An impact evaluation of the prison-based Thinking Skills Programme (TSP) on reoffending

Aimee Brinn, John Preston, Rosina Costello, Tyler Opoku, Emily Sampson, Ian Elliott and Annie Sorbie

Ministry of Justice

2023

Summary

About the Thinking Skills programme

The Thinking Skills Programme (TSP) is an accredited offending behaviour programme designed and delivered by His Majesty's Prison and Probation Service (HMPPS). TSP is suitable for adult men and women assessed to be at medium and above risk of reoffending. TSP is the highest volume accredited programme delivered in custody.

The TSP is designed to reduce general reoffending by supporting improvements in four ways:

1. Developing thinking skills (such as problem solving, flexible thinking, consequential thinking, critical reasoning)
2. Applying these skills to managing personal risk factors
3. Applying thinking skills to developing personally relevant protective factors
4. Applying thinking skills to setting pro-social goals that support relapse prevention.

The programme format comprises 19 sessions (15 group sessions and 4 individual sessions, resulting in around 38 hours of contact time (dose).

The Evaluation

The aim of this evaluation is to assess the impact of TSP delivered in prison on proven general reoffending within a two-year follow-up period.

The analysis involved a treatment group of 20,293 adults (18,555 males, 1,738 females) who participated in the TSP programme between 2010 and 2019 and this was compared to a matched comparison group of 375,647 adults (345,084 males, 30,563 females) who did not participate in the programme. Propensity score matching (PSM) was used to ensure comparable treatment and comparison groups. The evaluation used the largest number of PSM matching variables for a HMPPS accredited programme evaluation to date.

The evaluation also has a large sample size which means it is likely to be representative of the population of TSP participants. A larger sample generates more precise results and increases the power of statistical testing. This increases the likelihood of finding a statistically significant finding (i.e., not due to chance) even if the difference between the treatment group and the matched comparison group is small. All adults in this study were released from prison between 2010 and 2020.

The impact of TSP was evaluated against three proven general reoffending metrics over a two-year follow up period:

1. **Binary measure of reoffending** (reoffending rate) – did they re-offend?
2. **Frequency of reoffences committed** – How many re-offences over the two-year period?
3. **Time to first reoffence**

Males and females were analysed separately due to the known differences in reoffending behaviour. Headline results include all participants in the programme, separated by gender. Analyses were conducted to investigate the potentially differential effect of TSP participation on distinct subgroups and to provide information on how differences in TSP delivery may impact on its effectiveness. It was not always possible to conduct sub-analyses due to small sample sizes.

Four key sub-analyses (more details are in ‘Explanation of sub-analyses’) were identified as potentially important moderators of TSP effectiveness:

- **Suitability for TSP** (ideally suitable and not ideally suitable)
- **Completion of TSP** (completed and not completed)
- **Programme integrity** using the HMPPS 2016-2019 Interventions Integrity Framework (broadly maintained and compromised)
- **Risk of reoffending prior to TSP** (Offender Group Reconviction Score (OGRS): low, medium, or high risk of reoffending).

Additional sub-analyses were conducted to provide further context and explanation of results included:

- **Index offence group** (acquisitive offences, sexual offences, and OVP (OASys Violence Predictor) offences – based on grouping of Home Office offence codes)
- **Exclusivity of TSP** (participation in TSP only and in one or more other accredited programmes)
- **Ethnic group** (‘Asian and Asian British’, ‘Black, Black British, Caribbean, and African’, ‘mixed and multiple ethnic groups’, and ‘White’, as per Office for National Statistics aggregate categories)
- **Learning Disabilities and Challenges (LDC)** (more likely to present with characteristics associated with LDC and less likely to present with characteristics associated with LDC)
- **Age** (18-25, 26-30, 31-49 and 50+)

Results of the evaluation

Headline results - Male:

Results show that, over a two-year period from release, those who had participated in TSP were **less likely to reoffend, reoffended less frequently,** and **took longer to reoffend**, compared to males who did not participate in TSP. These results were statistically significant with mostly very small effect sizes.

Key sub-analyses - Male:

The results showed that males who participated in TSP and met any of the following conditions; (a) ideally suitable for TSP, (b) completed TSP, (c) participated in TSP in a prison between 2016 and 2019 where programme integrity was broadly maintained, or (d), were at medium and above risk of reoffending (OGRS3 risk score between 50 and 100), were **less likely to reoffend**, **reoffended less frequently,** and **took longer to reoffend** over a two year period, compared to similar males who did not participate in TSP. These results were statistically significant with mostly very small effect sizes.

Headline Results - Female:

Results showed that, over a two-year period from release, those who participated in TSP **reoffended less frequently,** compared to those who did not participate in the programme. These results were statistically significant and have mostly very small effect sizes.

Participation in TSP did not have a statistically significant effect on the two-year binary reoffending rate for females, or the amount of time before a female offender committed their first proven reoffence.

Key sub-analyses - Female:

Female sub-analyses were limited due to small sample sizes and therefore would be less likely to produce statistically reliable results. Of those conducted, results showed that female participants of TSP who were ideally suitable for the intervention were **less likely to reoffend** and **reoffended less frequently** over a two-year follow up period, compared to females who did not participate in TSP. These results were statistically significant and have mostly very small effect sizes**.** There was no statistically significant effect on the time taken to reoffend for this subgroup.

Females who completed TSP **reoffended less frequently** within a two-year period, compared to females who did not participate in TSP. These results were statistically significant. There was no statistically significant effect on the binary reoffending rate or time taken to reoffend for this subgroup.

Conclusion

Both effect sizes and whether the result is statistically significant (likelihood of findings due to chance) should be taken into consideration when interpreting the findings of this TSP impact evaluation.

For the male cohort, the results of the overall analysis and each of the four key sub-analyses (ideally suitable, completed TSP, programme integrity maintained, medium or high-risk of reoffending) were statistically significant in reducing reoffending across all outcome measures. For the smaller female cohort, there were some statistically significant results in reducing reoffending; those who participated in TSP reoffended less frequently, those who were ideally suitable were less likely to reoffend and reoffended less frequently, and those who completed TSP reoffended less frequently.

These consistent results are reflective of TSP theory and indicate that good programme delivery contributes to effective rehabilitation. In the field of criminal justice and offender interventions evaluations, effect sizes are typically found to be small to medium with robust evaluation designs tending to yield small effect sizes. Overall, across the analyses the effect sizes were mostly very small.

Key results for male cohort

Two-year proven general reoffending measures for males: Headline and key sub-analyses

Headline	46.5% of the treatment group reoffended with a general reoffence in the two years following release from prison. This is a 1.7%-point difference when compared to the comparison group or a 4% lower reoffending rate[1].	⬇	This is significantly[2] fewer than the comparison group (48.2%).
Participants who met the ideal suitability criteria[3]	53.7% of the treatment group reoffended with a general reoffence in the two years following release from prison. This is a 2.4%-point difference when compared to the comparison group or a 4% lower reoffending rate.	⬇	This is significantly fewer than the comparison group (56.1%).
Completed TSP	49.8% of the treatment group reoffended with a general reoffence in the two years following release from prison. This is a 1.7%-point difference when compared to the comparison group or a 3% lower reoffending rate	⬇	This is significantly fewer than the comparison group (51.5%).
Programme integrity broadly maintained[4] (2016-2019)	42.5% of the treatment group reoffended with a general reoffence in the two years following release from prison. This is a 2.4%-point difference when compared to the comparison group or a 5% lower reoffending rate.	⬇	This is significantly fewer than the comparison group (44.9%).
With OGRS3 score 50-74 (medium risk)	45.0% of the treatment group reoffended with a general reoffence in the two years following release from prison. This is a 1.9%-point difference when compared to the comparison group or a 4% lower reoffending rate.	⬇	This is significantly fewer than the comparison group (46.9%).
OGRS3[5] score 75+ (high risk)	65.3% of the treatment group reoffended with a general reoffence in the two years following release from prison. This is a 2.5%-point difference when compared to the comparison group or a 4% lower reoffending rate.	⬇	This is significantly fewer than the comparison group (67.8%).

Green arrow for significant finding, grey arrow for non-significant

[1] The **percentage change** in reoffending rate is the rate of change (i.e., how much a value has changed in relation to a previous value). In this context, it is calculated as ((treatment group % - comparison group %) / comparison group %)*100. Using the headline figure as an example: ((46.5%-48.2%)/48.2%)*100 = 4%). This is different to the **percentage point change,** which is the absolute numerical difference between two percentages, which is used to show the magnitude of change between the treatment and comparison group. It is calculated as (treatment group % - comparison group %). In this example, 46.5%-48.2% = -1.7 percentage point change.

[2] Statistical significance level set at $p < 0.05$. There are a range of reasons why an evaluation might not find a statistically significant effect. These include but are not limited to: there is no effect to be found, lower underlying rates of reoffending can make it harder to achieve significance, smaller sample sizes for some analyses or unobservable variables that were not accounted for in the evaluation approach.

[3] As defined by TSP accreditation panel report (The Correctional Services Accreditation Panel Report 2009-2010, Annex E (publishing.service.gov.uk))

[4] As quality assured by HMPSS using the Interventions Integrity Framework (IIF)

[5] An OGRS score is the percentage likelihood of committing any offence within 2 years leading to reconviction (proven reoffending). This is based on static factors such as age, gender, and criminal history. An OGRS score of 50% or more means that an offender is more likely than not to commit a proven reoffence within 2 years.

Headline	An average of 1.75 proven general reoffences were committed by each of the men in the treatment group.	↓	This is significantly fewer than the comparison group (2.00 proven general reoffences).
Participants who met the ideal suitability criteria	An average of 2.10 proven general reoffences were committed by each of the men in the treatment group.	↓	This is significantly fewer than the comparison group (2.38 proven general reoffences).
Completed TSP	An average of 1.86 proven general reoffences were committed by each of the men in the treatment group.	↓	This is significantly fewer than the comparison group (2.15 proven general reoffences).
Programme integrity broadly maintained (2016-2019)	An average of 1.65 proven general reoffences were committed by each of the men in the treatment group.	↓	This is significantly fewer than the comparison group (1.81 proven general reoffences).
With OGRS3 score 50-74 (medium risk)	An average of 1.49 proven general reoffences were committed by each of the men in the treatment group.	↓	This is significantly fewer than the comparison group (1.62 proven general reoffences).
OGRS3 score 75+ (high risk)	An average of 2.92 proven general reoffences were committed by each of the men in the treatment group.	↓	This is significantly fewer than the comparison group (3.37 proven general reoffences).
Headline	The average time before a reoffender in the treatment group committed their first proven general reoffence was 267 days.	↑	This is significantly later than the comparison group (257 days).
Participants who met the ideal suitability criteria	The average time before a reoffender in the treatment group committed their first proven general reoffence was 262 days.	↑	This is significantly later than the comparison group (252 days).
Completed TSP	The average time before a reoffender in the treatment group committed their first proven general reoffence was 262 days.	↑	This is significantly later than the comparison group (250 days).
Programme integrity broadly maintained (2016-2019)	The average time before a reoffender in the treatment group committed their first proven general reoffence was 275 days.	↑	This is significantly later than the comparison group (257 days).
OGRS3 score 50-74 (medium risk)	The average time before a reoffender in the treatment group committed their first proven general reoffence was 286 days.	↑	This is significantly later than the comparison group (277 days).
OGRS3 score 75+ (high risk)	The average time before a reoffender in the treatment group committed their first proven general reoffence was 240 days.	↑	This is significantly later than the comparison group (229 days).

Key results for female cohort

Two-year proven general reoffending measures for females: Headline and key sub-analyses[6]

Green arrow for significant finding, grey arrow for non-significant

Headline	42.6% of the treatment group reoffended with a general reoffence in the two years following release from prison. This is a 1.6%-point difference when compared to the comparison group or a 4% lower reoffending rate	⬇	This is not significantly fewer than the comparison group (44.2%).
Participants who met the ideal suitability criteria	48.7% of the treatment group reoffended with a general reoffence in the two years following release from prison. This is a 3.3%-point difference when compared to the comparison group or a 6% lower reoffending rate.	⬇	This is significantly fewer than the comparison group (52.0%).
Completed TSP	44.3% of the treatment group reoffended with a general reoffence in the two years following release from prison. This is a 1.8%-point difference when compared to the comparison group or a 4% lower reoffending rate.	⬇	This is not significantly fewer than the comparison group (46.1%).
Headline	An average of 2.17 proven general reoffences were committed by each of the women in the treatment group.	⬇	This is significantly fewer than the comparison group (2.43 proven general reoffences).
Participants who met the ideal suitability criteria	An average of 2.67 proven general reoffences were committed by each of the women in the treatment group.	⬇	This is significantly fewer than the comparison group (3.05 proven general reoffences).
Completed TSP	An average of 2.10 proven general reoffences were committed by each of the women in the treatment group.	⬇	This is significantly fewer than the comparison group (2.51 proven general reoffences).
Headline	The average time before a reoffender in the treatment group committed their first proven general reoffence was 262 days.	⬆	This is not significantly later than the comparison group (247 days).
Participants who met the ideal suitability criteria	The average time before a reoffender in the treatment group committed their first proven general reoffence was 250 days.	⬆	This is not significantly later than the comparison group (238 days).
Completed TSP	The average time before a reoffender in the treatment group committed their first proven general reoffence was 268 days.	⬆	This is not significantly later than the comparison group (258 days).

Green arrow for significant finding, grey arrow for non-significant

[6] There was insufficient power to conduct sub-analyses to investigate the effect of programme integrity or OGRS risk score for female participants.

Impact on reoffences for male cohort

For any **100** typical men who receive the intervention, compared with any **100** similar men who do not receive it:

The number of men who commit a proven general reoffence within ***two years*** could be **lower by between 1 and 2 men**. **This is a statistically significant result**.

The number of proven general reoffences committed within ***two years*** could be **lower by between 20 and 30 offences**. **This is a statistically significant result**.

On average, the time before an offender committed their first proven general reoffence within ***two years*** could be **longer by between 5 and 14 days**. **This is a statistically significant result**.

Impact on reoffences for female cohort

For any **100** typical women who receive the intervention, compared with any **100** similar women who do not receive it:

The number of women who commit a proven general reoffence within ***two years*** could be **lower by between 1 and 4 women**. **This is not a statistically significant result**.

The number of proven general reoffences committed within ***two years*** could be **lower by between 4 and 49 offences**. **This is a statistically significant result**.

On average, the time before an offender committed their first proven general reoffence within ***two years*** could be **between shorter by 1 day and longer by 30 days**. **This is not a statistically significant result**.

What you can and cannot say about these results can be found in Annex 1.

Table of contents

TSP description by the programme developer (HMPPS)

The Thinking Skills Programme (TSP) is an accredited cognitive skills programme for adult men and women aged 18 years and above, and is suitable for individuals assessed to be at medium and above risk of reoffending. It is suitable for people with any offence and is delivered across His Majesty's Prison and Probation Service (HMMPS). It is the highest volume accredited programme delivered in custody.

The programme is designed to help develop participants' skills in pro-social problem solving, perspective taking, developing, and managing relationships, and self-management. It encourages pro-social attitudes, behaviour, and goals for the future.

The aim of TSP is to support participants to develop skills which can help stop them from reoffending and encourage them to live a successful, pro-social life moving forward. It does this by targeting criminogenic need (i.e., dynamic risk factors) to develop participants' ability to manage their emotions, make effective decisions, solve problems, achieve their goals, manage the influence of anti-social relationships, and using pro-social interpersonal skills in their interactions with others.

More broadly, TSP aims to reduce reoffending in the following four ways:

- Developing participants' thinking skills;
- Coaching participants to apply new and existing thinking skills to identifying and managing their risk factors;
- Coaching participants to apply new and existing thinking skills to develop personally relevant protective factors;
- Coaching participants to apply new and existing thinking skills to achieving pro-social goals that support relapse prevention.

The key principles of TSP are:

- an explicit focus on risk factors, protective factors, and pro-social goals;
- a focus on engagement and motivation;
- ensuring that the programme is experienced by each participant as being personally relevant;
- a facilitation style best characterised as coaching; and
- promoting continuity within programme design and with case management.

TSP has been designed to incorporate maximum responsivity and flexibility of delivery format. The programme comprises 19 sessions (15 group sessions and 4 individual sessions), divided across three modules (Self-Control, Problem Solving, and Positive Relationships). TSP was recommended for accreditation by the Correctional Services Accreditation and Advice Panel (CSAAP), for more information on CSAAP, see Annex 2.

International meta-analyses have repeatedly found that cognitive skills programmes reduce reconviction rates for general reoffending when they are delivered as intended (Lipsey et al 2001; Landenberger & Lipsery, 2005; Lipsey & Landenberger, 2006; Tong & Farrington, 2008). Early evaluations of cognitive skills programmes delivered in England and Wales in HM Prisons (e.g., Cann et al, 2003; Falshaw et al, 2003; Friendship et al, 2002), and the community (e.g., Hollin et al. 2004; Palmer et al. 2007; Hollin et al. 2008; McGuire, 2008) reported mixed outcomes. The suggested causes for this included the expansion of programmes, challenges with implementation such as non-completion, and challenges with delivering high quality evaluation methodology.

Later, evaluations were more promising (Travers et al, 2013), and evaluations using more robust matching protocols (propensity score matching) have shown that TSP's predecessor, the Enhanced Thinking Skills (ETS) programme, significantly reduced reoffending by 6.3-percentage points (Sadlier, 2010), and for those suitable for the programme, starting the TSP in the community was associated with a 5-percentage point reduction (Travers, 2016). According to Travers (2016), the effect for suitable completers of the TSP was a 10-percentage point advantage over non-starters and non-completers.

Summary of methodology

The aim of this evaluation is to assess the impact of the TSP on proven reoffending outcomes. It is a good quality quasi-experimental evaluation[7] with a large sample. The study includes individuals who participated in the TSP between 2010 and 2019[8]. A long study period was chosen to increase the sample size of the study. Studies with larger sample sizes are often more representative of the population from which the sample is drawn. A larger sample also increases the power of statistical testing which can increase the likelihood of finding a statistically significant finding where one truly exists (i.e., not due to chance)[9].

Person-level intervention data from TSP was matched to the following datasets:

- **Police National Computer (PNC)** to provide criminal histories and to determine reoffending outcomes

- **Offender Assessment System (OASys)** to provide offending-related risks and needs information

- **Prison population data** to provide information on self-reported ethnicity

The linking of these datasets produced a comprehensive suite of data for each participant of the TSP. These individuals all served a custodial sentence between 2010 (when the TSP had passed the pilot stage and was delivered throughout prisons) and 2020 (a cut-off which ensures that all individuals have at least 2 years of follow-up data following their release). An additional comprehensive suite of data was extracted for a comparison group of offenders with similar characteristics and serving a custodial sentence during the same period.

This suite of data, comprising 92 matching variables (Annex 3), was used as the basis for building propensity score matching (PSM) models. This is the largest number of matching variables used by a HMPPS evaluation to date. Males and females were analysed in separate PSM models due to known differences in their reoffending behaviours[10]. To minimise the potential effect of differences in participation year on outcome measures, TSP participation year (or pseudo-TSP

[7] Equivalent to a Level 4 on The Maryland Scientific Methods Scale (SMS) (Farrington, D. P., Gottfredson, D. C., Sherman, L. W., & Welsh, B. C. (2003). The Maryland scientific methods scale. In *Evidence-based crime prevention* (pp. 13-21). Routledge.) For further discussion see Sherman, L. W. (1998). Preventing crime: What works, what doesn't, what's promising. US Department of Justice, Office of Justice Programs, National Institute of Justice.

[8] There have been some changes to the delivery of TSP during this period. Two changes in delivery were particularly relevant to this impact evaluation. First, based on commissioning policy, between May 2013 and February 2017 those with acquisitive index offences were only eligible via a clinical over-ride. Second, in January 2014 TSP was made available to those with an OGSR3 risk score of 25-49 as an additional risk override (where they were not taking the place of those with higher OGRS3 risk scores). To minimise any effects of differences in TSP delivery over time, participation year was included as a matching variable in our propensity score models.

[9] See for example, Weisburd, D., Lum, C.M and Petrosino, A (2021) Does Research Design Affect Study Outcomes in Criminal Justice? The American Academy of Political and Social Science, ANNALS, 578, November 2021.

