Effectiveness and cost-effectiveness of a group programme for men who are concerned about their abusive behaviour with women: A Randomised Controlled Trial

Part of the REPROVIDE Programme Grant for Applied Research (NIHR Ref: RP-PG-0614-20012)

reprovide
Reaching Everyone Programme of Research On Violence In diverse Domestic Environments

Statistical Analysis Plan

Principal Authors: Beverly A. Shirkey, PhD, Maximiliano Vazquez Morales, MS

Role	Name	Signature	Date
Authorising Statistician	Tim Peters, PhD	T. J. Peters	17 April 2024
Chief Investigator	Gene Feder, BSc, MB, BS., MD(Lond.), FRCGP	[illegible]	17 April 2024

Table of contents

List of Abbreviations

Acronym	Details
ABI	Abusive Behaviors Inventory
ABI-R	Abusive Behaviors Inventory-Revised
AdminDB	Administration Database
AE	Adverse Event
APR	Annual Progress Report
AQ-10	Autism Spectrum Quotient
AUDIT-C	Alcohol Use Disorders Identification Test-C
BAME	Black, Asian and Minority Ethnic
BANES	Bath and North East Somerset (CCG)
BNSSG	Bristol, North Somerset and South Gloucestershire (CCG)
BTC	Bristol Trials Centre (formerly Bristol Randomised Trials Collaboration)
CACE	Complier Average Causal Effect
CBT	Cognitive Behavioural Therapy
CCG	Clinical Commissioning Group (replaced by ICB)
CHU-9D	Child Health Utility
CI	Chief Investigator
CONSORT	CONsolidating Standards of Reporting Trials
CPQ-SF	Communication Patterns Questionnaire – Short form
CRF	Case Report Form
DAPP	Domestic Abuse Perpetrator Programme (new name replacing DVPP)
DMEC	Data Monitoring and Ethics Committee
DSA	Data Sharing Agreement
DUDIT	Drug Use Disorders Identification Test
DV(A)	Domestic Violence (and Abuse)
DVPP	Domestic Violence Perpetrator Programme
EQ-5D-5L	EuroQuol – 5D-5L version
GAD-7	Generalised Anxiety Disorder assessment – 7 item
GCP	Good Clinical Practice
GDPR	General Data Protection Regulation
HEAP	Health Economics Analysis Plan
HRA	Health Research Authority
ICECAP-A	ICEpop CAPability measure – Adult
ICB	Integrated Care Board
ICC	Intra Cluster Correlation
ICF	Informed Consent form

ISF	Investigator Site File
IDVA	Independent Domestic Violence Advisor
IPVRAS	Adapted Intimate Partner Violence Responsibility Attribution Scale
IQR	InterQuartile Range
MARAC	Multi Agency Risk Assessment Conference
ISRCTN	International Standard Randomised Controlled Trials Number
NHS R&D	National Health Service Research & Development
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health Research
PAS	Propensity for Abusiveness Scale
PC-PTSD-5	Primary Care Post Traumatic Stress Disorder screen – 5 item
PGfAR	Programme Grant for Applied Research
PHQ-9	Patient Health Questionnaire – 9 item
PI	Principal Investigator
PIS	Participant Information Sheet
PPI	Patient and Public Involvement
PSC	Programme Steering Committee
PSS	Public Sector and Societal (perspectives)
QALY	Quality Adjusted Life Years
RCT	Randomised Controlled Trial
REC	Research Ethics Committee
REPROVIDE	Reaching Everyone: Programme of Research on Violence in Diverse Domestic Environments
REDCap	Research Electronic Data Capture
RPG	Relapse Prevention Group
SAE	Serious Adverse Event
SD	Standard Deviation
SDV	Source Data Verification
SF-12	Short Form health survey – 12 items
SOP	Standard Operating Procedure
SSI	Site Specific Information
TMG	Trial Management Group
TSC	Trial Steering Committee
TMF	Trial Master File
UoB	University of Bristol

1. Introduction

1.1 Summary of document

1.1.1 Scope

This statistical analysis plan (SAP) for the REPROVIDE (effectiveness and cost-effectiveness of a group programme for men who are concerned about their abusive behaviour toward women) randomised controlled trial has been written following Bristol Trials Centre's standard operating procedures, the CONSORT (Schulz, Altman and Moher, 2010) statement, and the International Conference on Harmonisation (ICH) Statistical Principles for Clinical Trials E9(R1) (ICH 2019), by Beverly A. Shirkey, author, (Senior Research Associate in Medical Statistics), with assistance from Maximiliano Vazquez Morales, (Research Associate in Medical Statistics) under the supervision of Tim Peters PhD, SAP reviewer, Professor of Primary Care Health Services Research), all employed by the University of Bristol. The SAP covers all final statistical analyses to be performed, outlined in the study protocol. All other statistical analyses will be considered exploratory. There will be separate analysis plans for the health economics and the qualitative findings of the study.

Note that large sections of the trial protocol have been directly quoted to help the Statistical Analysis Plan be a stand-alone document; if any inadvertent differences exist, the latest protocol version prevails. A protocol paper has been published (Morgan *et al*, 2023), and the current version of the approved protocol is available online at https://www.isrctn.com under trial ISRCTN15804282 and is titled 261128_RCTofDVPP_Protocol_Vs_9.0_04.10.2023. Any revisions to the protocol will be published on this site before the primary analysis is undertaken.

PDF's of the Baseline and 12-month questionaries for the male participants will be published in due course, for example online at https://www.isrctn.com under trial ISRCTN15804282 or as appendices to this SAP.

1.1.2 Planned analyses and dissemination

Based on data collected up to the end of autumn 2024, and when all the pre-specified final analyses have been performed, a summary of the findings will be disseminated to the TMG (Trial Management Group).

An independent data monitoring and ethics committee (DMEC) will review the safety and ethics of the study. Study update reports will be reported on an overall basis, except for serious and non-serious adverse events, treatment adherence, retention, and PHQ-9 (Patient Health Questionnaire-9), which are reported by study arm. The reports to the DMEC are produced and disseminated on at least an annual basis. The DMEC, considering the interim reports and any advice or evidence they wish to request, will, if necessary, report to the Trial/Study Steering Committee (TSC/SSC) if there are any concerns regarding the safety of the intervention or ethics of the study. The DMEC have the authority to recommend that the study stops if deemed necessary based on the observed data.

Due to the lockdown(s) during the COVID-19 pandemic, groups were paused. The pause lasted for 135 days, and the follow-up collection times were shifted to account for the group sessions being paused. Specifically, outcomes for participants randomised before 1 July, 2020, due to the effect of COVID-19 shutdown on the intervention, were delayed by 135 days, and hence the 12-month time point would be ~16.5 months for these participants (see the Section on visit windows for further details). Note that this longer collection time of data also applies to the collection of the police data post randomisation. To account for the effects of the COVID-19 pandemic, a sensitivity analysis will be performed by including an interaction effect for those who were randomised before or after 1 July 2020 (so delineating those affected by the treatment pause).

1.2 Background to the study

Domestic violence and abuse (DVA) is defined as any incident, or pattern up toof incidents, of controlling-coercive, or threatening behaviour, violence or abuse between people aged 16 or over who are or have been intimate partners or family members, regardless of gender or sexuality. DVA poses a significant public health and clinical challenge to the NHS (Alberti, 2010; NICE, 2014). It is associated with health problems in victims, perpetrators, and their children, including poor physical health, long-term illness or disability, and poor mental health, at an annual cost to the NHS of 1.8% of the total budget with even more significant societal costs (Walby, 2009). The NHS (and health services internationally) have not responded adequately to this need (García-Moreno *et al*., 2014). There is growing recognition of its impact on women and children, but virtually no recognition by clinicians of men as victims or perpetrators and little research on effective interventions for men in healthcare settings. The evidence reviews in the NICE DVA guidelines (NICE, 2014) identify evidence gaps with regard to an integrated healthcare response and effective interventions targeted at perpetrators.

This trial forms part of an NIHR Programme grant called Reaching Everyone: Programme of Research on Violence in Diverse Domestic Environments (REPROVIDE) and builds on previous work undertaken as part of the PROVIDE Programme (see http://www.bristol.ac.uk/population-health-sciences/projects/provide/).

Study rationale

The rationale for this trial is that, despite the ubiquity of perpetrator programmes in the UK, Europe and North America, there is still uncertainty about their effectiveness. More experimental studies must be conducted internationally (Gondolf, 2012) and outside North America (Akoensi *et al*., 2013). The health impact of DVA makes the provision of effective perpetrator programmes to prevent further violence a legitimate part of healthcare services. With the move towards evidence-based commissioning of health services, we need to rigorously test programmes regarding safety and health outcomes for victims, survivors, and perpetrator behaviour. A significant research recommendation of the NICE DVA guidelines was to determine the effectiveness of perpetrator interventions regarding victims' safety across levels of risk and including diverse and marginalised groups.

1.3 Aims and objectives

1.3.1 Primary objective

To investigate the effectiveness of a group programme intervention in reducing men's abusive behaviour against women. This will be achieved by the recruitment and (as far as possible) retention of 316 male perpetrators who will be randomised to either a 23-week weekly community-based perpetrator programme or usual care control arm plus, wherever possible, recruitment of their partners/ex-partners with a 12-month follow-up.

1.3.2 Secondary objectives

1. Assess the effect of the perpetrator intervention on measures of DVA, health and wellbeing of the male participant, plus reports of police incidents.
2. Assess the effect of the intervention on measures of experience of DVA, health and wellbeing of female partners and ex-partners.
3. To compare the costs and consequences of the intervention from NHS and public and societal perspectives (PSS).
4. Determine the acceptability of the intervention to perpetrators, victims and professionals working with perpetrators and victims.
5. Through a mixed methods process evaluation, explore the extent to which the intervention was implemented, fidelity to the intervention, and how and why the intervention was or was not effective.

