STATISTICAL ANALYSIS PLAN

The London Young People Study: Evaluation of the Your Choice Programme Using a Cluster Randomised Controlled Trial

Institute for Fiscal Studies

Principal investigator: Imran Rasul

YOUTH ENDOWMENT FUND

The London Young People Study: An evaluation of the Your Choice intervention

Statistical analysis plan

Evaluating institution: Institute for Fiscal Studies and Anna Freud Centre

Principal investigator(s): Imran Rasul

Project title[1]	The London Young People Study: Evaluation of the Your Choice Programme Using a Cluster Randomised Controlled Trial
Developer (Institution)	London Innovation and Improvement Alliance (LIIA) Violence Reduction Unit (VRU)
Evaluator (led)	Institute for Fiscal Studies
Principal investigator(s)	Imran Rasul
Protocol author(s)	Sarah Cattan, Monica Costa Dias, Julian Edbrooke-Childs, Imran Rasul
Trial design	Two-armed cluster randomised controlled trial with random allocation at the youth practitioner level
Trial type	Efficacy
Evaluation setting	30 Local Authorities of London

[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).

Target group	Any child aged between 11-18 years old who is assessed as medium or high risk of harm / vulnerability as a result of extra-familial harm and has been considered by a multi-agency panel (typically MACE / Pre-MACE)
Number of participants	The number of participants was initially estimated to be between 1700 and 1800 young people across the internal pilot and efficacy stage. Across teams that have ever been randomised, this included an expected 1698 young people recruited between August 2023 and December 2024 and 161 young people who were recruited during the internal pilot phase (82 of which we have endline data for). We expect a 70% data completion rate, which sets the overall sample for impact evaluation at 1,272. Due to lower-than-expected recruitment and higher than expected attrition, in January 2025 the recruitment period was extended to the end of February 2025. To inform this decision, we re-estimated the number of participants based on recruitment to date and new information from the Local Authorities and revised the number of participants down to 1,563 young people recruited across the internal pilot and efficacy trial recruited between August 2023 and end of February 2025. Across teams that have ever been randomised, this included an expected 1,402 young people recruited between August 2023 and February 2025 and 161 young people who were recruited during the internal pilot phase (82 of whom we have endline data for). We expect a 70% data completion rate across the internal pilot and efficacy trials, which sets the overall sample for impact evaluation at 1,094.
Primary outcome and data source	Indicator for scoring in the high or very high range of the conduct problems subscale of the Strengths and Difficulties Questionnaires (SDQ). The high and very high range threshold is defined by the SDQ's four-fold categorisation. The SDQ is taken at the start and end of the young person's pathway through the programme, at weeks 1 and 20 after

	recruitment. In both cases, they are administered during a session with the practitioner, as part of the baseline and endline surveys. Some small variation in the timing of these surveys is allowed, to accommodate differing start and end dates for the programme and convenience in scheduling meetings between participants and practitioners. Below we will be referring to week 1 and 20 as expected times for these surveys.
Secondary outcome and data source	*Criminal activity*: Indicator for recorded arrest in Police National Computers during first 16 months after recruitment (allowing for 4-months treatment plus 1 year following treatment). *Self-reported and practitioner-reported perceptions of young person's safety*: Young person and practitioner versions of "Checkpoint. A safety scale for young people", which is an instrument to measure young people's perceptions of safety developed by the research team. Measured at baseline and endline as part of the participant's and practitioner's surveys (weeks 1 and 20 after recruitment, respectively). *Wellbeing*: The Short Warwick–Edinburgh Mental Well-being Scale (SWEMWBS). Measured at endline as part of the participant's survey administered in session with practitioner at week 20 after recruitment. *Emotional self-regulation*: Trait Emotional Intelligence Questionnaire – Adolescent Short Form (TEIQUE-ASF) – Self regulation subscale. Measured at endline as part of the participant's survey administered in session with practitioner at week 20 after recruitment. *Social connectedness*: Social Connectedness Scale – Revised (SCS-R). Measured at baseline and endline as part of the participant's surveys administered at weeks 1 and 20 after recruitment. *Internalising behaviours*: emotional difficulties and peer difficulties subscales of the SDQ measured at baseline and endline as part of the participant's surveys administered

	following recruitment and at week 20 after recruitment respectively. *Hyperactivity*: hyperactivity subscale of the SDQ measured at baseline and endline as part of the participant's surveys administered following recruitment and at week 20 after recruitment respectively. *Prosocial behaviours*: Strength and Difficulties prosocial behaviour subscale measured at baseline and endline as part of the participant's surveys administered following recruitment and at week 20 after recruitment respectively. *Prosocial identity*: Pro-social Identity Scale (PIDS) measured at baseline and endline as part of the participant's surveys administered following recruitment and at week 20 after recruitment respectively.

