Ministry of Justice

Evaluating the Building Better Relationships (BBR) programme

Feasibility study for an impact evaluation of proven reoffending

Analytical Priority Projects

Data and Analysis

Ministry of Justice Analytical Series

2023

Data and Analysis exists to improve policy making, decision taking and practice by the Ministry of Justice. It does this by providing robust, timely and relevant data and advice drawn from research and analysis undertaken by the department's analysts and by the wider research community.

Disclaimer

The views expressed are those of the authors and are not necessarily shared by the Ministry of Justice (nor do they represent Government policy).

First published 2023

Any enquiries regarding this publication should be sent to us at researchsupport@justice.gov.uk

This publication is available for download at http://www.justice.gov.uk/publications/research-and-analysis/moj

ISBN 978-1-84099-992-1

Acknowledgements

We would like to thank Olivia Hambly and Wura Gerasimov in Data and Analysis for their investigations of the operations of BBR. We would like to thank the various colleagues who reviewed and commented on this report, along with two anonymous academic peer reviewers, for their constructive and valuable feedback.

Authors

James Teasdale

Annie Sorbie

Justice Data Lab, Data and Analysis

Ian Elliott

Analytical Priority Projects, Data and Analysis

Contents

List of tables

List of figures

1. Executive Summary

Building Better Relationships (BBR) is a His Majesty's Prison and Probation (HMPPS) moderate-intensity cognitive-behavioural programme for adult men convicted of an Intimate Partner Violence (IPV) offence. The programme was accredited for use by CSAAP (Correctional Service Accreditation and Advisory Panel) in 2013. Domestic abuse (DA) covers a wide range of crimes, including domestic violence, stalking, coercive control, and IPV. BBR focuses on men convicted of IPV offences. Previous international meta-analyses that combine the results of prior individual studies have indicated promising outcomes for DA programmes on both general and violent reoffending.

Aims and objectives

Outcome/impact evaluations typically require significant resource investment and, thus, a feasibility study can help to ensure that such resources are utilised at the most suitable time and in the most efficient way possible. The aim of this report is to establish whether it is possible to undertake a robust evaluation to measure programme effects so that findings can inform operational practice and policy before resources are allocated to a full-scale evaluation. This investigation into the feasibility of an evaluation was led by the Ministry of Justice (MoJ) Justice Data Lab.

A 2018 HMI Probation report on the work undertaken by Community Rehabilitation Companies (CRCs) with perpetrators and victims of domestic abuse highlighted the need to focus on the quality of that work. It was recommended that the MoJ "evaluate and legitimise domestic abuse interventions and provide assurance that interventions are evidence-based" (p. 11). HMPPS committed to assessing appropriate methodological approaches to evaluating the impact of BBR on reoffending and to examining whether an evaluation would deliver accurate estimates of impact that can inform future development.

Methods

This feasibility study was conducted in two phases. Phase 1 involved identifying the appropriate methodology and design considerations that would need to be agreed to conduct a statistically rigorous impact evaluation for BBR. Since BBR is already

operational, this feasibility analysis was focused on retrospective evaluation methods (not prospective methods such as a randomised controlled trial). Previous successful evaluations of accredited programmes have used a propensity score matching (PSM) methodology (e.g., 2017 SOTP and 2021 RESOLVE evaluations) and this evaluation approach was prioritised as a technique to measure the impact of BBR. PSM uses statistical techniques to construct an artificial control group by matching each individual who accessed BBR with an equivalent non-programme comparator of similar characteristics. An estimate of the impact of an intervention is then produced by comparing outcomes for the intervention and comparison group. Phase 2 examined operational data to test the proposed methodology and define the necessary parameters regarding design considerations. This was achieved by simulating the development of a BBR intervention group, as well as simulating the process of matching that group to a comparison cohort.

Conclusions and recommendations

This report delivers a set of analytical recommendations for whether a reoffending impact evaluation of BBR is feasible. It is important to note that such evaluations are complex and pose several challenges, including identifying Intimate Partner Violence (IPV) offences (as there is no comprehensive DA offence), selecting follow up periods (since new domestic violence reconvictions are expected to take longer to occur as relationships would need to be re/established), dealing with data quality issues associated with using community offending behaviour programme data, and challenges delivering BBR (and some other accredited programmes) in the community during the time period from which this data was drawn.

The results of Phases 1 and 2, together with the sample size analyses, provide evidence that an impact evaluation is feasible, with sufficient sample sizes to be able to detect an intervention effect (presuming one is present) for BBR. However, several methodological concerns were also identified that make a reoffending impact evaluation substantially more challenging. These include not being able to account for the quality of programme delivery in the evaluation, which has been shown to be important to reduce reoffending. Given the time and resources required to deliver evaluations, an impact study would provide limited operational insight until the methodological issues could be resolved.

2. Context

2.1 Building Better Relationships

Building Better Relationships (BBR) is a His Majesty's Prison and Probation (HMPPS) moderate-intensity cognitive-behavioural programme for adult men convicted of an Intimate Partner Violence (IPV) offence. The programme was accredited for use by CSAAP (Correctional Service Accreditation and Advisory Panel)[1] in 2013. BBR also exists as part of a wider, multi-agency response to IPV and within that network supports work by other organisations who are providing services that also address IPV.

BBR is targeted at adult male perpetrators in custody and the community who are assessed as (a) having a history of partner abuse that has resulted in a current IPV-related conviction, and (b) being of moderate risk of re-offending or above (as assessed by the Spousal Assault Risk Assessment; SARA; Kropp et al., 1995). Intervention needs targeted by BBR can be classified into three domains: (1) pro-offending thinking styles; (2) emotional awareness and management; and (3) relationship problems. Overall, BBR aims to encourage:

- Better understandings of why the participant used violence and aggression in their intimate relationships; attitudes and beliefs that underpin that behaviour; and factors reinforcing and maintaining use of that behaviour in that context.
- Identification and use of existing strengths and skills to change their behaviour.
- Motivation to engage with the programme and effect change by providing a safe, respectful, stimulating, and challenging environment in which they can learn.
- The development of practical and sustainable strategies for maintaining change once they have departed from the programme.
- Better lives for all affected by the participant's violent and aggressive behaviour, including the men themselves.