[10] https://publications.parliament.uk/pa/cm5803/cmselect/cmjust/265/report.html

participation year for the comparison group)[11] was included as a matching variable in our propensity score matching models.

PSM is a statistical matching technique which uses factors theoretically and empirically associated with both receiving treatment and the outcome variable (reoffending) to predict a 'propensity score' [12]. This propensity score reflects the likelihood that an individual in custody received the intervention, given the recorded characteristics. Individuals in the treatment group were matched to similar individuals who did not receive treatment. Overall, the matching quality for the headline and sub-analyses was very good[13] (see standardised differences annex for more detail).

The reoffending rates for the treatment and comparison groups were then compared. The rates are calculated using the weighted values[14] for each person after matching. Three reoffending outcomes were used to estimate the impact of the intervention in a two-year period, and were applied for both the male and female cohorts. The outcomes are as follows:

1) A binary reoffending outcome: the number of people who commit a proven reoffence, expressed as a percentage of the group.

2) A frequency reoffending outcome: the number of proven reoffences committed, expressed per person

3) Time to reoffence: the average number of days between a person's prison release date and the date on which they commit their first proven reoffence, including only those who reoffend

The same general headline measures were applied to nine sub-analyses which examined the effect of TSP on specific sub-groups. For a summary, see 'Explanation of sub-analyses' below. Each analysis undergoes a different and unique PSM process.

The outcome measures in this study solely examine the effect of TSP participation on reoffending behaviour. An additional report examining the effect of TSP participation on prison adjudications is also available.

[11] See explanation of 'pseudo-start dates' in Annex 4 for how these were calculated.

[12] A propensity score is a value between 0 and 1 which represents the likelihood of receiving treatment. More details on the matching methodology are included in Annex 4.

[13] Matching quality in JDL analyses uses a traffic light scale (see standardised differences annex). The mean absolute standardised differences for all sub-analyses was less than 5%. Therefore, the matching quality achieved based on recorded factors was 'green' or 'very good'.

[14] As we use matching with replacement; each treatment group member is given a weight equal to 1. Each comparison group member is given a weight based on how many comparison units are matched to each treatment unit, and how many treatment units they are matched to.

Interpreting results

Both effect sizes and whether the result is statistically significant (likelihood of findings due to chance) should be taken into consideration when interpreting the findings of this impact evaluation.

The difference in reoffending outcomes between the treatment and comparison groups is compared using statistical significance testing, which returns a ‘p’ value. In this report, the results are examined using the standard 0.05 significance level. If less than 0.05, the difference between the two groups is considered to be statistically significant and unlikely to be due to chance. The direction of the difference in reoffending rates indicates whether the treatment effect is positive or negative. The estimated differences shown are the 95% confidence intervals[15] for the differences between the relevant treatment and comparison group outcomes. If the 95% confidence interval range crosses over 0 then the result is not statistically significant.

Smaller sample sizes and increased variability are factors which cause confidence intervals to widen. Thus, analyses with smaller sample sizes may be more likely to have statistically insignificant results. Larger sample sizes lead to increased power to find a statistically significant effect. In turn, this leads to an increased tendency to find a difference which is statistically significant, even when the clinical significance of such a difference is modest[16].

To aid the interpretation of the effect size, the Cohen’s *d* statistic is typically categorised as follows (Cohen, 1988):

- **Small**: denoting an effect size greater than or equal to 0.2 but below 0.5
- **Medium**: denoting an effect size greater than or equal to 0.5 but below 0.8
- **Large**: denoting an effect size greater than or equal to 0.8

Small, medium, and large categories act as a simple guide to interpretating the effect size and are relative to the area of behavioural science or specific research method being employed (Cohen, 1988). They are a rule of thumb, and application to specific social science outcomes must be tailored to context.

Published effect sizes can be inaccurately inflated. Publication bias (the phenomenon that studies with statistically significant results are more likely to be published than those with statistically insignificant results) and poor-quality research methodology (such as biased or non-robust methodologies) are likely responsible for the inflation of published effect sizes (Schafer & Schwarz, 2019). A recent project found that even when studies published in highly prestigious journals are replicated, their effect sizes can reduce by half (Camerer et al, 2018). It is speculated that small effect sizes found from studies with large sample sizes are the most likely the reflect the true state of nature (Funder et al, 2019). Due to these limitations, comparing effect sizes between studies can be difficult and it can be challenging to find an appropriate benchmark for

[15] A range of values that you can be 95% confident contains the true mean of the population.

[16] Clinical significance is the practical importance of a treatment effect (whether the intervention provides real, noticeable benefits which are palpable enough to be justified given associated costs/harms/inconveniences). Statistical significance implies whether there is mathematical difference between the two groups (treated and not treated), which for this study is set as $p < 0.05$.

effect sizes within specific research areas (e.g., reducing reoffending). Despite this limitation, it has been found that within the field of criminal justice and offender interventions evaluations, effect sizes are on average small to medium (see for example, Barnes, TenEyck, Pratt, & Cullen, 2020). Effects are often found to be smaller when evaluating routinised (real-world) programmes delivered at scale compared to small trial programmes. For example, Lipsey and Landenberger (2006) found the average reduction in recidivism was 11% lower for real-world large practise programmes than small research and demonstration projects (where there is likely high-fidelity to delivery-as-designed).

The results set out in this report should be interpreted using a combination of: (a) whether the statistical tests meet a standard threshold for "statistical significance" by considering the p-value and (b) the "effect size" associated with that statistical test which, in these tables, is the Cohen's *d*. Together, these tell you whether there appear to be genuine differences between the TSP treatment and comparison groups and the magnitude of that change.

For additional insight, odds ratios for the reoffending rate for males and females are included in Annex 15. These show the odds that an outcome will occur (in this case reoffending) given exposure to an intervention such as the TSP, compared to the odds of the outcome occurring if not exposed to the intervention.

Despite efforts to include all observed factors known to be predictive of selection onto the TSP and of reoffending risk into the PSM model, the importance of information that is not recorded cannot be known. As a result, there may be unobserved and unaccounted for factors which affect the results of this study. Other limitations include: smaller sample sizes for females compared to males, small sample sizes for certain sub-analyses, and unknown/non-proven reoffending which is not included in the analysis. For further detail on methodology, see Annex 4. A fuller list of limitations can be found in Annex 5.

Explanation of sub-analyses

Further analyses were undertaken to examine the specific effects of TSP for relevant subgroups. Four key sub-analyses (ideal suitability for TSP, completion of TSP, programme integrity, risk score) examined differential effects of TSP delivery, whilst additional sub-analyses (offence group, exclusivity of TSP, ethnic group, learning disabilities and challenges, and age) examined the relationship between TSP participation and individual factors. Each subgroup underwent a separate PSM process and therefore results are not comparable across the sub-analyses.

All sub-analyses were considered for both male and female cohorts. However, some subgroups were too small to be analysed. These can be found in Annex 6.

Key sub-analyses directly related to TSP theory and delivery

Ideal suitability for TSP

This sub-analysis sought to determine whether the effect of TSP was statistically significantly different for participants who were 'ideally suitable'. Those who met the 'risk' and 'need' criteria for the programme in full were identified as 'ideally suitable'. The criteria for 'ideal suitability' for TSP was modified in 2019. This study used the ideal suitability criteria defined by the TSP manual published in 2010 as it better aligned with the study time period.

Those who accessed the programme through a discretionary 'risk' override by the TSP treatment manager, or did not meet the suitability criteria, were regarded as 'not ideally suitable'. 'Ideal suitability' was measured using strict application of the TSP 'risk' and 'need' criteria as outlined in the TSP Management Manual.

For a candidate to be considered 'ideally suitable', they must have **both:**

1) an OGRS3 reoffending risk assessment score greater than or equal to 50 (medium and above risk)

and **one of either:**

2i) Offending needs assessment: score greater than or equal to 7 on the 7 (OASys)[17] items, or

2ii) Offending needs assessment: score equal to 5 on the 7 OASys items with a score of 2 on items 11.6 or 11.7

OASys items are scored from 0-2, where a higher score denotes a higher need. The 7 items scored from the OASys assessment as part of the ideal suitability criteria are:

[17] A system introduced in 2001 and built on the existing 'What Works' evidence base. It combines actuarial methods of prediction with structured professional judgement to provide standardised assessments of offenders' risks and needs, helping to link these risks and needs to individualised sentence plans and risk management plans.

Thinking Skills Programme Targets	**Target OASys item**
Stop and think	**11.7** awareness of consequences
Emotional Awareness	**11.4** temper control
Problem Solving	**11.6** problem solving
Perspective Taking	**2.6** recognises the impact and consequences of offending on victim, community/wider society **11.9** understands other people's point of view
Offence free relationships	**7.2** regular activities encourage offending
Goals and Values	**12.1** pro criminal attitudes

Those who met these criteria were compared to a matched comparison group of 'ideally suitable' individuals who did not receive the TSP.

All other participants did not fully meet the TSP 'risk' and 'need' criteria and were regarded as 'not-ideally suitable'. A proportion of these individuals would have been appropriately selected onto TSP because they were eligible for a 'risk override' at the discretion of a TSP Treatment Manager. For more information on this group see Annex 8. The remaining participants in the not-ideally suitable group would not have been eligible for a 'risk override' and were likely selected onto the TSP on an individual case-by-case basis in consultation with Interventions Services or were selected incorrectly.

All candidates in the not-ideally suitable group were matched to a comparison group of 'not ideally suitable' individuals who did not receive TSP. Further information on the profile of this 'not-ideally suitable' group and its proportions are provided in Annex 8.

Completion of TSP

This sub-analysis aimed to determine whether the effect of the TSP was different for those who completed the programme and those who started but did not complete it.

This analysis created two subgroups by dividing the treatment group into 'TSP completers' and 'TSP non-completers'. Subsequently, each subgroup was matched to a 'no treatment' comparison group.

Programme Integrity

This sub-analysis sought to evaluate the extent to which the quality of TSP delivery may have an impact on outcome. The 'Quality of Delivery' data was supplied by

HMPPS and refers to quality assurance of TSP completed through the Interventions Integrity Framework (IIF); this analysis covers the timeframe 2016 - 2019.

Using the quality assurance framework, two subgroups of the TSP treatment group were created by dividing the treatment group into ‘programme integrity broadly maintained 2016-19’ and ‘programme integrity compromised 2016-19’. When programme integrity could not be clearly categorised, those establishments were omitted from the analysis.

The two subgroups can be described as follows:

1) Programme integrity was broadly maintained when delivered in prison settings that met the guidelines outlined in programme and management manuals, compared to a matched comparison group.
2) Programme integrity was compromised when delivered in prison settings that did not meet the guidelines outlined in programme and management manuals, compared to a matched comparison group.

More information on how these groups were defined can be found in Annex 8.

Risk Scores

TSP is intended for individuals with medium to high (50-74) and high to very high (75+) OGRS3 risk scores; these groups are therefore of particular interest. This sub-analysis examined how the effect of the TSP may differ for individuals with different reoffending risk levels, using OGRS3 risk scores.

OGRS3 is defined as ‘percentage likelihood of committing any offence within 2 years leading to reconviction (proven reoffending)’. This is based on static factors such as age at current conviction, age at follow up (earliest opportunity to reoffend), age at first sanction, gender, number of previous sanctions and current offence type. An OGRS3 score of 50% or more means that an individual is more likely than not to commit a proven reoffence within 2 years.

Bands of OGRS scores were used to create three subgroups of increasing risk for the analysis: ‘25-49’ (low risk), ‘50-74’ (medium risk), and ‘75+’ (high risk). Each OGRS band was matched to a ‘no treatment’ comparison group.

Additional sub-analyses conducted to provide further context and explanation of results

Offence Group

This sub-analysis assessed the reoffending rates of individuals who have been charged with specific index offences[18]. The type of offence a participant has committed could affect their response to the programme.

Three index offence groups were considered for this sub-analysis: acquisitive offences, sexual offences, and OVP (OASys Violence Predictor) offences[19], based on groups of Home Office offence codes. Each index offence group was matched to a

[18] The index offence is the offence for which they are serving the current sentence.

[19] Available on request.

comparison group of individuals with the same index offence group who didn't participate in the TSP.

For a list of other offence groups which did not meet the sample size for analysis, see Annex 7.

Participation in TSP only (during the same sentence)

This sub-analysis measured the isolated treatment effect of the TSP accredited programme for those who did not participate in another accredited programme before starting TSP (during the same sentence). If offenders have participated in other intervention programmes, there could be combined effects of engaging in treatment from multiple programmes. This sub-analysis was conducted to partially control for any such effects.

This analysis created two subgroups of the TSP treatment group:

1. 'TSP only': these individuals did not participate in another accredited programme during this sentence **before** they participated in the TSP. They may, however, have participated in another accredited programme during the same custodial sentence **after** participation in the TSP, **or** during a different custodial sentence (before or after the current custodial sentence). This group was compared to a matched comparison group who did not take part in the TSP.
2. 'Other Accredited Programme': these individuals participated in another accredited programme before they participated in TSP and during the same custodial sentence. This group was compared to a matched comparison group who participated in another accredited programme prior to their TSP pseudo-start date who did not participate in TSP.

For a breakdown of which other accredited programmes were attended by individuals in the treatment group (during their index sentence but prior to their participation in the TSP), see Annex 9.

Ethnic groups

The sub-analysis sought to investigate the effectiveness of the TSP for different ethnic groups. Each ethnic group was compared to a matched 'no treatment' comparison group.

This analysis refers to self-reported ethnicity as obtained from the prison population data. Four subgroups are used for this sub-analysis: 'Asian and Asian British, 'Black, Black British, Caribbean, and African', 'Mixed and multiple ethnic groups' and 'White', as per the Office for National Statistics high-level aggregate categories.[20] Further breakdowns of self-reported ethnicities included within these groupings can be found in Annex 10.

[20] See Ethnic group, national identity and religion - Office for National Statistics.

Learning Disabilities and Challenges (LDC)

This sub-analysis sought to investigate how participants with characteristics associated with Learning Disabilities and Challenges (LDC) were impacted by the TSP. The two subgroups were those more likely to have presented with characteristics associated with LDC, and those less likely to have presented with characteristics associated with LDC. Each group was compared to a matched comparison group who did not participate in the TSP.

LDC is measured using the HMPPS Learning Screen Tool (LST) (Wakeling, 2018). An LST score of more than or equal to 3 is considered to represent "potential LDC identified" and is explored through further assessment. However, as part of the development of the screening tool, the author found that as the LST score increases, the rate of true negative scores increases (i.e., the number of individuals correctly identified as not having LDC increases). Consequently, for this sub-analysis an individual was identified as being more likely to present with LDC if their LST score was greater than or equal to 5 (a higher threshold than used routinely), and less likely if their score was less than 5.

Age

This sub-analysis sought to investigate the effect a participant's age has on the impact of the TSP. Each age band was matched to a 'no treatment' comparison group.

The age of an individual is measured at the release date of their corresponding sentence. Ages were banded into four subgroups: '18-25', '26-30', '31-49', and '50+', inclusive of the minimum and maximum of the range. The youngest age group was decided as 18–25-year-olds who are most likely to present with low psychosocial maturity and may also be sent to the young adult estate. The group 26–30-year-olds was selected based on evidence that some young adults take longer to develop psychosocial maturity, which may impact their likelihood to engage with accredited programmes, and their offending behaviours (Monahan et al, 2013). The age group '50+' was selected to allow a large enough sample size for sufficient power, which enabled us to investigate the impact of TSP for the older range of participants.

Male results in detail

Table 1 presents the sample sizes for both the treatment group and the comparison group for male participants. This includes the unweighted and weighted number of reoffenders in the comparison group, with the weighted numbers being used to calculate the reoffending rate in Table 2[21]. Where sample sizes are relatively small, they will be unlikely to produce a statistically significant result and therefore have a lower likelihood of supporting conclusions with an acceptable level of confidence.

Tables 2-4 show the two-year measures for proven reoffending for both the treatment group and the comparison group. Rates are expressed as percentages and frequencies expressed per person. Effect sizes (expressed as Cohen's *d* statistic) are also included to indicate the strength of the relationship between the two groups. The estimated differences shown are the 95% confidence intervals for the differences between the relevant treatment and comparison group measures.

The profile and descriptive characteristics of the male treatment group can be found in Annexes 11 and 14 respectively.

[21] The calculated reoffending rate uses the weighted values for each person and therefore does not necessarily correspond to the unweighted figures.

Table 1. Sample sizes for male cohort after matching for two-year reoffending analyses.

Analyses	Treatment group size	Comparison group size	Reoffenders in treatment group	Reoffenders in comparison group (weighted number)
Overall	18,555	345,084	8,631	209,138 (166,380)
Participants who met ideal suitability criteria	12,787	147,505	6,870	117,193 (82,808)
Participants who did not meet ideal suitability criteria	5,758	190,778	1,755	87,395 (58,353)
Completed TSP	9,972	175,259	4,971	93,936 (90,282)
Did not complete TSP	917	248,740	595	147,822 (155,783)
Programme integrity broadly maintained 2016-19	2,188	35,571	929	19,498 (15,969)
Programme integrity compromised 2016-19	559	8,021	268	4,428 (4,011)
With OGRS3 risk score 25-49 (low risk)	2,753	41,508	720	14,581 (11,406)
With OGRS3 risk score 50-74 (medium risk)	8,146	83,183	3,665	50,919 (38,979)
With OGRS3 risk score 75+ (high risk)	5,958	114,550	3,893	97,118 (77,684)
Index offence is a sexual offence	1,258	8,000	295	1,440 (1,866)
Index offence is an OASys Violence Predictor (OVP) offence	9,385	131,704	4,074	77,029 (59,633)
Index offence is an acquisitive offence	2,668	88,229	1,975	71,819 (66,588)
Participated in TSP only	16,261	309,826	7,523	185,435 (148,453)
Participated in another accredited programme prior to TSP	2,240	16,607	1,089	10,534 (8,214)
Asian and Asian British ethnicity	1,039	10,707	414	4,075 (4,344)
Black, Black British, Caribbean, and African ethnicity	2,208	20,503	1,052	10,551 (9,808)
Mixed and multiple ethnic groups	830	4,841	411	2,676 (2,508)
White	14,005	269,405	6,549	167,248 (131,078)

Analyses	Treatment group size	Comparison group size	Reoffenders in treatment group	Reoffenders in comparison group (weighted number)
More likely to present with characteristics associated with learning disabilities and challenges (LDC)	3,805	52,427	2,107	39,649 (29,643)
Less likely to present with characteristics associated with learning disabilities and challenges (LDC)	11,302	182,029	4,997	108,088 (83,837)
Aged between 18-25	6,708	97,760	3,813	62,556 (56,622)
Aged between 26-30	4,220	54,881	1,913	33,023 (26,645)
Aged between 31-49	6,504	143,696	2,677	88,439 (60,534)
Aged 50+	974	10,757	196	2,380 (2,106)

Results Summary

Statistically significant results of the male general reoffending measures

This table shows there were 42 statistically significant results among the 75 analyses. These provide evidence that:

Overall

- **Participants are less likely to commit a general reoffence** than non-participants.
- **Participants commit fewer general reoffences** than non-participants.
- **Participants who reoffend within a two-year period commit their first proven reoffence later** than non-participants.

Met the ideal suitability criteria

- **Participants are less likely to commit a general reoffence** than non-participants.
- **Participants commit fewer general reoffences** than non-participants.
- **Participants who reoffend within a two-year period commit their first proven reoffence later** than non-participants.

Did not meet the ideal suitability criteria

- **Participants commit fewer general reoffences** than non-participants.

Completed TSP

- **Participants are less likely to commit a general reoffence** than non-participants.
- **Participants commit fewer general reoffences** than non-participants.
- **Participants who reoffend within a two-year period commit their first proven reoffence later** than non-participants.

Programme integrity broadly maintained 2016-19

- **Participants are less likely to commit a general reoffence** than non-participants.
- **Participants commit fewer general reoffences** than non-participants.
- **Participants who reoffend within a two-year period commit their first proven reoffence later** than non-participants.

With OGRS3 risk score 50-74 (medium risk)

- **Participants are less likely to commit a general reoffence** than non-participants.
- **Participants commit fewer general reoffences** than non-participants.
- **Participants who reoffend within a two-year period commit their first proven reoffence later** than non-participants.