SAP version history

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

Any changes to the design or methods need to be discussed with the YEF Evaluation Manager and the developer team prior to any change(s) being finalised. Describe in the table above any agreed changes made to the evaluation design. Please ensure that these changes are also reflected in the SAP (CONSORT 3b, 6b).

Table of contents

Introduction

The Your Choice programme is a Cognitive Behavioural Therapy (CBT) enhanced approach to practice, delivered through high intensity contact within adolescent services to young people aged 11-18 at medium to high risk of contextual harm. The 12-18-week programme is delivered by specially trained practitioners, who are trained in CBT tools and techniques and are supported by regular clinical supervision. Training for practitioners is delivered through a train the trainer model by clinicians with experience in the delivery of CBT.

The impact evaluation is based on a two-armed cluster randomised controlled trial (RCT) where the unit of randomisation is teams of youth practitioners. Teams supporting young people eligible for Your Choice are randomly assigned to train in and deliver Your Choice (treatment group) or to supporting young people following Business As Usual (BAU) practices (control group). Teams randomised out of training during the efficacy trial will be offered to be trained in Your Choice later on.[2]

Randomisation is done within all participating Local Authorities (LA) of London. The trial therefore involves a small unit of randomisation (teams) within small strata (all but two LAs have 10 or fewer than 10 teams), with an expected number of observations over 10 young people recruited in team on average. De Chaisemartin and Ramirez-Cueller (2024)[3] show that, in such a case, the regular practice of clustering standard errors at the unit of randomisation can lead to downward bias in estimates of the variance of the treatment effect, resulting in over-rejecting the null hypothesis (of no effect). They advise that standard errors should be clustered at the strata level (LA) when RCTs have this type of configuration (and this result holds whether or not strata fixed effects are controlled for).[4] To accommodate for these results, we perform power calculations via simulations (instead of using traditional power calculations commands).

[2] This was explicitly agreed with LAs given the reservations they had about randomisation. It seems unlikely that control teams react to the possibility of future training so much in advance. LAs also have an obligation towards the young people they see, to deliver statutory services according to the BAU processes. For these reasons, this setting is unlikely to affect results.

[3] De Chaisemartin, C. and Ramirez-Cueller, J. (2022). "At what level should one cluster standard errors and small-strata experiments?", National Bureau of Economic Research WP 27609, http://www.nber.org/papers/w27609

[4] In cases where clusters are small (fewer than 10 observations per unit of randomisation on average), estimates of the standard errors should not adjust for degrees of freedom. This is not the case we expect even under the conservative projections in terms of participating young people.

Ultimately, the aim of the randomisation exercise is to ensure independence of the assignment of young people to trial arms and team treatment status. In this trial, however, the RCT design does not require that young people are randomly assigned to teams, as individual level randomisation of young people would interfere with the usual delivery of statutory services in ways that would be impractical for and unacceptable to Local Authorities. Instead, the proposed RCT design is drawn under the expectation that the assignment of young people to teams is based on team availability at the time of referral, in a system that works at capacity and that team availability and time of referral is random, and hence such system effectively acts to ensure the random matching of young people and teams. Teams being randomly assigned to training then ensures that the provision of training is not related to special characteristics of teams that could interfere with how they deliver the treatment.

The one exception to the capacity rule determining the assignment of young people to teams is for those young people returning to LA services after a brief interruption, who would be assigned to the same team that previously worked with them. Such an assignment rule, if rigorously followed, effectively guarantees that assignment is independent of the team status regarding Your Choice training.[5]

The implementation and process evaluation uses both quantitative and qualitative methods, including collection of process data on delivery of Your Choice and BAU support as well as qualitative interviews of LA staff and young people involved in the delivery of Your Choice. The aim is to establish the impact of Your Choice on participant's behaviours and attitudes towards violence, and to investigate possible mechanisms underlying observed responses.

