STATISTICAL ANALYSIS PLAN

Toward Sport - A randomised multi-site trial to evaluate a sports-based intervention aiming to enhance postive outcomes for children and young people in the context of youth offending

Alma Economics

Principal investigator: Nick Spyropoulas

YOUTH ENDOWMENT FUND

Toward Sport – A randomised multi-site trial to evaluate a sports-based intervention aiming to enhance positive outcomes for children and young people in the context of youth offending

Statistical analysis plan

Evaluating institution: Alma Economics

Principal investigator(s): Nick Spyropoulos

YEF statistical analysis plan

Project title[1]	Toward Sport – A randomised multi-site trial to evaluate a sports-based intervention aiming to enhance positive outcomes for children and young people in the context of youth offending.
Developer (Institution)	StreetGames
Evaluator (Institution)	Alma Economics
Principal investigator(s)	Nick Spyropoulos
SAP author(s)	Nick Spyropoulos, Suzie Harrison, Lucille McKnight
Trial design	Multi-site trial: Two-arm individual-level randomisation of CYP within local authorities. Intervention will take place across multiple (~50) Delivery Partner Organisations, which will adhere to a Shared Practice Model, ensuring a consistent intervention across sites.
Trial type	Efficacy trial with internal pilot

[1] Please make sure the title matches that in the header and that it is identified as a randomised trial as per the CONSORT requirements (CONSORT 1a).

Evaluation setting	Local Authorities and Delivery Partner Organisations (DPOs)
Target group	10- to 17-year-olds at a tertiary and secondary level of risk of offending (see detailed criteria below)
Number of participants	2,500
Primary outcome and data source	Offending (violent and non-violent, source = Police National Computer and local police force data)
Secondary outcome and data source	Conduct problems, hyperactivity/inattention, peer relationship problems, and prosocial behaviour (as measured by the Strengths and Difficulties questionnaire self-rated version for 11-17-year-olds)*. The outcome measure will be the total score, as well as each of the following subscales: ● Conduct problems subscale. ● Hyperactivity/inattention subscale. ● Peer relationships problem subscale. ● Prosocial behaviour subscale. Wellbeing as measured by the ONS 4 questions Physical Activity Participation** Transferable Skills and Knowledge*** Data will be collected through surveys with CYP participating in the evaluation. * Goodman R, Ford T, Corbin T, Meltzer H. Using the Strengths and Difficulties Questionnaire (SDQ) multi-informant algorithm to screen looked-after children for psychiatric disorders. Eur Child Adolesc Psychiatry. 2004;13 Suppl 2:II25-31. doi: 10.1007/s00787-004-2005-3. PMID: 15243783. For those aged 10 in the study sample, the case worker will instruct and work with the parent of the CYP to implement the One-sided SDQ for parents or teachers of 4-17-year-olds, found on the SDQ tool site here.

	**As measured by Milton K, Bull FC, Bauman A. Reliability and validity testing of a single-item physical activity measure. Br J Sports Med. 2011 Mar;45(3):203-8. doi: 10.1136/bjsm.2009.068395. Epub 2010 May 19. PMID: 20484314. ***As measured by the National Citizen Service Evaluation by DCMS or the Youth Rating of Socio-emotional Skills

SAP version history

Version	Date	Changes made and reason for revision
1.2 [*latest*]		
1.1		
1.0 [*original*]		

Table of contents

Introduction

This project involves carrying out a randomised multi-site trial designed to evaluate a sports-based intervention aiming to enhance positive outcomes for Children and Young People (CYP) in the context of youth offending. The objectives of the trial are:

- To estimate the impact of participation in voluntary sports programmes on youth offending rates (violent and non-violent offending).
- To estimate the impact of participation in voluntary sports programmes on secondary outcomes, such as conduct problems, hyperactivity/inattention, peer relationship problems, and prosocial behaviour (as measured by the Strengths and Difficulties tool), wellbeing, participation in physical activity, and transferable skills and knowledge, to assess the mechanisms underlying the efficacy of the intervention.
- To contribute to the evidence gap on the efficacy of such positive activities on offending and reoffending for children and young people from Black, Asian, and minority ethnic backgrounds.

The programme is offered on a voluntary basis, will be available for 24 weeks for each participating young person, and consists of weekly group-based sessions lasting two hours. The session will be delivered by Delivery Partner Organisations (DPOs) already working with vulnerable or at-risk children aged 10-17 years old with advanced safeguarding practices and risk assessments in place, or familiar with embedding them. Young people with a tertiary and upper-secondary level of need will be eligible for participation in the evaluation and will be referred to the programme by caseworkers in the local authority across Youth Justice, Supporting Families, and other Early Help teams. These local authority teams will also assist with identifying eligible young people and baseline data collection.

This trial is designed as a multi-site trial to: (i) leverage the large networks of DPOs delivering sports programmes with at-risk cohorts of CYP, providing sufficient sample sizes for the efficacy trial, and reflecting a delivery model consistent with widespread practice; and (ii) working with an Umbrella Organisation (StreetGames) to ensure a consistent model of delivery is being tested against business-as-usual across sites.