With OGRS3 risk score 75+ (high risk)

- **Participants are less likely to commit a general reoffence** than non-participants.
- **Participants commit fewer general reoffences** than non-participants.
- **Participants who reoffend within a two-year period commit their first proven reoffence later** than non-participants.

Index offence is an OASys Violence Predictor (OVP) offence

- **Participants are less likely to commit a general reoffence** than non-participants.
- **Participants commit fewer general reoffences** than non-participants.
- **Participants who reoffend within a two-year period commit their first proven reoffence later** than non-participants.

Index offence is an acquisitive offence

- **Participants commit fewer general reoffences** than non-participants.
- **Participants who reoffend within a two-year period commit their first proven reoffence later** than non-participants.

Participated in TSP only

- **Participants are less likely to commit a general reoffence** than non-participants.
- **Participants commit fewer general reoffences** than non-participants.
- **Participants who reoffend within a two-year period commit their first proven reoffence later** than non-participants.

Participated in other accredited programmes

- **Participants commit fewer general reoffences** than non-participants.

Black, Black British, Caribbean, and African ethnicity

- **Participants commit fewer general reoffences** than non-participants.
- **Participants who reoffend within a two-year period commit their first proven reoffence later** than non-participants.

White

- **Participants are less likely to commit a general reoffence** than non-participants.
- **Participants commit fewer general reoffences** than non-participants.
- **Participants who reoffend within a two-year period commit their first proven reoffence later** than non-participants.

More likely to present with characteristics associated with learning disabilities and challenges (LDC)

- **Participants commit fewer general reoffences** than non-participants.
- **Participants who reoffend within a two-year period commit their first proven reoffence later** than non-participants.

Less likely to present with characteristics associated with learning disabilities and challenges (LDC)

- **Participants are less likely to commit a general reoffence** than non-participants.
- **Participants commit fewer general reoffences** than non-participants.

Aged between 18 and 25

- **Participants commit fewer general reoffences** than non-participants.

Aged between 26 and 30

- **Participants are less likely to commit a general reoffence** than non-participants.
- **Participants commit fewer general reoffences** than non-participants.

Aged between 31 and 49

- **Participants commit fewer general reoffences** than non-participants.
- **Participants who reoffend within a two-year period commit their first proven reoffence later** than non-participants.

Table 2. Proportion of males who committed a proven general reoffence in a two-year period after support from the TSP, compared with matched comparison groups.

			Two-year proven general reoffending rates for males					
Analyses	**Number in treatment group**	**Number in comparison group**	**Treatment group rate (%)**	**Comparison group rate (%)**	**Estimated difference (% points)**	**Standardised effect size (Cohen's d)**	**Statistically significant difference?**	**p-value**
Overall	18,555	345,084	46.5	48.2	-2.4 to -1	-0.034	Yes	<0.01
Participants who met ideal suitability criteria	12,787	147,505	53.7	56.1	-3.3 to -1.5	-0.049	Yes	<0.01
Participants who did not meet ideal suitability criteria	5,758	190,778	30.5	30.6	-1.3 to 1.1	-0.002	No	0.86
Completed TSP	9,972	175,259	49.8	51.5	-2.7 to -0.7	-0.033	Yes	<0.01
Did not complete TSP	917	248,740	64.9	62.6	-0.8 to 5.4	0.047	No	0.15
Programme integrity broadly maintained 2016-19	2,188	35,571	42.5	44.9	-4.6 to -0.3	-0.049	Yes	0.03
Programme integrity compromised 2016-19	559	8,021	47.9	50.0	-6.4 to 2.2	-0.041	No	0.35
With OGRS3 risk score 25-49 (low risk)	2,753	41,508	26.2	27.5	-3 to 0.4	-0.030	No	0.13
With OGRS3 risk score 50-74 (medium risk)	8,146	83,183	45.0	46.9	-3 to -0.7	-0.037	Yes	<0.01
With OGRS3 risk score 75+ (high risk)	5,958	114,550	65.3	67.8	-3.7 to -1.2	-0.107	Yes	<0.01
Index offence is a sexual offence	1,258	8,000	23.4	23.3	-2.4 to 2.6	-0.119	No	0.92
Index offence is an OASys Violence Predictor (OVP) offence	9,385	131,704	43.4	45.3	-2.9 to -0.8	-0.137	Yes	<0.01

			Two-year proven general reoffending rates for males					
Analyses	**Number in treatment group**	**Number in comparison group**	**Treatment group rate (%)**	**Comparison group rate (%)**	**Estimated difference (% points)**	**Standardised effect size (Cohen's d)**	**Statistically significant difference?**	**p-value**
Index offence is an acquisitive offence	2,668	88,229	74.0	75.5	-3.1 to 0.2	-0.034	No	0.09
Participated in TSP only	16,261	309,826	46.3	47.9	-2.4 to -0.9	-0.033	Yes	<0.01
Participated in another accredited programme prior to TSP	2,240	16,607	48.6	49.5	-3.1 to 1.4	-0.017	No	0.45
Asian and Asian British	1,039	10,707	39.8	40.6	-3.9 to 2.4	-0.015	No	0.65
Black, Black British, Caribbean, and African	2,208	20,503	47.6	47.8	-2.4 to 2	-0.004	No	0.86
Mixed and multiple ethnic groups	830	4,841	49.5	51.8	-6 to 1.4	-0.046	No	0.22
White	14,005	269,405	46.8	48.7	-2.7 to -1	-0.038	Yes	<0.01
More likely to present with characteristics associated with learning disabilities and challenges (LDC)	3,805	52,427	55.4	56.5	-2.8 to 0.5	-0.024	No	0.16
Less likely to present with characteristics associated with learning disabilities and challenges (LDC)	11,302	182,029	44.2	46.1	-2.8 to -0.9	-0.037	Yes	<0.01
Aged between 18-25	6,708	97,760	56.8	57.9	-2.3 to 0.1	-0.022	No	0.08
Aged between 26-30	4,220	54,881	45.3	48.6	-4.8 to -1.7	-0.065	Yes	<0.01
Aged between 31-49	6,504	143,696	41.2	42.1	-2.2 to 0.3	-0.020	No	0.12
Aged 50+	974	10,757	20.1	19.6	-2.1 to 3.2	0.014	No	0.68

Table 3. Frequency of proven general reoffences for males in a two-year period after support from the TSP, compared with matched comparison groups.

	Two-year proven general reoffending frequencies (offences per person) for males							
Analyses	Number in treatment group	Number in comparison group	Treatment group frequency	Comparison group frequency	Estimated difference	Standardised effect size (Cohen's d)	Statistically significant difference?	p-value
Overall	18,555	345,084	1.75	2.00	-0.3 to -0.2	-0.072	Yes	<0.01
Participants who met ideal suitability criteria	12,787	147,505	2.10	2.38	-0.34 to -0.21	-0.074	Yes	<0.01
Participants who did not meet ideal suitability criteria	5,758	190,778	0.95	1.03	-0.15 to -0.02	-0.034	Yes	<0.01
Completed TSP	9,972	175,259	1.86	2.15	-0.36 to -0.23	-0.083	Yes	<0.01
Did not complete TSP	917	248,740	2.97	3.01	-0.35 to 0.26	-0.010	No	0.77
Programme integrity broadly maintained 2016-19	2,188	35,571	1.65	1.81	-0.3 to -0.02	-0.048	Yes	0.03
Programme integrity compromised 2016-19	559	8,021	1.83	2.04	-0.5 to 0.08	-0.059	No	0.16
With OGRS3 risk score 25-49 (low risk)	2,753	41,508	0.72	0.76	-0.11 to 0.03	-0.020	No	0.29
With OGRS3 risk score 50-74 (medium risk)	8,146	83,183	1.49	1.62	-0.2 to -0.08	-0.049	Yes	<0.01
With OGRS3 risk score 75+ (high risk)	5,958	114,550	2.92	3.37	-0.56 to -0.33	-0.224	Yes	<0.01
Index offence is a sexual offence	1,258	8,000	0.60	0.64	-0.15 to 0.06	-0.126	No	0.41
Index offence is an OASys Violence Predictor (OVP) offence	9,385	131,704	1.50	1.67	-0.23 to -0.11	-0.237	Yes	<0.01
Index offence is an acquisitive offence	2,668	88,229	3.38	3.91	-0.7 to -0.36	-0.103	Yes	<0.01

			Two-year proven general reoffending frequencies (offences per person) for males					
Analyses	**Number in treatment group**	**Number in comparison group**	**Treatment group frequency**	**Comparison group frequency**	**Estimated difference**	**Standardised effect size (Cohen's d)**	**Statistically significant difference?**	**p-value**
Participated in TSP only	16,261	309,826	1.72	1.97	-0.3 to -0.2	-0.072	Yes	<0.01
Participated in another accredited programme prior to TSP	2,240	16,607	1.93	2.15	-0.37 to -0.06	-0.060	Yes	<0.01
Asian and Asian British	1,039	10,707	1.33	1.36	-0.18 to 0.14	-0.008	No	0.80
Black, Black British, Caribbean, and African	2,208	20,503	1.51	1.69	-0.3 to -0.07	-0.064	Yes	<0.01
Mixed and multiple ethnic groups	830	4,841	1.85	1.88	-0.26 to 0.2	-0.009	No	0.82
White	14,005	269,405	1.82	2.11	-0.35 to -0.23	-0.079	Yes	<0.01
More likely to present with characteristics associated with learning disabilities and challenges (LDC)	3,805	52,427	2.28	2.54	-0.39 to -0.13	-0.064	Yes	<0.01
Less likely to present with characteristics associated with learning disabilities and challenges (LDC)	11,302	182,029	1.61	1.86	-0.31 to -0.19	-0.074	Yes	<0.01
Aged between 18-25	6,708	97,760	2.15	2.29	-0.23 to -0.07	-0.043	Yes	<0.01
Aged between 26-30	4,220	54,881	1.76	2.02	-0.37 to -0.15	-0.070	Yes	<0.01
Aged between 31-49	6,504	143,696	1.52	1.79	-0.34 to -0.18	-0.076	Yes	<0.01
Aged 50+	974	10,757	0.62	0.67	-0.21 to 0.11	-0.021	No	0.53

Table 4. Average time to first proven general reoffence for males in a two-year period after support from the TSP, compared with matched comparison groups.

	Average time to first proven general reoffence for males in a two-year period, for reoffenders only (days)							
Analyses	**Number in treatment group**	**Number in comparison group**	**Treatment group time**	**Comparison group time**	**Estimated difference**	**Standardised effect size (Cohen's d)**	**Statistically significant difference?**	**p-value**
Overall	8,631	209,138	267	257	5 to 14	0.049	Yes	<0.01
Participants who met ideal suitability criteria	6,870	117,193	262	252	5 to 15	0.050	Yes	<0.01
Participants who did not meet ideal suitability criteria	1,755	87,395	284	283	-9 to 10	0.004	No	0.88
Completed TSP	4,971	93,936	262	250	7 to 18	0.062	Yes	<0.01
Did not complete TSP	595	147,822	233	228	-11 to 21	0.025	No	0.55
Programme integrity broadly maintained 2016-19	929	19,498	275	257	4 to 31	0.087	Yes	0.01
Programme integrity compromised 2016-19	268	4,428	280	256	-1 to 48	0.121	No	0.06
With OGRS3 risk score 25-49 (low risk)	720	14,581	291	300	-24 to 7	-0.043	No	0.26
With OGRS3 risk score 50-74 (medium risk)	3,665	50,919	286	277	3 to 16	0.046	Yes	<0.01
With OGRS3 risk score 75+ (high risk)	3,893	97,118	240	229	5 to 18	0.042	Yes	<0.01
Index offence is a sexual offence	295	1,440	273	288	-40 to 11	0.068	No	0.26
Index offence is an OASys Violence Predictor (OVP) offence	4,074	77,029	288	281	0 to 13	0.016	Yes	0.04

			Average time to first proven general reoffence for males in a two-year period, for reoffenders only (days)					
Analyses	Number in treatment group	Number in comparison group	Treatment group time	Comparison group time	Estimated difference	Standardised effect size (Cohen's d)	Statistically significant difference?	p-value
Index offence is an acquisitive offence	1,975	71,819	212	201	3 to 19	0.106	Yes	<0.01
Participated in TSP only	7,523	185,435	268	255	8 to 18	0.066	Yes	<0.01
Participated in another accredited programme prior to TSP	1,089	10,534	253	259	-19 to 7	-0.030	No	0.35
Asian and Asian British	414	4,075	300	289	-9 to 30	0.055	No	0.28
Black, Black British, Caribbean, and African	1,052	10,551	283	270	1 to 25	0.066	Yes	0.04
Mixed and multiple ethnic groups	411	2,676	276	265	-9 to 31	0.056	No	0.29
White	6,549	167,248	261	251	5 to 15	0.049	Yes	<0.01
More likely to present with characteristics associated with learning disabilities and challenges (LDC)	2,107	39,649	250	241	0 to 18	0.044	Yes	0.05
Less likely to present with characteristics associated with learning disabilities and challenges (LDC)	4,997	108,088	269	264	0 to 11	0.027	No	0.06
Aged between 18-25	3,813	62,556	264	258	0 to 13	0.032	No	0.05
Aged between 26-30	1,913	33,023	269	262	-2 to 16	0.035	No	0.13
Aged between 31-49	2,677	88,439	269	257	5 to 21	0.063	Yes	<0.01
Aged 50+	196	2,380	252	244	-22 to 37	0.037	No	0.62

Female results in detail

Table 5 presents the sample sizes for both the treatment group and the comparison group for female participants. This includes the unweighted and weighted number of reoffenders in the comparison group, with the weighted numbers being used to calculate the reoffending rate in Table 6. Where sample sizes are relatively small, they will be unlikely to produce a statistically significant result and therefore have a lower likelihood of supporting conclusions with an acceptable level of confidence. Due to smaller sample sizes, there are fewer sub-analyses conducted for the female cohort.

Tables 6-8 show the two-year measures for proven reoffending for both the treatment group and the comparison group. Rates are expressed as percentages and frequencies expressed per person. Effect sizes (expressed as Cohen's *d* statistic) are also included to indicate the strength of the relationship between the two groups. The estimated differences shown are the 95% confidence intervals for the differences between the relevant treatment and comparison group measures.

The profile and descriptive statistics of the female treatment group included in the female headline analysis can be found in Annexes 12 and 14, respectively.

Table 5: Sample sizes for female cohort after matching for two-year reoffending analyses.

Analyses	Treatment group size	Comparison group size	Reoffenders in treatment group	Reoffenders in comparison group (weighted number)
Overall	1,738	30,563	741	19,583 (13,504)
Participants who met ideal suitability criteria	1,166	13,313	568	11,138 (6,920)
Participants who did not meet ideal suitability criteria	563	16,976	169	8,336 (5,494)
Completed TSP	940	14,087	416	8,292 (6,491)
Index offence is an OASys Violence Predictor (OVP) offence	1,065	8,425	365	4,436 (3,088)
Participated in TSP only	1,510	16,294	652	9,633 (7,247)

Results Summary

Statistically significant results of the female general reoffending measures.

This table shows there were 7 statistically significant results among the 18 analyses. These provide evidence that:

<u>**Overall**</u>

- **Participants commit fewer general reoffences than non-participants.**

<u>**Participants who met the ideal suitability criteria**</u>

- **Participants are less likely to commit a general offence** than non-participants.
- **Participants commit fewer general reoffences** than non-participants.

<u>**Participants who did not meet the ideal suitability criteria**</u>

- **Participants commit fewer general reoffences** than non-participants.

<u>**Participants who completed TSP**</u>

- **Participants commit fewer general reoffences** than non-participants.

<u>**Index offence is a OASys Violence Predictor (OVP) offence**</u>

- **Participants commit fewer general reoffences** than non-participants.

<u>**Participants who participated in TSP only**</u>

- **Participants commit their first proven reoffence later** than non-participants.

Table 6. Proportion of females who committed a proven general reoffence in a two-year period after support from the TSP, compared with matched comparison groups.

			Two-year proven general reoffending rates for females					
Analyses	**Number in treatment group**	**Number in comparison group**	**Treatment group rate (%)**	**Comparison group rate (%)**	**Estimated difference (% points)**	**Standardised effect size (Cohen's d)**	**Statistically significant difference?**	**p-value**
Overall	1,738	30,563	42.6	44.2	-3.9 to 0.8	-0.031	No	0.20
Participants who met ideal suitability criteria	1,166	13,313	48.7	52.0	-6.3 to -0.3	-0.065	Yes	0.03
Participants who did not meet ideal suitability criteria	563	16,976	30.0	32.4	-6.2 to 1.5	-0.051	No	0.23
Completed TSP	940	14,087	44.3	46.1	-5.1 to 1.5	-0.037	No	0.28
Index offence is an OASys Violence Predictor (OVP) offence	1,065	8,425	34.3	36.7	-5.4 to 0.6	-0.050	No	0.12
Participated in TSP only	1,510	16,294	43.2	44.5	-3.9 to 1.3	-0.026	No	0.33

Table 7. Frequency of proven general reoffences for females in a two-year period after support from the TSP, compared with matched comparison groups.

			Two-year proven general reoffending frequencies (offences per person) for females					
Analyses	Number in treatment group	Number in comparison group	Treatment group frequency	Comparison group frequency	Estimated difference	Standardised effect size (Cohen's d)	Statistically significant difference?	p-value
Overall	1,738	30,563	2.17	2.43	-0.49 to -0.04	-0.056	Yes	0.02
Participants who met ideal suitability criteria	1,166	13,313	2.67	3.05	-0.68 to -0.06	-0.071	Yes	0.02
Participants who did not meet ideal suitability criteria	563	16,976	1.14	1.46	-0.6 to -0.05	-0.094	Yes	0.02
Completed TSP	940	14,087	2.10	2.51	-0.7 to -0.1	-0.086	Yes	<0.01
Index offence is an OASys Violence Predictor (OVP) offence	1,065	8,425	1.35	1.58	-0.44 to -0.01	-0.066	Yes	0.04
Participated in TSP only	1,510	16,294	2.18	2.39	-0.46 to 0.05	-0.044	No	0.11

Table 8. Average time to first proven general reoffence for females in a two-year period after support from the TSP, compared with matched comparison groups.

			Average time to first proven general reoffence for females in a two-year period, for reoffenders only (days)					
Analyses	**Number in treatment group**	**Number in comparison group**	**Treatment group time**	**Comparison group time**	**Estimated difference**	**Standardised effect size (Cohen's d)**	**Statistically significant difference?**	**p-value**
Overall	741	19,583	262	247	-1 to 30	0.069	No	0.07
Participants who met ideal suitability criteria	568	11,138	250	238	-5 to 30	0.061	No	0.17
Participants who did not meet ideal suitability criteria	169	8,336	300	278	-11 to 55	0.103	No	0.19
Completed TSP	416	8,292	268	258	-11 to 32	0.048	No	0.35
Index offence is an OASys Violence Predictor (OVP) offence	365	4,436	285	279	-18 to 29	0.025	No	0.65
Participated in TSP only	652	9,633	267	247	3 to 37	0.096	Yes	0.02

Acknowledgements

From the Ministry of Justice, the authors would like to thank Lucy Cuppleditch, Alana Diamond, James Teasdale, Johannes Huber, Samuel Hales, and Aiden Mews.

The authors would also like to thank members of the CSAAP (Correctional Services Accreditation and Advice Panel) who provided advice for this evaluation, and to Ruth Hatcher and Ralph Serin who peer reviewed the report.

Final thanks are to HMPPS Intervention Services for their communication and support throughout the evaluation process.

Contact points

Press enquiries should be directed to the Ministry of Justice press office. Other enquiries about the analysis should be directed to:

Justice Data Lab and **Reducing Reoffending Analytical Priority Projects** teams

Ministry of Justice

10th Floor

102 Petty France

London

SW1H 9AJ

E-mail: justice.datalab@justice.gov.uk

General enquiries about the statistical work of the Ministry of Justice can be e-mailed to: statistics.enquiries@justice.gov.uk

General information about the official statistics system of the United Kingdom is available from: https://uksa.statisticsauthority.gov.uk/about-the-authority/

Produced by the Ministry of Justice

Alternative formats are available on request from: justice.datalab@justice.gov.uk

Annexes

Annex 1: What you can say about the results

Male results

✓ What you can say about the two-year general reoffending measures for males:

“This analysis **provides evidence** that support from the TSP may **decrease the number of proven male reoffenders** during a two-year period.”