[1] The Correctional Services Advice & Accreditation Panel (CSAAP) consists of academics and experts who assist HMPPS to accredit programmes by reviewing programme design, quality assurance procedures and findings, and programme evaluations.

Domestic abuse (DA) covers a wide range of crimes, including domestic violence, stalking, coercive control, and IPV. BBR focuses on men convicted of IPV offences. The term intimate partner abuse (IPA) may be more appropriate for the types of offences that are targeted by BBR, which do not just focus on violence. However, we use IPV throughout this paper to remain consistent with BBR manuals, which defines IPV as the use of violent and aggressive acts towards romantic partners (e.g., wives, girlfriends, civil partners, husbands, boyfriends).

BBR was preceded by three programmes: The Heathy Relationship Programme (HRP), delivered in custody, and the Integrated Domestic Abuse Programme (IDAP) and the Community Domestic Violence Programme (CDVP), both delivered in the community. A large-scale study found that the predecessors to BBR, the IDAP and the CDVP demonstrated statistically significant reductions in rates of reconvictions (Bloomfield & Dixon, 2015). However, a review of these programmes had identified areas in which their content could be improved in line with developments in the literature on the causes of domestic abuse. That review concluded that there was a need to develop one new programme to be delivered across both Prisons and Probation to replace IDAP, CDVP, and HRP.

Previous meta-analyses, which combine results from prior individual studies, have also indicated promising results for DA programmes across the world for both general and violent reoffending (Arce et al., 2020; Babcock et al., 2004; Cheng et al., 2019; Gannon et al., 2019[2]). However, these meta-analyses include a diverse range of programmes and the size of the impact is typically smaller in studies using more rigorous evaluation methods.

2.2 Aims and objectives

Outcome/impact evaluations typically require large financial, time, and human resource investment and, thus, a feasibility study can help to ensure that such resources are utilised at the most suitable time and in the most efficient way possible. The key is to establish that

[2] It was not possible given time and resource constraints, nor proportional to the task of producing estimates for power calculations, to conduct a full systematic review of studies and meta-analyses related to DA programmes. We acknowledge that other meta-analyses of the topic are likely to exist in the literature.

the intervention has a measurable effect, and that estimates of impact will benefit practice and policy before resources are allocated to a full-scale evaluation. Feasibility work has been referred to as "[the] extent to which an activity or project can be evaluated in a reliable and credible fashion" (OECD-DAC, 2010) and a "low-cost pre-evaluation activity to prepare better for conventional evaluations of programmes" (Leviton et al., 2010).

A 2018 HMI Probation report on the work undertaken by Community Rehabilitation Companies (CRCs) with perpetrators and victims of domestic abuse highlighted the need to focus on the quality of that work. It was recommended that the MoJ "evaluate and legitimise domestic abuse interventions and provide assurance that interventions are evidence-based" (p. 11). In response, the MoJ committed to assessing appropriate methodological approaches to evaluating the impact of BBR on reoffending and whether an evaluation would deliver accurate estimates of impact that can inform future development. This study was commissioned by the HMPPS Accredited Programmes and Interventions Delivery Strategy Board (APIDSB), who govern the nature and sequencing of accredited programme evaluations, for the assessment of BBR in the community context. BBR delivered in custody was not considered, as numbers of participants in prisons were too low.[3]

2.3 Feasibility methodology

This feasibility study was conducted in two phases. Phase 1 involved identifying the appropriate methodology and design considerations that would need to be defined to conduct a statistically rigorous retrospective impact evaluation for BBR. These included:

- What data are available for impact evaluation and from where?
- Over what time-period could we evaluate the programme?
- What outcome(s) would be considered evidence of reoffending?
- What moderator or sub-group analyses should and could be conducted?
- At what point in the process should we start counting reoffences?
- Who will be selected to form the intervention group?
- How do we select appropriate comparisons for the intervention group?

[3] Data collected from the HMPPS Annual Digest 2018-19 indicated that only 779 participants started BBR in custody between March 2010 to March 2019 (see https://www.gov.uk/government/statistics/hmpps-annual-digest-2018-to-2019).

- What sample sizes might be required to ensure statistical rigour?

Since BBR is already operational, this feasibility analysis was focused on retrospective evaluation methods (not prospective methods such as a randomised controlled trial). Alternative rigorous retrospective methods, in particular regression discontinuity design,[4] were examined for the 2021 RESOLVE evaluation and rejected on methodological grounds that would also apply to BBR. As the MoJ's Sex Offender Treatment Programme - SOTP (Mews et al., 2017) and RESOLVE (Robinson et al., 2021; Teasdale, 2021) evaluations used a propensity score matching (PSM) statistical technique,[5] PSM was the focus of this investigation. PSM is a statistical matching technique that uses factors theoretically and empirically associated with both receiving the intervention and the outcome variable (i.e., reoffending) to predict a "propensity score",[6] representing the likelihood of participating in BBR. This propensity score is then used to match treated individuals to non-treated individuals similar to them. Phase 2 involved the collection of operational data to test the proposed methodology and design considerations by simulating the development of a BBR group and simulating the process of matching that group to a comparison cohort via PSM.

This report delivers a set of analytical recommendations for whether a reoffending impact evaluation of BBR is feasible. Such evaluations are complex and pose challenges in identifying IPV offences (as there is no comprehensive DA offence), selecting follow up periods (since new domestic violence reconvictions are expected to take longer to occur, since relationships need to be re/established), dealing with data quality issues associated with using community offending behaviour programme data, and ongoing challenges delivering BBR (and other accredited programmes) in the community.

---

4 Regression discontinuity designs (RDD) compare individuals "just above" and "just below" a single threshold for eligibility (for example, a risk assessment score). It was rejected because eligibility for BBR is based on a combination of multiple variables (OASys classifications, etc.) and because results might not generalise to participants outside of RDD's narrow focus on those located close to the chosen threshold.