The trial was preceded by a pilot phase. The pilot included delivery of the intervention from November 2024 to May 2025 (with a review point in February 2025, prior to the start of the full efficacy trial). The full evaluation includes a delivery phase from May 2025 to April 2026.

Design overview

Trial design, including number of arms		Efficacy trial. Two-armed multi-site trial with randomisation at the individual (CYP) level. Within each local authority, CYP are randomised after referral, when they have provided their consent to participate in the evaluation. The randomisation occurs on a rolling basis after the eligible CYP engages with the practitioner and provides their consent. If young people do not consent to participate in the evaluation, they are not included in the trial.
Unit of randomisation		Individual CYP level, within local authorities on a rolling basis, on a 50-50 treatment-control basis to maximise power.
Stratification variables (if applicable)		Randomisation takes place within local authorities (stratification at the local authority level). Within local authorities, randomisation occurs on a rolling basis at the CYP level.
Primary outcome	variable	Prevalence of offending: Binary variable if an offence or multiple offences (violent and non-violent) occur in the data between baseline and follow-up (true for both follow-ups).
	measure (instrument, scale, source)	Recorded incidents to date, 0 upwards, Police National Computer.
Secondary outcome(s)	variable(s)	● Conduct problems, hyperactivity/inattention, peer relationship problems, and prosocial behaviour. ● ONS4 Wellbeing ● Physical Activity

		● Transferable skills and knowledge
	measure(s) (instrument, scale, source)	● Strengths and Difficulties questionnaire (one-sided self-rated SDQ for 11-17-year-olds), response scale is Not True/Somewhat True/Certainly True, scoring follows the SDQ scoring approach. For 10-year-olds in the study, the case worker will instruct and work with the parent of the CYP to implement the one-sided SDQ for parents or teachers of 4-17-year-olds, found on the SDQ tool site here. ● ONS4 Wellbeing Questions, Scale 0-10. ● Milton et al. (2010) single-item physical activity measure, Scale 0-7 ● Transferable skills and knowledge questions used in DCMS evaluation of National Citizen Service, Very Confident/Confident/Neither confident nor not confident/Not very confident/Not at all confident/Don't Know items converted to 0-6 Scale.
Baseline for primary outcome	variable	Prevalence of offending: Binary variable equal to 1 if an offence or multiple offences (violent and non-violent) occur in the data at any point between baseline and follow-up (true for both six-month and 12-month follow-ups).
	measure (instrument, scale, source)	Recorded incidents before referral into the programme using PNC. At baseline, this variable is equal to 1 if any offence occurs prior to baseline (e.g., if there have been any offences in the CYP's record); at follow-up, this is equal to 1 if any offence occurs between baseline and follow-up (true for both follow-ups).

Baseline for secondary outcome	variable	• Conduct problems, hyperactivity/inattention, peer relationship problems, and prosocial behaviour. • ONS4 Wellbeing • Physical Activity • Transferable skills and knowledge
	measure (instrument, scale, source)	The variables listed above are measured at baseline when CYP consent to participate in the programme using the following instruments: • Strengths and Difficulties questionnaire (one-sided self-rated SDQ for 11-17-year-olds), response scale is Not True/Somewhat True/Certainly True, scoring follows the SDQ scoring approach. • ONS4 Wellbeing Questions, Scale 0-10. • Milton et al. (2010) single-item physical activity measure, Scale 0-7. • Transferable skills and knowledge questions used in DCMS evaluation of National Citizen Service, Very Confident/Confident/Neither confident nor not confident/Not very confident/Not at all confident/Don't know items converted to 0-6 Scale.

A multi-site trial is required for this evaluation to gather sufficient sample sizes of CYP receiving comparable support through community-based sports organisations. Since such organisations are typically small, it is infeasible to design an RCT within a single organisation. The multi-site trial allows a larger number of CYP to be recruited for the evaluation by partnering with multiple organisations. To ensure that the treatment being tested is consistent across CYP and organisations, a Shared Practice Model has been developed for organisations to deliver common sports programme components. Multi-site trials also improve the external validity of evaluations versus single-site settings[2].

Randomisation will be conducted at the individual CYP level on a rolling basis within each local authority to maximise statistical power, ensuring that local authority-specific variation does not absorb or confound treatment/control variation. Once local authority-specific stratification is accounted for, the treatment and control groups will be similarly representative of the population of CYP who are eligible and consent to participate in the evaluation—there will be no differential characteristics between treatment and control CYP, driven by the treatment-control allocation differing across sites.

Two arms are chosen because of the trial's design relative to the core research questions: comparing the impact of Toward Sport against business-as-usual activities.

The trial will be delivered across eight local authorities, ~50 DPOs, with a minimum of ~2500 individual CYP randomised into treatment and control at referral. The delivery of the programme and evaluation will be conducted through practitioners across different teams in the local authority (e.g., Supporting Families teams), with the data team in the local authority supporting the identification of eligible CYP and data-sharing, and the case worker engaging with CYP to assess interest in sports and consent to participate in the evaluation, as well as collecting primary data.