“This analysis **provides evidence** that support from the TSP may **decrease the number of proven reoffences** committed by its male participants during a two-year period.”

“This analysis **provides evidence** that support from the TSP may **lengthen the average time to first proven reoffence** for its male participants during a two-year period.”

✗ What you cannot say about the two-year general reoffending measures for males:

“This analysis provides evidence that support from the TSP increases/has no effect on the **reoffending rate for male participants** during a two-year period.”

“This analysis provides evidence that support from the TSP increases/has no effect on the **number of proven reoffences** committed by its male participants during a two-year period.”

“This analysis provides evidence that support from the TSP shortens/has no effect on the **average time to first proven reoffence** for its male participants during a two-year period.”

Female results

✓ What you can say about the two-year general reoffending measures for females:

“This analysis **does not provide clear evidence** on whether support from the TSP increases or decreases **the number of female participants who commit a proven reoffence** in a two-year period”

“This analysis **provides evidence** that support from the TSP may **decrease the number of proven reoffences** committed by its female participants during a two-year period.”

“This analysis **does not provide clear evidence** on whether support from the TSP shortens or lengthens **the average time to first proven reoffence** for its female participants during a two-year period.”

✗ **What you cannot say about the two-year general reoffending measures for females:**

“This analysis provides evidence that support from the TSP increases/decreases/has no effect on **the number of proven female reoffenders** during a two-year period.”

“This analysis provides evidence that support from the TSP increases/has no effect on the **number of proven reoffences** committed by its female participants during a two-year period.”

“This analysis provides evidence that support from the TSP lengthens/shortens/has no effect on the **average time to first proven reoffence** for its female participants during a two-year period.”

Annex 2: Description of CSAAP

The Correctional Services Accreditation and Advice Panel (CSAAP) comprises independent international academics and expert practitioners who advise HMPPS on accrediting programmes for use across prisons and probation. CSAAP also provide independent, evidence-based advice on programme development and practice. The Ministry of Justice uses accreditation to provide confidence that its offending behaviour programmes are designed based on the best available evidence, will be delivered as intended, and will be evaluated to show the outcomes that are being met. The HMPPS Rehabilitation Strategy Board accredit programmes for implementation across prisons and probation.

Once an accredited programme has been running for a sufficient amount of time, CSAAP considers the impact of the programme when deciding whether to recommend that the programme maintains accreditation. If CSAAP do not recommend that the programme maintains accreditation, HMPPS may consider withdrawing the programme.

Programmes are assessed using the evidence-based principles for effective interventions. The Accreditation Criteria are laid out below.

The requirements for accreditation state that programmes and services must demonstrate that they:

1. Are evidence-based and/or have a credible rationale

2. Address factors relevant to reoffending and desistance

3. Are targeted at appropriate participants

4. Develop new skills (as opposed to only raising awareness)

5. Motivate, engage, and retain participants

6. Are delivered as intended by staff with appropriate skills and quality assured,

via:

  a. a quality assurance plan,
  b. by providing quality assurance findings

7. Are evaluated, via:

  a. an evaluation plan, and
  b. by providing results of evaluation

Annex 3: Details of matching criteria

Below is a table of variables that were used for propensity score matching (PSM). The name of each variable, its type and categories are shown.

Table A3.1: Matching variables used in propensity score matching model.

Variable	Type	Categories
	Demographics	
Ethnicity (self-reported)	Categorical	Asian and Asian British; Black, Black British, Caribbean, and African; Mixed and multiple ethnic groups; White; Unknown
UK Nationality	Categorical	UK; Non-UK; Unknown
Age at index (release) date	Continuous (integer)	
	Criminal history	
Age at first contact with criminal justice system	Continuous (integer)	-
Primary index offence group	Categorical	Violence against the person; Sexual offences; Robbery; Theft offences; Criminal damage and arson; Drug offences; Possession of weapons; Public order offences; Miscellaneous crimes against society; Fraud offences; Summary offences excluding motoring; Summary motoring offences; Unknown
Primary index offence severity	Categorical	Indictable only; Triable either way; Summary only
Index custodial sentence length	Categorical	Less than or equal to 6 months; More than 6 months to less than 12 months; 12 months to less than 4 years; 4 years to 10 years; More than 10 years; Mandatory Life sentence; Other Life sentence; Imprisonment for Public Protection
Year of release from prison from index offence	Categorical	2010; 2011; 2012; 2013; 2014; 2015; 2016; 2018; 2019; 2020

Number of previous prison events	Continuous (integer)	-
Number of previous convictions	Continuous (integer)	-
Number of previous court orders	Continuous (integer)	-
Number of previous offences	Continuous (integer)	-
Number of previous indictable only offence	Continuous (integer)	-
Number of previous triable either way offences	Continuous (integer)	-
Number of previous summary offences	Continuous (integer)	-
Number of previous violent offences	Continuous (integer)	-
Number of previous robbery offences	Continuous (integer)	-
Number of previous public order offences	Continuous (integer)	-
Number of previous domestic burglary offences	Continuous (integer)	-
Number of previous other burglary offences	Continuous (integer)	-
Number of previous theft offences	Continuous (integer)	-
Number of previous handling offences	Continuous (integer)	-
Number of previous fraud or forgery offences	Continuous (integer)	-
Number of previous theft of vehicles offences	Continuous (integer)	-
Number of previous drink driving offences	Continuous (integer)	-
Number of previous criminal damage offences	Continuous (integer)	-
Number of previous drug import/export/production/supply offences	Continuous (integer)	-
Number of previous drug possession or supply offences	Continuous (integer)	-
Number of previous sexual offences	Continuous (integer)	-
Number of previous breach offences	Continuous (integer)	-
Copas rate (logarithmic rate of convictions and cautions over time)	Continuous (numerical)	-
Employment and benefits		
Any Pay As You Earn (PAYE) employment within one month before conviction	Categorical	Unknown; No; Yes

Any PAYE employment within one year before conviction	Categorical	Unknown; No; Yes
Any out-of-work benefits received within one year before conviction	Categorical	Unknown; No; Yes
Any Job Seeker's Allowance received within one year before conviction	Categorical	Unknown; No; Yes
Any Incapacity Benefit or Income Support received within one year before conviction	Categorical	Unknown; No; Yes
Accredited Programmes		
Year of participation in TSP (start date)	Categorical	2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017: 2018; 2019
Any other Accredited Programme taken during the same sentence, prior to starting TSP	Categorical (binary)	No; Yes
OASys Assessment (between 12 months before and 1 month after starting TSP)		
OVP Score	Continuous (integer)	For the purposes of matching, these have been banded as follows: 0-9; 10-19; 20-29; 30-39; 40-49; 50-59; 60-69; 70-79; 80-89; 90-100; Unknown
OGRS3 Score	Continuous (integer)	For the purposes of matching, these have been banded as follows: 0-9; 10-19; 20-29; 30-39; 40-49; 50-59; 60-69; 70-79; 80-89; 90-100; Unknown
Does the offender have either reading, writing, or numeracy problems?	Categorical	Unknown; None; Some; Severe
Does the offender have problems with numeracy?	Categorical	Unknown; No; Some; Significant
Does the offender have problems with reading?	Categorical	Unknown; No; Some; Significant
Does the offender have problems with writing?	Categorical	Unknown; No; Some; Significant
Does the offender have learning difficulties?	Categorical	Unknown; No; Some; Significant
Does the offender recognise the impact and consequences of offending on their victim /community/wider society?	Categorical	Unknown; No; Some; Significant
Does the offender currently have a permanent place of accommodation?	Categorical	Unknown; No; Some; Significant

Is the offender unemployed, or will they be unemployed on release?	Categorical	Unknown; No; Some; Significant
Does the offender have any problems with their financial situation?	Categorical	Unknown; No; Some; Significant
Does the offender currently have a relationship with their close family members?	Categorical	Unknown; No; Some; Significant
Is there evidence that the offender has even been a victim of domestic violence/partner abuse?	Categorical	Unknown; No; Yes
Is there evidence that the offender has ever been a perpetrator of domestic violence/partner abuse?	Categorical	Unknown; No; Yes
Does the offender's regular activities encourage offending?	Categorical	Unknown; No; Some; Significant
Is the offender easily influenced by criminal associates?	Categorical	Unknown; No; Some; Significant
Does the offender have a manipulative or predatory lifestyle?	Categorical	Unknown; No; Some; Significant
Does the offender engage in recklessness and risk-taking behaviour?	Categorical	Unknown; No; Yes
Does the offender have drugs misuse issues that are linked to their offending behaviour?	Categorical	Unknown; No; Yes
Does the offender have drugs misuse issues that are linked to their risk of serious harm, risks to the individual, and other risks?	Categorical	Unknown; No; Yes
Has the offender ever misused drugs, either in custody or the community?	Categorical	Unknown; No; Some; Significant
Is the offender motivated to tackle their drug misuse?	Categorical	Unknown; No; Some; Significant
Are drug use or obtaining drugs a major activity or occupation for the offender?	Categorical	Unknown; No; Some; Significant
Does the offender have alcohol misuse issues that are linked to their offending behaviour?	Categorical	Unknown; No; Yes
Does the offender have alcohol misuse issues that are linked to their risk of serious harm, risks to the individual, and other risks?	Categorical	Unknown; No; Yes

Does the offender currently have issues with alcohol?	Categorical	Unknown; No; Some; Significant
Has the offender engaged in binge drinking or excessive use of alcohol in the last 6 months?	Categorical	Unknown; No; Some; Significant
Has the offender frequently or seriously misused alcohol in the past?	Categorical	Unknown; No; Some; Significant
Does the offender have a history of violent behaviour related to alcohol use at any time?	Categorical	Unknown; No; Some; Significant
Is the offender motivated to tackle their alcohol misuse?	Categorical	Unknown; No; Some; Significant
Does the offender have difficulties coping with everyday life?	Categorical	Unknown; No; Some; Significant
Does the offender currently have psychological problems, including depression?	Categorical	Unknown; No; Some; Significant
Does the offender self harm, have they attempted suicide, or do they possess suicidal thoughts or feelings?	Categorical	Unknown; No; Some; Significant
Does the offender currently have psychiatric problems?	Categorical	Unknown; No; Some; Significant
What level of interpersonal skills does the offender possess?	Categorical	Unknown; No; Some; Significant
Does the offender have issues with impulsivity?	Categorical	Unknown; No; Some; Significant
Does the offender demonstrate aggressive or controlling behaviour?	Categorical	Unknown; No; Some; Significant
Can the offender appropriately control their temper?	Categorical	Unknown; No; Some; Significant
Does the offender possess the ability to recognise problems?	Categorical	Unknown; No; Yes
Does the offender possess appropriate problem solving skills?	Categorical	Unknown; No; Some; Significant
Is the offender aware of the consequences of their actions?	Categorical	Unknown; No; Some; Significant
Is the offender able to understand other people's point of view?	Categorical	Unknown; No; Some; Significant
Is the offender able to engage in concrete/abstract thinking?	Categorical	Unknown; No; Some; Significant
Does the offender possess pro-criminal or offence-supportive attitudes?	Categorical	Unknown; No; Some; Significant

Does the offender have positive attitudes towards staff?	Categorical	Unknown; No; Some; Significant
Does the offender have positive attitudes towards supervision and/or their licence?	Categorical	Unknown; No; Some; Significant
Does the offender have positive attitudes towards their community and/or wider society?	Categorical	Unknown; No; Some; Significant
Does the offender understand their motivation for offending?	Categorical	Unknown; No; Some; Significant
Is the offender motivated to address their offending behaviour?	Categorical	Unknown; No; Some; Significant
Does the offender possess any physical or mental health conditions?	Categorical	Unknown; No; Some; Significant
Does the offender understand the importance of completing programmes?	Categorical	Unknown; No; Some; Significant
On the basis that they could be released imminently back into the community, what risk does the offender currently pose to children?	Categorical	Low; Medium; High; Very High; Unknown
On the basis that they could be released imminently back into the community, what risk does the offender currently pose to known adults?	Categorical	Low; Medium; High; Very High; Unknown
On the basis that they could be released imminently back into the community, what risk does the offender currently pose to the public?	Categorical	Low; Medium; High; Very High; Unknown

Annex 4: Methodological Approaches

This Justice Data Lab (JDL) study evaluates the reoffending patterns of a cohort of treated and comparison (untreated) offenders after their release from a prison sentence, to estimate the impact of the intervention on proven reoffending. The treatment group is comprised of those who started the TSP during a prison sentence. This includes participants where there was intention-to-treat (ITT)[22] but did not necessarily complete the full programme requirements. The comparison group includes those who did not attend (i.e., start) TSP during their sentence.

The pre-matched treatment group comprised 20,302 records (18,564 male records, 1,738 female records). Each record relates to a distinct sentence (not a distinct individual) and there are cases in both the treatment and comparison group where individuals had multiple 'eligible'[23] sentences and are therefore represented by more than one record. For a comprehensive explanation of how we dealt with those who participated in the TSP more than once, and those in the comparison group who had multiple 'eligible' sentences, please see Annex 13 – sentence selection methodology.

Pseudo-start dates

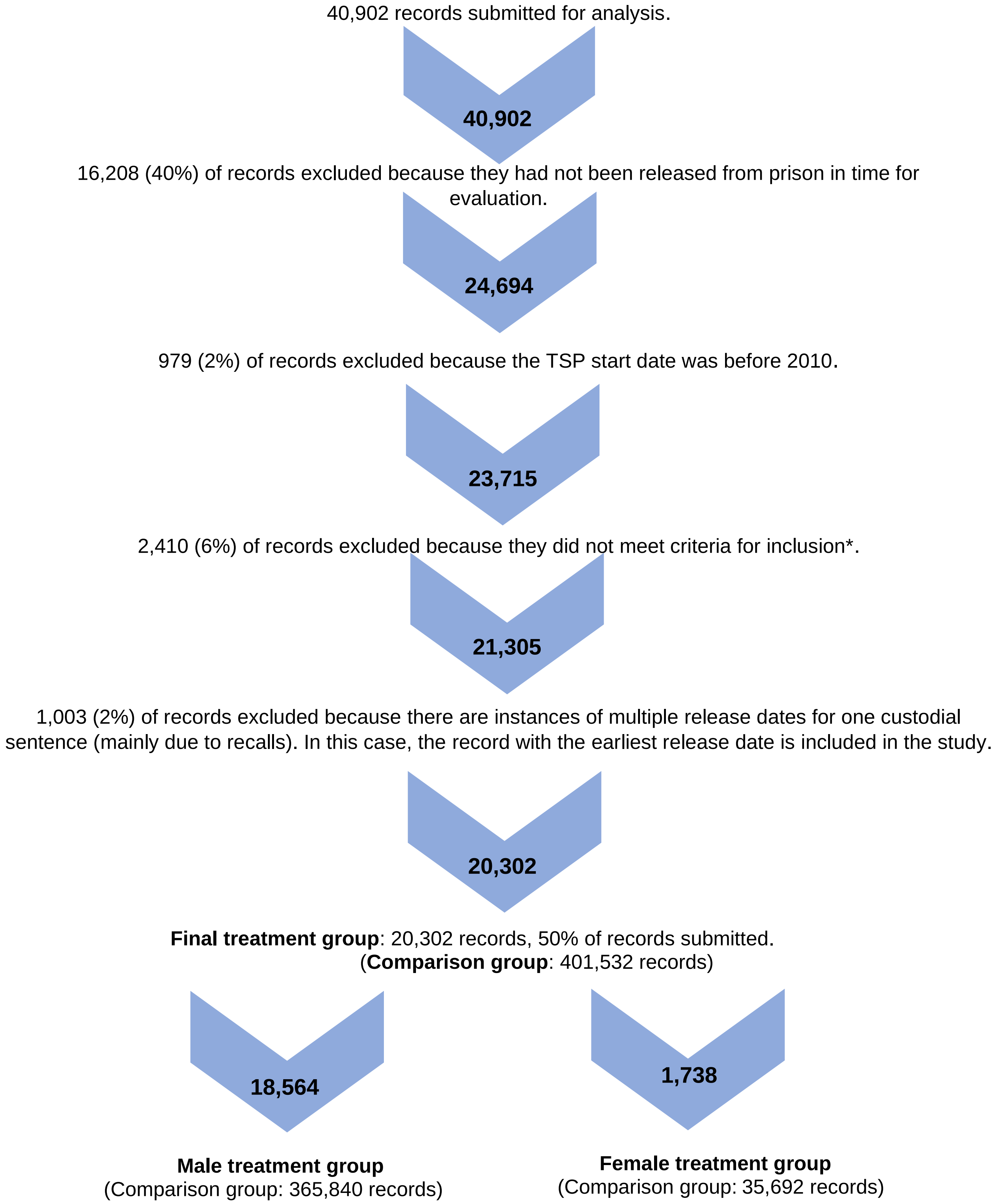
The date at which an individual in the treatment group started the TSP (the TSP start date) is an important variable which enables the extraction of the timeliest data from other sources (e.g., the OASys assessment/ prison population data closest to TSP participation). This data is readily available for those in the treatment group. The comparison group do not have a TSP start date, so a TSP pseudo-start date is imputed.

The imputation process involves an algorithm which utilises individual sentencing and demographic information to estimate a (pseudo) TSP start date for individuals in the comparison group. In other words, the hypothetical date at which an individual is predicted to have started the TSP if they had participated in the programme. The algorithm uses the treatment group as training data to create its predictions for the comparison group.

[22] Intention-to-treat analysis is a method for analysing results in a prospective randomized study where all participants who are randomized are included in the statistical analysis and analysed according to the group they were originally assigned, regardless of what treatment (if any) they received

[23] Eligible sentences are custodial sentences between 2010 and 2020, which meet the defined criteria for inclusion (age >= 18 years, OASys record 12 months before or 1 month after TSP start date, prison population record within 1 week of TSP start date).

Figure A4:1 Attrition from treatment group to create final pre-matched[24] cohort



40,902 records submitted for analysis.

16,208 (40%) of records excluded because they had not been released from prison in time for evaluation.

979 (2%) of records excluded because the TSP start date was before 2010.

2,410 (6%) of records excluded because they did not meet criteria for inclusion*.

1,003 (2%) of records excluded because there are instances of multiple release dates for one custodial sentence (mainly due to recalls). In this case, the record with the earliest release date is included in the study.

Final treatment group: 20,302 records, 50% of records submitted.
(**Comparison group**: 401,532 records)

Male treatment group
(Comparison group: 365,840 records)

Female treatment group
(Comparison group: 35,692 records)

*Age >=18 years, OASys record within 12m before 1m after the TSP start date, prison population record within 1 week window of the TSP start date.

[24] This chevron presents pre-matched figures as a unique PSM model was created for each headline (male/female) and each sub-analysis. Therefore, each analysis contained a different number of records.

Propensity score matching

Offenders in the treatment group were matched to untreated offenders using propensity score matching (PSM). PSM is a statistical matching technique which uses factors theoretically and empirically associated with both receiving the treatment and the outcome variable (reoffending) to predict a 'propensity score' (see Annex 3 for variables). This propensity score reflects the likelihood that an offender received the intervention, given the recorded characteristics. It is a value between 0 and 1. Treatment group members were matched to similar untreated offenders, where their propensity scores were within a certain tolerance level. Where several comparison group members had propensity scores within the required tolerance for a given treatment group member, the comparison group records all received the same weighting factor. For example, if 10 comparison records were matched to a single treatment group record, each comparison group record would have a weight of 1/10 applied, with the treatment group record having a weight of 1. Where treatment group records had no corresponding comparison group record within the tolerance level, they were excluded from the analysis (their weight was set to 0). Using the post-matched groups, the weighted reoffending rates for the treatment and comparison groups were compared. PSM can provide a robust quasi-experimental approach, although offenders can only be matched on observable variables. While extensive efforts were undertaken in identifying relevant factors, it is possible that unobserved factors could influence the results that emerge from this research.

Imputation of OASys variables

In statistics, imputation is the process of replacing missing data with substituted values. Imputation was used to deal with a small proportion of missing OASys data.