1.3.3 Objectives and outcomes

Table 1: Objectives and outcomes

Objectives	**Outcome Measures**	**Timepoint(s) of collection of this outcome measure**
To investigate the effectiveness of the group programme intervention in reducing men's abusive behaviour against women.	Primary outcome: Abusive Behavior Inventory-revised (ABI-29)	**Baseline, 4 months, 8 months, and 12 months:** ABI-29
To assess the effect of the perpetrator intervention on measures of DVA, health and wellbeing of the male participants, plus reports of police incidents	ABI-29; IMPACT toolkit questions; criminal justice questions; mental health questions; Patient Health Questionnaire -9 (PHQ-9), Generalised Anxiety Disorder assessment (GAD7); Primary Care-Post Traumatic Stress Disorder 5 (PC-PTSD-5); Propensity for Abusiveness Scale (PAS); Adapted communications patterns questionnaire- short form (CPQ-SF); adapted Intimate Partner violence Responsibility Attribution Scale (IPVRAS); Alcohol Use Disorders Identification test-C (AUDITC); Drug Use Disorders Identification Test (DUDIT); childhood experiences questionnaire; Reflective Functioning Questionnaire (RFQ); capability (ICECAP-A); quality of life (SF-12); Police data	**Baseline only:** mental health questions; childhood experiences questionnaire **Baseline and 12 months only:** criminal justice questions; AUDITC; DUDIT; RFQ, SF-12 **Baseline, 4, 8 and 12 months:** ABI; IMPACT toolkit;; PHQ-9; GAD7; PC-PTSD-5; PAS; Adapted CPQ-SF; adapted-IPVRAS; ICECAP-A **12 months post randomisation and 12 months pre baseline:** Police data
To assess the effect of the intervention on measures of experience of DVA, health and wellbeing of female partners and ex-partners	IMPACT toolkit; ABI-R*; criminal justice questions; PHQ-9; mental health questions; GAD7; PC-PTSD-5; AUDITC; DUDIT; childhood experiences questionnaire; SF-12	**Baseline only:** mental health questions; childhood experiences questionnaire **Baseline and 12 months only:** AUDIT-C; DUDIT; SF-12 **Baseline, 4, 8 and 12 months:** IMPACT toolkit; ABI-R; PHQ-9; GAD7; PC-PTSD-5
To determine the cost-effectiveness of the intervention to the individual and society	Resource use questionnaire; police data; EQ-5D-5L; CHU-9D (see the HEAP for details)	**Baseline, 4, 8 and 12 months:** Resource use questionnaire; EQ-5D-5L; CHU-9D **12 months post randomisation and 12 months pre baseline:** Police data

* Section 4.3 explains the difference between ABI-29 and ABI-R

2. Study methods

2.1 Synopsis of the study

Table 2: Trial Summary

Trial Title	The effectiveness and cost-effectiveness of a group programme for men who are concerned about their abusive behaviour in relationships with women: A randomised controlled trial	
Internal ref. No. (or short title)	Group programme for men who are concerned about their abusive behaviour in relationships with women	
Phase	Main study	
Trial Design	Individually randomised controlled trial	
Trial Participants	Men aged 21 and over and their female partners or ex-partners (aged 18 or over)	
Planned Sample Size	316 men (reduced from the original sample size of 366 men).	
Treatment duration	The intervention group programme for men lasts 23 weeks (with additional individual sessions based on need and a monthly Relapse Prevention Group (RPG) for a further six months following completion of the programme). The intervention for linked female partners and/or ex-partners is support from a women's safety worker for the same duration as the men as needed (23 weeks with a further six months).	
Follow-up duration	12 months	
Planned Trial Period	April 2019 – December 2024	
	Objectives	Outcome Measures
Primary	To investigate the effectiveness of the group programme intervention in reducing men's abusive behaviour against women	Men's self-reported measure of abusive behaviours (ABI-29) at 12 months
Secondary	To investigate the effectiveness and cost-effectiveness of the group programme intervention on men's abusive behaviour and the wellbeing of partners and ex-partners' experience of abusive behaviours and wellbeing	Men's self-reported measures of abusive behaviours and wellbeing, police reports of incidents and partners/ex-partners' self-reported measures of the experience of abusive behaviours and wellbeing

2.2 Description of the Intervention

The DVPP consists of a 23-week programme incorporating additional individual sessions based on need and a monthly RPG for an additional six months following completion of the programme and are delivered by services that are RESPECT-accredited or working towards such accreditation.

The group sessions are delivered by two experienced DVPP facilitators (one male and one female, where possible, to model good gender role behaviours). The programme runs as a rolling programme, allowing new intakes of participants to join at specified intervals.

The weekly group sessions incorporate most of the elements that exist in standard DVPPs. These include: goal identification and goal setting; recognising abuse; denial and minimisation; intents of violence; essential anger management; identifying urges to perpetrate abuse and cooling-down strategies; basic Cognitive Behavioural Therapy; effects of DVA on partners and children; participant’s own childhood experiences; impacts on children; active listening; conflict resolution; masculinity; beliefs and expectations; sexual respect; attachment styles; building empathy; loving relationships; emotional abuse; and accountability.

The individual sessions are tailored to participants’ needs following the initial and ongoing assessment. Possible individual interventions include: deconstructing specific incidents of abuse; accountability letters or planning discussions with partner or children; relaxation, or emotional regulation work. The delivery team refer and signpost men to specialist services as part of their normal DVPP conduct.

2.3 Partner (women’s intervention)

Women partners or ex-partners of men allocated to the intervention arm are contacted by a designated women’s safety worker as part of the intervention. Participants can signpost more than one partner or ex-partner. It is the woman’s decision whether she engages with the women’s safety worker. Women can engage with the women’s safety worker and decline to take part in the research, or they can take part in the research and decline the women’s safety worker supporter, or they can accept or decline both.

Women’s safety workers offer support to women who may want to remain at home, feel unsafe at home and need to go to a safehouse, or want to stay home and are not ready or do not wish to leave their abusive partner. Practical and emotional support is given to help victims keep safe, help with any court proceedings, connect with the community, and plan for the future.

2.4 Control men and partners

Men allocated to the usual care control arm will not receive any intervention or referrals from the research team; however, they can access any other services available as part of their usual care. The research team may signpost to other appropriate services (e.g. mental health services) if deemed appropriate and necessary. Regardless of their partner’s allocation, all women will be signposted to women’s support services.

2.5 Relapse Prevention Group (RPG)

Men who are allocated to the intervention arm are able to access the RPG upon completing the DVPP. The RPG meets monthly and is run by the local service provider team (facilitator or DVPP coordinator). These meetings are less structured than the DVPP programme, with an emphasis on ‘checking in’ on how the participants manage their behaviours.

2.6 Study setting

This study is community-based. Initial meetings are arranged in a mutually convenient location for the researcher, DVPP coordinator and potential participant or on an online platform such as Zoom. This is within a community well-being organisation, health or social care building, or university building.

2.7 Trial areas

Services that are either already RESPECT-accredited or working towards such accreditation deliver the intervention. The group programme intervention is delivered in community settings in five areas: three in southwest England i) Bristol, North Somerset and South Gloucestershire ii) Somerset, iii) Wiltshire; and the fourth and fifth in Wales iv) Blaenau Gwent and v) Neath Port Talbot. (Note that area and site have the same meaning in the protocol, but for the SAP we will use the former term throughout.)

2.8 Design

This is a pragmatic, parallel-group, individually randomised controlled trial. We will investigate the effectiveness and acceptability of the DVPP (including a Relapse Prevention Group; RPG) and integrated women’s support worker service as the trial intervention. We will use a mixed methods process evaluation to inform acceptability and barriers to implementation and conduct an economic cost-effectiveness analysis. Neither the economic evaluation nor the mixed methods process evaluation is covered in this SAP.

2.9 Randomisation

Bristol Trials Centre (BTC) provides an automated randomisation procedure whereby participants are randomly allocated in a 2:1 ratio to the intervention and control arms respectively, via a computer programme accessed remotely by the recruiting researcher.

Randomisation is stratified by area (Bristol/North Somerset/South Gloucestershire (BNSSG), Somerset, Wiltshire, Blaenau Gwent and Neath Port Talbot) and minimised by relationship status. This will ensure similar distribution of selected participant factors between study arms. The first participant is independently randomly allocated; for each subsequent participant, the treatment allocation that minimises the imbalance in the relationship status (whether the participant is still living all or most of the time with the abused partner) between arms at that time is selected, albeit with a probabilistic element retained. The probability of being assigned to a group providing balance is 80%.

The allocation is confirmed via an email from the randomisation system to the research team. This information will then be recorded on the trial database, although not revealed to the authorizing statistician and health economists, who will be blinded to allocation until the SAP is completed and all principal data-related decisions have been made.

2.10 Framework

This is a randomised superiority trial.

2.11 Sample size

Due to the design of this study (including that each intervention group has a rolling intake of participants), the power calculation for the primary analysis of comparing the mean participant ABI-29 between the two treatment groups was calculated both with and without taking account of the potential clustering within the intervention group.

All power calculations assumed that a 2:1 allocation ratio (intervention:control) would be used, a total of 219 participants would be available for analysis, and a two-sided significance level of 5% applied. The original target sample size for recruitment was inflated to account for 40% attrition giving 366 participants to be recruited (244 intervention and 122 control).

The sample size unadjusted for clustering was calculated using the power command in Stata 15.1, with an effect size of 0.4 of a standard deviation (0.4SD). This would give 79% power.

The sample size accounting for clustering was calculated using the clsampsi command in Stata 15.1. The mean cluster sizes for the intervention treatment groups were 9.125 participants for each (16 groups in total), and no clustering was assumed in the control group. A range of intervention intra-cluster correlation coefficients (ICCs) were considered (0.025 to 0.05), due to the uncertain (and at this stage unavoidably unknown) effect of clustering within this rolling-intake group structure.

Notwithstanding this uncertainty, with an (arguably conservative) ICC of 0.05, the above sample size would yield a power of 73% to detect an effect size of 0.4SD and a power of 80% to detect a 0.435SD effect size. For an ICC of 0.025, there will be 76% power to detect an effect size of 0.4SD and 80% power to detect an effect size of 0.42SD. In summary, our sample size will therefore have between 73% and 80% power to detect an effect size of between 0.4 and 0.435 SD for a range of plausible ICCs.

A sample size calculation was not conducted for the partners/ex partners. Each participant listed partners and/or ex partners who were invited to join the study. In the feasibility study, the number of partners/ex partners joining the study was half the number of participants randomised.

2.12 Reduction in sample size

Difficulties were experienced in meeting the original target for recruitment, primarily because the first 18 months of recruitment were conducted during the COVID-19 pandemic. Hence, along with an extended recruitment period, a reduction of 50 in this target was subsequently approved by the Programme Steering Committee (PSC), Data Monitoring and Ethics Committee (DMEC), and the funding body. Using the SD of 10.7 derived from 201 baseline measurements in this trial and again assuming 40% attrition, the revised sample size of 316 to be recruited will provide approximately 80% power to detect an effect size of 0.43SD under the primary assumption of no clustering. This slightly revised target difference, corresponding to 4.6 ABI-29 scale points (compared with the original 4.4), was still deemed both plausible and worth detecting to inform policy and commissioning of services.

2.13 Blinding

This study is not blinded, except for the authorizing statistician and health economists prior to the signing of the SAP and HEAP, respectively.

3. Populations

3.1 Study populations

3.1.1 Inclusion criteria for male participants

- ≥ 21 years of age.
- Use of abusive behaviour in current or previous relationships with women partner(s) or ex-partner(s) and concerned about that behaviour.
- Ability to complete outcome questionnaires with or without the assistance of the researcher.
- Need to be able to understand and participate in an English-speaking group setting.
- Must have contact with an abused partner or ex-partner within the last twelve months at the time of recruitment and/or anticipate having contact with an abused partner or ex-partner within the next twelve months.