[2] William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Boston: Houghton Mifflin Company, 2002); Graeme Blair and Gwyneth McClendon, "Conducting Experiments in Multiple Contexts," in *Handbook of Advances in Experimental Political Science*, eds. James N. Druckman and Donald P. Green (Cambridge: Cambridge University Press, 2021).

Sample size calculations overview

		Protocol	Randomisation
Minimum Detectable Effect Size (MDES)		19%[3] (0.11 in Cohen’s h)[4]	
Pre-test/ post-test correlations	level 1 (participant)	0	
	level 2 (cluster)	N/A	
Intracluster correlations (ICCs)	level 1 (participant)	0	
	level 3 (cluster)	N/A	
Alpha		0.05	
Power		0.8	
One-sided or two-sided?		Two-sided	
Average cluster size		N/A	
Number of clusters	intervention	N/A	
	control	N/A	
	total	N/A	

[3] The MDES is measured as a percentage point change.

[4] Cohen’s h can be interpreted in the same way as Cohen’s d with respect to the magnitude of the effect. See Cohen (1988) for more details on the use of Cohen’s h in the context of differences between proportions.

		Protocol	Randomisation
Number of participants	intervention	1,250	
	control	1,250	
	total	2,500[5]	

Sample size calculation

The sample size was determined in collaboration with the project team, StreetGames, based on information gathered from local authorities that are involved in the efficacy study and insights that were collected during the pilot. Alma Economics conducted power calculations to estimate the minimum sample required to detect an effect size in line with the literature (i.e., a reduction in offending rates between 30% and 50%)[6]. For the power calculations, we assumed a significance level (alpha) of 0.05 and a statistical power of 80%. Additionally, we assumed a pre-post-test correlation equal to zero[7].

The evaluation team then discussed the results of the power calculations as well as the assumptions underpinning the calculations with StreetGames to assess the feasibility of achieving the minimum sample size from an operational perspective. Based on predictions of cohorts of CYP shared by the local authorities participating in the trial to StreetGames, StreetGames suggested that a minimum sample of approximately 2,500 CYPs being randomised is achievable.

The table below presents sample size estimates for different levels of the Minimum Detectable Effect Size (MDES), expressed both as a percentage reduction in the probability of offending and as a standardised effect. The MDES ranges from 50%, which is the approximate effect estimated in the YEF Sports Toolkit, to 19%, which corresponds to the effect associated with the sample size that is operationally feasible, according to StreetGames and predictions

[5] 2,500 CYP is the required sample size at endline.

[6] Power calculations were carried out in STATA. The full code can be found in Appendix B.

[7] We assumed a pre-test post-test correlation equal to zero for several reasons: (i) we do not have a prior of what the correlation may be, (ii) there is a debate in the literature on whether one should control for such correlation when calculating the sample, and it is not clear that one approach is better than another one, and (iii) there are additional factors that, once accounted for, may impact the statistical power and therefore the sample size in different ways (e.g., clustering may increase the sample size needed, controlling for other confounding factors may increase/decrease the sample size, etc.). For these reasons, we believe it is prudent to follow a conservative approach and assume a correlation of 0 at this stage.

shared by local authorities. The calculation has been carried out under an assumed 25% offending rate in the control group. This is an approximation based on evidence from a Department for Education report and recent Youth Justice Statistics on offending and re-offending amongst young people who have risk factors aligned with the eligibility criteria for the trial. An assumed control group offending rate of 25% is also used by the YEF Sports Toolkit in their calculation of effect size.

The proposed sample size (in bold) allows for the identification of an impact of a 19% reduction in offending rates for the whole sample, with the standardised effect size detectable below the 0.2 threshold, and a 33% reduction for CYP from Black, Asian, and minority ethnic backgrounds. In both cases, the sample sizes will be large enough to ensure a well-powered trial and detect an effect in line or smaller than the average impact estimated in similar studies.

Table 1 Effect size for a range of sample sizes

MDES %	Treatment effect size in Cohen's h[8]	Sample size
50%	0.32	304
45%	0.29	386
40%	0.25	500
35%	0.22	670
30%	0.18	932
25%	0.15	1372
19%	**0.11**	**2500**

[8] Binary effects are translated into a standardised effect size using the arcsin transformation (Cohen, 1988): insert the two proportions into this function: 2*asin(sqrt(p1))-2*asin(sqrt(p2)). In each of the above, the pre-post-test correlation is assumed to be zero for conservativeness.

We use Cohen's h (rather than Cohen's d) because our primary outcome (offending) is binary. Cohen's d is calculated as $\frac{X_t - X_c}{S_{pooled}}$ or the control group mean subtracted from the treatment group mean, divided by the pooled standard deviation[9]. While this approach works for comparing the difference between two means, when working with proportions, simply subtracting one percentage from another can be misleading because the given difference depends on the value of both proportions.