For the following variables, missing entries could sometimes be logically inferred:

- Section 6, Question 7: Evidence of domestic violence/partner abuse
- Section 6, Question 7: Evidence of domestic violence/partner abuse - Victim
- Section 6, Question 7: Evidence of domestic violence/partner abuse - Perpetrator
- Section 4, Question 7: Has problems with reading, writing or numeracy

In these cases, entries were logically imputed' based on corresponding OASys variables. E.g., if the individual has a missing entry for 'Section 6, Question 7: Evidence of domestic violence/partner abuse – Victim' but a '0 – no domestic violence' entry for 'Section 6, Question 7: Evidence of domestic violence/partner abuse', we can logically impute a '0 – no domestic violence' entry for the 'Section 6, Question 7: Evidence of domestic violence/partner abuse – Victim' variable.

Sensitivity Analyses

A series of sensitivity analyses were run on the male headline analysis, to measure the possible effect of certain methodological decisions on the results. It should be noted that the chosen method was selected on the basis that the model should include variables both theoretically and empirically associated with selection and outcome. Given other theoretical considerations, having a lower mean absolute standardised difference does not necessarily mean that the matching is better. The following table provides an explanation of selected sensitivity analyses and their results, with reference to the two-year general reoffending male headline analysis.

Table A4.1: Sensitivity analyses conducted on the male headline analysis and their findings.

Sensitivity	Explanation	Findings
Standard Approach	The chosen approach is displayed here for comparison against other sensitivities. Radius matching (with replacement) with a uniform kernel was applied.	206 variables were included in the final model with a mean absolute standardised difference of 0.57%.
Parsimonious	To explore the effect of having fewer variables in the model (tougher constraints imposed when determining model variables).	141 variables were included in the final model with a mean absolute standardised difference of 0.66%. Only a small number of OASys variables (53 compared to 107 in std approach) made it into the final model. The results and matching quality were very similar to the standard approach.
Non-parsimonious	To explore the effect of having more variables in the model (looser constraints imposed when determining model variables)	225 variables were included in the final model with a mean absolute standardised difference of 0.58%. The results and matching quality were very similar to the standard approach taken.
Common Support	To explore the effect of having a restriction on common support. Treatment group members were automatically excluded where their propensity scores were outside the overall range of propensity scores of the comparison group, and comparison group members were automatically excluded where their propensity scores were outside the	206 variables were included in the final model with a mean absolute standardised difference of 0.58%. Almost no difference to the number matched in the treatment group. Results very similar to the standard approach, showing any outliers

	overall range of propensity scores of the treatment group.	have been appropriately excluded in our matching process.
Epanechnikov Kernel	This explores using an alternative type of kernel matching sometimes used for PSM models.	206 variables were included in the final model with a mean absolute standardised difference of 0.58%. The matching quality and the results were very similar to the standard approach taken.
Matching on propensity scores	To explore the effect of matching on propensity scores rather than the logit of propensity scores.	206 variables were included in the final model with a mean absolute standardised difference of 0.61%. Matching quality and results were very similar to the standard approach.
Exclusion of OASys (including OVP and OGRS) variables from the model	To explore the effect of including OASys data in the model.	82 variables were included in the final model with a mean absolute standardised difference of 0.44%. The results were not significant for the binary reoffending rate at 5% significance level, but results for the other two measures were similar to the standard approach.

Annex 5: Limitations and caveats to our findings

Whilst this study uses a recognised evaluation methodology (propensity score matching), which is considered level 4 on the SMS (Scientific Methods Scale)[25], it is not as robust as a randomised control trial or a prospectively matched evaluation. For a detailed discussion of the strengths and limitations of propensity score matching, see Mews, Hillier, McHugh, & Coxon (2013), and Ministry of Justice (2015).

As such, there are several limitations and caveats that should be considered when observing the results of this study.

- While propensity score matching can provide a robust quasi-experimental approach, it can only match, and therefore reduce bias, on **observed factors** (information that is recorded). Despite efforts to include all observed factors known to be predictive of selection onto the TSP and of reoffending risk, the importance of information that is not recorded cannot be known. As such, it is possible that unobserved factors could influence these results.

- For the completers analysis, it is not possible to match on an observable "completion" counterfactual filter in the comparison group. As a result, we compared TSP completers to the entire comparison group. It is therefore possible that the analysis did not fully control for unobserved factors relating to the likelihood that an individual would complete the TSP if they were assigned to the intervention.

- These analyses only concern reoffending outcome data. Although outside the scope of Justice Data Lab analyses, there may be other important outcomes to consider for rehabilitation interventions. Examples of this might include increased employability, improved mental health, healthier relationships, or positive attitudes.

- A sub-analysis was performed to isolate the effect of the TSP from other accredited programmes. To ensure methodological robustness, we operationalised our 'participated in TSP only' subgroup as individuals who did not participate in another accredited programme **before** TSP during that index sentence. However, it is possible that those individuals participated in another accredited programme during the same sentence **after** participation in the TSP. The effect of the TSP and any effect of other accredited programmes would be hard to disentangle and not accounted for in these results. Moreover, it is possible that individuals in the comparison group (those who did not participate in the TSP during the index sentence) instead undertook a non-accredited cognitive behavioural programme during the index sentence, which might have had an impact on reoffending rates.

- Exploratory work for the 'exclusivity of TSP' sub-analysis indicated that a small number of comparison group individuals were enrolled on other accredited

[25] This is a five-point scale ranging from 1, for evaluations based on simple cross-sectional correlations, to 5 for randomised control trials. Systematic reviews and meta-analyses typically include impact evaluations scored 3 or above on the SMS to attempt to understand what works. (See Sherman et al. 1998).

programmes on their TSP pseudo-start date. It should therefore be noted that as the TSP pseudo-start date variable is imputed, it doesn’t take account of comparison group individuals being on other accredited programmes.

• This evaluation measures a treatment effect using proven reoffending outcomes in accordance with the standard Ministry of Justice definition as used in Proven Reoffending National Statistics. As such, the study only accounts for proven reoffending. This does not measure treatment effects on crimes that are committed but not recorded by the police or do not lead to a caution or conviction.

• This evaluation does not adjust for any restriction on the time individuals are at liberty to reoffend in the community. Such restrictions include recall, additional sentences and time spent outside the UK.

• All female analyses were considerably smaller than male analyses, and this should be considered when reviewing the results. Small sample sizes lead to a reduced likelihood of achieving statistical significance and may account for some of our insignificant findings. It is therefore more difficult to conclude with an acceptable level of confidence that any difference in reoffending between the treatment and control groups was real rather than the result of chance.

• Statistical significance as defined in this report means that if no real differences exist there is a 5% chance of each result nonetheless being found to be statistically significant. On the same basis though, the chance of at least one of the many results being found to be statistically significant is much higher than 5%. Given the number of analyses, sub-analyses and outcome measures involved in this evaluation, care should therefore be taken when interpreting the findings. While multiple correction methods can be applied to reduce the risk of incorrectly finding a positive treatment effect, they can also increase the likelihood that real differences will not be detected. The results presented in this report have therefore not undergone multiple correction methods.

• For a few sub-analyses (‘Mixed male’ and ‘Aged 50+ male’), the drop-out rate of the treatment group at the matching stage was higher than usual and it should be noted that the impact estimate reported can only apply to the group of individuals who were matched. Characteristics of the matched and unmatched treatment groups for these analyses were compared and there were a much higher proportion of individuals with IPP sentences who were unmatched than matched. Therefore, care must be taken when considering who the impact estimates for these sub-analyses apply to; in particular, the sentence type characteristics of those not included.

• As per the sentence selection methodology (see Annex 13), there were several methodological approaches available to conduct this study. Following evaluation (as outlined in the proposal in Annex 13), the method ‘all participations all non-participations’ (APAN) was the option which best balanced methodological robustness and sufficient sample size. However, there are still limitations to the method. Firstly, there are several individuals in the treatment group who have participated in the TSP more than once. It is possible that participating in the TSP multiple times has a different effect compared to a single participation. Despite this, we treat all participations in the TSP as equal. Secondly, randomised controlled trials

(RCTs) are widely considered to be the ‘gold standard’ methodology for examining the efficacy of an intervention (Hariton & Locascio, 2018). The ‘RCT-like’ methodology was the most comparable to a real RCT. However, as the ‘RCT-like’ method minimised our sample size and the APAN dataset could be reasonably fit to the RCT-like method (using IPF fitting), the APAN methodology was chosen.

- Sexual offenders are known to have relatively low recidivism rates (Falshaw et al, 2003, Hanson, 2018). The TSP manual states that sexual offenders are suitable for the intervention so it was decided that they should be included in the analysis. Moreover, presence of sexual offenders in both the treatment and comparison groups and PSM matching should minimise any potential skew caused by inclusion of sexual offenders.

- During data cleaning, it was noted that there were a small number of cases where an individual had multiple index (release) dates for one conviction date. An investigation using prison discharge data indicated that this is mainly due to prison recalls. Recalls create a competing risk problem when measuring reoffending. Choosing the earliest release date represents the earliest date from which an individual is at liberty to reoffend and the earliest date from which an offender is included in the proven reoffending statistics. However, if there are multiple recalls during the following two-year period, there are periods where the individual is no longer at liberty to reoffend. Choosing the latest index date results in no further recalls during the two-year follow-up period but allows for missed reoffences prior to this date. This study used the earliest index date, which was thought to be the most consistent and accurate approach to finding all reoffences whilst noting the aforementioned limitations.

- Work and employment data (sourced from the Department of Work and Pensions and His Majesty’s Revenue and Customs) was only available until 2017. As a result, individuals in the treatment or comparison group who entered prison after 2017 were not matched on employment data in the PSM stage.

Annex 6: Full list of analyses undertaken

The full list of analyses undertaken are listed below in table A6.1 and were conducted for the two-year reoffending measures. Male and female sub-analyses were run separately. Sub-analyses which were considered but did not reach sufficient power are listed in Annex 7, along with their power analysis RAG rating.

For each of the following sub-analyses, the treatment group was matched to offenders in England and Wales using demographics, criminal history and individual offending-related risks and needs.

Table A6.1: List of all sub-analyses conducted.

Sub-analysis	**Male**	**Female**
Headline	✓	✓
Participants who met the programme's ideal suitability criteria	✓	✓
Participants who did not meet the programme's ideal suitability criteria	✓	✓
Participants who completed TSP	✓	✓
Participants who started but did not complete TSP	✓	✗
Participants who participated in TSP in prisons where the programme integrity was broadly maintained (2016-2019 assessment)	✓	✗
Participants who participated in TSP in prisons where the programme integrity was compromised (2016-2019 assessment)	✓	✗
Participants with an OGRS3 risk score between 25-49 (low risk)	✓	✗
Participants with an OGRS3 risk score between 50-74 (medium risk)	✓	✗
Participants with an OGRS3 risk score 75+ (high risk)	✓	✗
Participants whose index offence is a sexual offence	✓	✗
Participants whose index offence is an OVP offence	✓	✓
Participants whose index offence is an acquisitive offence	✓	✗
Participants who participated in TSP only and not in any other accredited programmes (during the TSP sentence)	✓	✓
Participants who completed another accredited programme prior to TSP during the same sentence	✓	✗
Participants whose self-reported ethnicity was Asian and Asian British	✓	✗
Participants whose self-reported ethnicity was Black, Black British, Caribbean, and African	✓	✗

Participants whose self-reported ethnicity was Mixed and multiple ethnic groups	✓	✘
Participants whose self-reported ethnicity was White	✓	✘
Participants who were more likely to present with characteristics associated with learning disabilities and challenges (LDC)	✓	✘
Participants who were less likely to present with characteristics associated with learning disabilities and challenges (LDC)	✓	✘
Participants aged between 18-25 years old	✓	✘
Participants aged between 26-30 years old	✓	✘
Participants aged between 31-49 years old	✓	✘
Participants aged 50+	✓	✘

Annex 7: Power analysis

A power analysis is the calculation used to estimate the smallest sample size needed for an experiment, given a required significance level, statistical power, and effect size. Power analysis was conducted on all sub-analyses for both the male and female cohorts, to determine whether the statistical power was large enough given the sample size of each sub-analysis.

Power analysis was conducted using the epiR package in R, testing across a range of effect sizes (odds ratios from 0.65 to 0.80) that broadly represent reductions in reoffending rates between approximately 5 and 10 percentage points in the treatment groups compared to the comparison groups. Statistical power will also depend on the baseline rate of reoffending in the population, which has been calculated as 0.473. Finally, we presumed statistical tests would require a threshold for statistical significance of $p < 0.05$.

Based on its statistical power, we have assigned each sub-analysis a RAG rating that reflects whether it is likely to generate reliable findings. These can be interpreted as follows:

- **GREEN:** Statistical power has been estimated to be greater than or equal to 0.80 (the standard academic benchmark for adequate statistical power). It is highly likely that results will be reliable and not due to chance.
- **AMBER:** Statistical power is greater than and equal to 0.70 and less than 0.80. Results are unlikely to be due to chance, but reliability is not guaranteed.
- **RED:** Statistical power is lower than 0.70. There is a strong likelihood that results will be spurious and not reliable.

Below is a list of analyses which were excluded due to their statistical power being less than sufficient (i.e., categorised as green as described above). It was agreed that if there was only one ‘green’ rating for a subgroup, these wouldn’t be analysed. For example, for the female age sub-analyses only the age group ‘31-49’ had a green RAG rating. As a result, this sub-analysis was not conducted.

Table A7.1: Power analysis RAG rating of excluded sub-analyses.

Sub-analysis (excluded)	**RAG rating**
Female: did not complete TSP	**RED**
Female: programme integrity broadly maintained 2016-2019	**RED**
Female: programme integrity compromised 2016-2019	**RED**
Female: OGRS3 risk score 25-49 (low risk)	**RED**
Female: OGRS3 risk score 50-74 (medium risk)	**GREEN**
Female: OGRS3 risk score 75+ (high risk)	**AMBER**
Male: index offence CSE (child sexual exploitation)	**RED**
Male: index offence Stalking	**RED**
Male: index offence Substance Misuse: Possession/Small-scale Supply/Drink Driving	**RED**

Male: index offence Substance Misuse: Import/Export/Production/Supply	RED
Female: Index offence a sexual offence	RED
Female: index offence CSE (child sexual exploitation)	RED
Female: index offence Stalking	RED
Female: index offence an acquisitive offence	RED
Female: index offence Substance Misuse: Possession/Small-scale Supply/Drink Driving	RED
Female: index offence Substance Misuse: Import/Export/Production/Supply	RED
Female: Participated in another accredited programme prior to TSP	RED
Male: Arab or other ethnicity	RED
Female: Arab or other ethnicity	RED
Female: Asian and Asian British ethnicity	RED
Female: Black, Black British, Caribbean, and African ethnicity	RED
Female: Mixed and multiple ethnic groups	RED
Female: White ethnicity	GREEN
Female: More likely to present with characteristics associated with learning disabilities and challenges (LDC)	RED
Female: Less likely to present with characteristics associated with learning disabilities and challenges (LDC)	GREEN
Female: age between 18-25	RED
Female: age between 26-30	RED
Female: age between 31-49	GREEN
Female: age 50+	RED

Annex 8: Additional information on ideal suitability and programme integrity sub-analyses

Ideal Suitability

Selection onto TSP as part of the TSP programme manager ‘risk override’ group

Not all the participants who were considered “not ideally suitable” for the ideal suitability sub-analysis were ineligible for participation in TSP.

Individuals who meet the OASys needs criteria but not the OGRS risk criteria do not meet the standard eligibility criteria for TSP but are potentially able to access the programme based on pre-defined characteristics of risk, making TSP a potentially suitable offer for support. These additional eligibility criteria are referred to as “risk overrides” and allow Treatment Managers to use their experience and discretion to offer additional places on TSP to individuals who may benefit from it.

Additionally, changes to the eligibility criteria for TSP over time mean that some individuals who did not meet the criteria for our ideal suitability sub-analysis (as defined at the time of the study) did meet different risk eligibility criteria at the time their place was allocated.

Those eligible for a ‘risk override’ were:

1. Individuals with an index or prior sexual offence(s) with a low OGRS3 score, assessed as medium risk and above using, the Risk Matrix 2000[26].
2. Indeterminate sentence prisoners with a low OGRS3 score and a high risk of harm or above, on one or more items assessed within OASys (e.g., risk to children, public, etc.).
3. Those who fall within three points of the OGRS3 cut-off score of 50 (i.e., scores of 47-49).
4. Between 2014-19, those with an OGRS3 score of 25-49.

Those eligible for a risk override must also meet the TSP need criteria.

[26] Risk Matrix 2000 (RM2000) is a statistically derived risk classification process intended for males aged at least 18 who have been convicted of a sex offence (Thornton, D. (2007). Scoring guide for risk matrix 2000.9/SVC. *Unpublished document*.)

Table A8.1: Not-ideally suitable sub-group (males).

Risk override group	Frequency	Proportion of not-ideally suitable sub-group
Group 1	540	9.4%
Group 2	12	0.2%
Group 3	552	9.6%
Group 4	1,558	27.1%
Not eligible for risk override	3,678	63.8%

Total not-ideally suitable male TSP participants: **5,759** (**31**% of male cohort).

Total not-ideally suitable male TSP participants who meet the risk override criteria: **2,081 (36.2%)**

Table A8.2: Not-ideally suitable sub-group (females).

Risk override group	Frequency	Proportion of not-ideally suitable sub-group
Group 1	4	0.7%
Group 2	-	-
Group 3	46	8.0%
Group 4	182	31.8%
Not eligible for risk override	361	63.1%

Total not-ideally suitable female TSP participants: **572** (**32**% of female cohort).

Total not-ideally suitable female TSP participants who meet the risk override criteria: **211 (36.9%)**

Note 1: Some individuals may be present in multiple groups (e.g., Group 1 as well as Group 2) and so the frequency column will not add to give the total not-ideally suitable figure.

The participants included in the “not ideally-suitable” sub-analysis could be placed into one of five groups (see Table A8.2):

- **Group 1**: Individuals with an index offence or prior sexual offence(s), a low OGRS3 score, assessed as medium risk or above using the Risk Matrix 2000/s.

- **Group 2**: Indeterminate sentenced prisoners with a low OGRS3 score, and a high risk of harm or above, on one or more relevant items assessed within OASys (e.g., risk to children, risk to public, etc).

- **Group 3**: Those who fall within three points of the OGRS3 cut-off score of 50 (scores of 47-49).

- **Group 4**: Between the years of 2014-19, those with an OGRS3 score of 25-49.

- **Not eligible for risk override**: Those who were not eligible for a risk override because:

  (a) They did not meet the TSP need criteria in full.
  (b) They did not meet the exception to the TSP risk criteria as per group 1-4 above.
  (c) They did not meet the TSP need criteria in full and were also not eligible for a risk override as per group 1-4 above.

Programme Integrity Classification

Quality Assurance Approach Summary

HMPPS Intervention Services oversees the Interventions Integrity Framework (IIF), with the main aim of supporting and developing practice to ensure effective accredited offending behaviour programme delivery. 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 and 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. For this evaluation, the second iteration – '2016-2019' has been used.

The IIF is divided into four key components; these are referred to as Key Lines of Enquiry (KLOE). These four KLOEs underpin the effective delivery of all our programmes. The definitions of KLOEs are below.

KLOE 1: Is the intervention(s) 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, scheduling and attendance 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 treatment 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 practice new skills as appropriate. Group dynamics and boundaries were also reviewed to support an effective learning environment.

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

Facilitation of effective rehabilitative activities require 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 treatment 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.

KLOE scores are scored from 1 to 4 whereby a score of 1 is the lowest score and 4 is the top score.

For this evaluation, to group establishments into the categories 'Programme integrity broadly maintained 2016-19' and 'Programme integrity compromised 2016-19', the sum of all four KLOE scores for each prison was used to give an overall 'Quality of Delivery score'. The criteria for classifying quality of delivery for the overall outcome measure was as follows.

Prisons were classified as "integrity broadly maintained" if:

- Overall QoD score of 13 or greater, **and**
- No scores of 1 or 2 on any of the 4 individual KLOE metrics.

Prisons were classified as "integrity compromised" if:

- Overall QoD score of 11 or less.

Establishments with scores of 12, or those with any individual KLOE scores of 1 or 2, were excluded from this analysis as it was not possible to classify them in either the programme integrity 'broadly maintained' or 'compromised' subgroups.

Annex 9: Participation in other accredited offending behaviour programmes

Table A9.1. Other accredited programmes which were participated in by individuals in the 'other accredited programme' sub-analysis pre-matched treatment group.