5 The MoJ Justice Data Lab provides a free service for organisations working with offenders and provides an opportunity for them to assess the impact of their intervention on proven reoffending. It also helps develop a collaborative understanding of effective rehabilitation. Further details can be found here: https://www.gov.uk/government/publications/justice-data-lab

6 Please refer to Justice Data Lab's methodology paper for more information on PSM: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/392929/justice-data-lab-methodology.pdf

3. Phase 1: Design considerations

3.1 Programme integrity/Quality of delivery

Findings from the 2021 MoJ evaluation of the RESOLVE programme, as well as the international evidence base indicate that quality of programme delivery can be an important factor for interventions aimed at reducing proven reoffending (e.g., Dowden & Andrews, 2004; Gannon et al., 2019; Papalia et al., 2019). Analyses addressing programme delivery would aim to evaluate the extent to which the quality of delivery may have an impact on the outcome of BBR. BBR in the community was, for the evaluation date period, delivered by 21 Community Rehabilitation Companies[7] (CRCs), with HMPPS quality assurance of delivery completed over two timeframes: 2014-2016 and 2016-2019.

HMPPS accredited programme quality assurance for accredited programmes is completed through the Interventions Integrity Framework (IIF). This explores evidence of practice in relation to whether the programme delivered met the guidelines set out in the programme and management manual. The quality assurance sample sizes were higher in the 2016-2019 period than in 2014-2016 and therefore, our analysis focussed on this period. The 21 CRCs were assessed based on four Key Lines of Enquiry (KLOEs: see Appendix A) which are rated from 1 (programme integrity is compromised) to 4 (programme integrity is maintained effectively). All four KLOEs are important for effective programme delivery. An acceptable score is for CRCs to have a total score across all four KLOEs of at least 12 (out of a possible score of 16) and a score of at least three for each of four KLOEs.

To maintain consistency with previous and currently active accredited programme impact studies, the criteria for classifying quality of delivery for the overall outcome measure (modelling as expected vs. not as expected delivery) was that used in the RESOLVE evaluation. This approach classified prisons as “broadly meeting delivery standards” when

[7] In June 2014 the MoJ introduced its Transforming Rehabilitation reforms. It dissolved 35 self-governing probation trusts and created a public sector National Probation Service and 21 Community Rehabilitation Companies (CRCs). CRCs supervise offenders who present a low or medium risk of harm, while the National Probation Service manages offenders who present higher risks. In February 2015, the CRCs were transferred to eight, mainly private sector, suppliers working under contracts managed by HMPPS. In June 2021, this operational model was replaced by the newly formed Probation Service.

(1) their overall quality of delivery score was 13 or greater, and (2) no scores on any of the four individual KLOE metrics were less than 3. The key observations for the 21 CRCs during the 2016-19 assessment period for BBR were as follows:

- Only one CRC (Hampshire & Isle of Wight, total score 12) scored at least 3 for all four KLOEs, including 280 participants (3% of the pre-matched intervention group for this assessment period).
- The total score for the four KLOEs ranged from 5 to 12 (out of a maximum possible score of 16), with a mean of 9.2 and median of 9.
- Three CRCs scored 12, including 1,475 participants (15% of the pre-matched intervention group for this period), all of which scored at least 3 for KLOE 1 and no more than one score of 2 for KLOEs 2-4.
- Eight CRCs (including approximately half of the 9,874 pre-matched intervention group records for this period) scored 8 or less.

Given that only one out of 21 CRCs scored at least 3 or more in all four KLOEs, the overall performance of CRCs delivering BBR provides a substantial challenge in identifying a cohort of participants from sites where programme integrity is broadly maintained. Therefore, MoJ Data and Analysis do not recommend performing this analysis, owing to issues relating to CRC performance and our ability to perform analyses using the HMPPS quality of delivery data. To fully understand the relationship between quality of delivery and outcome, more differentiation between the scores would be required, particularly at the higher end. An alternative approach could consider relative quality between CRCs, rather than focusing on better performing CRCs. However, given that 20 out of the 21 CRCs include at least one KLOE rated with a score of 2 or less, the issue of CRCs not reaching the required standard remains.

3.2 Evaluation data

Various data sources were considered for this feasibility study evaluation. These included administrative data extracts from NDelius and Interventions Manager (IM). IM can provide more detail in some areas – such as providing access to recorded sessions and to written materials – but does not provide national coverage (it was not mandated for use by all community providers) and some variables were found not to be consistently recorded or complete. While NDelius has similar issues across some variables, it is a rich data source,

providing a complete view of probation proposals, as well as programmes given, started, and completed by an offender, and has broader coverage (i.e., NDelius data are available from all community providers). Consequently, this report focuses on NDelius data.

Table 1: NDelius data extracts considered for BBR evaluation feasibility study

Dataset	Brief Description	Coverage period	Extract date	Number of records
Programme starts	All offenders recorded as having started the BBR programme in the community	Programme starts over financial years 2013/14 to 2019/20	June 2021	33,063
Programme referrals	All offenders recorded as having been referred to BBR in the community, regardless of whether they started on the programme	Programme referrals over financial years 2015/16 to 2018/19	February 2020	32,935 (approximate split of 22k programme starts, 11k with no programme start date)

Based on advice from relevant Data and Analysis teams and analysis conducted on the data, the key variables necessary for an impact evaluation of BBR were considered of sufficient quality and fit for purpose. Data validation established that, to a reasonable degree, the final cohort dataset included adult males, data from all providers, and accurate data on index offences, key dates, completion status, and suitability assessment metrics.

3.3 Types of reoffending to measure

There are challenges in defining a proven reoffending outcome measure that captures the type of behaviour/offences that the BBR programme is aimed at reducing. These include:

- Determining exactly who the programme is for, and agreeing on definitions and terminology, and interpretations of the programme manuals.
- Identifying offences where IPV is involved (there is not a single offence code or offence category that fully captures DA-related offences), given data availability.
- Accepting that not all IPV offences are reported or convicted.

Various approaches to defining an IPV reoffending measure were explored. Further work will be needed to determine the most appropriate IPV measure(s) for impact evaluation, with approval from key stakeholders. Six potential outcome measures[8] were considered.