Cohen's h uses a non-linear transformation to standardise the difference between proportions, such that the effect size is comparable across the entire range of values. Therefore, Cohen's h gives values that do not depend on whether the proportion falls in the middle or on one side of a range. A full mathematical explanation is provided in Cohen (1988)[10].

The sample size is indicative at this stage and represents the minimum number, rather than a final target. Indeed, referrals will continue until the end of the six-month referral period, or until 3,000 young people are randomised into the trial. Additionally, if CYP who are randomised in the treatment group decide not to engage with the programme (i.e., they do not attend sports sessions for eight consecutive weeks), while they will continue being part of the evaluation and will be included in the analysis, they will be replaced by new referrals until a maximum of 3,000 randomised in the trial is reached. This is illustrated through the following examples:

- Child 1: Randomised into the treatment group. They take up a place in a DPO and attend consistently. This child will be included in the analysis at endline.
- Child 2: Randomised into the treatment group, but does not take up their place in a DPO. If they have not engaged for eight weeks, they are replaced within the programme by another young person. However, as they were randomised in, they are still included in the treatment group during analysis to estimate the intention-to-treat basis (ITT).

Population of interest

The children and young people eligible for the programme will be those with a tertiary or secondary level of need:

[9] https://pmc.ncbi.nlm.nih.gov/articles/PMC5133225/pdf/kjae-69-555.pdf

[10] See pages 180-182 in https://www.utstat.toronto.edu/brunner/oldclass/378f16/readings/CohenPower.pdf

- Tertiary level of need: Young people who have already been involved in crime or anti-social behaviour. This does not include CYP living in the secure estate. This includes CYP aged 10-17 years[11] who meet any of the following criteria:
  - CYP who have been provided with a warning or caution.
  - CYP who have been arrested but not convicted.
  - CYP who have been arrested and convicted.
  - CYP who have been involved in anti-social behaviour, defined as conduct that has caused or is likely to cause harassment, alarm, or distress to any person.
  - CYP who are violent or abusive in their home, are involved in gangs, serious violence, weapons carrying, or other high-risk-taking behaviour.
- Secondary level of need: Young people aged 10-17 years who meet any of the following criteria:
  - At risk of or experiencing criminal or pre-criminal exploitation.
  - Experiencing harm outside the family (e.g., peer-to-peer abuse, online harassment, or sexual harassment or offences).
  - Currently or historically affected by domestic abuse.
  - Identified as being at risk of or affected by radicalisation.
  - Lives with an adult (18+) who is involved in crime and/or ASB (at least one: offence/arrest/named as a suspect/ASB incident in the last 12 months).
  - Excluded from school and not engaging in education (and not employed).

Based on referrals from the pilot, we anticipate that approximately 50% of CYP referred to the evaluation will have a tertiary level of need, while 50% will have a secondary level of need. The eligibility criteria for the evaluation were selected to align with the Supporting Families eligibility criteria, to ensure a similar cohort of young people were referred across areas, and practitioners were already familiar with the criteria, lessening the burden placed on those making referrals. The criteria were tailored to the trial and finalised based on discussions with local authority practitioners, service managers, and StreetGames.

[11] A small number of young people referred to the trial are expected to turn 18 during the course of the evaluation. These young people will be included in the analysis, on the basis that there were 17 when they were referred in, and therefore have met the eligibility criteria.

Analysis

The analysis of the data will be on an ITT basis. The ITT parameter will be estimated based on a regression of the follow-up outcome on the treatment indicator, the baseline level of the outcome, and local authority (strata) fixed effects. This approach follows the 'Conditional inference' YEF analysis guidance[12]. The confidence intervals will be based on heteroskedasticity-robust standard errors at the individual level[13].

Primary outcome analysis

The main regression model specification is as follows:

$$(1) \qquad YF_{il} = \propto + \beta 1 Treatment_{il} + \beta 2 YB_{il} + \beta 3 LA_{il} + \varepsilon_{il}$$

- *i* indicates the young person and *l* the local authority they belong to;
- YF_{il} is a binary indicator of the probability of offending following participation in the programme, and it is equal to 1 if the young person has offended between the start of the programme and the follow-up and 0 if they have not offended;
- $Treatment_{il}$ estimates the treatment effect on the programme on the likelihood of offending, and it is equal to 1 if the young person was assigned to the treatment group and 0 as assigned to control;
- YB_{il} is a binary variable indicating whether the young person has offended before participating in the programme (once or multiple times); this control variable accounts for pre-existing differences between young people;
- LA_{il} is a dummy variable capturing the local authority fixed effects;
- ε_{il} is the error term, which captures any unobserved factors affecting the probability of offending.

If the randomisation is less effective or the sample size is significantly lower than expected, we could include additional covariates, such as age, gender, and ethnicity in our model to increase the precision of our estimates.

Equation (1) will be estimated using Ordinary Least Squares (OLS), effectively representing a linear probability model. Heteroskedasticity-robust standard errors will be employed, as the

[12] As highlighted in the guide, conditional inference is more appropriate when we do not attempt to generalise beyond the sites within a trial; this approach is more appropriate for efficacy trials and requires the use of a fixed effects model.