Treatment Group (n=20,302)		
Accredited Programme	**Number of treatment group participants**	**Percentage of treatment group records**
Total of 'other accredited programmes'	*2,818*	*13.9%*
Prisons Addressing Substance Related Offending (PASRO)	753	3.7%
Sex Offender Treatment Programme (SOTP) Core Programme (CP)	278	1.4%
Short Duration Programme (SDP)	274	1.3%
Building Skills for Recovery (BSR)	260	1.3%
Controlling Anger and Learning to Manage it (CALM)	224	1.1%
Twelve Step Programme	133	0.7%
Control of Violence for Angry Impulsive Drinkers (COVAID)	107	0.5%
Enhanced Thinking Skills (ETS)	100	0.5%
Alcohol Related Violence (ARV)	96	0.5%
RESOLVE	73	0.4%
Therapeutic Community (TC)	67	0.3%
Democratic Therapeutic Community Model (DTC)	44	0.2%
Alcohol Dependence Treatment Programme (ADTP)	40	0.2%
Focus on Resettlement (FOR)	35	0.2%
Healthy Relationships Programme (HRP)	33	0.2%
Cognitive Skills Booster (CSB)	29	0.1%
STOP	29	0.1%
(Adapted) Better Lives Booster (BLB/ABLB)	25	0.1%
Prison Partnership Twelve Step Programme (PPTSP)	24	0.1%
Becoming New Me (BNM)	23	0.1%
FOCUS	23	0.1%
Kainos Challenge to Change (KAINOS CTC)	34	0.1%
Bridges (short version of RAPT)	22	0.1%
Juvenile Enhance Thinking Skills (JETS)	19	0.1%
EP	18	0.1%

SOTP Rolling Programme (RP)	15	0.1%
Building Better Relationships (BBR)	13	0.1%
Choices, Actions, Relationships and Emotions (CARE)	6	0.0%
Self-Change Programme (SCP)	6	0.0%
Timewise	6	0.0%
RAPt 12 Step	4	0.0%
AP	1	0.0%
Cognitive Self Change Programme (CSCP)	1	0.0%
Healthy Relationships Programme (HRP) High Intensity (HI)	1	0.0%
KAIZEN GV (General Violence)	1	0.0%
New Me Strengths (NMS)	1	0.0%

Annex 10: Ethnic Groups

Ethnic groups were created using the Self-Defined Ethnicity – 18+1 Standard as per the Office for National Statistics categories[27].

Table A10.1: Ethnic groups created using self-defined ethnicity codes.

Ethnic group	18+1 Self-defined ethnicity code	18+1 Self-defined ethnicity
Asian and Asian British	A1	Indian
	A2	Pakistani
	A3	Bangladeshi
	A4	Chinese
	A9	Any other Asian background
Black, Black British, Caribbean, and African	B1	Caribbean
	B2	African
	B9	Any other Black background
Mixed and multiple ethnic groups	M1	White and Black Caribbean
	M2	White and Black African
	M3	White and Asian
	M9	Any other mixed background
Arab and other ethnic groups	O2	Arab
	O9	Any other background
White	W1	British
	W2	Irish
	W3	Gypsy or Irish Traveller
	W9	Any other White background
Not stated	NS	Not stated

[27] self-defined-ethnicity-18plus1.pdf (publishing.service.gov.uk)

Annex 11: Profile of male treatment group

The following descriptive statistics aim to provide an overview of the characteristics of 18,555 male treatment group offenders included in the headline male analysis. The treatment group included males with an age range from 18 to 84 years old. The tables below contain information on demographics, offence history, offending-related risks/needs, and participation in other accredited programmes.

Table A11.1: Demographic information for the PSM-matched male treatment sample.

Variable	**Frequency** (or mean where stated)
Mean age	31 (IQR 23-36)
Ethnic Group	
White	76%
Black, Black British, Caribbean, and African	12%
Asian and Asian British	6%
Mixed and multiple ethnic groups	5%
Unknown	1%
Nationality	
UK national	96%
Non-UK national	3%
Unknown	1%
Learning difficulties	
No problems	76%
Some problems	10%
Significant problems	4%
Unknown	10%
Difficulties with either numeracy, reading or writing	
No problems	70%
Some problems	22%
Significant problems	6%
Unknown	2%
Participation in other accredited programmes (APs)	
Participation in TSP only	88%
Participated in another accredited programme prior to TSP	12%

Offence-history information for the treatment sample.

Variable	**Frequency** (or mean/average where stated)
Sentence length	
Less than or equal to 6 months	1%
Between 6 and 12 months	1%
12 months to less than 4 years	46%
4 to 10 years	40%
More than 10 years	4%
Indeterminate or life sentence	8%
Index offences	
Violence against the person	26%
Sexual offences	12%
Robbery	12%
Theft offences	17%
Possession of weapons	5%
Drug offences	16%
Summary offences excluding motoring	2%
Fraud offences	0%
Public order offences	3%
Criminal damage and arson	3%
Miscellaneous crimes against society	4%
Prior criminal appearances	
Mean number of previous offences	31 (IQR 10-42)
Mean number of previous convictions	13 (IQR 5-18)
Mean number of previous custodial sentences	4 (IQR 1-6)
Risk assessment	
Mean Offender Violence Predictor (OVP) score	43 (IQR 27-56)
Mean Offender Group Reconviction Scale (OGRS3) score	62 (IQR 51-78)

Annex 12: Profile of female treatment group

The following descriptive statistics aim to provide an overview of the characteristics of 1,738 female treatment group offenders included in the headline female analysis. The treatment group included females with an age range from 18 to 70 years old. The tables below contain information on demographics, offence history, offending-related risks/needs, and participation in other accredited programmes.

Table A12.1: Demographic information for the PSM-matched female treatment sample.

Variable	**Frequency** (or mean where stated)
Mean age	32 (IQR 25-40)
Ethnicity	
White	88%
Black, Black British, Caribbean, and African	6%
Asian and Asian British	1%
Mixed and multiple ethnic groups	5%
Unknown	1%
Other	0%
Nationality	
UK national	97%
Non-UK national	2%
Unknown	1%
Learning difficulties	
No problems	79%
Some problems	7%
Significant problems	2%
Unknown	12%
Difficulties with either numeracy, reading or writing	
No problems	73%
Some problems	20%
Significant problems	5%
Unknown	2%
Participation in other accredited programmes (APs)	
Participation in TSP only	91%
Participated in another accredited programme prior to TSP	9%

Offence-history information for the treatment sample.

Variable	**Frequency** (or mean/average where stated)
Sentence length	
Less than or equal to 6 months	2%
Between 6 and 12 months	3%
12 months to less than 4 years	65%
4 to 10 years	25%
More than 10 years	1%
Indeterminate or life sentence	4%
Index offences	
Violence against the person	31%
Sexual offences	1%
Robbery	17%
Theft offences	21%
Possession of weapons	2%
Drug offences	13%
Fraud offences	1%
Summary offences excluding motoring	2%
Public order offences	2%
Criminal damage and arson	5%
Miscellaneous crimes against society	3%
Prior criminal appearances	
Mean number of previous offences	40 (IQR 13-53)
Mean number of previous convictions	17 (IQR 6-23)
Mean number of previous custodial sentences	4 (IQR 0-6)
Risk assessment	
Mean Offender Violence Predictor (OVP) score	36 (IQR 21-49)
Mean Offender Group Reconviction Scale (OGRS3) score	66 (IQR 54-81)

Annex 13: Sentence selection methodology

Background to why investigation was required

TSP differs from prior Accredited Programmes that have been evaluated using Propensity Score Matching (PSM), due to the presence of *multiple participation of the programme*. This means that there are a non-negligible number of individuals in the treatment group who have participated in the TSP during multiple distinct prison stays. In a Randomised Controlled Trial (RCT) design, it would be possible to prevent multiple participation by making lack of prior participation a requirement for being included in the trial in either the treatment or control groups. However, because the evaluation of the TSP is retrospective and because the number of participants with multiple participation is so large, this presents challenges for quasi-experimental evaluation.

If participants with multiple participations are excluded, then this reduces the sample size and selects a specific subset of the prison population, which may not be representative of the overall population who received the TSP, and this in turn affects the usefulness of any results.

Alternatively, if participants with multiple participations are included, then it is necessary to decide when an individual should appear in the treatment or the comparison group, in order to produce a reliable estimate of the causal effect of programme participation on the outcomes of interest. This is of particular concern for Propensity Score Matching, because it is necessary to provide *selection criteria* which sorts observations into treatment and comparison groups in a way as similar to an RCT as possible.

Due to the complexity of this evaluation and these different trade-offs, the analytical team developed a set of possible sentence selection approaches, along with a decision protocol to enable them to make an unbiased decision about how the treatment and comparison groups should be constructed for this study.

Approaches investigated

Table A13.1 details the four different approaches to sentence selection which were developed by the analytical team, along with their advantages and disadvantages.

Table A13.1: Comparison of different sentence selection methodologies.

	All Participations with All Non-Participations (APAN)	RCT-like (RCT)	First Participation with All Non-Participations (FPAN)	First Participation with First Non-Participation (FPFN)
Treatment group	Formed by selecting all sentences since 2010 where people participated in TSP.	Formed by selecting the first sentence since 2010 for every individual, and then choosing the subset of sentences where an individual participated in TSP in that sentence.	Formed by selecting the first sentence since 2010 where people participated in TSP.	Formed by selecting the first sentence since 2010 where people participated in TSP.
Comparison group	Formed by selecting all sentences since 2010 for all people who have never participated in TSP.	Formed by selecting the first sentence for every individual since 2010, and then choosing the subset of sentences where an individual did not participate in TSP in that sentence.	Formed by selecting all sentences since 2010 for all people who have never participated in TSP.	Formed by selecting the first sentence since 2010 for all people who never participated in TSP.
Previous analyses using this method	2021 RESOLVE reoffending impact evaluation	2017 SOTP reoffending impact evaluation	None	None
Unit of analysis	Sentence	Individual	Mixed: individuals in treatment group and sentences in comparison group	Individual

Advantages	Largest possible sample size. Representative of the true population of all individuals who participated in TSP.	Treatment assignment time is consistent for all individuals: treatment is assigned once, on their first sentence since 2010. Closest design to a real RCT.	Representative of the true population of all individuals who participated in TSP.	Representative of the true population of all individuals who participated in TSP.
Disadvantages	No consistent treatment assignment time across treatment and comparison groups. For individuals who do receive TSP, their sentences prior to their first participation do not appear in the comparison group.	Discards a large quantity of the sample. Selects a non-representative treatment group, particularly in later programme years.	No consistent treatment assignment time across treatment and comparison groups. Unclear how the causal estimate should be interpreted.	No consistent treatment assignment time across treatment and comparison groups.

Although the RCT dataset is the most theoretically valid design, it selects a very specific sub-sample of the population of individuals who have participated in the TSP. With the RCT dataset, it is only possible for an individual to be in the treatment group if their first sentence since 2010 included TSP participation. This distorts the composition of later programme years because it requires a gap in the individual's offending history. For example, for an individual who participated in the TSP in 2017 to be included in the treatment group for the RCT dataset, they would have to have had no sentences between 2010 and 2017. Otherwise, had they had a sentence during that time and during which they did not participate in the TSP, then they would be added to the comparison group instead. Thus, although the RCT dataset provides a more theoretically valid design, it does not select a sample which is representative of the population who have actually participated in the TSP, and therefore it is not likely that this analysis would have yielded findings which would be as useful for operational or policy purposes.

Alternatively, the APAN dataset would provide the most representative dataset, but it also violates the principle that treatment assignment should be consistent for all observations in both the treatment and comparison groups, as would be the case in an RCT.

In order to choose between these approaches, additional investigation of the dataset was required. However, this came with the risk of *researcher bias*, because ideally any methodological decisions by the analytical team should be independent of how those decisions affect the findings of the analysis.

Making a decision

To mitigate researcher bias, a decision protocol was developed and signed off by the analytical team and the independent MoJ Statistical Methodology team, prior to undertaking any further analytical work. Table A13.2 provides on outline of the decision protocol. In summary, the analytical team constructed each of the four datasets and performed a series of comparisons between them.

Table A13.2: Decision protocol for choosing sentence selection methodology.

1. Are there differences in propensity and prognostic factors between the treatment groups selected by the different methodologies?
    a. For each factor, compute Cohen’s d for each pair of treatment groups, to compute the difference between the two treatment groups.
    b. Is the mean of the effects less than 0.05? Are any of the individual effects greater than 0.075?
    c. If there are any differences, then we need to do further investigation on the sample composition.
        i. *Proceed to step 2*
    d. If all methods produce similar treatment groups, then we can conclude that the choice of selection methodology does not create significant differences in the sample composition. However, we still need to compare the outcomes before a decision is made.
        i. *Proceed to step 3*
2. Can we correct for differences in the composition of treatment groups across methodologies?
    a. Take two methodologies which have different propensity/prognostic factors per step 1
    b. Use Iterative Proportional Fitting to resample one of the pre-matched treatment groups to be closer to the pre-matched group for the other methodology.
    c. Perform step 1 again using the weighted datasets. Is the mean of the effects less than 0.05? Are any of the individual effects greater than 0.075?
    d. If all datasets are now selecting the same sample, we can examine the propensity score matching and the outcomes.
        i. *Proceed to step 3*
    e. If there are some datasets where we still cannot adjust the sample, then we will not be able to investigate equivalent causal effects in those datasets.
        i. The RCT dataset is our gold standard for causal estimates.

        ii. If we can fit one or more of APAN, FPAN, or FPFN to RCT, then we can discard the subset of datasets which cannot be fitted to the RCT dataset, since we are able to verify the causal estimates of the remaining datasets through comparison to the RCT dataset.
            1. *Discard the datasets which cannot be reweighted, and proceed to step 3*

        iii. If the sample composition differences have split the datasets into the RCT dataset and the other datasets, and if IPF cannot allow us to reweight any of these datasets (APAN, FPAN, FPFN) to fit the RCT dataset per step 3, then we are not able to determine if those other datasets are producing reliable causal estimates.
            1. *Choose RCT*

3. Does each methodology achieve good matching between treatment and comparison groups?
    a. Run the propensity score models, and look at the standardised differences for the listed propensity and prognostic factors.
    b. Is the mean of the effects less than 0.05? Are any of the individual effects greater than 0.075?
    c. Are there a large number of discarded observations? (Subject to JDL standard methodological approach)
    d. If any method does not produce well-matched treatment and comparison groups on these factors, or discards too many individuals, that method must be discarded as it is not possible to use it with PSM.
        i. *Discard unfeasible methodologies and proceed to step 4*
4. For methodologies which selected similar treatment groups, are there any differences in the one year reoffending rate between treatment and comparison groups?
    a. Take two methodologies which passed the Cohen’s d tests in step 1, or were IPF-reweighted to correct for sampling differences in step 2, and compute the one year reoffending rates for the treatment and comparison groups.
    b. Now calculate Cohen’s d for each pair of outcomes between the two datasets, comparing treatment group to treatment group and comparison group to comparison group.
    c. Are any of the effects greater than 0.05?
    d. If all datasets produce similar treatment groups and similar outcome, then we can conclude that the choice between these selection methodology does not introduce significant bias, and we can choose the methodology which will give the greatest power for sub-analyses.
        i. *Choose APAN, FPAN, or FPFN, in this preference order, to maximise sample size*
    e. If one or more of the APAN, FPAN, or FPFN datasets (or their IPF-reweighted equivalents) produce similar treatment groups and similar outcomes to the RCT dataset (or IPF-reweighted equivalents), but not all of them, we can choose the largest dataset which produces estimates consistent with the RCT dataset.

        i. *Choose APAN, FPAN, or FPFN, in this preference order, to maximise sample size, but only if they align with the RCT dataset*

    f. If none of the APAN, FPAN, or FPFN datasets (or their IPF-reweighted equivalents) produce similar treatment groups and similar outcomes to the RCT dataset (or IPF-reweighted equivalents), then we can conclude that the choice of selection methodology is introducing fundamental structural differences between different datasets, and we must choose the most theoretically robust approach.

        i. *Choose RCT*

In order to provide a thorough investigation, it was necessary to complete a full analysis on each of these datasets and to compare outcomes from those analysis. To reduce the risk of researcher bias, the analytical team only calculated and compared the one year proven reoffending rates. This outcome was chosen because it was expected to correlate with, but be distinct from, the two-year proven reoffending rate, which is one of the target outcomes of the overall study.

First, the treatment groups were compared across a set of variables between the different datasets. The analytical team hypothesised that the APAN, FPAN, and FPFN datasets would select similar treatment groups, but that the RCT dataset would select a different treatment group, and this was confirmed in testing. Since it was determined that APAN, FPAN, and FPFN were leading to equivalent estimates, FPAN and FPFN were discarded from later steps of the decision protocol in favour of APAN due to the larger sample size it provided.

In order to determine if the causal estimate from the APAN dataset would be reliable, the analytical team used Iterative Proportional Fitting (IPF) to resample the APAN treatment group and reweight it to fit the same distribution of characteristics as the RCT dataset. This was necessary to be able to compare outcomes between RCT and APAN in the next stage of the decision protocol. For example, if the RCT dataset had selected a treatment group with a lower mean number of previous offences, compared to the APAN dataset, then it would not be possible to directly compare the one year proven reoffending rates between those treatment groups. IPF is most often used to build synthetic populations by combining survey data with a marginal population data such as Census cross-tabulations. IPF iteratively reweights the individuals in the survey data to bring that dataset in alignment with the cross-tabulations from the population. In this case, IPF was used to reweight observations in the treatment group from the APAN dataset to match the marginal distributions of characteristics in the treatment group from the RCT dataset. The key assumption made is that if the distribution of characteristics within the treatment group is the same between two different datasets, and the outcomes are also similar, then both datasets would reliably identify the same causal effect in the underlying population from which they were sampled. Successful application of IPF to the APAN dataset resulted in a fifth dataset.

From this point, the standard JDL PSM methodology was followed. This involved selecting models for the propensity score for each dataset, predicting the logit of the propensity score from these models, performing Propensity Score Matching, computing outcomes for the treatment and comparison groups, testing for statistically significant difference between these groups, and computing standardised differences between these groups across the different variables present in the datasets. Finally, the one year reoffending rate for the treatment and comparison groups were compared across all datasets.

Results

The analytical team found that the IPF-adjusted APAN dataset and the RCT dataset led to similar estimates for the outcome, although these were very close to the threshold for differences specified in the decision protocol. Further investigation revealed some issues with model specification, and so once the variables had been finalised the decision protocol was carried out a second time. After correcting these issues, the outcomes for the IPF-corrected APAN dataset and the RCT dataset were similar within the threshold specified by the decision protocol. As a result, the analytical team used the APAN approach for sentence selection for the final analyses included in this report, and the FPAN, FPFN, and RCT datasets were discarded.

Annex 14: Summary of descriptive statistics

The Descriptive Statistics Excel annex provides further insight into the types of offenders that participated in TSP and explores participant characteristics. These descriptive statistics have been presented separately for males and females and are reflective of the pre-matched treatment group.

As these statistics include all participants prior to undergoing PSM, these figures include individuals who may not appear in the final treatment group for each distinct headline and sub-analysis.

For specific statistics on the make-up of each distinct analysis based on the matched treatment and comparison groups, please see the accompanying Standardised Differences Excel annex.

The statistics include:

1) Distribution of age (banded) at release from custody (in relation to the prison sentence during which offenders took part in TSP).
2) Characteristics of the index offence (including sentence length, offence group, severity and common offences).
3) Distribution of Offender Group Reconviction Scale (OGRS3) scores (banded).
4) Distribution of time period (banded) between treatment (the TSP end date) and release from custody.
5) Characteristics of reoffences, for those who go on to reoffend following release from prison.
6) Profile of the treatment groups, by criteria used to determine the ideal suitability cohorts.

The figures showed that whilst violence against the person was the most common index offence group (26%) for the male treatment group, it only accounted for 6% of reoffences. The most common index offence groups were violence against the person (26%), theft (17%), drug offences (16%), robbery (12%), and sexual offences (12%). The most common reoffences were theft (27%), summary offences excluding motoring (22%), and summary motoring offences (13%).