Design overview

Trial design, including number of arms	**Two-arm, cluster randomised**
Unit of randomisation	Teams of youth practitioners
Stratification variables	Local Authority

[5] Evidence collected during the pilot trial showed that the observed characteristics of young people assigned to treated and control teams were well balanced, and that capacity and historical assignment were major drivers of assignment of young people to teams, as described in the next point below.

(if applicable)		
Primary outcome	variable	Conduct problems
	measure (instrument, scale, source)	A binary indicator for scoring in the high and very high range of the conduct problems subscale of the Strengths and Difficulties Questionnaires, taking the value 1 if the score is in the high or very high range and 0 otherwise. Measured as part of the endline young person survey, administered at week 20 after recruitment during a session with the practitioner.
Secondary outcome(s)	variable(s)	Offending activity, mental wellbeing, emotional self-regulation, social connectedness, internalising behaviours, hyperactivity, self-reported and practitioner-reported perceptions of young person's safety, prosocial behaviour and prosocial identity. Continuous scores will be used for all outcomes except offending activity, which will be measured as a binary indicator taking the value 1 if the young person offended and 0 otherwise. For internalising behaviours, hyperactivity and prosocial behaviour, we will also use binary indicators taking value 1 if the young person scores in the high or very high range and 0 otherwise.
	measure(s) (instrument, scale, source)	*Criminal activity*: recorded arrest in Police National Computers during the period of 16 months after recruitment. *Self-reported and practitioner-reported perceptions of young person's safety*: Young person and practitioner versions of "Checkpoint. A safety scale for young people", which is an instrument to measure young people's perceptions of safety developed by the research team. Part of the endline

		young person questionnaire, administered at week 20 after recruitment. *Wellbeing*: The Short Warwick–Edinburgh Mental Well-being Scale (SWEMWBS). Part of the endline young person questionnaire, which is administered at some point between weeks 14 and 20 after recruitment. *Emotional self-regulation*: Trait Emotional Intelligence Questionnaire – Adolescent Short Form (TEIQUE-ASF) – Self regulation subscale. Part of the endline young person questionnaire, administered at week 20 after recruitment. *Social connectedness*: Social Connectedness Scale – Revised (SCS-R). Part of the endline young person questionnaire, administered at week 20 after recruitment. *Internalising behaviours*: emotional difficulties and peer difficulties subscales of the SDQ measured at baseline and endline as part of the participant's surveys administered following recruitment and at week 20 after recruitment respectively. *Hyperactivity*: hyperactivity subscale of the SDQ measured at baseline and endline as part of the participant's surveys administered following recruitment and at week 20 after recruitment respectively. *Prosocial behaviours*: Strength and Difficulties prosocial behaviour subscale measured at baseline and endline as part of the participant's surveys administered following recruitment and at week 20 after recruitment respectively. *Prosocial identity*: Pro-social Identity Scale (PIDS) measured at baseline and endline as part of the participant's surveys administered following

		recruitment and at week 20 after recruitment respectively.
Baseline for primary outcome	variable	Conduct problems
	measure (instrument, scale, source)	Indicator for scoring in the high or very high range of conduct problems scale from the Strengths and Difficulties Questionnaires. Measured as part of the baseline participant's survey, administered in week 1 after recruitment, before the start of Your Choice.
Baseline for secondary outcomes	variable	Social connectedness score, internalising behaviour, hyperactivity, prosocial behaviour and prosocial identity, self-reported and practitioner-reported perceptions of young person's safety.
	measure (instrument, scale, source)	All measured as part of the baseline participant's survey, administered in week 1 after recruitment, before the start of Your Choice. *Social connectedness*: The Social Connectedness Scale – Revised (SCS-R), survey of young person following consent before Your Choice starts (in some cases, some form of BAU work/support will have already been taking place). *Internalising behaviours*: emotional difficulties and peer difficulties subscales of the SDQ. *Hyperactivity*: hyperactivity subscale of the SDQ. *Prosocial behaviours*: Strength and Difficulties prosocial behaviour subscale. *Prosocial identity*: Pro-social Identity Scale (PIDS) measured. *Self-reported and practitioner-reported perceptions of young person's safety*: Young person and practitioner versions of "Checkpoint. A safety scale

		for young people", which is an instrument to measure young people's perceptions of safety developed by the research team.