3.1.2 Exclusion criteria for male participants

- Court-mandated referral to the perpetrator programme.
- Men who are deemed too high risk as assessed by a DVPP coordinator or by the research team.
- Men whom the DVPP coordinator deems as not willing to engage with the intervention.
- Men with known previous violence or aggression towards professionals.
- Participants who cannot understand the English language sufficiently well to give informed consent and to complete the questionnaires (with or without assistance) or to participate in a group setting.
- Participants unable to consent to and engage with a group programme (this will include, but is not limited to, persons with severe mental health difficulty, serious learning disability or unstable substance misuse difficulties.
- Men who are currently in Child Arrangement Order (CAO) proceedings with an open Children and Family Court Advisory and Support Service (CAFCASS) case, who have been in such proceedings in the last 12 months, or who state they intend to open such proceedings in the next 12 months. This exclusion criterion may be adapted to accord with the guidance from Respect.
- Men who have ongoing criminal justice investigations for a DVA incident towards a partner or ex-partner (i.e. waiting to hear if they will be going to court or waiting for a court date).

- Men unwilling or unable to provide partner/ex-partner details to enable the research team to contact them. Men who fall outside the catchment areas (for the purposes of collecting data on police records).
- With the exception of attending a group programme of any length while in prison, men who have already participated in a group perpetrator programme which was longer than 10 weekly sessions or 10 days, within the last 12 months.

3.1.3 Inclusion criteria for partners/ex-partners

- Female partners or ex-partners of men using violence/abuse in their relationships.
- $\geq$18 years.
- Ability to complete outcome questionnaires with or without assistance of the researcher.

The difference in minimum ages between men and women has been discussed and the expert opinion is that younger men who abuse (pre-21) are often less ready to change their abusive behaviour and the younger age group of men report having qualitatively different types of relationships with women. For example, younger men can use the internet and social media to be abusive much more and therefore do not relate as well to men of older ages in group settings. In addition, younger men are more likely to be groomed (in terms of potentially abusive and/or criminal behaviours) and it could therefore be problematic having very young men and older men together in the same group.

3.1.4 Exclusion criteria for partners/ex-partners

- Participants who cannot understand English sufficiently well to give informed consent and to complete the questionnaires (with or without assistance).
- Women who are deemed (by the women's safety worker, DAPP coordinator or research team) to be put at greater risk if they take part in the study.

3.2 Data sources

Most study data will be collected on case report forms (CRFs), with the exception of withdrawal information, the attendance of the group sessions police data, and individual sessions and telephone call contact with the participants. CRFs (paper or electronically collected) will be recorded in REDCAP. The withdrawal information, the attendance of group sessions and individual sessions and phone call contact were collected on spreadsheets.

3.2.1 Police data

The key value of the police data is that it gives an *external* source of data with which to evaluate the incidents of abuse in the two trial arms. The police data also give us information about men who did not complete their questionnaire.

For each man with consent for police data collection, we will identify domestic violence incidents and crimes (in which he might have been the perpetrator and referred to as 'involved person' or 'suspect' in the 12 months prior to recruitment, and the 12 months since recruitment, including the pause period. We will request for each male participant (who has consented) where he was likely to be the perpetrator: a) a count of the number of police incidents/crimes flagged as domestic violence and abuse in which he was likely to be the perpetrator; b) the date(s) of the incident(s)/crimes; c) the police case ‘outcome’ for each incident/crime (e.g. No Further Action, Charge, Domestic Incident only etc); d) a count of the number of entries on the Log of Enquiries for each of these incidents; e) risk scores/ratings for each of these incidents where available and whether referred to Multi Agency Risk Assessment Conference (MARAC).

There were four relevant police forces covering the study population. The police force records associated with the man’s primary address were the ones that will be initially searched. If no records were found for a man on the initial police force search, their names were added to a ‘miscellaneous’ list and then subsequently searched in the records of the remaining three forces. Searching primarily for incidents and crimes in one force for most men may undercount the number of incidents and crimes if men committed incidents and crimes in several areas. However, record searching for the police is burdensome, (with some smaller forces searching records by hand). To successfully negotiate data sharing agreements with each force it was considered pragmatic to request smaller numbers of men primarily associated with the force area.

Only the analysis of the count of police attended incidences post randomisation will be covered in this SAP, the rest of the data will be used in the health economics and mixed methods process evaluation.

3.3 Analysis populations

3.3.1 As randomised – with available data (excluding any without proper consent)

All summaries and main analyses will be conducted on the observed data on an as randomised basis i.e. data will be analysed in the groups to which participants were randomised, regardless of intervention received. This will consist of all participants with the outcome available, included according to their randomised intervention, regardless of whether they are found to be ineligible post-randomisation, the amount of intervention received, or otherwise have protocol deviations. All analysis will exclude any participants randomised who did not provide documentation of consent.

3.3.2 Complier Average Causal Effect (CACE) population

A per-protocol analysis will not be conducted. A single CACE analysis at a cut off level of 6 group sessions for compliance will be carried out, along with summary statistics at two further cut off levels (2 and 12, so one either side of the threshold of 6 in the CACE analysis) as a descriptive sensitivity analysis. As in per-protocol analysis, CACE analysis is subject to bias, as similar participants cannot be matched to those in the control arm.

3.3.3 Safety

All Adverse Events and Serious Adverse Events will be reported, including potentially for people not randomised. The number of events (by serious/non-serious categories and per randomised arm and the count of participants who had serious adverse events will be reported by randomised arm.

3.3.4 Withdrawals from intervention (group sessions)

All participants who are excluded from their group or inform us of their decision to no longer attend their intervention group will continue to be followed up until the last data collection, unless they withdraw full consent for further follow-up data to be collected.

A table showing the number per arm, and reasons for exclusions or withdrawal from the group intervention, will be included.

3.3.5 Withdrawals from the trial (implicit or explicit withdrawal of consent)

In this study, withdrawal from the trial means implicit or explicit withdrawal of consent for future contact or follow up. The study does not have an official withdrawal form, and most withdrawals of consent are made verbally. A table showing the number per arm of withdrawals from the trial, and descriptions of reasons for withdrawal description and/or withdrawal circumstances (if available) will be included. For some of the participants expressing implicit or explicit withdrawal of consent for future contact or follow up, a letter was sent providing opportunity to withdraw consent for police data. A table showing the number of men contacted and the number who replied and withdrew consent to collect police data will be provided.

All data previously collected by the participant will be used in the analysis unless they specifically withdraw consent for previously collected data to be used.

4. Statistical analyses and report content

4.1 General considerations

All principal analyses will be carried out on an *as randomised* and a complete case basis as determined by the variables included in the relevant analysis. All confidence intervals and statistical tests will be two-sided, and 95% confidence intervals will be reported.

The trial has been designed with a two-sided alpha of 0.05, with no formal adjustment designated for multiple testing in relation to the numerous secondary outcomes; due consideration will need to be taken in the interpretation of secondary outcome results to reflect the number of statistical tests performed.

4.2 General calculations

Unless otherwise stated, all percentages will be calculated using the total number of participants from the relevant sample as the denominator regardless of whether or not they are missing for that variable.

4.3 Outcomes

4.3.1 Primary outcome

Abusive Behavior Inventory – 29 (ABI-29)

The primary outcome will be abuse reported by men based on the Abusive Behavior Inventory – 29 (ABI-29) measure of abuse. ABI-29 is an updated version of the ABI (Shepard and Campbell, 1992), containing twenty-nine of the thirty original items. ABI-29 is divided into two sections, the psychological and physical items containing twenty and nine items, respectively. Participants are required to choose between options 1 (Never), 2 (Rarely), 3 (Occasionally), 4 (Frequently), and 5 (Very Frequently). The total score ranges from 29 to 145, with higher scores indicating greater frequencies of abuse. The questionnaire will be applied at baseline, 4, 8 and 12 months' follow-up. The questionnaire is scored by summing all the items with equal weight to each question.

Our questionnaire has been modified as follows: 1) we removed the question regarding spanking, as recommended by Postmus, Stylianou and McMahon (2015); 2) in question 21, we changed the word "parent" to "person"; 3) starting from 31 January, 2020, participants were instructed to tick 1 (Never) if any question did not apply to them.

4.3.2 Secondary outcomes

Abusive Behavior Inventory – Revised (ABI-R) for women

The Abusive Behavior Inventory – Revised (ABI-R) (Postmus, Stylianou and McMahon, 2015) will be used to assess victims' abuse experiences from their partners. ABI-R is an updated and 'fit-to-victims' version of the ABI (Shepard and Campbell, 1992) containing twenty-five questions. ABI-R is divided into three sections, psychological, physical, and sexual, containing thirteen, nine and three items, respectively. Participants are required to choose between options 1 (Never), 2 (Rarely), 3 (Occasionally), 4 (Frequently), and 5 (Very Frequently). The total score ranges from 25 to 125, with higher scores indicating greater frequencies of abuse. The questionnaire will be applied at baseline, 4, 8 and 12 months' follow-up. The questionnaire is scored by summing all the items with equal weight to each question.

GAD-7

The Generalized Anxiety Disorder – 7 (GAD-7; Spitzer, R *et al*. 2006) was chosen to assess the severity of generalised anxiety disorder symptoms in participants. It consists of seven items that capture the frequency and intensity of common anxiety symptoms experienced over the past two weeks. Participants are required to choose between the options 0 (Not at all), 1 (Several days), 2 (Over half the days), and 3 (Nearly every day). The total score ranges from 0 to 21, with higher scores indicating higher anxiety levels. The questionnaire will be applied at baseline, 4, 8 and 12 months' follow-up. The questionnaire is scored by summing the 7 items with equal weight to each question.

Note that there is an additional (eighth) question about difficulty in work, home or getting along with people. This question is not part of the score and hence will not be part of the main analyses for this outcome; it will, however, be reported separately in using descriptive statistics.

PHQ-9

The Patient Health Questionnaire – 9 (PHQ-9; Kroenke, Spitzer and Williams, 2001) was chosen to assess the severity of depressive symptoms in participants. It consists of nine items that capture the frequency and intensity of common depression-related symptoms experienced over the past two weeks. Participants are required to choose between the options 0 (Not at all), 1 (Several Days), 2 (More than half the days), and 3 (Nearly every day). The questionnaire is scored by summing the 9 items with equal weight to each question. The total score ranges from 0 to 27, with higher scores indicating more depressive symptoms. The questionnaire will be applied at baseline, 4, 8 and 12 months' follow-up.

Note that there is an additional (tenth) question about difficulty in work, home or getting along with people. This question is not part of the score and hence will not be part of the main analyses for this outcome; it will, however, be reported separately using descriptive statistics.