[13] The analysis will be conducted using Stata 18.5.

binary nature of the dependent variable will induce heteroskedasticity. Additionally, we will estimate Equation (1) using a logit model via Maximum Likelihood. The logit model accommodates non-linear effects and ensures predicted probabilities lie between 0 and 1, unlike the linear probability model, which may produce estimates outside this range. However, this limitation is less of a concern in the present context, as our primary interest lies in the coefficient on the treatment variable—for which the linear probability model yields unbiased and consistent estimates.

Secondary outcome analysis

The secondary outcome variables measured through the Strengths and Difficulties questionnaire will be the total score, and the comparison in the analysis will be the mean total score in the treatment group versus the mean total score in the control group. We will also analyse the impact of the treatment on each of the following subscales of the Strengths and Difficulties: conduct problems subscale, hyperactivity/inattention subscale, peer relationships problem subscale, and prosocial behaviour subscale.

The trial will not be powered to these outcomes, and the secondary outcome analysis is exploratory. These analyses are included to test key mechanisms identified in our Theory of Change (ToC). In addition to hypothesising that participation in sport may reduce the probability of offending or reoffending, our ToC also hypothesises that there is a link between participation in sport and other positive outcomes, including the development of prosocial identities, and positive contribution of CYP to their communities.

To estimate the impact of the programme on secondary outcomes, we will follow the same approach adopted for primary outcomes by estimating the following equation:

$$SDQF_{il} = \propto + \beta 1T_{il} + \beta 2SDQB_{il} + \beta 3LA_{il} + \varepsilon_{il}$$

$SDQF_{il}$ indicates the SDQ score at follow-up and $SDQB_{il}$ is the SDQ score at baseline. As mentioned above, we will estimate the impact using the Total Difficulties score (0 to 40) as a general measure of mental health, as well as the impact on individual SDQ subscales.

Other secondary outcomes will include personal wellbeing, physical activity, and transferrable skills and knowledge. We will adopt the same model as used for the SDQ, wherein our outcomes variable is the relevant score at follow-up, with the baseline score as a regressor, comparing the mean scores in the treatment group to the mean scores for the control.

The data-collection tools consist of the Strengths and Difficulties Questionnaire[14] for the 11-17 age group;[15] the ONS 4 Wellbeing questions[16]; the single-item measure for physical activity[17], and Transferrable Skills, and Knowledge questions taken from the DCMS' Evaluation of the National Citizen Service[18], collected at baseline and following completion of the intervention.

Subgroup analyses

We will examine whether the effectiveness of the programme differs for young people from Black, Asian, and minority ethnic backgrounds. This analysis is motivated by the gap in the literature around the role of sport in reducing offending for young people from Black, Asian, and minority ethnic backgrounds. It is plausible that the programme has a stronger effect on CYP from Black, Asian, and minority ethnic groups due to differing baseline risks, access to or engagement with services, or responsiveness to intervention. In addition to testing for heterogeneous effects based on race and ethnicity, we will conduct additional analysis focused specifically on race. This analysis separates white from non-white CYP to explore a shared experience of racism or racialisation among non-white CYP, which may impact their outcomes or interaction with police services.

To assess this, we will augment Equation (1) by including an interaction term between the treatment variable and an ethnicity indicator identifying CYP from Black, Asian, and minority ethnic backgrounds. We will also include an additive dummy to account for potential differences in baseline offending rates between white and non-white CYP[19].

We will also conduct additional subgroup analyses, examining whether the programme's impact varies by: (i) more granular ethnic categories—distinguishing between Black, Asian, and other minority ethnic groups; (ii) special educational needs (SEN) status; and (iii) level of risk (i.e., secondary versus tertiary risk levels). However, the results from these analyses

[14] https://www.sdqinfo.org/py/sdqinfo/b3.py?language=Englishqz(UK)

[15] For those aged ten in the study sample, the case worker will instruct and work with the parent of the CYP to implement the One-sided SDQ for parents or teachers of 4-17-year-olds, found on the SDQ tool site here.

[16] https://evaluationframework.sportengland.org/media/1333/sport-england-child-question-bank.pdf

[17] Milton K, Bull FC, Bauman A. Reliability and validity testing of a single-item physical activity measure. Br J Sports Med. 2011 Mar;45(3):203-8. doi: 10.1136/bjsm.2009.068395. Epub 2010 May 19. PMID: 20484314.

[18] https://assets.publishing.service.gov.uk/media/61323c95d3bf7f05b3fbd767/NCS_2019_Evaluation_Technical_Report.pdf

[19] An alternative approach would be to estimate separate models for each group. However, using a single model with additive and interaction terms offers greater flexibility, particularly by allowing us to impose common restrictions, such as local authority fixed effects across groups, thereby increasing the statistical power of the analysis.

should be interpreted as indicative, given the potential limitations in statistical power that may affect the precision of the estimated impacts.