- **General (all offences):** All proven reoffences. The impact of the intervention on general reoffending is an important consideration for all offending behaviour programmes, irrespective of the type of offenders that the programme is aimed at.
- **OASys Violence Predictor (OVP):** Reoffences with a Home Office offence code that are used for the OASys Violence Predictor.[9] It includes a broad range of index offences, including 94% of the offences for which participants were referred to BBR.
- **Top 10 index offences:** Those reoffences with a Home Office offence code that feature in the 10 most common index offences for which offenders were referred to the BBR programme. These account for 82% of all index offences.
- **Violence to partner offences based on OASys:** Reoffences where a subsequent OASys assessment identified the offence as involving violence to partner (question 2.3D[10] in the OASys assessment). We believe this to be recording physical violence (not a comprehensive assessment of DA) but explored as an additional proxy.
- **Domestic abuse offences based on the CPS "DA flag":** Reoffences that have been received by the CPS and flagged as a domestic abuse (DA) case. Prior work by Data and Analysis to support the Domestic Abuse Bill demonstrated it is feasible to use this data for evaluation purposes. However, considerable new resource (from both MoJ and CPS) would be required to provide/process these data.
- **Violence to partner offences based on Spousal Assault Risk Assessment (SARA):** While SARA[11] assessments are useful in providing detail on risk and

8 The measures considered are not mutually exclusive. All proven reoffences will be included in the general reoffending measure, with some reoffences counted in all measures.

9 The OASys Violence Predictor (OVP) is an actuarial violence risk measure, first implemented within the Ministry of Justice in 2009, which predicts the likelihood of (non-sexual) violent reoffending over a two-year period.

10 Question 2.3 states, 'Did any of the following occur?' Option D states, 'Physical violence towards partner'.

11 The SARA is a clinical checklist of risk factors for spousal assault, including an imminent risk of violence towards partner rating.

need, when considering eligibility criteria for BBR, they cannot corroborate whether IPV was present in a specific reoffence.[12] Using these data in an IPV outcome measure is not recommended.

Being able to identify offences where IPV is involved is difficult, given that there are few offences directly linked to IPV. By far the most common index offence for BBR referrals is common assault and battery. However, this offence is not exclusively associated with IPV, since the victim's relation to the offender can include partners, family members, friends, and strangers. Additional data are required to corroborate whether an offence involved IPV.

For Phase 2, general reoffending and OVP were explored. While further work is required, our findings suggested that proxy measures exist that can help to identify offences related to IPV that are frequently used for impact evaluations in the U.K. If an impact evaluation is commissioned, an OASys Violence to Partner indicator will be explored as a potential proxy measure. If different measures were adopted for an impact evaluation, the matching quality and sample sizes will differ, but using these initial two measures to assess feasibility should be sufficient to make a recommendation.

3.4 Starting point for measuring reoffending

A further factor in defining the outcome measure is the point in time from which to measure reoffending, and for how long to measure it. For programmes delivered in custody, the standard approach for Data and Analysis has been to measure reoffending from the point of release from prison (although lags between participation on BBR and release from custody are considered). For evaluation of BBR, we also considered that:

- Most participants can reoffend from the point they are referred to the programme.
- The BBR programme often has long waiting lists (29% of programme starts have a start date over 6 months after their referral date).
- The BBR programme is generally delivered over a long time-period (75% of BBR completers spent over 6 months on the programme).

[12] Based on our understanding, neither the presence of a SARA nor any of the questions in the assessment in relation to the index offence (the most recent offence) can identify whether IPV was present in a specific reoffence

Three time points were explored as the starting point to measure reoffending:

1. **Referral date:** The date is available for both the intervention and comparison group. However, where there is a long period between referral and programme start, a significant portion of the follow-up period may not measure an intervention effect.
2. **Start date:** It avoids measuring any non-intervention effects during the waiting period and captures those reoffences that trigger terminations. However, no programme start date is available for the comparison group so a pseudo date would need to be calculated. As the follow-up period begins with the intervention phase, effects found during this initial period will only include a partial intervention dose.
3. **End/termination date:** A "full dose" of intervention (at least up to the point of termination for non-completers) is measured. However, no programme end date is available for the comparison group so a pseudo date would need to be calculated. If the reason for termination is a reoffence then this will have occurred prior to termination, so will not be captured.

For Phase 2, the follow-up period was assumed to start from the programme start date. This date was considered most appropriate and enabled us to adequately determine our feasibility recommendations. Though there are methodological issues with each time point, when weighing up the advantages and disadvantages, using programme start date appears to be the most appropriate and allows for defensible feasibility recommendations. To apply this approach, a "pseudo" programme start date was calculated for the comparison group.

Many offenders with BBR referrals who did not start the programme may have had their community order revoked, their suspended sentence activated, or were recalled to prison, having been on licence, for breaches/non-compliance (including reoffences). This was likely to disproportionately affect the waiting-list group, since few – if any – of the programme starters would have committed such a breach while waiting to start the programme. To be consistent, records were excluded for both the intervention and comparison groups where a reoffence or termination date was recorded in the period

between referral date and (pseudo) programme start date. This resulted in 12% of the original intervention group and 36% of the original comparison group being excluded.

3.5 Period over which reoffending is measured

The follow-up period over which reoffences are measured was considered with reference to the following objectives:

- It should have regard to the volume and distribution of IPV reoffences over time (e.g., if there are proportionately fewer in the first two years of the follow-up period, but many more in later years, then a longer follow-up period would be appropriate).
- It should be long enough to capture a sufficiently large volume of IPV reoffences.
- It should be short enough to ensure that:
  - the findings relate to programme delivery within a relatively recent period.
  - any intervention effect is still observable (as effects may wane over time).
  - a sufficiently large sample size is retained to be able to detect an intervention effect (this is a secondary objective – the longer the follow-up period, the fewer offenders we can track over the whole period).

Analysis of reoffences in the three-year periods following referral date, programme start date, and programme end date was carried out on BBR participants using four different definitions of reoffending: (1) any reoffence, (2) a violence-to-partner reoffence, (3) a reoffence defined as violent for the purpose of the OVP assessment, and (4) the top ten index offences. While overall reoffending rates differ by definition, the reduction in the proportion with no reoffences slows over time. Equivalently, the proportion of participants with an initial reoffence in the first year of the follow-up period is higher than the proportion of participants with an initial reoffence in the second year, and so on.

Except for violence-to-partner offences (where the reoffending rate is lower and the sample size reduces to factor in an additional period to capture OASys assessments), for Phase 2, a two-year follow-up period best meets the objectives set out above, and this period has therefore been assumed in making the feasibility recommendations. Further consultation on approach will be necessary if an impact evaluation is commissioned.