Sample size calculations overview

The expected analytical sample will have two features that are important to consider for inference:

- *There will be a small number of units of randomisation (teams) within strata (LA)*: Among the 29 LAs participating in the whole evaluation (internal pilot + efficacy), the median number of teams per LA is 3 (the average 4.6), with all but 2 LAs having 10 or fewer teams involved in the trial (Brent has 12 teams and Barking and Dagenham has 21 teams).[6]
- *Units of randomisation (teams) have more than 10 observations on average*: Based on the conservative predictions of participating young people provided by Local Authorities so far, the average expected number of participants per team is 13.[7]

De Chaisemartin and Ramirez-Cueller (2024) show that in stratified cluster RCT where the number of units of randomisation (teams) is small (10 or fewer) within strata (Local Authorities), the regular practice of clustering standard errors at the unit of randomisation can lead to downward bias in estimates of the variance of the treatment effect, resulting in over-rejecting the null hypothesis (of no effect). They advise that standard errors should be clustered at the strata level (LA) when RCTs have this type of configuration (and this result holds whether or not strata fixed effects are controlled for).[8]

We are not aware of any software allowing us to perform power calculations that take these issues into account. In order to account for the exact nature of our dataset and for these technical complexities, we therefore opted for calculating the Minimum Detectable Effect associated with 0.8 power via simulations programmed in STATA and adapting the procedure

[6] Even after removing the 6 teams that recruited in the pilot but did not collect any endline data, the median number of teams per LA is also 3 (the average if 4.5).

[7] After considering the actual attrition at endline in the pilot (54%) and the expected 30% attrition rate at endline in the efficacy, we would have 9.4 observations per team on average.

[8] In cases where clusters are small (fewer than 10 observations per unit of randomisation on average), estimates of the standard errors should not adjust for degrees of freedom. This is not the case we expect even under the conservative projections in terms of participating young people.

suggested in McConnell and Vera-Hernandez (2015). The simulation code used to perform these calculations is enclosed but we describe the steps used to perform these simulations below.

Each simulation is characterised by five parameters:

- Standard deviation of the outcome (σ)
- Intra-cluster correlation (ρ)
- Minimum Detectable Effect (β)
- Attrition rate (r)
- A threshold on the continuous outcome above which the binary indicator takes the value 1 and 0 otherwise (η)

Each simulation is based the following steps:

1. Randomise teams to be newly randomised to treated or control within each LA with 50-50 split.
2. In LAs that have an odd number of teams to randomise, we randomly select whether (n+1)/2 or (n-1)/2 teams would get treated (where n is the number of teams to randomise)
3. Generate a normally distributed random variable at the Local Authority level, θ_j, of mean 0 and variance $\rho\sigma^2$
4. Generate a normally distributed random variable at the young person level, ϵ_{ij}, of mean 0 and variance $(1-\rho\ \ )\sigma^2$
5. Implement the following data generation process for the outcome Y:
$$Y_{ij} = \beta T_{ij} + \theta_j + \epsilon_{ij}$$
where T_{ij} =1 if participant *i* is recruited in a treated team and 0 in a control team
6. Create a binary outcome D_{ij} such that it takes the value 1 if the score above is above a particular threshold η[9] and 0 otherwise[10]
7. Estimate OLS regressions of the chosen outcome (continuous Y_{ij} or binary D_{ij}) on the treatment dummy, clustering the standard errors at the LA level, with or without controlling for Local Authority fixed effects, on a randomly selected sample of observations of size (1-*r*) of the expected number of study participants in order to

[9] This threshold is chosen such that the mean of D is equal to the probability of observing our primary outcome in the baseline pilot data.

[10] This step is to report power calculations for the binary outcome based on the continuous score, though the data generating process assumed here only enables us to simulate average impacts on the continuous score and hence may miss possibly stronger impacts the intervention may have at the top of the distribution

simulate a random attrition rate *r* to which we add the 83 observations from pilot that have endline data.

8. Repeat steps 1 through 6 1000 times and compute the power as the proportion of times the coefficient β is significant at the 95% level.