Sensitivity analysis and robustness checks

We will conduct the following further sensitivity analyses:

- **Covariate adjustment**: For both primary and secondary outcomes, we will also explore the impact of mediating factors in the efficacy of the intervention by running the same regression specification presented in the "Primary outcomes analysis" section, but interacting the following variables with the treatment indicator:
    - an indicator for whether the CYP is male.
    - an indicator for whether the CYP has a tertiary level of risk.
    - an indicator for whether the CYP has special education needs.
    - whether the CYP is the same sex as the coach of the sports sessions.
    - whether the CYP is the same ethnicity as the coach of the sports sessions.
- **Dosage**: We will replace the treatment indicator with a variable for the number of sessions attended to assess the extent to which efficacy varies by attendance. Monitoring data on attendance will be captured weekly by DPOs and shared with the evaluation team.
- **Differential impact between LAs and DPOs**: we will interact local authority fixed effects and DPO fixed effects with the treatment indicator to explore whether there is evidence of variation in impacts across areas and organisations.

To mitigate the risk associated with 'data mining,' whereby some models might indicate statistical significance by chance, we use the following strategies:

- Our chosen specifications, subgroup analyses, and sensitivity checks are grounded in economic theory and informed by causal chains identified within our ToC and previous evaluations or research, meaning our results can be tested against the ToC and existing literature.
- Our interpretation of the results will be transparent and thorough, reporting findings for all estimated specifications rather than cherry-picking.
- We will place more emphasis on parsimonious models. For example, we will identify the preferred models using the Bayesian Information Criteria (BIC), which penalises large models and tends to select parsimonious models.

Interim analysis

In addition to analysing data from the efficacy trial, we will analyse results from the pilot. This will include regression analysis of (i) offending outcomes (i.e., our primary outcome) via offending data provided by the Local Authority[20] and (ii) secondary outcomes collected through follow-up surveys with CYP. Primary and secondary outcome analysis for the pilot will be conducted in line with the specifications provided above for the main efficacy trial. Analysis of pilot data will be purely exploratory, and largely to test data collection processes, due to it being a very small sample.

Longitudinal follow-up analyses

Data for both the treatment and control participants will be collected at the following intervals:

- Baseline (at the stage of referral, after the CYP consents to participate in the evaluation); this will include the collection of demographic data and data necessary for data archiving.
- At the end of the 24-week timeframe, consistent with referral into the evaluation.
- Six months after the 24-week timeframe has ended.

The two follow-ups will allow for analysis of the impact of participation in Toward Sport in the short run, as well as assessing whether the programme has a longer-term impact. We will conduct longitudinal follow-up analyses using the same equation as our ITT (Equation 1), where $YFollow - up2_{il}$ is the outcome at the second follow-up (e.g., 12 months after baseline data collection):

$$(2) YFollow - up2_{il} = \propto + \beta_1 Treatment_{il} + \beta_2 YBaseline_{il} + \beta_3 LA_{il} + \varepsilon_{il}$$

Imbalance at baseline

We will summarise the following characteristics of the treatment and control groups at baseline and both follow-up points:

- Ethnicity
- Age
- Gender
- Level of need (i.e., tertiary or secondary)

[20] PNC data will not be available for the analysis of the pilot.

- Looked-after status
- SEND status
- Baseline offending record
- Baseline SDQ score

Descriptive statistics presented at baseline will include all CYPs who were randomised and will show whether the randomisation resulted in a balanced sample. We will carry out t-tests to document whether there are significant differences in the characteristics listed above between treatment and control groups. While t-tests are not always necessary in the context of an RCT, because we expect some attrition at follow-up (especially in terms of the secondary outcomes captured by the survey), t-tests help to demonstrate that the sample engaged with at follow-up is similarly balanced to baseline observable characteristics and provides a further check that our randomisation has been successful.

Summaries of characteristics at follow-up will include all CYP included in the analysis and will indicate whether attrition was higher for CYP with specific characteristics or across treatment versus control, creating an unbalanced sample. We have selected the above characteristics based on their relevance to our research questions and their likelihood of affecting our primary outcome measure.

Missing data

Collecting high-quality and comprehensive data will be a priority of the trial. One of the key findings from the pilot was that baseline data collected by practitioners was generally high-quality and complete. Despite the pilot taking place in only one local authority, because practitioners are already familiar with many of the secondary outcome measures (e.g., SDQ), and we received positive feedback from practitioners on the survey's ease of use, we are confident that missing items will not be a significant problem in the efficacy trial.

As we plan to use administrative data on offending, we expect missingness in our primary outcome to be very low. We do anticipate some missingness in the trial, as data on secondary outcomes are collected through self-report survey data. This could be all observations missing based on non-response to follow-up surveys, as well as missing data on covariates based on some questions being left blank.