3.6 Creation of intervention and comparison groups

In deciding which of the two datasets to use ("starts" or "referrals"), the main consideration was how to construct the comparison group. The aim is to create a comparison group that is as similar to the intervention group as possible, with the only difference being that the intervention group participated in the programme, while the comparison group did not.

Data and Analysis' standard approach starts with a pre-matched comparison group comprising many offenders with characteristics broadly spanning those of the intervention group, and then uses propensity score matching (PSM) to produce post-matched groups that are as well matched as possible on all the key variables. Programme starts would be used to create the pre-matched intervention group (given that it spans a longer period and includes more starter records), with the pre-matched comparison group generated using the Police National Computer (PNC) and linking various other datasets to it.

Part of this approach aims to select a comparison group that is as similarly suited to the BBR programme as those who officially participated in the programme. However, programme referrals already include a sizeable cohort of offenders who were referred to BBR but did not start the programme and, by being referred, this cohort has been (at least partially) assessed to be suitable for the programme. So, as an alternative approach, by starting with this smaller but more suited cohort of "non-starters" for the pre-matched comparison group, the PSM process and other filters could then be used to control for any other key differences between "starters" and "non-starters". This can therefore be used to select both the pre-matched intervention and comparison group records, before linking with other datasets to pull in all the required variables.

We identified two comparison group approaches that could be used to run an impact evaluation for BBR.

1. Use a much larger and broader cohort of offenders and let the PSM process create matched groups. A BBR referral is not required to be included.
2. Filter the pre-matched comparison group to only include those offenders who were referred onto BBR but did not start the programme.

So long as the PSM process yields sufficiently good matching and high enough sample sizes, our recommendation would be to use Option 2 (i.e., referred onto BBR but did not start the programme). The main reason for this is that there is greater confidence that the post-matched comparison group will include only those offenders suitable for the programme, given that this option only includes programme referrals. Nevertheless, it would be important to ensure that any individuals in the referral group who were suitable, but who had refused to take part, were assigned to the appropriate group.[13]

The above recommendation is conditional on sufficiently good matching and sample sizes. To investigate whether Option 2 is viable, Phase 2 has been carried out using this approach, mirroring the key steps in an impact evaluation up to the point where the post-matched intervention and comparison group weightings, as well as the key metrics of sample sizes and overall matching quality, are calculated. The purpose was to determine whether sufficient matching quality and sample sizes are likely to be achieved using this approach as the basis for creating the intervention and comparison groups.

Using a programme's "referrals" or "waiting list" as a comparison group is a recognised approach for impact studies. It has been used in evaluations of U.K. public sector interventions, including a HMPPS evaluation of the BBR predecessor programme. Nevertheless, should an impact evaluation be agreed, this approach would require further consideration and consultation with operational, subject matter, and analytical experts to ensure the analysis controls for the main factors that determine whether offenders referred to BBR start the programme. If this was not deemed possible, Option 1 would be used instead.

[13] In a randomised controlled trial (RCT), the intervention and control groups are defined by the "intention to treat" (i.e., the allocation to the intervention or comparison group), not whether they were eventually exposed to the intervention. Technically, there is an intention to treat individuals on a waiting list. However, the "waiting list study design" presumes those on a waiting list are effectively the same as a comparison group created by a process of randomisation to intervention and comparison groups. In an RCT, an individual refusing to participate would need to have been assigned to the intervention group to refuse it. Therefore, in a quasi-experimental study, they should also be considered part of the intervention group.

3.7 Sample sizes required

A key component in scoping the feasibility of an evaluation is to establish the sample size required to achieve statistical power. Statistical power is the probability of detecting an effect with a statistical test (presuming there is an effect to be found). The standard for acceptable power in relevant academic fields is a probability of 80% or more of detecting a true effect. A power analysis can be used to ensure that the study characteristics are sufficient to generate the desired amount of power. Estimates of these characteristics need not be perfect but should be plausible and defensible. In this instance, we sought an estimate of the minimum sample size required to achieve at least 80% statistical power.

To establish a plausible expected minimum effect size, four recent meta-analyses of interventions for domestic violence were reviewed (Arce et al., 2020; Babcock et al., 2004; Cheng et al., 2019; Gannon et al., 2019). A meta-analysis was used to calculate aggregate effect sizes for general, violent, and IPV reconvictions across all studies included in those meta-analyses that used PSM or better as the methodology.

A plausible effect size of 1.43 (odds ratio) was estimated for general reoffending. Analysis of comparison group reoffending rates predicted a 2-year baseline reoffending rate of 36.7% for general reoffending and 17.9% for violent reoffending. Too few studies were available to estimate specific effect sizes for violence or domestic violence. Twelve power analyses were conducted. This showed that, to detect an effect of 1.43 [95% CI: 1.13, 1.83] on general reoffending, a minimum intervention group of 509 to 851 participants would be required. For an effect of the same magnitude with a lower baseline (e.g., for violent reoffending), an intervention group of 751 to 1,275 participants would be required. Thus, an evaluation of BBR is likely to detect a true difference in general and violent reoffending rates.

4. Phase 2: Simulating an impact study

In Phase 2 of this report, we outline the steps carried out to complete the simplified analysis on the overall BBR cohort.

4.1 Selecting variables for the feasibility analysis

A smaller number of variables were selected for this simplified analysis than would typically be used for an impact evaluation, as the purpose is only to assess feasibility. This does not include outcome measure variables, which are considered later in this section. Variables that were considered included:

- **Programme data and offender demographics:** Mainly NDelius, but also some demographic data from the PNC (approximately 10 variables).
- **Criminal history and index offence:** PNC and Reoffending database data (approximately 30 standard variables used in Data and Analysis evaluations).
- **Risk and need:** OASys and SARA data (approximately 20 variables, taken from programme manuals, etc.)

4.2 Data validation and manipulation

The programme referrals dataset was the starting point for constructing the pre-matched intervention and comparison groups (Table 2). The steps involved in developing it were:

1. Carry out validation and sense checks on the data.
2. Link data to other datasets to bring in the required variables.
3. Remove duplicate or suspect records, or cases with multiple “starter” and/or “non-starter” records.
4. Calculate pseudo programme start dates for the comparison group, to use as the starting point for measuring the reoffending outcomes.
5. Create reoffending outcome metrics: here, two-year binary reoffending indicators based on the general and OVP outcome measures.
6. Remove records with a termination or reoffence prior to the real or “pseudo” programme start, to ensure intervention and comparison groups are balanced.