As part of this exercise, we wanted to explore the power implications of controlling for a lagged (baseline) outcome. Instead of making additional assumptions to simulate such baseline outcome, we instead performed the simulation above with smaller values of the variance of the outcome variable and of the ICC in order to mimic the effect of controlling for a lagged outcome on these parameters as estimated using data from the pilot period.

Table 4: Values picked for parameters in simulations and justification

	Scenario 1 **(original assumptions held at the beginning of the trial)**	**Scenario 2** **(assumptions updated as part of the January 2025 revision)**
Attrition Rate	**30%**	**30%**
Mean of primary outcome	0.3 *(estimated on relevant baseline and endline pilot data)*	0.6 *(estimated on baseline efficacy and pilot data available at the time of revision)*
Intra-cluster correlation (ρ)	0.12 *(estimated on relevant baseline and endline pilot data)*	0.09 *(estimated on baseline efficacy and pilot data available at the time of revision)*
Minimum Detectable Effect	A grid of values between 10% and 40% reduction in the primary outcome (likelihood to have high/very high SDQ conduct problem score) *(This is obtained by simulating the continuous outcome with SD σ= 1 and using a grid of MDES β on the continuous SDQ score appropriate so that the MDES on the binary primary outcome varies between 10% and 40% of the primary outcome mean).*	

Primary population of interest

The primary population of interest will be young people aged 11-18 referred to their Local Authorities for support and at medium to high risk of contextual harm.

Table 5 provides information about the number of young people and teams, and their distribution across treated and control groups, which were used to perform the power calculations. These numbers reflect the best information available at the time of writing, drawn from the Local Authorities' plans in terms of teams and expected numbers of participants in each team, before the trial started (sample A) and when the decision to extend the recruitment period was considered (Sample B).

Sample A reflects the size and composition of the final sample as estimated before the trial started. It totals 1857 young people across 138 teams.

Sample B reflects the size and composition of the final sample as re-estimated as of January 2025. It totals 1563 young people across 132 teams. This number was re-estimated based on recruitment to date and new information from Local Authorities about the number of young people they would expect to recruit if the recruitment period was extended to the end of February 2025.

Table 5: Structure of the two datasets used to perform power calculations

		Sample A (1.0 original version)	**Sample B (1.1 January 2025 update)**
Average cluster size (if clustered)		Average number of participants per cluster (team): 13	Average number of participants per cluster (team): 11.8
Number of clusters (teams)	Intervention	70	83
	Control	70	49
	Total	138	132
Number of participants	Intervention	960	937
	Control	897	626
	Total	1857	1563

Although randomisation during the efficacy phase allocates teams to training and control on a 50-50 basis, legacy randomisation from the Home Office and pilot phases means more teams are trained than control. The power calculations reported above account for this imbalance.

Our power calculations account for a 30% attrition rate. This attrition rate is calculated based on all participants ever recruited in randomised teams in the pilot and efficacy trials (regardless of not they collected any endline data).

Table 6 report the MDE estimated using the simulation methodology described above, under the below assumptions about power, significance level and the type of test we want to perform and under different assumptions about the attrition rate and about whether Local Authority fixed effects are controlled for in the regression. Note that, when we revised the expected sample size in January 2025, we also revised our assumptions about the ICC and mean of the baseline based on estimates of these parameters in the baseline sample recruited during efficacy, which we deemed more reliable than the pilot sample due to its much greater size.

These assumptions turned out to be less conservative than those made when we initially ran power calculations: the ICC was lower (0.09 as opposed to 0.12) and the mean of the primary outcome variable was higher (0.6 as opposed to 0.3).This meant that, even though the revised sample size in scenario 2 is lower than the initial sample size in scenario 1 the MDE is slightly lower (22% instead of 24%).

Table 6: Minimum Detectable Effect on primary outcome (binary indicator for high to very high range of conduct problems) under the assumption we do not control for lagged outcome in the larger dataset

Scenario	Scenario 1: Sample A with original priors for simulation parameters (1.0 original version)	Scenario 2: Sample B with revised priors for simulation parameters (1.1 January 2025 update)
Alpha [11]	0.05	0.05
Power	0.8	0.8

[11] Please adjust as necessary for trials with multiple primary outcomes, 3-arm trials, etc., when a Bonferroni correction is used to account for family-wise errors.