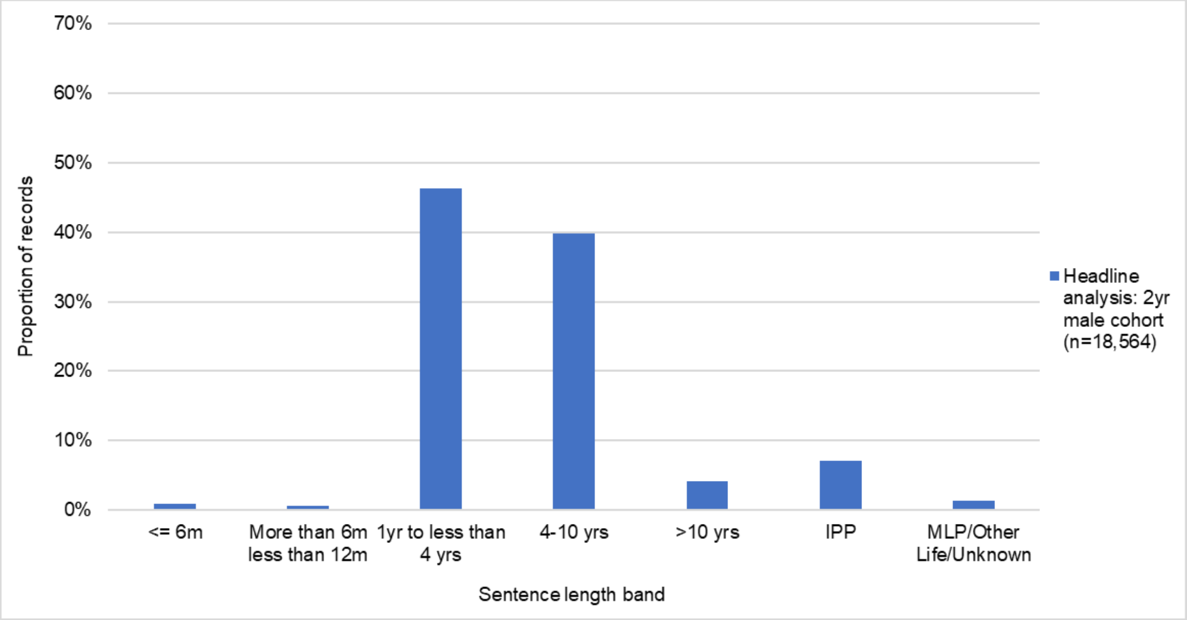
Full details of all the statistics outlined above are included in the Descriptive Statistics Excel annex.

As an example, charts showing the distribution of custodial sentence length and index offences and reoffences split by offence group are reproduced below.

Male pre-matched treatment group

Sentence length (male)

Chart A14.1 Distribution of custodial sentence length for index offences, pre-matched male treatment group.



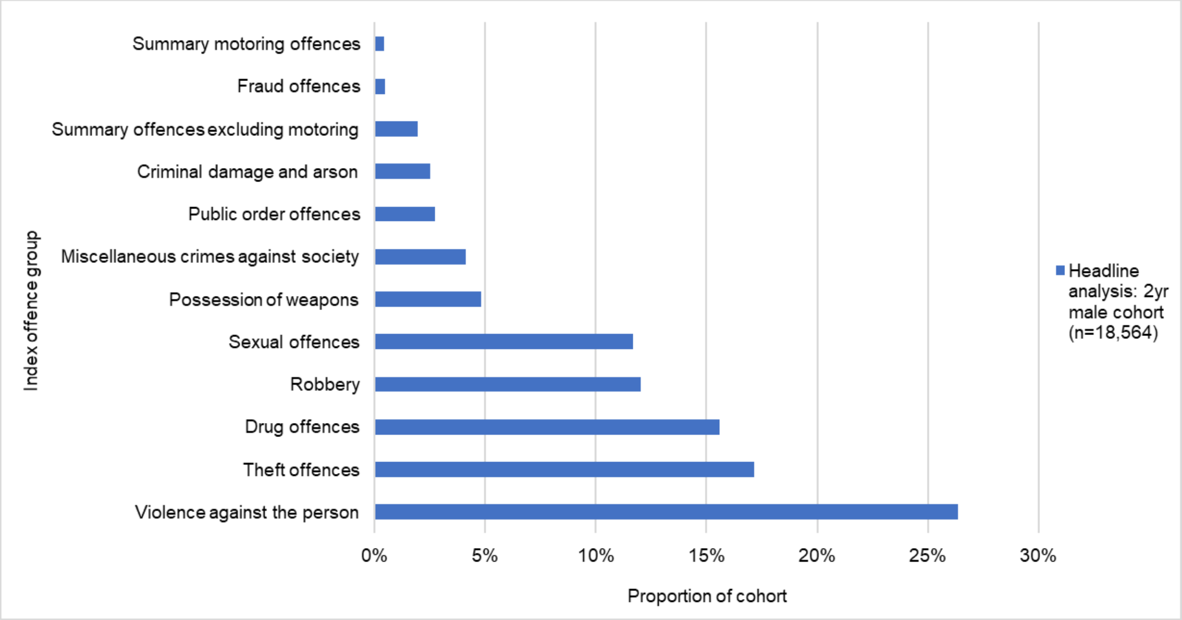
Source: Table A2.1 (male), Descriptive Statistics Excel annex

The sentence length band relates to the index offence. The index offence is the primary offence for which the individual received a custodial sentence, during which they participated in TSP or the equivalent sentence in the comparison group. The MLP (Mandatory Life sentence), Other Life and Unknown bands are aggregated.

The highest proportions of male treatment group participants had sentences between 1 and 4 years (46%) and between 4 and 10 years (40%), whilst 7% were on IPP (Imprisonment for Public Protection).

Offences and reoffences (male)

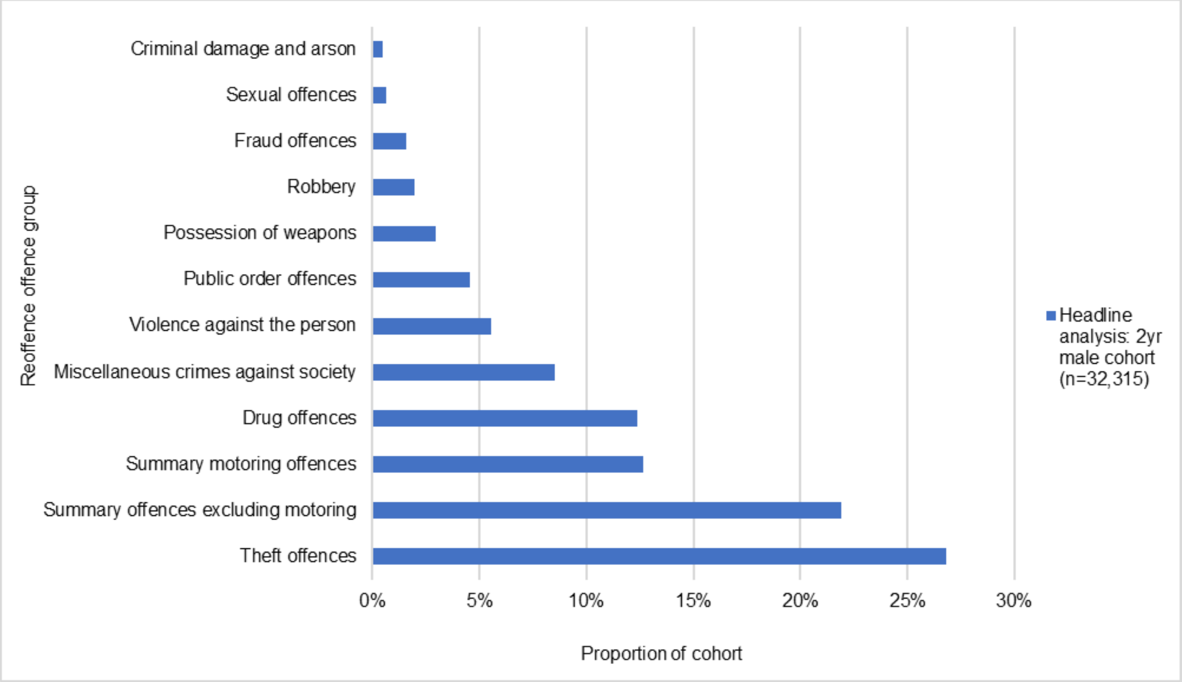
Chart A14.2: Distribution of offence groups for index offence, pre-matched male treatment group.



Source: Table A3.1 (male), Descriptive Statistics Excel annex

The index offence is the primary offence for which offenders received a custodial sentence, during which they participated in TSP.

Chart A14.3: Distribution of offence groups for reoffences, pre-matched male treatment group.

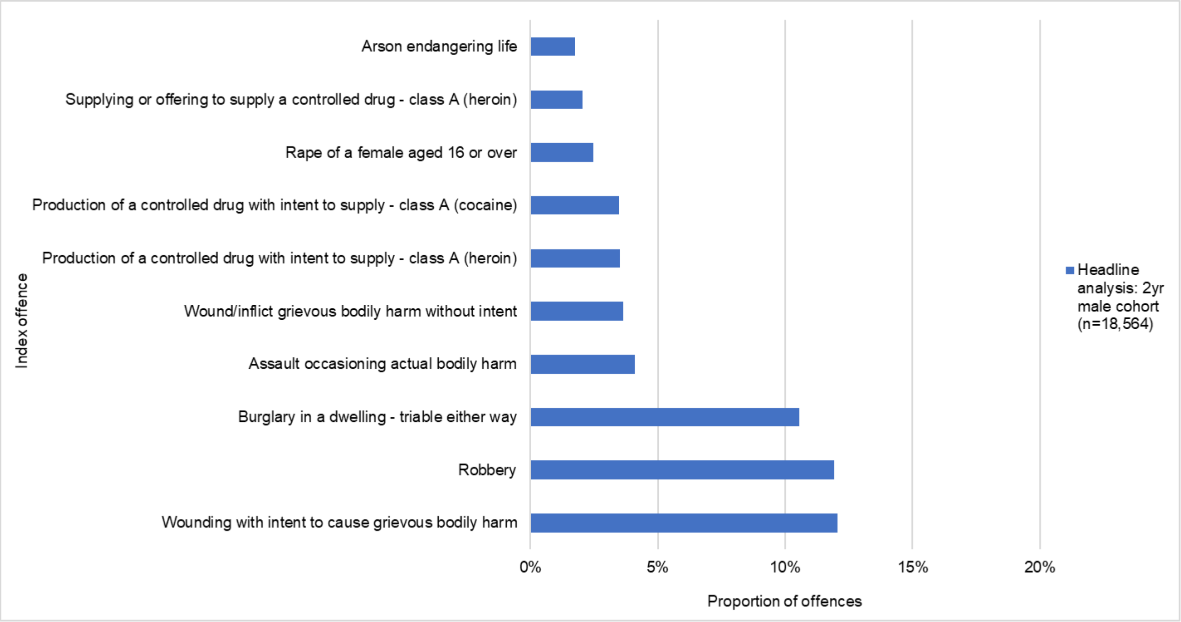


Source: Table A8.1 (male), Descriptive Statistics Excel annex

Note 1: The figures above relate to proven reoffences in the two-year follow-up period after release from custody period in which the offender participated in the TSP programme.

In the male headline analysis, comprising 18,564 records, there were 32,315 reoffences in the two-year follow-up period, relating to 8,632 records with at least one reoffence.

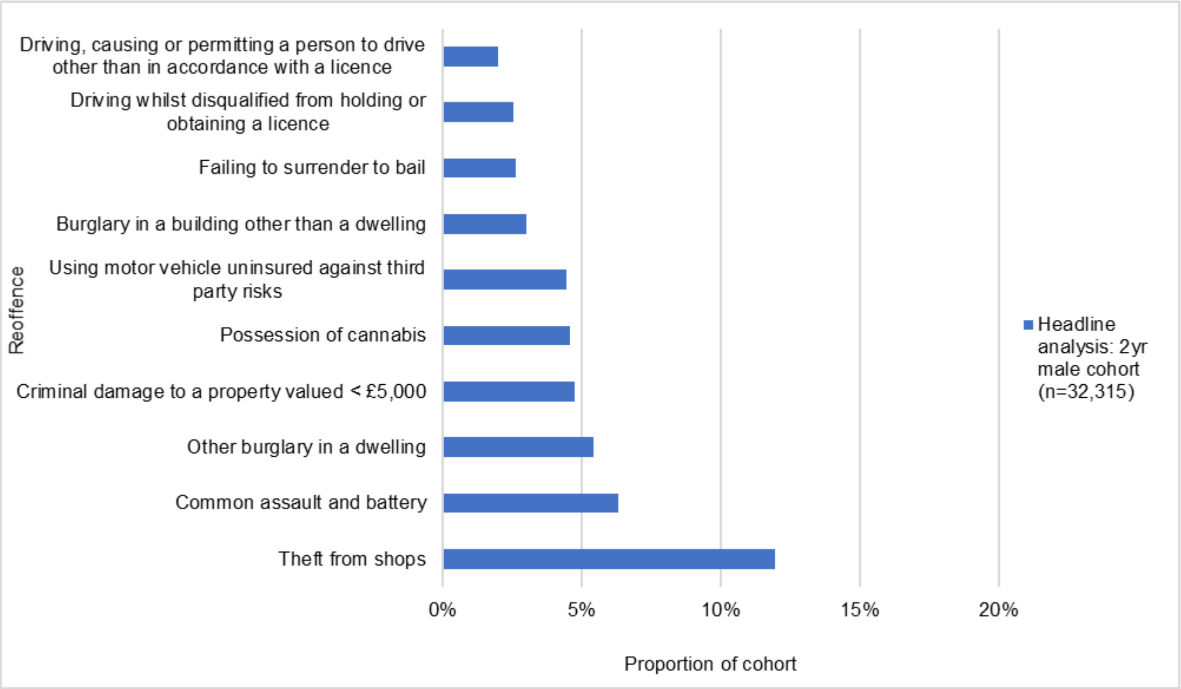
Chart A14.4: Distribution of most common index offences, pre-matched male treatment group.



Source: Table A5.1 (male), Descriptive Statistics Excel annex

The most common index offences amongst the treatment group were wounding with intent to cause grievous bodily harm (12%), robbery (12%), and burglary in a dwelling – triable either way (11%), whilst the most common reoffences were theft from shops (12%), common assault and battery (6%) and other burglary in a dwelling (5%).

Chart A14.5: Distribution of most common reoffences, pre-matched male treatment group.

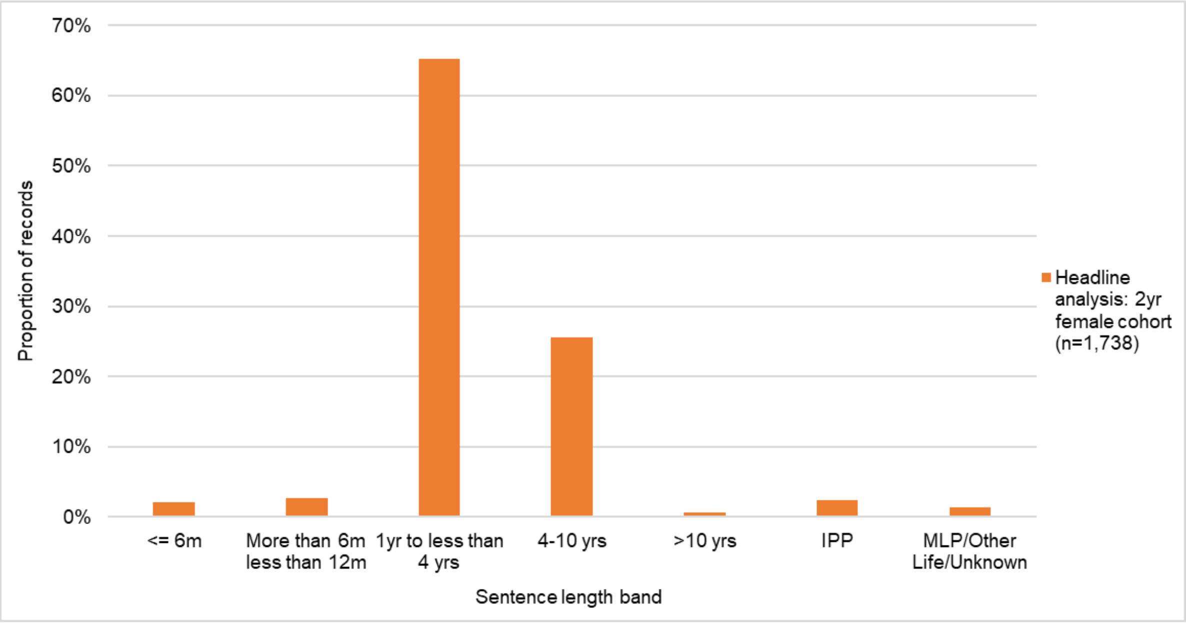


Source: Table A9.1 (male), Descriptive Statistics Excel annex

Female pre-matched treatment group

<u>Sentence length (female)</u>

Chart A14.6 Distribution of custodial sentence length for index offences, pre-matched female treatment group.



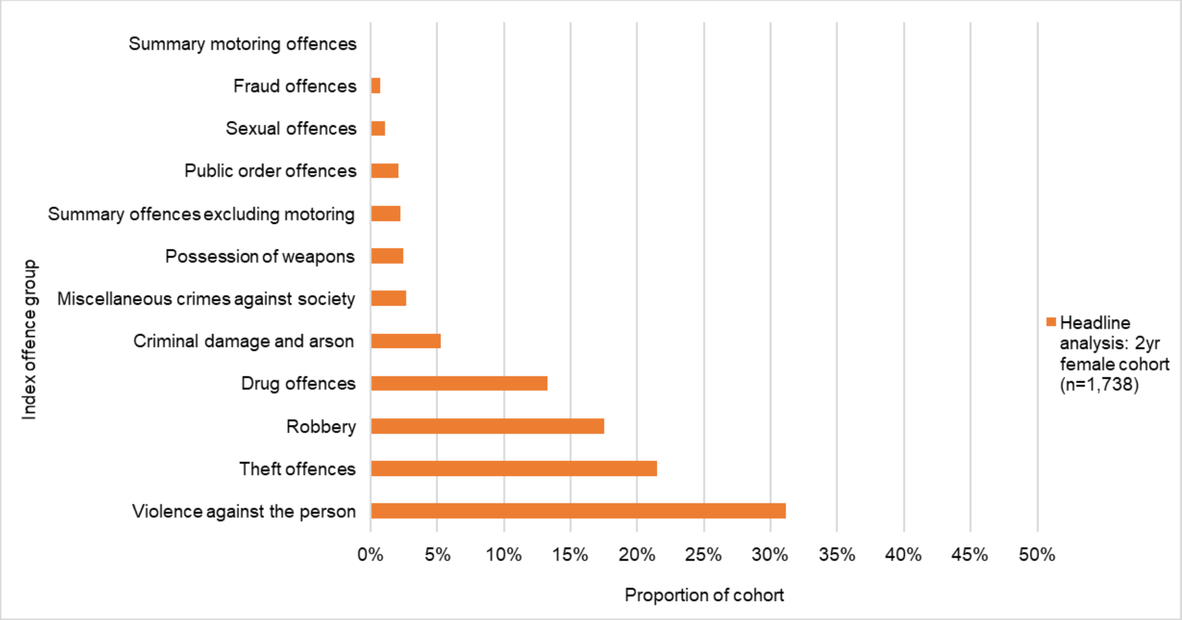
Source: Table A2.2 (female), Descriptive Statistics Excel annex

The sentence length band relates to the index offences, the primary offence for which offenders received a custodial sentence, during which they participated in TSP. The MLP (Mandatory Life sentence), Other Life and Unknown bands are aggregated.

The highest proportion of treatment group female participants had sentences between 1 and 4 years (65%) and between 4 and 10 years (26%).