PC-PTSD-5

The Primary Care Post Traumatic Syndrome Disease – 5 (PC-PTSD-5) (Prins *et al*., 2016) was chosen to assess the severity of PTSD symptoms in participants. The questionnaire begins with an item to assess whether the participant has experienced a traumatic event. If the participant denies this, then the questionnaire is scored with 0, and additional questions are not shown if completing the questionnaire online. Otherwise (or on paper), the participant must answer five additional yes/no questions about how that trauma has affected them over the past month. The total score ranges from 0 to 5, with higher scores indicating higher PTSD symptoms. The questionnaire will be applied at baseline, 4, 8, and 12 months' follow-up.

AUDIT-C

The Alcohol Use Disorders Identification-Concise (AUDIT-C) (Bush *et al*., 1998) was chosen to measure and identify participants with active alcohol use disorders or hazardous drinkers. The instrument has three questions and is scored on a scale of 0-12. Each question has four answers ranging from 0-4 points. Higher scores indicate the more likely a person's drinking affects their life. The instrument will be used at baseline and 12 months follow-up.

DUDIT

The Drug Use Disorder Identification Test (DUDIT) (Berman *et al*., 2005) was selected to collect information about drug use among the participants. This instrument is effective in filtering and categorising participants with drug problems. The instrument consists of 11 questions covering the patient's mental and health picture. Higher scores indicate higher drug abuse. The instrument will be used at baseline and 12 months' follow-up.

In contrast to the original questionnaire in Berman *et al* (2005), a yes/no question was placed in the beginning to ask participants if they have ever used drugs. In the online form, DUDIT's original questions are not displayed if participants answered no. Additionally, neither our paper questionnaire nor online form show the page of example drugs that is part of the official question. These differences could lead to the underreporting of drug use.

IPVRAS – Adapted

The Intimate Partner Violence Responsibility Attribution Scale (IPVRAS) (Lila, et al. 2014) – Adapted will be used to assess attribution of responsibility in male abusive behaviour in an adapted form. The instrument comprises twelve questions attributing responsibility for their abusive behaviour to the legal system, the victim or the offender's context. Each question has five answers ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). The questionnaire will be applied at baseline, 4, 8 and 12 months' follow-up.

The adapted version used in the trial encompasses the following changes from the original instrument. All questions regarding legal matters were removed (1. "I am here because of an injustice." 3. "An unfair legal system (laws, judges, etc.) is the reason why I am in this situation." 7. "The reason why I am here is because the Law gets involved in private matters." 9. "I am here because nowadays, "domestic violence" is a label applied to trivial things."). Additionally, some questions in the IPVRAS-Adapted changed the wording. Questions 2, 8, and 9 changed the phrase “I am here because…” to “the abuse in my relationship.” Furthermore, question 4 will be added to IPVRAS-Adapted (“My partner’s alcohol or substance abuse is the reason for the abuse in my relationship”). Individual answers to questions will be presented, as this set of questions are not a validated instrument.

RFQ

The Reflective Functioning Questionnaire (RFQ) (Fonagy *et al*., 2016) will measure the participant’s capacity for reflective functioning. Reflective functioning refers to a person’s ability to understand and interpret their thoughts, feelings and behaviours, as well as those of others, in terms of underlying mental states, such as desires, intentions and beliefs. The instrument comprises eight items, and the participant must tick a box on a scale from 1 (Strongly Disagree) to 7 (Strongly Agree), depending on their experiences.

There are some differences in the version of the questionnaire used versus the published version (Fonagy *et al*., 2016). The scale bar presented in the original instrument only shows “Strongly Disagree” and “Strongly Agree” at both ends of the scale. In contrast, the scale presented in our questionnaire divides the numbers into three sections. Sections 1 and 2 represent “Strongly Agree”, 3, 4, and 5 represent “Neither Agree or Disagree”, and 6 and 7 represent “Strongly Agree.”

We will be using the RFQ-6 (Müller, *et al*., 2021) method to score the questionnaire, where questions 4 (“When I get angry, I say things that I later regret”) and 7 (“I always know what I feel”) are omitted. The final score is the mean of the six questions using actual values, not the recoding of the original scoring method. The scores range from 6 to 42, with higher scores are indicator of hypomentalising behaviour. The instruments will be applied at baseline and 12 months’ follow-up.

Propensity for Abusiveness Scale (PAS)

The Propensity for Abusiveness Scale (PAS) (Dutton 1995; Dutton *et al*., 2001) comprises 30 questions, divided into three subscales: Affective Inability, Trauma Symptoms and Recalled Negative Parental Treatment. We are only asking questions regarding Affective Inability, which is twelve statements from this instrument. Participants are required to select between the options 1 (Completely Undescriptive of You), 2 (Mostly Undescriptive of You), 3 (Partly Undescriptive, Partly Descriptive of You), 4 (Mostly Descriptive of You), 5 (Completely Descriptive of You), depending on their experiences. The score is the sum of the answers to the 12 items, equally weighted. The score ranges from 12 to 60, with a higher score indicating a higher propensity for abusiveness. The questionnaire will be applied at baseline, 4, 8 and 12 months’ follow-up.

Communication Patterns Questionnaire – Short Form (CPQ-SF)

The Communication Patterns Questionnaire – Short Form (CPQ-SF) (Futris *et al*., 2010) will measure communication patterns within interpersonal relationships. It is designed to assess how individuals communicate and interact with others in various contexts. The CPQ-SF consists of eleven questions that evaluate different aspects of communication, such as assertiveness, expressiveness, and responsiveness. It measures positive and negative communication patterns, providing insight into how individuals express their thoughts, feelings, and needs to their partner. The questionnaire will be implemented at baseline, 4, 8 and 12 months’ follow-up.

The questionnaire is divided into two sections, which start the statement of each sentence (A. “When some problem in my relationship arises”; B. ”During a discussion of this issue or problem”).

Participants are required to select a number from 1 (Very Unlikely) to 9 (Very Likely) to rate their behaviours while dealing with problems with their partner.

There are six subscales to score the questionnaire, each of them focuses on a specific pattern of communication that is represented by a subgroup of the eleven questions. The subscales are the following: (a) female demand/male withdraw (Items 4, 9,and 11); (b) male demand/female withdraw (Items 3, 8, and 10); (c) original total demand/withdraw (Items 3, 4, 8–11); (d) alternate demand/withdraw (Items 1, 3, 4, 8, and 9); (e) criticize/defend (Items 6, 10, and 11); and (f) positive interaction (Items 2, 5, and 7).

Furthermore, there have been some modifications to the questionnaire. Wording has been changed to refer to spouses as partners in our questionnaire. Statements of both sections have been reworded (A. “When issues or problems arise, how likely is it that…”. B. “During a discussion of issues or problems, how likely is that…”)

Controlling Behaviours Scale – Isolation (CBS-I) modified

The Controlling Behaviours Scale – Isolation (CBS-I) will be used to measure males’ controlling behaviours to isolate their partners from their personal and social life. CBS-I is a section of the CBS questionnaire (Graham-Kevan and Archer, 2003).

Our questionnaire has some modifications. In the original questionnaire, all items are written as questions, in ours as statements. Question three changed the word “limit” to “restrict”. Question four changed the words “jealous” and “suspicious” to “going mad”. Question five, “If yes, was this used as a reason to monitor and control the other’s activities?” was omitted.

The resulting four-item questionnaire was asked, and participants are required to answer between options 0 (Never), 1 (Rarely), 2 (Sometimes), 3 (Often), and 4 (Always). CBS-I was measured at baseline and 12 months in men and at baseline, 4, 8 and 12 months for ex-partner or partner.

No formal scoring or validation of these 4 questions has been published. Answers to these questions will be reported separately.

Other Behaviours

The Other Behaviours instrument assesses controlling and abusive tendencies. It includes statements on repeated contact attempts, surveillance, verbal abuse, manipulation of children, blaming, confinement, concerns about child safety, and animal mistreatment. The questions come from three sources Gilchrist *et al*., 2020; Woodlock, 2017; Kelly and Westmarland, 2015). This tool aims to identify problematic behaviours exhibited by an individual towards their partner or former partner in order to provide valuable insights for intervention and support. The questionnaire contains ten statements, where participants are required to select an option between Never (1), Rarely (2), Occasionally (3), Frequently (4), Very Frequently (5). However, no formal scoring or validation of these 10 questions has been published. Answers to these questions will be reported separately.

IMPACT Toolkit

The IMPACT Toolkit (https://www.work-with-perpetrators.eu/fileadmin/WWP_Network/redakteure/IMPACT/WWP_ImpactToolkit_A5_publication_web.pdf) has not yet been validated. Unless and until a suitable scoring system has been developed (even if not fully validated) before the relevant analysis commences, responses to this

Toolkit will not be covered in this SAP, except for the scores and means by arm for the three sections noted below. The question at 12 months has been modified to say “How often have you done the following to your partner/most recent ex-partner? (The one you have been abusive toward) in the past 4 months. The answer header is also changed to “WITHIN the last 4 months”. The answers to items in the sections Emotional Behaviour (first 13 items), Physical Behaviour (first 14 items) and Sexual Behaviour (first 8 items) will be scored by scoring “Never” as one, “Sometimes” as two, and “Often” as three (Vall- personal communication). Note that all items in a section have equal value. In addition, note that the last item, which asks for something else related to the behaviour, will be ignored for the scoring purposes. Note that these scores may be lower than those in other studies in that they are asking for behaviour within the last 4 months, not over the last 12 months.

The IMPACT monitoring toolkit will evaluate the changes in male and female participants’ behaviour. The questionnaire has been adapted to be used in a clinical setting. The questionnaire has five different sections and subsections.

1. How you came to the programme.
2. Behaviour you have used towards your partner/ex-partner.
    a. Emotional behaviour
    b. Physical behaviour
    c. Sexual behaviour
    d. Impact for abusive behaviour on your partner
3. Your children.
4. Your partner/ex-partner (the one you were abusive to) and your relationship.
5. Final thoughts.

The first section is regarding the participant’s knowledge about the study. It contains baseline questions such as their age, gender, education, socioeconomic status, accommodation, religion, ethnicity, and partner status. These baseline questions are similar to the demographic questions asked in this study.

The questionnaire will be implemented at baseline and the 12 month follow-up. At baseline, participants are required to judge their behaviour within the last 12 months and before the last 12 months. Therefore, each question has two answers. At the 12 month follow-up, participants judge their behaviour just within the last 4 months. Each period has three options, “Never”, “Sometimes”, and “Often”. Furthermore, each section has two open-ended questions to mention other behaviours in the same time periods.

Conversely, section 2d questionnaire used in our study only uses the first two of the three questions. Question 1 presents statements that describe the impact their behaviour had on their partner. In question 2, presents reasons that made that behaviour.

Section 1 (Emotional Behaviour) is missing three questions (“Made her worried you might leave”, “Made her defend self/child/pets”, “Made her feel afraid about you”) in baseline questionnaire, which appear in the 12 month questionnaire.