One-sided or two-sided?	Two-sided	Two-sided
ICC	0.12	0.09
Mean of the primary outcome in control group	0.3	0.6
Minimum Detectable Impact on Binary score for scoring in high or very high range on conduct problems	MDE = 24% reduction in the mean of the primary outcome	MDE = 22% reduction in the mean of the primary outcome

Discussion of MDEs in the context of the pilot trial and of the literature

The power calculations in Table 6 above indicate that, under a 30% attrition rate and based on the best information we have at the beginning of the trial about key parameters (Scenario 1), we could detect with 80% power a minimum effect of 24% of baseline probability to be in the high to very high range of conduct problems in the sample expected based on information provided by Local Authorities before the start of the trial (whereby 138 teams recruit 1857 young people). Of a 0.3 mean probability to be in the high to very high range of conduct problems, this would be equivalent to a reduction of 0.072 percentage points.

Revised power calculations based on the updated values for the simulation parameters as of January 2025 (Scenario 2) indicate that we could detect with 80% power a minimum effect of 22% of baseline probability to be in the high to very high range of conduct problems. For a 0.6 mean probability to be in the high to very high range of conduct problems, this would be equivalent to a reduction of 10.8 percentage points.

Under both Scenario 1 and Scenario 2, the MDE is within the range of impacts suggested by the endline data, well below the effects of CBT interventions on externalising behaviours, such as Attention Deficit Hyperactivity Disorder, Conduct Disorder and Oppositional Defiant Disorder, which are reported in the literature and which range around 60-70% (Gaffney, Farrington and White, 2021).[12] Given that Your Choice is implemented on a large scale and delivered by non-clinical practitioners, it would be reasonable to expect a lower impact than

[12] Gaffney, H., Farrington, D. and White, H. (2021). "Cognitive Behavioural Therapy: YEF Toolkit technical report"

the impact of CBT interventions reported in most of the literature. Yet, an effect size of 22% or 24% reduction (or a third or the impact sizes reported in the literature) still seem within the range of reasonable effects to expect from the Your Choice intervention.

Analysis

All codes and data needed to replicate the results from the efficacy trial will be provided. In preparing these replication materials, we will follow standard guidelines set out by the American Economic Association (American Economic Association (aeaweb.org)).

The level of analysis for the core analysis will be the individual young person. The parameter that we will estimate in our main analysis with be an Intention-to-Treat (ITT) treatment effect measuring the effect on a young person's primary and secondary outcomes of being supported with a team that has trained in Your Choice (as opposed to a team that has not trained in Your Choice). The main Intention-to-Treat (ITT) specification related to our primary and secondary outcomes will be a linear regression, where we will control for baseline values of the outcome and baseline characteristics of the young person and/or their team that present statistically significant imbalances at baseline. In line with De Chaisemartin and Ramirez-Cueller (2022), we will present specifications with and without strata fixed effects. Some of the secondary analysis will be at the level of the meetings of the young-person and practitioner, and we will also consider some outcomes at the practitioner level – such as their assessment of the young person's needs and risks.

All methods were determined a priori and given what we have learned during the pilot phase.

We will use STATA version 16 or higher for the analysis.

Primary outcome analysis

We will estimate the Intention-to-Treat (ITT) effect by estimating the following equation for young person i will be of the form:

$$Y_{i1} = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 Y_{i0} + strata\ fixed\ effects + u_i$$

where Y_{i1} is the primary outcome of interest, which is an indicator that takes the value 1 if the young person scores in the high to very high conduct problems range and 0 otherwise, and Y_{i0} is the baseline value for that same outcome, D_i is an indicator for treatment assignment, X_i are any baseline characteristics of the young person and/or their team for which we find statistically significant imbalances between treated and control groups, and u_i is an error term. Standard errors will be clustered at the Local Authority level, following Chaisemartin and Ramirez-Cuellar (2022).