Offences and reoffences (female)

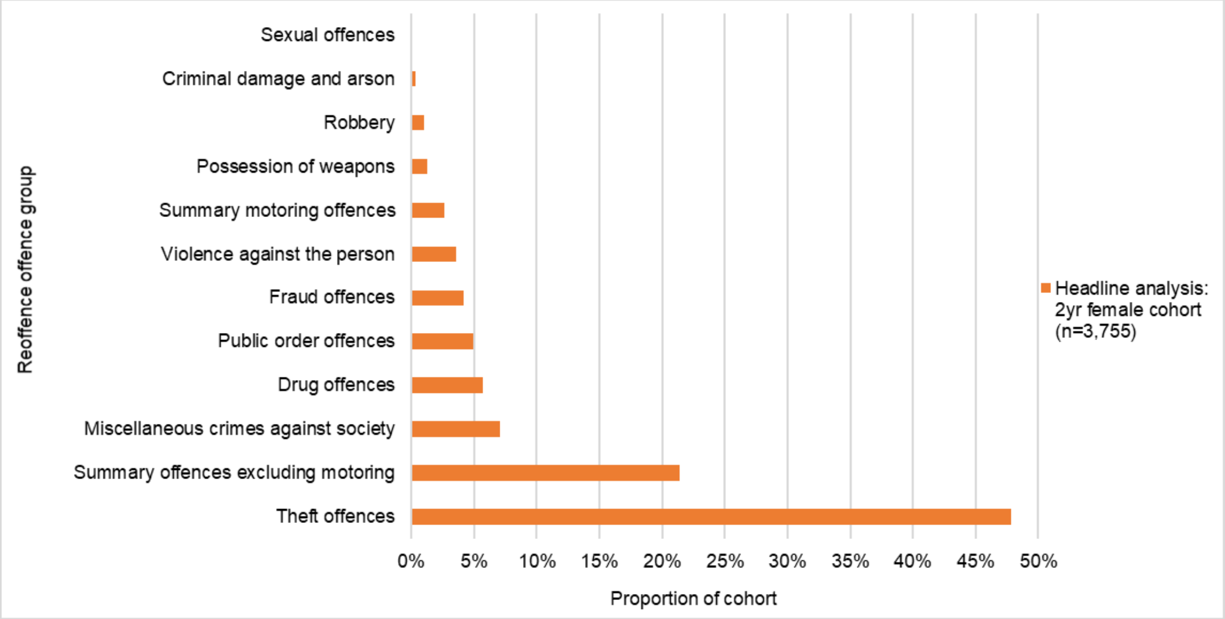
Chart A14.7: Distribution of offence groups for index offence, pre-matched female treatment group.



Source: Table A3.2 (female), Descriptive Statistics Excel annex

The index offence is the primary offence for which offenders were convicted and received a custodial sentence, during which they participated in TSP.

Chart A14.8: Distribution of offence groups for reoffences, pre-matched female treatment group.



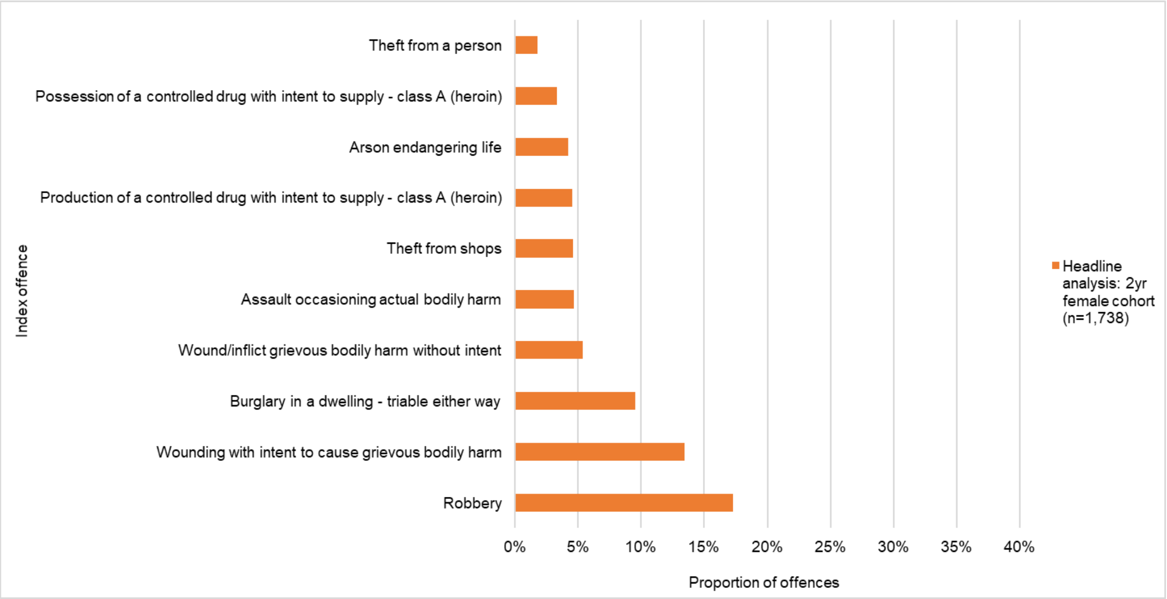
Source: Table A8.2 (female), Descriptive Statistics Excel annex

The figures above relate to proven reoffences in the two-year follow-up period after release from custody period in which the offender participated in the TSP programme.

In the female treatment group, comprising 1,738 records, there were 3,755 reoffences in the two-year follow-up period, in relation to 741 records with at least one reoffence.

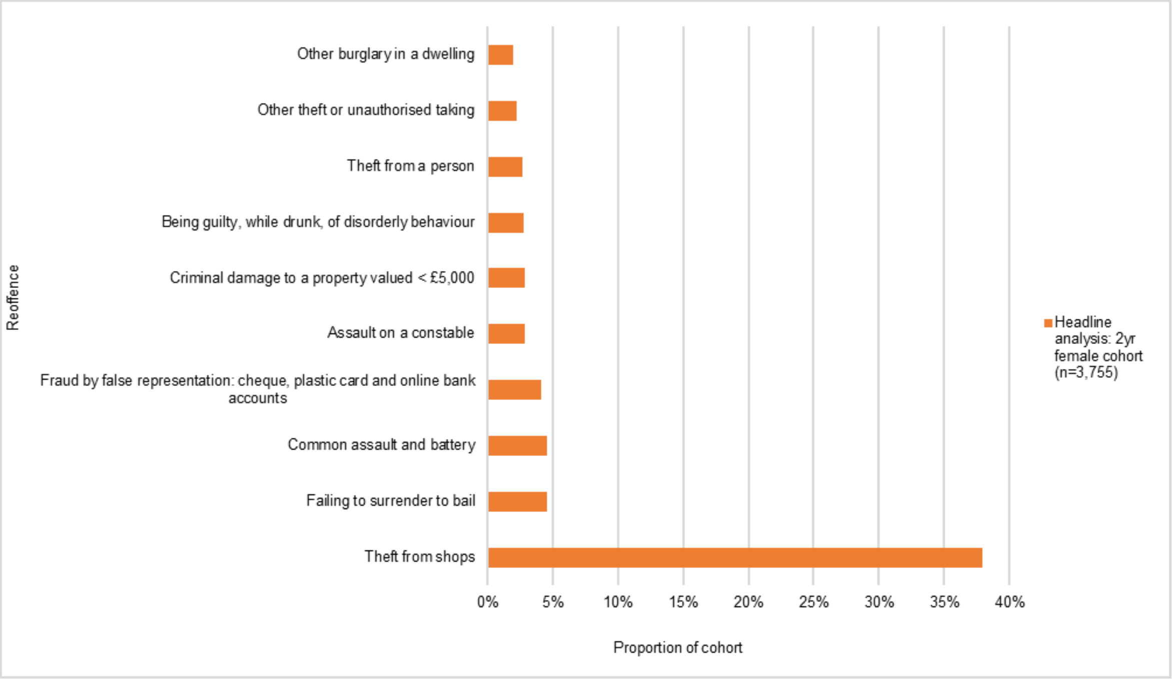
The most common index offence groups amongst the treatment group were violence against the person (31%), theft (22%), robbery (18%) and drug offences (13%), whilst the most common reoffences were theft (48%), summary offences excluding motoring (21%) and miscellaneous crimes against society (7%). The most common index offence of violence against the person (31%) only accounted for 4% of reoffences.

Chart A14.9: Distribution of most common index offences, pre-matched female treatment group.



Source: Table A5.2 (female), Descriptive Statistics Excel annex

Chart A14.10: Distribution of most common reoffences, pre-matched female treatment group.



Source: Table A9.2 (female), Descriptive Statistics Excel annex

The most common index offences amongst the treatment group were robbery (17%), wounding with intent to cause grievous bodily harm (13%) and burglary in a dwelling -triable either way (10%), whilst the most common reoffences were theft from shops (38%), failing to surrender to bail (5%) and common assault and battery (5%).

Annex 15: Odds ratios for reoffending rates

An odds ratio (OR) is the odds that an outcome will occur (like reoffending) given exposure to an intervention (like TSP), compared to the odds of the outcome occurring if not exposed to an intervention.

An OR greater than 1 indicates that the outcome (reoffending) is more likely to occur in the group that participated in TSP.

An odds ratio of 0.5 would mean that TSP participation group has half (50%) the odds of reoffending of the comparison group who didn’t participate in TSP.

Note, odds are not probabilities; a probability of 0.5 (equally likely to reoffend given participation in TSP than those who didn't participate in TSP) is equivalent to an OR of 1.

Table A15.1 shows the odds ratios (OR) for the binary measures by analysis and were calculated using the “questionr” package in R (version 0.7.3). ORs based on the mean treatment and comparison group rates are presented, alongside ORs for the upper and lower 95% confidence interval for treatment and comparison group rates.

Table A15.1: Odds ratios for males who committed a proven general reoffence in a two-year period after support from TSP, relative to matched comparison groups.

	Odds ratios for two-year proven general reoffending rates for males
Analyses	**Odds ratio**
Overall	0.93 (0.91 - 0.96)
Participants who met ideal suitability criteria	0.91 (0.89 - 0.93)
Participants who did not meet ideal suitability criteria	0.99 (0.95 - 1.04)
Completed TSP	0.94 (0.91 - 0.96)
Did not complete TSP	1.1 (0.97 - 1.26)
Programme integrity broadly maintained 2016-19	0.91 (0.85 - 0.97)
Programme integrity compromised 2016-19	0.92 (0.81 - 1.04)
With OGRS3 risk score 25-49 (low risk)	0.93 (0.88 - 0.99)
With OGRS3 risk score 50-74 (medium risk)	0.93 (0.9 - 0.96)
With OGRS3 risk score 75+ (high risk)	0.89 (0.86 - 0.93)
Index offence is a sexual offence	1.01 (0.93 - 1.09)
Index offence is an OASys Violence Predictor (OVP) offence	0.93 (0.90 - 0.96)
Index offence is an acquisitive offence	0.93 (0.86 – 1.00)

Participated in TSP only	0.94 (0.91 - 0.96)
Participated in another accredited programme prior to TSP	0.97 (0.92 - 1.02)
Asian and Asian British	0.97 (0.89 - 1.06)
Black, Black British, Caribbean, and African	0.99 (0.94 - 1.05)
Mixed and multiple ethnic groups	0.91 (0.84 - 0.99)
White	0.93 (0.90 - 0.95)
More likely to present with characteristics associated with learning disabilities and challenges (LDC)	0.95 (0.91 – 1.00)
Less likely to present with characteristics associated with learning disabilities and challenges (LDC)	0.93 (0.90 - 0.95)
Aged between 18-25	0.96 (0.92 - 0.99)
Aged between 26-30	0.88 (0.84 - 0.92)
Aged between 31-49	0.96 (0.92 - 1.00)
Aged 50+	1.03 (0.92 - 1.15)

Table A15.2: Odds ratios for females who committed a proven general reoffence in a two-year period after support from the TSP, relative to matched comparison groups.

	Odds ratios for two-year proven general reoffending rates for females
Analyses	**Odds ratio**
Overall	0.94 (0.87 - 1.01)
Participants who met ideal suitability criteria	0.88 (0.81 - 0.95)
Participants who did not meet ideal suitability criteria	0.90 (0.77 - 1.03)
Completed TSP	0.93 (0.84 - 1.02)
Index offence is an OASys Violence Predictor (OVP) offence	0.90 (0.83 - 0.98)
Participated in TSP only	0.95 (0.88 - 1.02)

Glossary of Terms

Average time to first reoffence: The average number of days between a person's index date and the date on which they commit their first proven reoffence, including only those who reoffend.

Clinical significance: The practical importance of a treatment effect (whether the intervention provides real, noticeable benefits which are palpable enough to be justified given associated costs/harms/inconveniences).

Comparison group: A group of offenders who did not receive the intervention being analysed. The comparison group is made up of offenders with similar characteristics to those in the treatment group.

Effect size: A value measuring the strength of the relationship between two variables in a statistical population.

Index date: The prison release date and the date from which the follow up period for measuring reoffending begins.

Index offence: The primary offence for which the offender was convicted and received a custodial sentence (specifically, the index sentence).

Interquartile range (IQR): A measure of variability that divides the dataset into quartiles. It is defined as the range of values between the first and third quartile. It is often used to show a more representative spread of values around a given variable as the IQR is resistant to outliers that may skew the mean of the treatment group.

Level of confidence: A range of values within an upper and lower bound. A 95% level of confidence would mean you could be 95% confident that the real value for a population of interest lies within the upper and lower bound. Levels of confidence (otherwise known as confidence intervals) are a key output for Justice Data Lab analyses as the reoffending rates for the treatment and control groups are essentially samples of larger populations.

Mean: This is a measure of the average in the dataset. It is calculated by adding all the values of a dataset and dividing it by the number of values in the set.

No significant difference – This means that, based on this analysis, it is not possible to say for sure whether the intervention had any effect (either positive or negative) on the outcome. There is a greater than 5% possibility that any differences between the groups were due to chance.

OASys Violence Predictor (OVP): Percentage likelihood of committing any violent proven reoffence within 2 years. This is based on static and dynamic factors including age, gender and criminal history. This includes minor violent offences like common assault, harassment and criminal damage and more serious violent offences. An OVP score of 30%+ is the criterion for accredited programmes that address violent offending behaviour. The more intensive programmes specify an OVP score of 60% or above.

Odds ratio: A measure of association between exposure and an outcome. The odds ratio represents the odds than an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure. Odds ratios more than 1 indicate increased occurrence of an event. Odds ratios less than 1 indicates decreased occurrence of an event.

Offender Assessment System (OASys): A system introduced in 2001 and built on the existing 'What Works' evidence base. It combines actuarial methods of prediction with structured professional judgement to provide standardised assessments of offenders' risks and needs, helping to link these risks and needs to individualised sentence plans and risk management plans.

Offender Group Reconviction Scale (OGRS3): Percentage likelihood of committing any offence within 2 years leading to reconviction (proven reoffending). This is based on static factors such as age, gender and criminal history. An OGRS3 score of 50% or more means that an offender is more likely than not to commit a proven reoffence within 2 years. OGRS scores can be used to target those resources designed to reduce reoffending. Accredited offending behaviour programmes often require particular OGRS scores as part of their eligibility criteria.

Police National Computer (PNC): An administrative data system used by all police forces in England and Wales, managed by the Home Office. The PNC records offender, crime and disposal details.

Propensity score matching (PSM): The methodology used for constructing a matched control group in Justice Data Lab analyses. Uses logistic regression to predict the likelihood of each offender receiving treatment; these predicted probabilities are called propensity scores. Treated and non-treated offenders are matched based on the closeness of their propensity scores.

P-value: The p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct.

Reoffending frequency: The number of proven reoffences committed, expressed per person.

Significant difference – This means the difference between groups is statistically not due to chance. The significance level used in this analysis is 5%, meaning there is a 95% certainty that the difference is due to the intervention, and not to chance.

Standardised mean difference: The standardised difference in means between the treatment and control groups, for an individual variable. The standardised mean difference is expressed as a percentage; the smaller the percentage the more similar the groups are on that variable.

Treatment group: The group of offenders that the provider delivered their intervention to. In other words, the offenders who received ‘the treatment’.

Two-year proven reoffending rate: The proportion of offenders in a cohort who committed an offence during a 24-month period starting on the index date and that resulted in a court conviction, caution, reprimand or warning in England or Wales during the same period or a further six-month waiting period.

References

Barnes, J. C., TenEyck, M. F., Pratt, T. C., & Cullen, F. T. (2020). How powerful is the evidence in criminology? On whether we should fear a coming crisis of confidence. Justice Quarterly, 37(3), 383-409.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature human behaviour, 2(9), 637-644.

Cann, J., Falshaw, L., Nugent, F. & Friendship, C. (2003). Understanding What Works: Accredited Cognitive Skills Programmes for Adult Men and Young Offenders. Home Office Research, Development & Statistics Directorate Research Findings No. 226. London: Home Office.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Routledge. ISBN 978-1-134-74270-7.

Falshaw, L., Bates, A., Patel, V., Corbett, C., & Friendship, C. (2003). Assessing reconviction, reoffending and recidivism in a sample of UK sexual offenders. Legal and Criminological Psychology, 8(2), 207-215.

Falshaw, L., Friendship, C., Travers, R. & Nugent, F. (2003). Searching for 'What Works': An Evaluation of Cognitive Skills Programmes. Home Office Research, Development & Statistics Directorate Research Findings No. 206. London: Home Office.

Friendship, C., Blud, L., Erikson, M., & Travers, R. (2002). An evaluation of cognitive behavioural treatment for prisoners. Great Britain, Home Office, Research, Development and Statistics Directorate.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. Advances in Methods and Practices in Psychological Science, 2(2), 156-168.

Hanson, R. K. (2018). Long-term recidivism studies show that desistance is the norm. Criminal Justice and Behavior, 45(9), 1340-1346.

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. BJOG: an international journal of obstetrics and gynaecology, 125(13), 1716.

Hollin, C. R., Palmer, E. J., McGuire, J., Hounsome, J., Hatcher, R., Bilby, C., & Clark, C. (2004). Pathfinder programmes in the Probation Service: A retrospective analysis (Home Office Online Report 66/04). London: Home Office.

Hollin, C. R., McGuire, J., Hounsome, J. C., Hatcher, R. M., Bilby, C. A. L., & Palmer, E. J. (2008). Cognitive skills offending behavior programs in the community: A reconviction analysis. Criminal Justice and Behavior, 35, 269-283.

Landenberger, N. A., & Lipsey, M. W. (2005). The positive effects of cognitive-behavioral programs for offenders: A meta-analysis of factors associated with effective treatment. Journal of Experimental Criminology, 1, 451-476.

Lipsey, M. W., Chapman, G., & Landenberger, N. A. (2001). Cognitive-behavioral programs for offenders. The Annals of the American Academy of Political and Social Science, 578, 144-

157.

Lipsey, M.W. & Landenberger, N.A. (2006) Cognitive-behavioral interventions. In. B. C. Welsh & D. P. Farrington (Eds.). Preventing Crime: What Works for Children, Offenders, Victims, and Places. Dordrecht, The Netherlands: Springer, 57-71.

McGuire, J., Bilby, C. A. L., Hatcher, R. M., Hollin, C. R., Hounsome, J. C., & Palmer, E. J. (2008). Evaluation of structured cognitive-behavioral treatment programs in reducing criminal recidivism. *Journal of Experimental Criminology*, 4, 21-40.

Mews, A., Hillier, J., McHugh, M. & Coxon, C. (2013). The impact of short custodial sentences, community orders and suspended sentence orders on reoffending. London: Ministry of Justice.

Ministry of Justice (2015). The impact of short custodial sentences, community orders and suspended sentence orders on reoffending. London: Ministry of Justice.

Monahan, K. C., Steinberg, L., & Cauffman, E. (2013). Age differences in the impact of employment on antisocial behavior. Child Development, 84(3), 791-801.

Palmer, E. J., McGuire, J., Hounsome, J. C., Hatcher, R. M., Bilby, C. A. L., & Hollin, C. R. (2007). Offending behaviour programmes in the community: The effects on reconviction of three programmes with adult male offenders. *Legal and Criminological Psychology*, 12, 251-264

Sadlier, G. (2010). Evaluation of the Impact of the HM Prison Service Enhanced Thinking Skills Programme. Outcomes of the Surveying Prisoner Crime Reduction (SPCR) Sample. Ministry of Justice Research Series 19/10. London: Ministry of Justice.

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. Frontiers in psychology, 10, 813.

Sherman, L.W., Gottfredson, D.C., Mackenzie, D.L., Eck, J., Reuter, P., & Bushway, S.D. (1998). Preventing Crime: What works, what doesn't, what's promising. Washington, DC: National Institute of Justice

Travers, R., Wakeling, H.C., Mann. R.E. & Hollin. C.R. (2013). Reconviction Following a Cognitive Skills Intervention: An Alternative Quasi-Experimental Methodology. *Legal and Criminological Psychology,* 18, 48-65.

Wakeling, H. (2018). The development of a screen to identify individuals who may need support with their learning. Analytic Summary.