ~ 0.21
Assuming 35 schools per arm (70 in each comparison);
$MDES2LCRT$=3.1408√0.00438 =0.208 ~ 0.21
Correcting for 2 comparisons increases the MDES estimate to 0.21.

Assuming three comparisons (OB v CJ; OB v control; CJ v control) and a 2-tailed test; Bonferroni correction used for α, per comparison = α /3 = 0.0167, two-tailed = 0.0083; Statistical power of 0.80. $M62$ = 3.1446.
$M62$ = 3.3081. ; $M66$ = 3.3036.
Therefore, assuming 33 schools per arm (66 in each comparison);
$MDES2LCRT$=3.3081√0.00464 =0.225 ~ 0.23
Assuming 35 schools per arm (70 in each comparison);
$MDES2LCRT$=3.3036√0.00438 =0.218 ~ 0.22
Correcting for 3 comparisons increases the MDES estimate to between 0.22 and 0.23.

For the planned follow-on subgroup analyses involving pupils ever classed as FSM, the same equation is used but with a reduced number of FSM pupils per school:
Km is the total number of individual (FSM pupils=396x2=792; 420x2=840)
m is the number of pupils per school (=12)

- • Assuming 1:1 comparison: MDES estimate = 0.22 (70 schools) - 0.23 (66 schools)
- • Correcting for 2 comparisons: MDES estimate = 0.24 (70 schools) - 0.25 (66 schools)
- • Correcting for 3 comparisons: MDES estimate = 0.26 (70 schools) - 0.27 (66 schools)

Summary

For the headline ITT analyses we estimate that the proposed design will be able to detect an effect size of 0.21 sds for all pupils and an effect size of between 0.24 and 0.25 for FSM pupils.

Appendix 2

Scoring the Strengths & Difficulties Questionnaire for age 4-17 or 18+

The 25 items in the SDQ comprise 5 scales of 5 items each. It is usually easiest to score all 5 scales first before working out the total difficulties score. 'Somewhat True' is always scored as 1, but the scoring of 'Not True' and 'Certainly True' varies with the item, as shown below scale by scale. For each of the 5 scales the score can range from 0 to 10 if all items were completed. These scores can be scaled up pro-rata if at least 3 items were completed, e.g. a score of 4 based on 3 completed items can be scaled up to a score of 7 (6.67 rounded up) for 5 items.

Note that the items listed below are for 4-17-year-olds, but the scoring instructions are identical for the similarly-worded '18+' SDQ

Table 1: Scoring symptom scores on the SDQ for 4-17 year olds

	Not True	Somewhat True	Certainly True
Emotional problems scale			
ITEM 3: Often complains of headaches… *(I get a lot of headaches…)*	0	1	2
ITEM 8: Many worries… *(I worry a lot)*	0	1	2
ITEM 13: Often unhappy, downhearted… *(I am often unhappy….)*	0	1	2
ITEM 16: Nervous or clingy in new situations… *(I am nervous in new situations…)*	0	1	2
ITEM 24: Many fears, easily scared *(I have many fears…)*	0	1	2
Conduct problems Scale			
ITEM 5: Often has temper tantrums or hot tempers *(I get very angry)*	0	1	2
ITEM 7: Generally obedient… *(I usually do as I am told)*	2	1	0
ITEM 12: Often fights with other children… *(I fight a lot)*	0	1	2
ITEM 18: Often lies or cheats *(I am often accused of lying or cheating)*	0	1	2
ITEM 22: Steals from home, school or elsewhere *(I take things that are not mine)*	0	1	2
Hyperactivity scale			
ITEM 2: Restless, overactive… *(I am restless…)*	0	1	2
ITEM 10: Constantly fidgeting or squirming *(I am constantly fidgeting….)*	0	1	2
ITEM 15: Easily distracted, concentration wanders *(I am easily distracted)*	0	1	2
ITEM 21: Thinks things out before acting *(I think before I do things)*	2	1	0
ITEM 25: Sees tasks through to the end… *(I finish the work I am doing)*	2	1	0
Peer problems scale			
ITEM 6: Rather solitary, tends to play alone *(I am usually on my own)*	0	1	2
ITEM 11: Has at least one good friend *(I have one goof friend or more)*	2	1	0
ITEM 14: Generally liked by other children *(Other people my age generally like me)*	2	1	0
ITEM 19: Picked on or bullied by other children… *(Other children or young people pick on me)*	0	1	2
ITEM 23: Gets on better with adults than with other children *(I get on better with adults than with people my age)*	0	1	2
Prosocial scale			
ITEM 1: Considerate of other people's feelings *(I try to be nice to other people)*	0	1	2
ITEM 4: Shares readily with other children… *(I usually share with others)*	0	1	2
ITEM 9: Helpful if someone is hurt… *(I am helpful is someone is hurt…)*	0	1	2
ITEM 17: Kind to younger children *(I am kind to younger children)*	0	1	2
ITEM 20: Often volunteers to help others… *(I often volunteer to help others)*	0	1	2