Secondary outcome analysis

Analysis of secondary outcomes will follow the same approach as for primary outcomes described above. To account for multiple hypothesis testing given the large number of secondary outcomes, we will calculate p-values using the step-down procedure of Romano and Wolf [2016]. The Romano-Wolf correction (asymptotically) controls the familywise error rate (FWER), that is, the probability of rejecting at least one true null hypothesis in a family of hypotheses under test. This correction is considerably more powerful than earlier multiple testing procedures such as the Bonferroni and Holm corrections, given that it takes into account the dependence structure of the test statistics by resampling from the original data. To implement it, we will use Stata's rwolf2 command, report unadjusted and adjusted *p*-values, and determine statistical significance of the results based on adjusted *p*-values below 0.05.

The estimating equation for ITT impacts for young period i will be of the form:

$$Y_{i1} = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 Y_{i0} + strata\ fixed\ effects + u_i$$

where, as before, Y_{i1} is the secondary outcome of interest, and Y_{i0} is the baseline value for that same outcome (if any), D_i is an indicator for treatment assignment, X_i are any baseline characteristics of the young person for which we find statistically significant imbalances between treated and control groups, and u_i is an error term.

Subgroup analyses

We will explore heterogeneous treatment effects by age, race/ethnicity, gender and practitioner risk assessment (when sample sizes allow this to be meaningfully done). Specifically, we will run separate models and test for statistically significant differences in effects across the following subgroups:

- Age 11-15 vs Age 16-18.
- White vs Non-White - specifically, we would expect to compare those of Black backgrounds to those of White backgrounds and compare those of South Asian backgrounds to those of White backgrounds. This is on the basis that those from minoritised backgrounds may face more structural challenges that may impede the impact of Your Choice.
- Males vs Females.
- Medium vs High Risk based on baseline practitioner risk assessment (as per the baseline practitioner survey).

Note that the trial will not be powered to detect differences in impacts between any of these subgroups, so this analysis will be considered exploratory rather than definitive.

Robustness checks

We will run additional secondary specifications to test the robustness of the results to altering the covariates included in the model. In expectation, due to the randomisation of treatment, doing so should not change the point estimates of impact, but may change the confidence intervals.

Specifically, we will estimate the following specifications:

- A simple model, controlling only for the treatment assignment and strata fixed effect as covariates
- A model controlling for covariates selected using a post-double selection LASSO procedure. This procedure uses LASSO to select variables that are most predictive of the treatment status (i.e. that are most imbalanced between control and treatment groups) and another LASSO to select variables that are most predictive of the outcome variables (to decrease the unexplained variance). The procedure is therefore a data-driven procedure to select both sets of covariates and estimates the treatment impacts controlling for those.

Our main ITT parameters will be estimated using Ordinary Least Squares (i.e. linear probability models for binary outcomes, such as our primary outcome). When used for binary outcomes, the linear regression model is the least demanding model in terms of distributional assumptions and behaves well for probabilities away from the bounds 0 and 1. Its main drawback (of potentially generating out-of-sample prediction outside the 0-1 range) does not apply to this application as we are not using the model for predictions. The coefficients of the linear probability model can be directly interpreted as marginal effects. For binary outcomes, we will also estimate probit and/or logit models in robustness checks. These models behave better than the linear probability model at the tails. We will present marginal effects based on these models. We do not expect our findings to vary substantially with the choice of model.

Descriptive analysis of intervention content

Additional analysis will be presented at the level of the Your Choice session meetings between the young person and their practitioner. This analysis will describe various aspects of sessions, both in the control and treatment groups, including: content covered, duration, frequency and level of engagement of the young person.

Longitudinal follow-up analyses

The endline data will be collected at week 20 after the baseline data. This is when primary and most secondary outcomes will be measured. If we are able to secure UPN for individuals, then we will attempt to apply for an extract of the linked NPD-PNC data for the sample

covered in the efficacy trial. This will be used to provide a long-term follow-up, for both outcomes related to education and criminal offences.

Imbalance at baseline

We will present a table of baseline balance between treated and control groups, for all individuals recruited in the teams that were randomised and for the sample of individuals with endline primary outcome data. This will report means and standard deviations for continuous variables (and only means for binary outcomes). This will use data both the workbooks and the baseline surveys fielded to young people and practitioners. The basic characteristics to be shown are the demographic characteristics of the young person, other background characteristics of the young person (their needs, activities and support), their SDQ measures, self-assessment of risk, and those of their practitioner.