Table 2: Production of intervention and comparison groups from NDelius extracts[14]

BBR referrals with programme start date (for intervention group)		**BBR referrals without programme start date (for comparison group)**	
Criterion applied	**Sample size**	**Criterion applied**	**Sample size**
21,504 records were submitted for analysis	21,504	11,431 records were submitted for analysis	11,431
688 records (3%) were excluded because they were duplicate records or there were inconsistencies identified when comparing records with a separate data extract including programme starters only	20,816	422 records (4%) were excluded because they were duplicate records or there were inconsistencies identified when comparing records with a separate data extract including programme starters only	11,009
1,149 records (5%) were excluded because they did not have a matching record in the offending data for the sentence in which the referral was made	19,667	800 records (7%) were excluded because they did not have a matching record in the offending data for the sentence in which the referral was made	10,209
1,074 records (5%) were excluded because a record was identified for the same person in both the ‘starter’ and ‘non-starter’ cohorts	18,593	1,162 records (10%) were excluded because a record was identified for the same person in both the ‘starter’ and ‘non-starter’ cohorts	9,047
2,669 records (12%) were excluded because of a reoffence or a referral termination in the period between referral date and programme start date	15,924	4,139 records (36%) were excluded because of a reoffence or a referral termination in the period between referral date and pseudo programme start date	4,908
1,314 records (6%) were excluded where start date was after April 2019 or gender = F (2 cases)	14,610	374 records (3%) were excluded if their pseudo start date was after April 2019	4,534
Overall pre-matched intervention group: 68% of records submitted for analysis included in final cohort	14,610	Overall pre-matched comparison group: 40% of records submitted for analysis included in final cohort	4,534

[14] Taken from the programme referrals dataset, comprising 32,935 records (see Table 1).

4.3 Descriptive statistics on the intervention and comparison groups

As part of further validation work, and before running the simplified analysis, a set of descriptive statistics were produced to determine:

- The key characteristics of BBR referrals and participants.
- Any key differences between the original cohort of 'starters' in the referrals data and the final cohort of 'starters' used for the simplified analysis.
- How well-suited are programme referrals (both starters and non-starters) to the programme, based on the eligibility and suitability requirements set out in the programme manual.
- Any key differences between the pre-matched starters and non-starters.
- The most common offences associated with programme referrals and participants.
- Whether data quality issues are present to such an extent that the data are not fit for purpose.

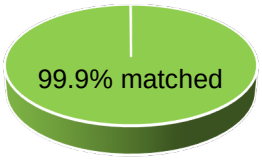
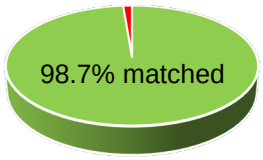
4.4 Running the logistic regression and PSM models

Following the data manipulation, as described above, the final cohorts for the overall analysis comprised 14,610 programme starters (the pre-matched intervention group) and 4,534 non-starters (the pre-matched comparison group). The standard Data and Analysis approach of building logistic regression models (one with intervention and one with reoffending as the binary response variables), simplifying the models, and then running the PSM process was completed on the final datasets for both the two-year general and OVP reoffending measures. The typical final step of running statistical tests to determine any intervention effect was not carried out.

The matching quality was excellent for all variables (see Table 3). The aggregate standardised mean difference in values between the intervention and comparison groups across all variables was less than 0.001 (or less than 1%). The proportion of the pre-matched intervention group matched was also very high for both outcome measures considered, resulting in large post-matched sample sizes for the overall analyses.

Although the two groups are not the same size in absolute terms (i.e., the intervention group is larger than the comparison group), the matching quality indicates that the two groups are balanced (i.e., the two groups are, on average, comparable on all known variables). This introduces the possibility that comparison units are matched to more than one treated participant. In that situation, however, the outcome for that comparison unit is weighted to ensure that any reconvictions they receive are not counted multiple times in the final analysis.

Table 3: Matching quality for simplified analysis (2-year reoffending follow-up period)

Analysis (cohort and outcome measure)	Post-matched sample size (number of BBR participants)	Proportion of pre-matched intervention group included	Overall Matching quality	Matching quality by variables
Overall: **2-year general reoffending**	 14,592 records ✓		excellent (0.64%) ✓	 0% poorly matched 0% reasonably matched 100% closely matched
Overall: **2-year OVP reoffending**	 14,423 records ✓		excellent (0.63%) ✓	 0% poorly matched 0% reasonably matched 100% closely matched

Note 1: *Overall matching quality* is based on the mean absolute standardised differences for all variables in the final logistic regression model, for each analysis. For each variable on which groups were matched, a standardised absolute mean difference was calculated representing the distance between two group means (Austin, 2009). This process of standardisation allowed us to compare the quality of matching for variables that have different scales (e.g., age measured in years vs. the proportion of individuals treated by CRCs or the National Probation Service, measured as a percentage). Standardising presents distances as a value between 0 (most perfect matching) and 1 (least perfect matching). This can be expressed as a percentage for ease of interpretation (i.e., the smaller the percentage, the more similar the groups are on that variable).

Note 2: *Matching quality by variables* is categorised using traffic light criteria based on the absolute standardised difference in the means of the matched intervention and comparison groups for each variable (where **green** = less than or equal to 5%, **amber** = between 6% and 10%, and **red** = greater than 10%). No agreed threshold exists for "substantial" imbalance, but less than 0.1 (or 10.0%) has been proposed to indicate a "negligible" difference between groups (Normand et al., 2001).

5. Conclusions and recommendations

The results of Phases 1 and 2, together with the initial power analyses, provide evidence that an impact evaluation is feasible, with sufficient sample sizes to be able to detect an intervention effect (presuming one is present) for BBR delivered in the community. This evaluation would provide an estimate of whether BBR, as delivered, influences rates of reoffending.