The third section is regarding their children’s status. The questionnaire aims to retrieve information about their legal status, social care plans and emotional relationship with their children. There is the possibility to give more detail about their relationship with their children in an open question. There is a question enquiring about children's disabilities or special educational needs. Participants are required to tick a box if any of the statements applies to one or more of their children. The questionnaire is applied at baseline and 12 months' follow-up.

The fourth section is a six-item questionnaire with Likert response, regarding the relationship status with the partner and the degree to which participants are afraid of their abusive behaviour. The instruments is completed at baseline and 12 months' follow-up.

The fifth section is an open text section.

Police data

The number of police incidents post randomization (12 months plus pause period) will be collected from the regional police stations.

ICECAP-A

The ICEpop CAPability measure for Adults (ICECAP-A) (Al-Janabi, Flynn, and Coast, 2012) was chosen to assess an individual's capability to enjoy a meaningful and fulfilling life. It consists of five attributes: stability, attachment, autonomy, achievement, and enjoyment. Respondents rate their level of capability in each attribute based on their own experiences. ICECAP-A contains five items regarding aspects of capability beyond health. Each item describes four different scenarios which score from 1 (indicating lowest level of capability) to 4 (highest level of capability). Tariffs for each answer for each of the items are given in Flynn *et al*., (2015) and the ICECAP-A Scoring document (https://www.bristol.ac.uk/population-health-sciences/projects/icecap/icecap-a/) will be used. The instrument is employed at baseline, 4, 8 and 12 months' follow-up.

SF-12

The Short Form Health Survey – 12 (SF-12) (Ware, Kosinki and Keller, 1996) questionnaire was chosen to measure participant's general health status and overall well-being. It consists of 12 questions that assess various aspects of physical and mental health in the last 4 weeks. The SF-12 covers eight health domains, including physical functioning, role limitations due to physical health problems, bodily pain, general health perception, vitality, social functioning, role limitations due to emotional problems, and mental health. The instrument is employed at baseline and 12 months' follow-up and will be scored using commercial software PRO-CORE (version 2.2, 2009 US Norms).

4.4 Health Economics Outcomes

EQ-5D-5L

The EuroQuol – 5D-5L (EQ-5D-5L) (Herdman *et al*., 2011) questionnaire assessed an individual's health-related quality of life. It has five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Details will be given in the Health Economics Analysis Plan.

Child Health Utility 9D (CHU-9D)

We will use the Child Health Utility – 9D (CHU-9D) (Stevens, 2012) questionnaire to measure the partners children quality of life. The assessment comprises a concise survey and a collection of preference weights based on general population values. Since partners or ex-partners will answer the survey, we will be using the proxy version.

These items will not be covered in this SAP; however, details of the analysis of these outcomes will be given in the Health Economics Analysis Plan.

4.5 Other survey or instruments collected only at baseline

AQ-10

The Autism Spectrum Quotient – 10 (AQ-10) (Allison, Auyeung and Baron-Cohen, 2012) will be used at baseline to assess autistic traits in male participants. It consists of ten questions designed to identify specific behaviours and preferences commonly associated with autism spectrum disorder (ASD). The scoring system is as follows: for questions 1, 7, 8, and 10, participants score 1 point for "Definitely Agree or Slightly Agree"; for questions 2, 3, 4, 5, 6, and 9, participants score 1 point for "Definitely Disagree" or "Slightly Disagree." The score ranges from 0-10 points, with a higher score indicating that the individual is more likely to be autistic.

Childhood experiences questionnaire

The childhood experiences questionnaire will retrieve information about abusive and traumatic experiences participants have had with their parents. The questions were adapted from the PROVIDE study (Hester*, et al*., 2015; Yakubovich, *et al*., 2019). It comprises eight dichotomous questions regarding sexual and physical violence committed by their parents. The questionnaire is composed of two sections of four Yes (1) / No (0) questions in each section. The first section recalls experiences happening at any time of their life, whereas the second section recalls experiences before the age of 16. The questionnaire will be applied at baseline. As no validated scoring method is currently available, each question will be listed separately in appropriate tables.

4.6 Analysis of the outcomes

All of the outcomes listed below are from the male participant, except for the ABI-R, which is provided by the partner/ex-partner.

4.6.1 Primary outcome

The primary analysis will be performed on a complete case basis and according to the arm to which the participant was allocated. The primary outcome of the ABI-29 at 12 months, adjusted for the pause period, will be analysed using a linear regression model that will include the baseline ABI-29 score, the stratification factor of centre and minimisation factor of relationship. For those missing baseline ABI-29, an indicator variable will be included and the mean baseline ABI-29 will be imputed.

Due to amount of missing outcome data (at least 30%) and the high likelihood of differential attrition, this method, as are all methods in the presence of missing data and/or differential attrition, is likely to be biased. It is nonetheless the most reproducible method, with no further assumptions and decisions on variables to be included. This method assumes that those who report are like those who do not, and the data are missing at random. Extensive sensitivity analysis will be conducted, including a best case/worst case analysis, a table of estimated intervention effects based on differences in means of those who reported and did not report the primary outcome and an analysis adjusting for variables observed to be associated with missingness.

4.6.2 Secondary outcome

Continuous secondary outcomes will be analysed in a similar manner as the primary outcome, except the baseline value of the secondary outcome will be used instead of the ABI-29.

The p values for the secondary outcomes will not be adjusted for multiple outcomes but we will emphasise the caution needed in interpreting them and will focus on patterns of effects rather than isolated outcomes with apparent evidence of differences between the trial arms.

Summary statistics for items and questions not listed in the secondary outcomes section will be reported by arm.

4.6.3 Mis-randomised patients

In general, participants will be analysed as randomised. In one case, verbal consent was given, but written consent was never obtained. The CONSORT (Schulz, Altman and Moher, 2010) chart will include this participant including their trial arm but, as indicated in Section 3.3.1 above they will not be included in any data analyses, including at baseline.

4.6.4 Sensitivity analyses

Several sensitivity analyses will be performed.

To show the possible effects of the missing data

A best case/worst case analysis will be conducted for the primary outcome at 12 months to reflect the range of intervention effects that could be seen based on these assumptions.

A two-by-two table indicating the difference in means between those with primary outcomes and those without by trial arm will be completed to indicate the intervention effect and estimated p value. This analysis will be use a simple substitution of the mean plus the indicated value for each arm as indicated for the missing men. This will allow evaluation of the effect of different assumptions about the missing data. The evaluation of this table will be informed by reference to the difference in intervention effect used for the sample size.

To estimate the effect of the COVID-19 pause period

An analysis will be conducted including a variable indicating whether or not the participant was randomised by 1 July 2020 (that is, affected by the COVID-19 pause) and an interaction term added between the intervention and this indicator variable. Interpretation will be cautious since the power for such an interaction term will be very limited.

To estimate any effects of the variables unbalanced at randomisation

If differences in the baseline demographic or key baseline data (police called question, disability question, AQ-10, AUDIT-C, DUDIT, GAD7, PC-PTSD -5, PHQ9 are greater than 10 percentage points for categorical data or 0.5 standard deviations for continuous data in comparing the randomisation arms at baseline, the variables exhibiting such lack of balance will be added to the main regression analysis for the primary outcome. For purposes of comparison, the baseline demographics will be collapsed into reasonable categories for comparison and will be reported in the collapsed categories.

To estimate the effect of including baseline variables that predict missingness on the intervention effect of the primary outcome

If baseline demographic or key baseline data (police called question, disability question, AQ-10, AUDIT-C, DUDIT, GAD7, PC-PTSD -5, PHQ9) are found to differ between the participants with missing and completed primary outcome by either 10 percentage points in a category (as collapsed) or 0.5 standard deviations in means, then such variables will be added to main regression analysis for the primary outcome.

To estimate the effect of intervention in presence of crossover/imperfect compliance

A CACE (Complier Average Causal Effect) estimate using instrumental variables regression will investigate the efficacy of the intervention while accounting for the degree of adherence to the programme. The mean ABI-29 and its standard deviation will be presented for the control group, and for the intervention group in the following categories of sessions attended: 0-1, 2-5, 6-11, 12+. A complier average causal effect (CACE) analysis will estimate the treatment effect in those men attending at least six group sessions, under the assumptions that randomisation has ensured an equal proportion of non-attenders (no more than five sessions) in the two groups, and a comparable outcome for those non-attenders irrespective of the group to which they were allocated.

Assumptions needed for CACE analysis include:

- Randomisation must affect outcome only through *receipt* of the intervention – that is:
  - Randomisation has no effect on “never-takers” (resentful demoralisation, or happy not to have to do group).
  - May not hold if “refusing a treatment when offered” could have an effect on outcome.
  - Treatment assignment itself may have some effect on the outcome (e.g., it may have a positive or a negative effect, in particular, on psychological outcomes).

These and other assumptions that are needed for the CACE analysis will be described in the report of the analysis and discussed. In addition, a table and graphic presentation of the number of sessions versus the ABI-29 will show the distribution of data used to estimate the level of effectiveness at each level of ‘dose’.

There is no previous evidence on the minimum number of sessions likely to be effective. Therefore, to inform a sensitivity analysis based on a minimum number of sessions, we sought expert opinion from a Respect senior manager and an experienced trainer of DAPP facilitators. Based on their opinions, our CACE analysis of the primary outcome will be based on the threshold of men attending at least 6 sessions. Because the 6-session threshold is uncertain, we will use summary descriptive statistics to consider potential effects of two further thresholds – namely, attending at least 2 sessions (a permissive threshold) and attending at least 12 sessions (a conservative threshold).

While our CACE (Complier Average Causal Effect) analyses will attempt to minimise such biases, these analyses will most likely be overestimates because we cannot exclude men from the control group who might not have attended the threshold number of sessions, which we will acknowledge when reporting these analyses.

In cases where there is no evidence of group sessions, we will assume that no sessions were attended by those participants. For cases where participants have had group sessions not following the REPROVIDE manual, we will include those sessions in the count of the sessions. Examples of this are individual sessions that were given in exceptional circumstances in place of group sessions and where some of the REPROVIDE manual content was followed. A footnote will state the numbers of sessions and number of participants who were with the same providers for this study, although not following the REPROVIDE manual, will be noted.

To estimate effects at different time points

To assess the stability of any intervention effect on the primary outcome, we will attempt to fit a mixed model for the primary outcome at 4, 8 and 12 months, adjusted for the baseline measure, if sufficient intermediate data are available. If over 50% of the intermediate values are missing then this analysis will not be attempted, and only means, SD, and number missing at each time point by arm will be reported.

To estimate the effect of excluding those who skipped items on the primary outcome

A sensitivity analysis including outcomes from those who skipped items in the ABI-29 will be conducted by assuming the missed item was answered at the highest level (score 5 – Very Frequently) in the intervention arm and at the lowest level (score zero – Never) in the control arm. Note that these men are excluded from the primary analysis, and included in the effects of non-random missingness table as if their primary outcome is missing.