Missing data

We will examine the extent of missingness, the patterns of missingness, and whether it correlates to treatment assignment. First, we will specify the number of complete cases (i.e. those without any data missing) and will attempt to establish the missingness mechanism (i.e. what variables in the data are predictive of non-response). To do the latter, we will run a logistic regression model (accounting for strata fixed effects and clustering standard errors at the LA level) of an indicator taking the value 1 if a variable is missing, and 0 otherwise, on information that might be predictive of missingness. This will be done separately for outcome variables and covariates included in the headline model.

We do not propose to impute missing outcome data for our main estimates. We will follow YEF's guidance on missing data analyses and perform relevant sensitivity analyses if there is more than 10% of data missing. If covariates are missing conditional on other covariates or outcomes, we will run Multiple Imputation and compare the results with complete cases and with MI. If the results are similar, this will suggest the complete cases are unlikely to be biased (but underpowered). If the results are not similar, we will discuss implications of this analysis clearly in the report.

Compliance

Compliance will be measured at several levels: at the young person level and at the practitioner level. At the young person level – there can be full compliance/non-compliance, as well as partial compliance.

Full compliance/non-compliance: In the Your Choice RCT, treatment is assigned at the level of the team of practitioners. However, it may be that teams deviate from the protocol they were given young people allocated to a Your Choice trained team may receive another form of support, and young people allocated to a Business As Usual team end up receiving the Your

Choice programme if they end up being supported by a Your Choice trained practitioner from another team. If that is the case, the ITT analysis may underestimate the impact of the intervention on those who actually received it.

Partial compliance: the intervention is supposed to involve 3 meetings per week over the course of 12 weeks, but there can be partial compliance if young people receive a share of the designated Your Choice sessions. As with full non-compliance, in the presence of partial compliance the ITT analysis will generally underestimate the effect of the intervention among those who actually received the full treatment.

To consider these possibilities, we will use monitoring data to measure the extent of non-compliance in the treatment and the control groups and the extent of partial compliance in the treatment group. We will then test whether the characteristics of young people non-compliers are different from those of compliers in statistically significant ways.

We will further extend our analysis by using an Instrumental Variable (IV) approach to give an indication of the effects a) of receiving Your Choice and b) of the intensity of treatment. To implement a) and b), we will use a Two Stage Least Squares (2SLS) approach with group allocation as the instrumental variable for compliance. In a), we will define compliance as a binary indicator that takes the value 1 if the young person receives any of the Your Choice programme and 0 otherwise. In b), we will define compliance as a continuous variable measuring the proportion of Your Choice sessions that the young person received out of the total planned for participants (=12x3). The first stage will model the compliance variable using the same explanatory variables used for the headline ITT analyses and the instrument. The second stage model will use predicted compliance in place of the group identifier variable in the ITT analyses specified above to generate Complier Average Causal Effect (CACE) estimates.

At the practitioner level – practitioners are supposed to received 3 + 1 days of training and to receive clinical supervision at least once monthly. Using information from the workbook on training and clinical supervision sessions, we will document the extent to which this happened across teams involved in the trial. We will not take this form of compliance into consideration in the estimation of the treatment effects.

Intra-cluster correlations (ICCs)

Clusters are teams of practitioners but given the discussion above about power, we will cluster standard errors at the LA level. We will calculate ICCs for both team and LA using the `estat icc' command in Stata.

Presentation of outcomes

Impact estimates will be reported as Effect Sizes (ES) standardised using the outcome variance for the control group. This choice is robust to the possibility that the treatment changes the variance of the outcome, which we see as a likely event if, as expected, the impact of treatment is heterogeneous. Where an outcome is defined as a binary variable, we will present ES as risk ratios and natural frequencies. We will show the 95% confidence interval for the estimated effect.

References

de Chaisemartin, Clément Ramirez-Cuellar, Jaime At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments? American Economic Journal: Applied Economics 16 1 193–212 2024 10.1257/app.20210252 https://www.aeaweb.org/articles?id=10.1257/app.20210252

McConnell, B and Vera-Hernandez, M. (2015). Going beyond simple sample size calculations: a practitioner's guide. London: Institute for Fiscal Studies. Available at: https://ifs.org.uk/publications/going-beyond-simple-sample-size-calculations-practitioners-guide

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413