However, serious methodological concerns were also identified:

- Considering expert advice from HMPPS, we do not recommend performing a quality of delivery/programme integrity analysis owing to issues relating to BBR delivery at Community Rehabilitation Companies (CRCs) (where only 1 out of 21 CRCs met the required BBR delivery standards). Additionally, inconsistencies in the way in which CRCs managed data mean it is not possible to meaningfully analyse the effect of programme integrity on the impact of BBR on reconviction rates.
- The cohorts used would include data from approximately 2013 to 2019, and therefore includes individuals who received BBR a long time prior to this feasibility study (or any future impact evaluation). It is possible that BBR itself, as well as the quality of its delivery, may have changed over that time-period, which would need to be investigated prior to an impact evaluation.
- A proxy IPV outcome measure is still to be agreed. Being able to identify offences involving IPV is problematic, given that there are very few offences directly linked to IPV. As noted in the report, the most common index offence for BBR is common assault and battery, which is not exclusively associated with IPV (e.g., the victim may be a stranger). Additional data are required to corroborate whether an offence involved IPV.

The cohorts used to determine the feasibility recommendations in this report are partially affected by the COVID-19 pandemic (see Appendix B). Although the proportion of the overall cohort of BBR starters with a termination date recorded after the first lockdown restrictions were imposed is very low, a portion of the reoffending follow-up period took

place during the pandemic for a subset of records. This means our ability to detect reoffending could be affected by changes in recorded crime during that time, limitations on court activity and outcomes, and needing to wait for offences to appear on official records and be counted in an evaluation. Some of the consequences of the pandemic may be mitigated by matching between the BBR and comparison group, so affect the intervention and comparison groups equally. However, they remain a source of uncertainty and have the potential to introduce omitted variable bias (i.e., could have effects that we are unaware of and are therefore unable to control for in our analyses.)

A feasibility study was conducted to assess whether a robust and operationally useful reoffending impact evaluation for BBR delivered in the community could be undertaken. The findings highlight that, whilst it is technically possible to deliver an impact evaluation, several methodological issues exist that could substantially limit the operational value of the information it would provide compared to the resources it would require.

The most critical example of these is that it is not possible to consider quality of delivery in the analysis, which has been shown to be important to reduce reoffending. A 2021 MoJ evaluation for the accredited violence programme RESOLVE showed that programme delivery is important to reduce violent reoffending (Robinson et al.,2021) based on a broad measure of violence similar to those being considered for BBR. Not being able to estimate the impact of BBR when it is delivered to an acceptable standard substantially limits the operational value of evaluation findings. Therefore, given the time and resources required to deliver impact evaluations and the methodological issues cited, an impact study would provide limited insight until issues of delivery and data quality could be resolved.

References

Arce, R., Arias, E., Novo, M., & Fariña, F. (2020). Are interventions with batterers effective? A meta-analytical review. Psychosocial Intervention, 29(3), 153-164.

Austin P. C. (2009). The relative ability of different propensity-score methods to balance measured covariates between treated and untreated subjects in observational studies. Medical Decision Making, 29, 661–677.

Babcock, J. C., Green, C. E., & Robie, C. (2004). Does batterers' treatment work? A meta-analytic review of domestic violence treatment. Clinical Psychology Review, 23(8), 1023-1053.

Bloomfield., S., & Dixon, L. (2015). An outcome evaluation of the Integrated Domestic Abuse Programme (IDAP) and Community Domestic Violence Programme (CDVP). London, U.K.: National Offender Management Service. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/449008/outcome-evaluation-idap-cdvp.pdf

Cheng, S. Y., Davis, M., Jonson-Reid, M., & Yaeger, L. (2021). Compared to what? A meta-analysis of batterer intervention studies using nontreated controls or comparisons. Trauma, Violence & Abuse, 22(3), 496-511.

Dowden, C., & Andrews, D. A. (2004). The importance of staff practice in delivering effective correctional treatment: A meta-analytic review of core correctional practice. International Journal of Offender Therapy and Comparative Criminology, 48(2), 203-214.

Gannon, T. A., Olver, M. E., Mallion, J. S., & James, M. (2019). Does specialized psychological treatment for offending reduce recidivism? A meta-analysis examining staff and program variables as predictors of treatment effectiveness. Clinical Psychology Review, 73, 101752.

Leviton, L. C., Khan, L. K., Rog, D., Dawkins, N., & Cotton, D. (2010). Evaluability Assessment to Improve Public Health Policies, Programs, and Practices. Annual Review of Public Health, 31(1), 213–233.

Mews, A., Di Bella, L., & Purver, M. (2017). Impact evaluation of the prison-based core sex offender treatment programme. London, U.K.: Ministry of Justice.

OECD-DAC. (2010). Glossary of key terms in evaluation and results-based management. Paris: OECD-DAC. http://www.oecd.org/development/peer-reviews/2754804.pdf

Normand S. L. T., Landrum M. B., Guadagnoli E., Ayanian J. Z., Ryan T. J., Cleary P. D., & McNeil B. J. (2001). Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: A matched analysis using propensity scores. Journal of Clinical Epidemiology, 54, 387–398.

Papalia, N., Spivak, B., Daffern, M., & Ogloff, J. R. (2019). A meta-analytic review of the efficacy of psychological treatments for violent offenders in correctional and forensic mental health settings. Clinical Psychology: Science and Practice, 26(2), e12282.

Robinson, C., Sorbie, A., Huber, J., Teasdale, T., Scott, K., Purver, M., & Elliott., I. (2021). Reoffending impact evaluation of the prison-based RESOLVE offending behaviour programme. London, U.K.: Ministry of Justice. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957855/RESOLVE_report.pdf

Teasdale, J., Sorbie, A., & Scott, K. (2021). Reoffending impact evaluation of the prison-based RESOLVE offending behaviour programme: Additional analyses and descriptive statistics. London, U.K.: Ministry of Justice. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1025042/RESOLVE_supplementary_appendix.pdf

Appendix A
Quality of delivery: KLOE definitions

Quality Assurance Approach Summary

Quality assurance for accredited programmes is completed through the Interventions Integrity Framework (IIF). This explores evidence of practice in relation to whether the programme delivered met the guidelines set out in the programme and management manual. Evidence is collected from a variety of sources, including self-assessment and questionnaires; centrally held data such as starts, completions, and training records; video recordings of sessions; and clinical evidence, such as supervision notes and post-programme reports. Two iterations of the IIF have been used since it was first introduced in 2014. Whilst similar evidence for BBR is used in both iterations, the headings under which the evidence is reported and scored are different. Due to these changes and a change in how scores are applied, the cycles are not comparable.