In line with YEF analysis guidance[21], the primary ITT regression model will be based on complete cases, thus assuming that data is missing at random. We will specify the number of

[21] Available at: https://res.cloudinary.com/yef/images/v1623145483/cdn/6.-YEF-Analysis-Guidance/6.-YEF-Analysis-Guidance.pdf.

complete cases in our analyses. However, if more than 5% of the data must be excluded from the model due to missing data, we will investigate the nature of the missing data following the steps highlighted in the flow chart (see Appendix A) in the YEF analysis guidance. We anticipate that there will be very little missing data in our primary outcome, as it will be sourced from administrative data (e.g., the Police National Computer). The first step in investigating the nature of the missing data will be to use logistic regression to analyse the extent to which missing data is attributable to observable characteristics. Depending on the results of the analysis, we will decide whether multiple imputation (MI) or sensitivity analyses will be carried out.

If the logit model determines that missing data is not attributable to specific observable characteristics, we will undertake MI and compare the results of MI to our analysis using only complete cases. The differences between the analysis using complete cases and imputed values will be clearly outlined in the report, alongside the implications of our findings. If missing data is attributable to covariates or observable characteristics, we will conduct additional sensitivity analyses.

Compliance

In line with YEF analysis guidance, our analysis will be on an ITT basis. However, ITT may underestimate the true effect of Toward Sport on those who take up the offer, as attendance is voluntary, and some young people who are assigned to treatment may still choose not to attend sports sessions. To understand the effect of the treatment on those who comply with assignment to the treatment group (e.g., the Local Average Treatment Effect (LATE)), we need to undertake additional modelling to understand (i) the rate of non-compliance and (ii) whether non-compliance is non-random or associated with other observed factors, potentially biasing our specification. To understand the potential bias, we will calculate whether any specific characteristics are associated with non-compliance. To mitigate against this bias, we will use an Instrumental Variable (IV) approach (see Angrist, Imbens, and Rubin, 1996). Bias is relevant in this context as individuals who believe they will benefit more from the programme may be more likely to attend sessions, while those who are less motivated or expect little benefit may drop out, which would create an upward bias in the LATE estimate.

In the first step, the IV approach estimates a regression model between compliance and the initial assignment to treatment or control. In the second step, it uses the fitted values from the first-stage regression as the “treatment” variable. The IV estimator can be thought of as an adjusted ITT effect, where the ITT effect is divided by the proportion of individuals who actually receive the treatment.

This approach relies on the assumption that the only difference between compliers and non-compliers is their compliance with the treatment assignment. If there are unobserved factors that influence the likelihood of compliance—i.e., whether an individual chooses to participate in the programme—then there is a risk of selection bias, and the LATE may be biased. If the decision to participate is a function of observable characteristics, then including these characteristics in the model may help mitigate the bias. The interpretation of the results and the limitations of this approach will be documented in the final report.

In order to measure the rate of non-compliance, we will need to set a definition for compliance. Ideally, we would base this definition on evidence that shows the programme is most impactful on children after they attend for X number of weeks. However, there is scarce existing literature on how variance in attendance rates in multi-week voluntary sports programmes impacts their outcomes and no evidence directly relevant to the high-risk cohorts of young people in the trial. Setting a definition of compliance at this stage would require establishing an arbitrary cutoff.

We will therefore calculate descriptive statistics on different measures of attendance to shed additional light on the programme design and what attendance rate is likely in the context of voluntary programmes with this cohort. We will use this analysis to establish various definitions for compliance, for which we will estimate the LATE analysis. One possible example of compliance could be participation in at least eight consecutive sport sessions. This is consistent with our delivery parameter for the trial that young people who do not attend sessions at a DPOs for eight consecutive sessions will be replaced with a new referral.

Presentation of outcomes

For the binary primary outcome (i.e., probability of offending), we will convert the estimate into the relative risk ratio[22], comparing the control mean probability of offending (adjusted for local authority fixed effects) in the treatment and control group post-intervention:

$$Relative\ Risk\ Ratio\ (RR) = \frac{\hat{Y}_{Treatment}}{\hat{Y}_{Control}}$$

Where $\hat{Y}_{Treatment}$ is the estimated probability of offending in the treatment group after the intervention, with adjustment for covariates, while $\hat{Y}_{Control}$ is the estimated probability of offending in the control group with adjustment for the same covariates.

For continuous secondary outcomes, the effect sizes will be calculated using Hedges' g, as specified in the following equation:

$$ES = \frac{\hat{Y}_t - \hat{Y}_c}{s}$$

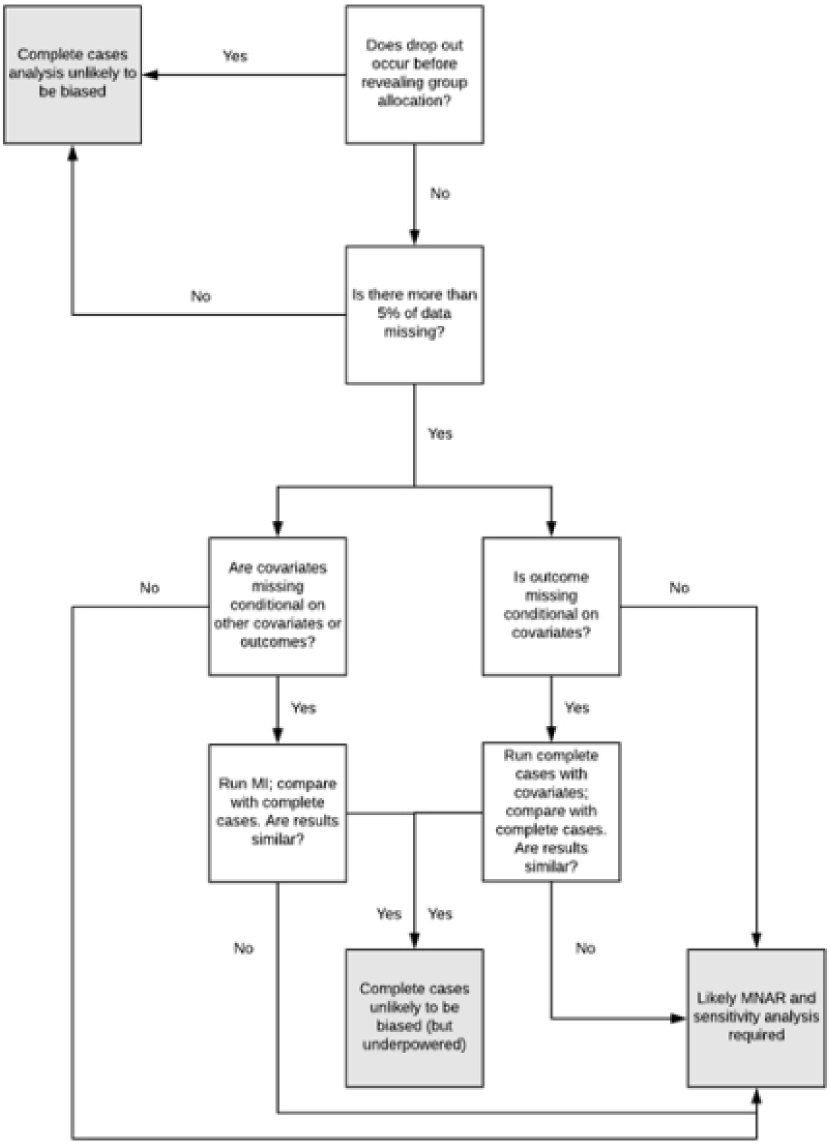
[22] See Tenny and Hoffman (2023). Available at: https://www.ncbi.nlm.nih.gov/books/NBK430824/

Where $\widehat{Y_t}$ and $\hat{Y}_c$ are the regression-adjusted mean for the treatment and control group, respectively, and s is the pooled standard deviation of both groups.

Alongside reporting of effect sizes, we will also report confidence intervals and p-values in full (rather than only reporting significance levels) to provide a measure of the statistical uncertainty of our estimates.

Appendix A

The following figure (taken from the YEF Analysis Guidance) shows the steps taken to investigate missing data in the analyses. The figure is taken from the YEF Analysis Guidance.


Complete cases analysis unlikely to be biased
Yes
Does drop out occur before revealing group allocation?
No
No
Is there more than 5% of data missing?
Yes
No
Are covariates missing conditional on other covariates or outcomes?
Is outcome missing conditional on covariates?
No
Yes
Yes
Run MI; compare with complete cases. Are results similar?
Run complete cases with covariates; compare with complete cases. Are results similar?
Yes
Yes
No
No
Complete cases unlikely to be biased (but underpowered)
Likely MNAR and sensitivity analysis required

Appendix B

The following code was executed in STATA for our power calculations:

```
//* Power calculation assuming pre-post test correlation equal to zero
//* Likelihood of offending pre programme (i.e. control group) = 25%
//* Power = 0.80
//* sample is balanced between treatment and control
//* (1) MDES 50%
di 0.25 * 0.5
power twoproportions 0.25, test(chi2) diff(-0.125)
*n = 304
//* (2) MDES 45%
di 0.25 * 0.45
power twoproportions 0.25, test(chi2) diff(-0.1125)
*n = 386
//* (3) MDES 40%
di 0.25 * 0.4
power twoproportions 0.25, test(chi2) diff(-0.1)
*n = 500
//* (4) MDES 35%
di 0.25 * 0.35
power twoproportions 0.25, test(chi2) diff(-0.0875)
*n = 670
//* (5) MDES 30%
di 0.25 * .3
power twoproportions 0.25, test(chi2) diff(-0.075)
```

```
*n = 932

//* (6) MDES 25%

di 0.25 * 0.25

power twoproportions 0.25, test(chi2) diff(-0.0625)

*n = 1372

//* (7) MDES 19%

di 0.25 * 0.19

powertwoproprotions 0.25, test(chi2) diff(-0.0475)

//* (8) MDES 20%

di 0.25 * 0.2

power twoproportions 0.25, test(chi2) diff(-0.05)

* n= 2188
```

YOUTH ENDOWMENT FUND

Impetus

Home Office

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413