4.6.5 Subgroup analyses

Below are the prespecified subgroup analyses, each of which will be conducted by adding an interaction term between the relevant (subgroup) variable and arm into the model for the primary outcome. These analyses will be conducted separately for each of these four variables, and we note that these will be underpowered and should be considered exploratory.

Age

Age will be included as a continuous variable. A scatterplot of age versus the difference in baseline and follow-up primary outcome by arm will be presented as well as a Lowess (Locally weighted regression, Stata command lowess) curve with bandwidth of 30%.

Area

Area will be represented by the above-defined five categories.

Living or not living with partner

This will be a binary variable according to whether or not the man is living with their partner at baseline.

Referral method (self-referral or other referral)

The last pre-specified subgroup analysis will involve a binary variable indicating whether the man was self-referred into the study or referral was by some other route.

4.6.6 Partner Data Analysis and comparison with male responses

The analysis of the partner data is complicated by the fact that the men were randomised to control and intervention, then the partners were contacted to see if they were willing to participate. Up to two partners were contacted and asked to participant for each of the male partners, resulting in some men having two partners in the study. Approximately half of the partners did not participate in the trial. Due to the non-randomisation/self-selection of partners, it is unlikely that the partner groups will be similar even at baseline.

In addition, one purpose of the partner data is to provide some validation/comparison of the men's report of abuse.

Note that the partner(s) who have responded may be current or former partners of the participant, so there is not a direct tie between the male report of abuse and the responses of the current or former partner. Also, the current and former designation is made at baseline; therefore, the current partner indicated may or may not be still in contact with the participant at the end of the study. We also note that the relationship status (whether the participant is still living all or most of the time with the abused partner) indicated at baseline may have changed as well.

Scatterplots, and where appropriate correlation coefficients, between the ABI-29 in the men and ABI-R in the partners/ex-partners will be produced at baseline and at 12 months, and on the change from baseline to 12 months. Note that the participant-to-partner ratio is not one-to-one; in several cases multiple partners were recruited for a specific male. Given that there is no one-to-one link between the participant and partner, averages of the female responses will be used for comparison with male responses when more than one female response is available. Further analyses of these data are beyond the scope of this SAP and will be covered in a separate plan.

4.6.7 Comparison of male report and police report of calls to police.

In the male 12 month survey the males were asked the following question: "In the last twelve months, how often have the police been called because of violence/abuse towards your partner/ex?".

We will compare the answer to the above question and the count from the police data of the number of police incidents/crimes flagged as domestic violence and abuse in which the participant was the perpetrator by showing paired answers in a table. If illuminating, a scatterplot and/or bar graph may also be produced.

We do not expect one-to-one correspondence, as a call may or may not have resulted in a police incident, but these data can provide some support for the men's self-report.

To compare the ABI-29 to the police reported data, a table will be produced comparing the number of police callouts at 12 months to the ABI-29 at 12 months. If illuminating, a scatterplot and/or bar graph comparing these data may also be produced.

4.6.8 Graph showing baseline and 12 month values of ABI-29 per participant, with indications of missing data.

A graph showing the baseline and 12 month value (if available) per participant by arm will be used as a data-checking exercise, and as an aid to understanding the results of the trial. Dotted lines will indicate the missing 12 month data, showing that it could theoretically range from 29 to 145.

4.6.9 Tables exploring the relationship between intervention and missing data.

A table showing the number of sessions attended and the number and proportion of missing primary outcomes will be provided. In addition, a table listing reasons for exclusion from group with the number of missing outcomes will also be provided.

4.6.10 Other considerations

Note that several instruments may have not shown all the questions on the online version, and this may lead to bias when comparing these with answers from paper versions showing all of the questions. Seeing all the questions may prompt the participant to change the initial yes/no answer provided. While it is not possible to test or measure this effect for each of the instruments where the complete instrument was not shown online, it should be considered when comparing scores from paper versions and online versions of tests where not all the items are shown. A table of the number of the forms completed by each method will therefore be included in the study report.

4.6.11 Further exploratory analyses

Any further exploratory analysis will not be covered by the SAP.

4.7 Analysis of Safety

Adverse event summaries will be reported by arm and participant type (male, female, non-consented) and will be split by serious and non-serious events. In addition, the counts of suicide, attempted suicide and other descriptions of suicide intentions will be presented by arm, including both the number of events and the number of those reporting events.

4.8 Missing data

Missing data will not be imputed for the primary analysis but, as noted above, sensitivity of the results to data missing not at random will be explored by best-case worst-case methods, and a two-by-two table showing the effect on the intervention difference when assuming differences in mean value of the missing from those reporting in each arm.

Unless procedures for imputing missing items are noted below, instruments (surveys) will not be scored if items are missing. For the primary outcome (ABI-29) the rationale is that the items are not likely to be missing at random, but potentially due to unwillingness to answer the item. For the primary analysis, those with any missing items will be excluded.

When published instructions exist to impute items missing in a scored instrument, they will be followed and listed in the appendix to the final report. If no published instructions exist, only instruments with no items/questions unanswered will be scored.

4.9 Outliers

All key measures are questionnaires with constrained values, so no values will be considered outliers. The number of primary outcomes at the upper and lower limits will be examined and reported for the baseline and 12 month outcome.

4.10 Visit windows

Note that the term *visit* in this instance means the on paper or online completion of the various measures in the participant follow-up questionnaires. The visit window for the 12 month follow-up (delayed by 135 days for those randomised before 1 July 2020) will be 2 months before to 6 months after the date of expected 12 month follow-up (see Table 3 for other visit windows).

The primary and secondary outcomes will be determined to be belonging to the following time periods:

Table 3: Visit windows

Time period	In Pause Group Randomised before 1 July, 2020		Not in Pause Group Randomised on or after 1 July, 2020		Explanation
	Expected follow-up time (days from randomisation	Windows for Forms (days from randomisation)	Expected follow-up time (days from randomisation)	Windows for Forms (days from randomisation)	
Baseline	(-7, 0)	(-61, 61)	(-7, 0)	(-61, 61)	2 months before to 2 months after
4 Months	257	(62,318)	122	(62,183)	2 months before to 2 months after
8 Months	378	(319,439)	243	(184,304)	2 months before to 2 months after
12 Months	500	(440,561)	365	(305,426)	2 months before to 6 months after expected date.

The study follow-up was paused for participants randomised before 1 July 2020 for 135 days due to COVID-19 (pause period 23 March 2020 to 4 August 2020). Baseline observations are expected to be completed on or before the date of randomisation (noting that in the case of paper completed forms, we may only have date of receipt).

For time periods that have more than one observation, the first observation after the expected time collection will be used. If all observations are before the expected collection time, the observation closest to the expected time will be used.

5. References

Alberti G, (2010). "Responding to violence against women and children: The role of the NHS." The report of the Taskforce on the Health Aspects of Violence Against Women and Children. Available online: https://www.health.org.uk/sites/default/files/RespondingtoViolenceAgainstWomenAndChildrenTheRoleofTheNHS_guide.pdf Accessed on 20 November 2020.

Akoensi TD, Koehler JA, Lösel F, Humphreys DK, (2013). Domestic violence perpetrator programs in Europe, Part II: A systematic review of the state of evidence. *International Journal of Offender Therapy and Comparative Criminology* **57:** 1206-1225.

Al-Janabi H, Flynn TN, Coast J, (2012). Development of a self-report measure of capability wellbeing for adults: the ICECAP-A. *Quality of Life Research* **21:** 167-176.

Allison C, Auyeung B, Baron-Cohen S, (2012). Toward brief “red flags” for autism screening: the short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry* **51:** 202-212.

Berman AH, Bergman H, Palmstierna T, Schlyter F, (2004). Evaluation of the Drug Use Disorders Identification Test (DUDIT) in criminal justice and detoxification settings and in a Swedish population sample. *European Addiction Research* **11:** 22-31.

Bush K, Kivlahan DR, McDonell MB *et al*, (1998). The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. *Archives of Internal Medicine* **158**: 1789-1795.

Dutton DG, (1995). A scale for measuring propensity for abusiveness. *Journal of Family Violence*. **10**: 203.

Dutton DG, Landolt MA, Starzomski A, Bodnarchuk M, (2001). Validation of the propensity for abusiveness scale in diverse male populations. *Journal of Family Violence* **16**: 59–73.

Flynn TN, Huynh E, Peters TJ, *et al*, (2015). Scoring the ICECAP-A capability instrument. Estimation of a UK general population tariff. *Health Economics* **24**: 258-269.

Fonagy P, Luyten P, Moulton-Perkins A, *et al*, (2016). Development and validation of a self-report measure of mentalizing: The reflective functioning questionnaire. *PloS One* **11**: e0158678.

Futris TG, Campbell K, Nielsen RB, Burwell SR, (2010). The communication patterns questionnaire-short form: A review and assessment. *Family Journal* **18**: 275-287.

García-Moreno C, Hegarty K, d' Oliveira AFL, *et al*, (2015). The health-systems response to violence against women. *Lancet* **385**: 1567-1579.

Gilchrist G, Potts LC, Connolly DJ, *et al*, (2023). Experience and perpetration of intimate partner violence and abuse by gender of respondent and their current partner before and during COVID-19 restrictions in 2020: A cross-sectional study in 13 countries. *BMC Public Health* **23**: 316.

Gondolf EW (2012). The future of batterer programs: Reassessing evidence-based practice. Boston: North Eastern University Press.

Graham-Kevan N, Archer J, (2003). Intimate terrorism and common couple violence: A test of Johnson's predictions in four British samples. *Journal of Interpersonal Violence* **18**: 1247-1270.

Herdman M, Gudex C, Lloyd A, *et al*, (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research* **20**: 1727-1736.

Hester M, Ferrari G, Jones SK, *et al*, (2015). Occurrence and impact of negative behaviour, including domestic violence and abuse, in men attending UK primary care health clinics: a cross-sectional survey. *BMJ Open* **5**: e007141.

ICH (2019). ICH Harmonised guideline addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials E9(R1) Final Version. Geneva: International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Available from: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf. Accessed 27 March 2024.

Kelly L, Westmarland N (2015). Domestic violence perpetrator programmes : steps towards change. Project Mirabal final report. London & Durham: London Metropolitan University and Durham University. Available from: https://anrows.intersearch.com.au/anrowsjspui/bitstream/1/20091/1/ProjectMirabalfinalreport.pdf Accessed 27 March 2024.

Kroenke K, Spitzer RL, Williams JBW, (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine* **16**: 606-613.

Lila M, Oliver A, Catalá-Miñana A, *et al*, (2014). The Intimate Partner Violence Responsibility Attribution Scale (IPVRAS). *European Journal of Psychology Applied to Legal Context* **6**: 29-36.