2014 – 2016 IIF

Part 1: Quality Completions

A specified dataset outlined the information required from providers to allow HMPPS Intervention Services (IS) to provide feedback on each offender attending the intervention. For example, whether the offender received the intervention by trained facilitators; whether the intervention was at the appropriate dosage and frequency; and whether the offender was in a group of an appropriate size and met the selection criteria for the programme. Sites provided pre/post-programme completion data (collated in a completions dataset), as well as session-level data on delivery and programme attendance (analysed by IS).

Part 2: Quality of Delivery

This component was assessed for each programme being delivered by specialist clinicians within IS, who reviewed quality of programme delivery in the group room as well as quality of intervention management. A sample of products was requested based on a site's volume of delivery. The products selected depended upon the requirements of the programme, but included a sample of session recordings, reports, and other material to provide a holistic assessment of performance once every two years. The quality of delivery

section was also underpinned by an element of self-assessment focused on the key processes which support programme delivery; for example, the composition of groups and the frequency of supervision.

Part 3: Rehabilitative Culture

This component of the IIF was aimed at measuring the rehabilitative environment that should have supported the aims and values of the intervention, thus allowing staff to feel supported in their delivery of the programme and offenders to feel supported to address their offending and reach their potential. This supports one of HMPPS' key overarching commissioning intentions which is to 'Enhance public protection and ensure a safe, decent environment and rehabilitative culture'. This intention focuses on the active development of an environment that is safe, secure and decent, and assists offenders towards rehabilitation. This requires a culture where authority is exercised confidently, consistently, and fairly in order to build trust and improve safety. Providing a safe and decent delivery environment is fundamental to achieving outcomes and is an essential foundation for building a supportive and rehabilitative culture that motivates and enables offenders to make positive changes in their lives. A number of key processes were reviewed which, if followed, would help to support a rehabilitative culture; for example, having the appropriate management structure. Staff and offender questionnaires further supported this.

2016 – 2019 IIF

Key Line of Enquiry (KLOE) 1: Is the intervention being delivered as designed?

This reviewed selection, attrition, and rate and dosage of delivery from central and local data sources. Research shows that the effectiveness of interventions is related to careful matching of the intervention to the assessed risks of reoffending, criminogenic needs, and learning styles of those who participate. To maintain momentum in learning and ensure motivation, attendance and scheduling must be at the appropriate dosage and rate.

KLOE 2: Is the learning environment safe, constructive, and effective?

In order for learning to be effective, the delivery style should be engaging, motivational, and supportive, and in line with the core competency framework. Materials including session recordings, supervision notes, and intervention planning information were reviewed to ensure the programme was delivered with integrity and responsively in a way that all individuals could understand the key learning points and practise new skills as appropriate. Group dynamics and boundaries were also reviewed to support an effective learning environment.

KLOE 3: Is the team enabled to effectively deliver the programme?

Facilitation of effective rehabilitative activities requires well-trained and appropriately supported staff. Delivery staff should be supervised and encouraged to maintain and continually develop their skills. This KLOE reviewed evidence including self-assessment, session monitoring reports, supervision notes, and post-programme reports to assess the quality of intervention management.

KLOE 4: Does the culture/environment support and enable change?

Providing a safe and decent delivery environment is fundamental to achieving outcomes and is an essential foundation for building a supportive and rehabilitative culture that motivates and enables individuals to make positive changes in their lives. The rehabilitative environment should authenticate the aims and values of the intervention so that participants feel fully supported to address their offending and reach their potential. This reinforces one of HMPPS' key overarching commissioning intentions, which is to 'Enhance public protection and ensure a safe, decent environment and rehabilitative culture'. To review this, self-assessment and staff and participant questionnaires were used.

RAG Rating (Red – Amber – Green)

For both cycles (2014-16 and 2016-19), following a review of the evidence, each section was awarded a RAG rating indicative of the integrity of the programme in each of the key areas (see Figure A1). All scores, except Rehabilitative Culture, are allocated by programme. An overall score is also given for the site. Rehabilitative Culture is scored only for the site as a whole.

Figure A1: RAG ratings used to designate quality of delivery

 Green	Programme integrity is maintained effectively. There are no risks to programme integrity – minor development areas may be noted.
 Amber/Green	Improvements are required to meet the required standard, although programme integrity is not threatened to a significant degree.
 Amber/Red	Programme integrity is under significant threat. Significant improvements need to be made in key areas to meet the required standard.
 Red	Programme integrity is compromised. Critical areas of improvement have been identified. Immediate remedial action is required to minimise the threat and restore programme integrity.

Appendix B
Impact of the COVID-19 pandemic

It should be noted that the cohorts used to determine the feasibility recommendations in this report are partially affected by the COVID-19 pandemic. Any effect during the intervention phase will be minimal because, to ensure a long enough follow-up period, the latest programme start date included was at the end of April 2019 and the proportion of the overall cohort of BBR starters with a termination date recorded after 23 March 2020 (when the first set of lockdown restrictions were imposed) is very low (less than 3%).

However, a portion of the reoffending follow-up period took place during the pandemic for a larger subset of records. Following the first set of lockdown restrictions, limits on court activity led to sharp decreases in the number of cases processed at the criminal court. While there were no court closures during subsequent lockdowns, police recorded crime data have shown overall reductions in the reporting and recording of many crime types over these periods, although for DA-related crime, the number of police-recorded crimes across England and Wales rose by 6% in the year ending March 2021. This follows increases seen in previous years and may reflect improved recording by the police alongside increased reporting by victims. However, referrals of suspects of DA-flagged cases from the police to the Crown Prosecution Service (CPS) for a charging decision decreased by 3% during the same period.

Proven reoffending statistics continue to use a six-month waiting period to allow the offence to be proven in court, so the same approach has been used for this feasibility analysis. On that basis, and due to the use of a two-year follow up period, any programme starts from late September 2017 will be partially affected (this affects 45% of the final cohort, although only 9% have more than half of their follow-up period on or after 23 March 2020). However, given that a consistent approach has been applied to the intervention and comparison groups, this does not present a methodological problem so long as any evaluation results are appropriately caveated to acknowledge this feature. Further analysis on the effect of the pandemic on evaluation results would also be carried out and reported on should an impact evaluation go ahead.