Morgan, K, Man M-S, Bloomer R, Cochrane M, *et a*l, (2023). The effectiveness and cost-effectiveness of a group domestic abuse perpetrator programme: protocol for a randomised controlled trial. *Trials* **24**: 617.

Müller S, Wendt LP, Spitzer C, *et al*, (2022). A critical evaluation of the Reflective Functioning Questionnaire (RFQ). *Journal of Personality Assessment* **104**: 613-627.

NICE (2014). Domestic violence and abuse: multi-agency working. London: NICE (National Institute for Health and Care Excellence). Available from: https://www.nice.org.uk/guidance/ph50 Accessed 27 March 2024.

Postmus JL, Stylianou AM, McMahon S, (2016). The abusive behavior inventory–revised. *Journal of Interpersonal Violence* **31**: 2867-2888.

Prins A, Bovin MJ, Smolenski DJ, *et al* (2016). The primary care PTSD screen for DSM-5 (PC-PTSD-5): development and evaluation within a veteran primary care sample. *Journal of General Internal Medicine* **31**: 1206-1211.

Schulz KF, Altman DG, Moher D, CONSORT Group, (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* **340**: c869.

Shepard MF, Campbell JA, (1992). The Abusive Behavior Inventory: A measure of psychological and physical abuse. *Journal of Interpersonal Violence* **7**: 291-305

Spitzer RL, Kroenke K, Williams JBW, Löwe B, (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine* **166**: 1092-1097.

Stevens K, (2012). Valuation of the child health utility 9D index. *Pharmacoeconomics* **30**: 729-747.

Walby S, (2009). The cost of domestic violence: up-date 2009. Lancaster, UK: Lancaster University. Available from: https://eprints.lancs.ac.uk/id/eprint/88449/1/Cost_of_domestic_violence_update_4_.pdf Accessed on 27 March 2024.

Ware JE, Kosinki M, Keller SD, (1996). A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary Tests of Reliability and Validity. *Medical Care* **34**: 220-33.

Woodlock D, (2017). The abuse of technology in domestic violence and stalking. *Violence against Women* **23**: 584-602.

Yakubovich AR, Heron J, Feder G, *et al*, (2019). Intimate partner violence victimisation in early adulthood: psychometric properties of a new measure and gender differences in the Avon Longitudinal Study of Parents and Children. *BMJ Open* **9**: e025621.

6. Revision history

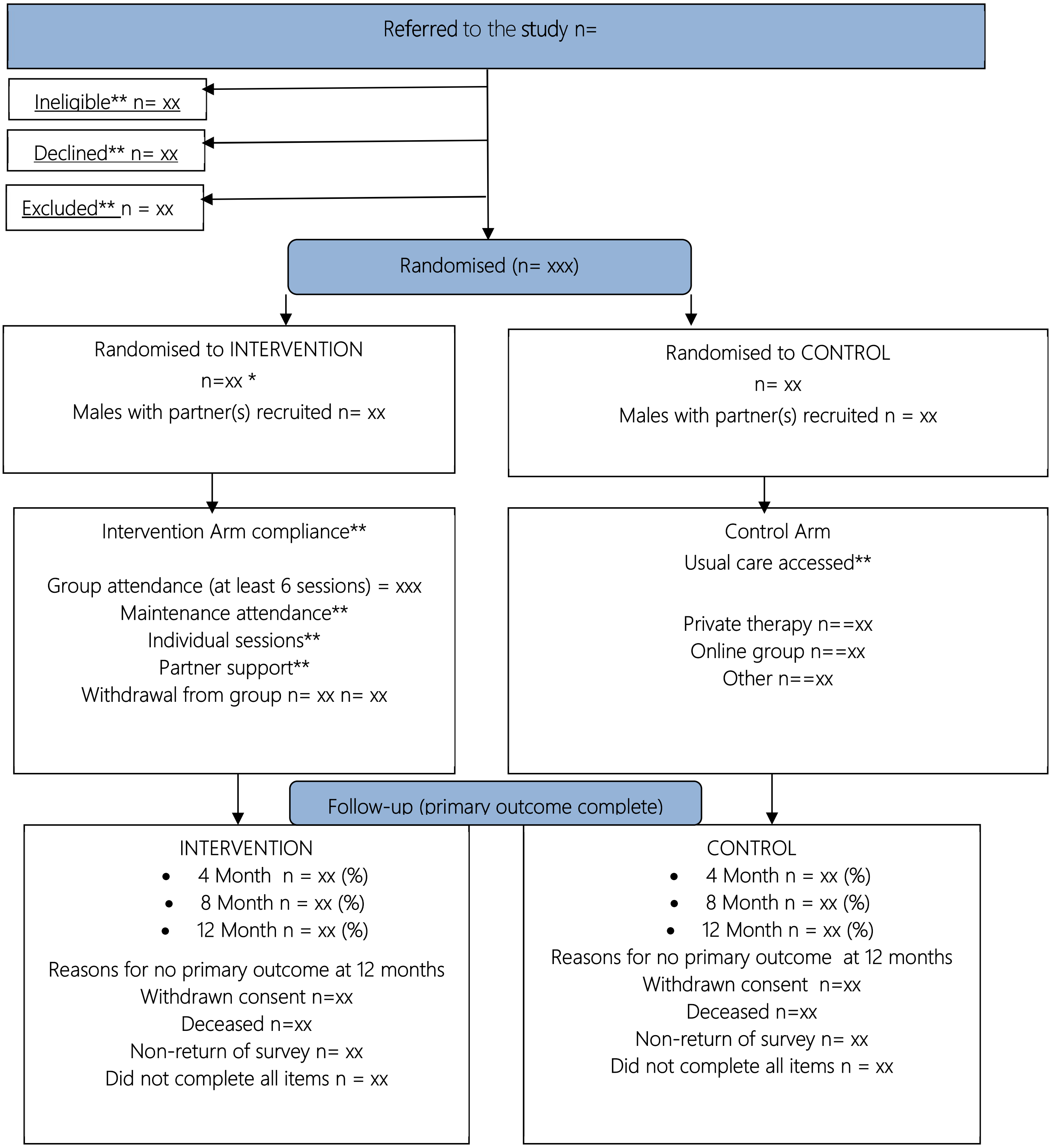
Version 1 of the SAP should be signed off by relevant personnel before any data analysis is carried out. If changes need to be made to v1.0 before this time, possibly due to emerging methodologies, these changes should be documented in Table below, with new version number, date and a summary of the changes with justification(s). If any changes to the methodologies are required after data analysis has begun, these should be documented in the final analysis report in a chronological manner, documenting all decisions made and their justification(s).

Table 4 : SAP revision history

Version number	Revision date	Justification for revision
		[Include details of timing of revision in relation to formal analyses]

7. Proposed tables and figures

Figure 1: CONSORT Flowchart



* One additional participant gave verbal consent, was randomised then failed to provide written consent, and was withdrawn on failing to provide this consent.

** Further details provided in the report.

Table C!: Reasons ineligible

*Ineligible	N
No recognition of abuse	
Says victim not perp	
No ex-partner contact	
No partner details	
Overlapping DVPP	
Court mandated	
Same sex relationship	
Prison recall	
Out of area	
No contact info	

Table C2: Reasons declined

Declined reasons	N
Unable to contact	
Doesn't need course	
Found other support	
Doesn't like group format	
Doesn't want to be in study	
Not guaranteed a place	
Not interested	
Unwilling to travel	
Unwilling to take time off	
Group times	
Too busy	
Doesn't feel ready	

Table C3: Reasons Excluded

***Excluded reasons	N
Can't attend due to work	
Can't attend due to transport	
Can't attend all sessions	
Multiple DNAs	
Aggressive	
Too high risk	
Abusive to professionals	
Serving police officer	
Not suitable for group	
No consent	

Table XX. Participants excluded from further follow-up *(report within CONORT or text or this table*)

Reason For Withdrawal from Study	**Control**	**DAPP**	**Total**
Consent issues			
Death			
Participant request			
Total			

Table XX. Participants excluded from police data collection *(report within CONORT or text or this table)*

Reason For Withdrawal from Study	**Control**	**DAPP**	**Total**
Consent issue			
Participant request			
Total			

7.1 Tables for main body of primary results paper

Table 1: Baseline Characteristics – Demographics

	Control (n=)	DVPP (n=)	Total
Mean age (SD), n			
Do you consider yourself disabled? n (%) Yes No Missing			
Ethnicity: n (%) White Other Missing			
Highest level of academic qualification: n (%) GCSEs or equivalent A levels or equivalent Degree or higher degree or equivalent No academic qualifications Missing			
Employment status: n (%) Employed Looking after your home/family Unemployed and looking for work Unable to work due to long term sickness Other (e.g. in education, retired) Missing			
Total household annual income before tax and benefits: n (%) Up to £11,999 £12,000 up to £37,999 £38,000 and above Prefer not to say/do not know Missing			
What is your religion?: n (%) Christian Other religion No religion Prefer not to say Missing			

Table 1 continued: Baseline Characteristics – Living situation

	Control n (%)	DVPP n (%)	Total
Current accommodation Housed in own tenancy Housed in someone else tenancy Housed in a property I own (with or without a mortgage) Other (sleeping rough, hostel, squatting, sleeping on sofa/floor, emergency accommodation or temporary accommodation) Missing			
Number of people in household including respondent *How many people are there in your household? (All children & adults that live with you for 3 or more days a week)* 1 2 3 4 5 to 8 Missing **Number of Children 0-17 years:** 0 1 2 3 or more Missing **Number of adults 18+ years (not respondent):** 0 1 2 3 or more Missing			
In the last 12 months, how often have the police been called because of violence/abuse towards your partner/ex-partner? None at all Once 2-5 times 6-10 times More than 10 times Missing			

Table 1 continued. Baseline questionnaire measures

	Control Mean (SD), n	DVPP Mean (SD), n	Total
ABI-29 Abusive Behavior Inventory (male participants) Total score Psychological abuse subscale Physical abuse subscale			
ABI-R Abusive Behavior Inventory Revised (female (ex)partners)* **Total score**			
Propensity for Abusiveness (Affective Inability subscale)			
CPQ-SF Communication Patterns Questionnaire – Short Form (modified) Male demand/female withdraw Female demand/male withdraw Original total demand/withdraw Alternate demand/withdraw criticize/defend Positive interaction			
RFQ SF Reflective Functioning Questionnaire (modified) (scored using Muller method)			
PHQ9 Patient Health Questionnaire			
GAD7 Generalized Anxiety Disorder			
PC-PTSD-5 Primary Care Post Traumatic Syndrome Disease			
AQ-10 Autism Spectrum Quotient			
AUDIT-C Alcohol Use Disorders Identification-Concise			
DUDIT Drug Use Disorder Identification Test			
SF-12 Short-Form health survey Physical health subscale Mental health subscale			
ICECAP-A ICEpop CAPability measure for Adults			
EQ-5D EuroQol 5 Dimension 5 Level Five-item score Visual analogue scale			

* xx men had two partners who provided ABI-R; these have been averaged, before including in the mean; note that this variable is included here for completeness but, given the additional selection factors involved, it will not be included in the analyses relating to baseline comparability in the trial for male participants.

Note that a similar baseline table will be repeated in the supplementary material that is split by primary outcome status (missing/non-missing).

Table 2 Analysis of the Abusive Behavior Inventory-29 (ABI-29)

Abusive Behavior Inventory-29 (ABI-29) **Range 29-145, from best to worst**	Control Mean (SD), N	DVPP Mean (SD), N	Difference in means (95% Confidence Interval)
Baseline Baseline for those completing 12 months			
Primary analysis (12 months)			
Sensitivity analyses:			
Indicator added for COVID pause			
Additional baseline balancing*			
Adjusted for variables predicting missing*			
Missing items imputed			
Missing data imputation			
Worst case best case favouring control			
Worst case best case favouring intervention			
CACE analysis			
0 to 5 sessions attended			
6 or more sessions attended			
Repeated measures**			
Baseline *Baseline for those with 4 month, 8 month or 12 month assessment*			
4 months			
8 months			
12 months			
* If needed, ** If enough intermediate data			

Table 3 Analysis of secondary outcomes at 12 months

	Range of score, best to worst unless noted	Control Mean (SD), n	DVPP Mean (SD), n	Difference in means (95% CI)	p-value*
ABI-29 Abusive Behavior Inventory (Males) Psychological abuse subscale – 17 items Physical abuse subscale 12 items	17-85 12-60				
ABI-R Abusive Behavior Inventory –(Female (ex)partners)**	25-125				
Propensity for abusiveness Affective Inability subscale (12 items)	12-60				
RFQ Reflective Functioning Questionnaire (modified) (scored using Muller method)	6 to 42				
Criminal Justice **Police reported** **Number of incidences** **# of participants with at least 1 incidence (%)** **Mean # of incidences per participant** **Missing (non-consented to police data) n(%)** **Participant reported** **In the last 12 months, how often have the police been called because of violence/abuse toward your partner/ex-partner** **None at all** **Once** **2-5 times** **6-10 times** **More than 10 times**					
PHQ-9 Patient Health Questionnaire Total Score	0 -27				
GAD-7 Generalized Anxiety Disorder	0-21				
PC-PTSD Primary Care Post Traumatic Syndrome Disease	0-5				
AUDIT-C Alcohol Use Disorders Identification-Concise	0-12				
DUDIT Drug Use Disorders Identification Test	0-44				
SF-12 Short Form health survey – 12 items Physical health Mental health	0-100, worst to best 0-100, worst to best				
ICECAP-A ICEpop CAPability measure – Adult	-0.001 to 1, worst to best				

* p-values are not adjusted for multiple outcomes
** xx men had two partners who provided ABI-R, these have been averaged, before including in the mean

Table 4 Group sessions attended, with corresponding missing primary outcomes and ABI-29 values.

Group Sessions Attended*	N (%)	Missing primary outcome N (%)	Baseline mean (sd), n	12 months mean (sd), n	Change from baseline to 12 months mean (sd), n
Control					
None					
DVPP					
None					
1					
2-5					
6-11					
12-17					
18+					
Unknown					

* Individual sessions that were given in exceptional circumstances in place of group sessions and where some of the REPROVIDE manual content was followed, are included in the count of the group sessions. For cases where participants have had group sessions not following the REPROVIDE manual, we will include those sessions in the count of the sessions. A footnote will state the numbers of sessions and participants who were with the same providers for this study, although not following the REPROVIDE manual.

Where participants have had group sessions with the same providers for this study, although not following the REPROVIDE manual, we will include those sessions in the count of the group sessions and the number of participants and sessions involved will be footnoted. Individual sessions that were given in exceptional circumstances in place of group sessions and where some of the REPROVIDE manual content was followed, are included in the count of the group sessions.

The exception to this inclusion is individual sessions with participants during the COVID-19 pandemic (even if they covered REPROVIDE manual content) when the main groups were paused. In addition to group sessions, participants typically had some additional one-to-one sessions to ensure that the programme was tailored to their needs and circumstances.

The number of these additional individual sessions are not easily available; however, the research team will look investigate case files for a representative sample and report the findings on the one-to-one tailored sessions in the mail trial paper. This activity is not covered by this SAP.

In addition, descriptions of the relapse prevention groups and work with the partners will be described in text in the study reports but will not be covered in this SAP.

Table 5 Analysis of ABI-29 at 12 months, comparing pre-defined subgroups

	Control Mean (SD), N	DVPP Mean (SD), N	Subgroup effect Interaction term (95% CI)	Interaction p-value*
Age				
Below median*				
Median or above*				
Area				
1				
2				
3				
4				
5				
Living arrangements				
Living with partner				
Not living with partner				
Referral Method				
Self				
Other				

*Means represented above and below median are for illustrative purposes, p value based on continuous age

Table 6 Effects of assuming different means for the missing and non-missing ABI-29 at 12 months

DIFFERENCE In Means Of Missing And Non-Missing		Control	Worst case XX	-30	-25	-20	-15	-10	-5	0	5	10	15	20	25	30
	Value substituted for missing primary outcome	Estimated Intervention Effect (~p value)								**						
Intervention																
Worst case xx																
-30																
-25																
-20																
-15																
-10																
-5																
0																
5	*															
10																
15																
20																
25																
Best Case xx																

Note: Increments of reporting will be chosen to be illustrative of differences in means needed to change direction and strength of evidence.
~p – Rough estimate of p value of intervention derived from a substitution of mean listed for the missing value in the arms as indicated. Further refining this p value would require additional assumptions about the distribution of the missing data.
* mean of responders in the Intervention group, ** mean of responders in the Control group

7.2 Tables and figures for primary results paper supplementary material

Supplementary Table S1 Area and Relationship Status (Randomisation Factors)

	Control (n=)	DVPP (n=)	Total
Area: n (%) 1 2 3 4 5			
Relationship Status (participant is still living all or most of the time with the abused partner): n(%) Yes No			

*** as reported at Randomisation – note xx participants received group therapy in a different area.**

Supplementary Table S2 Serious Adverse Events affecting participants, partners or ex-partners, and others.

	Male Participants		**Partners/Ex-Partners**		**Others effected**	
	Control	**DVPP**	**Control**	**DVPP**	**Control**	**DVPP**
Total number of events						
Number of participants with 1+ events						
Number of specific events: Deaths Suicide deaths Attempted Suicides *Others to be added*						

Supplementary Table S3 Adverse Events affecting participants, partners or ex-partners, and others.

Adverse Events	**Male Participants**		**Partners/Ex-Partners**		**Others effected**	
	Control	**DVPP**	**Control**	**DVPP**	**Control**	**DVPP**
Total number of events						
Number of participants with 1+ events						

Supplementary Table S4 Reasons for Exclusions from intervention group sessions, with mean, minimum and maximum number sessions attended.

Reason For exclusion*	Number	Mean (Min, Max) number of sessions attended
Aggressive to staff		
Alcohol/drug use		
CJS involvement		
Covid concerns		
Denies abuse		
Failing to disclose new relationship		
Found alternative		
Lack of motivation/attendance		
Learning disability		
Low group numbers		
Mental health probs		
No longer interested		
No valid ID		
Non-engagement		
Out of area		
Time commitment		
Too high risk		
Unhappy with course content		
Work commitments		
Total number of individuals		

*principal reason given at time of exclusion (xx had multiple exclusions)

Supplementary Table S5 Number of participants reporting Additional Treatments at baseline, 4, 8 and 12 months by arm.

Additional Treatments	Control at BL, 4, 8, 12 months	DVPP at BL, 4, 8, 12 months
Private Therapy		
NHS Therapy		
Online programs		

Supplementary Table S6 Table comparing police vs male report of incidences

# of Incidences at 12 months	Participant report of police calls	Missing	None at all	Once	2-5 times	More than 10 times
Police report of incidences						
Not included in police data						
0						
1						
2-5						
6-10						
More than 10 times						

*** The participant is answering question "In the last 12 months, how often have the police been called because of violence/abuse toward your partner/ex-partner".**

Supplementary Table S7 Table comparing police vs male report of ABI -29 at 12 months

Police report of incidences at 12 months	ABI-29 Mean (SD)	ABI-29 Min, Max
Not included in police data		
0		
1		
2-5		
6-10		
More than 10 times		

Supplementary Table S8 Protocol deviations

Protocol Violation	Description

Supplementary TableS9 IMPACT at 12 months (asked "within last 4 months")

IMPACT* (within last 4 months)	Range, Direction	Control Mean (SD), n	DVPP Mean (SD), n
Emotional Behaviour			
Physical Behaviour			
Sexual Behaviour			

Figure S1: Graph showing baseline and 12 month ABI values for each participant, by arm, also indicating missing data.

Figure S2: Scattergrams of male ABI-29 vs female ABI-R, overall and by study arm, at baseline and 12 months. (Comparison of male report to female report of abuse)

Figure S3: Scattergrams of male vs police report of incidences, overall and by study arm, at 12 months. (Comparison of male vs police report of incidence)

Figure S4: graph reflecting the # of sessions attended and ABI-29 and change in ABI-29

7.3 Other tables and figures for end of study report

Additional Figure A1: Histograms of primary outcomes at each timepoint per arm.

Additional Table A1 – Number of male participants with ABI-29 at lowest and highest values

	ABI-29 at lowest value	ABI-29 at highest value	ABI-29 missing
Control group, n (%)			
Baseline			
12 months			
Intervention group, n (%)			
Baseline			
12 months			

Additional Table A2 Additional questions of PHQ9, GAD7 at Baseline and 12 months

Instrument	**Control Baseline**	**DAPP Baseline**	**Control 12 months**	**DAPP 12 months**
PHQ9: Difficulty Question if checked any problems (n, %) No problems checked Not at all Somewhat Very Extremely Missing				
GAD7: Difficulty Question if checked any problems (n, %) No problems checked Not at all Somewhat Very Extremely Missing				

Additional Table A3 Table of items participants skipped on the ABI-29 at 12 months

Items not answered	Question(s) not answered

Additional Table A4 Form completion method

	Control				DVPP			
	Baseline	4 Month	8 Month	12 Month	Baseline	4 Month	8 Month	12 Month
Online								
Paper								
Phone call with researcher								
Mixed / phone call plus online or paper								

Additional Table A5 Questionnaire items reported where an established scoring system cannot be identified.

Survey – Item	Control	DVPP
Other Behaviours – 10 items (list questions)		
Controlling behaviours – 4 items (list questions)		
Childhood Experiences Questionnaire – adapted (list questions)		
IPVRAS – Adapted – 9 items (list questions)		