Ministry of Justice

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

Ian A. Elliott, Johannes Huber, Samuel Hales, Jonathan Carling & Jordyn Perry

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 individuals assessed to be at medium and above risk of reoffending.

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 custody on prison misconduct within a 6-month follow-up period. Prison misconduct was defined as any recorded proven adjudications in custody. Adjudications are part of the prison disciplinary system across England and Wales. Any rule breaking activity or accusation of rule breaking that occurs within the prison system can be tried and punished through the internal prison disciplinary system. The hearings in which evidence is presented and decisions are made are called adjudications.

The analysis involved a treatment group of 13,891 adults (12,938 males and 953 females) who participated in the TSP programme between 2011 and 2019. This was compared to a matched comparison group of 85,527 adults (82,784 males, 2,743 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 even if the difference between the treatment group and the matched comparison group is small.

The impact of TSP was evaluated against two prison misconduct metrics over the 6-month follow-up period:

1. **Binary measure of proven adjudications** (adjudication rate) – did they receive a proven adjudication or not during the follow-up period.
2. **Frequency of proven adjudications received** – how many adjudications were received during the follow-up period.

Where sample sizes allowed, two separate prison misconduct outcomes were also tested:

- **Any** type of proven adjudication.
- A subset of proven adjudications that had been classified as being **violent**.

Males and females were analysed separately due to potential differences in misconduct behaviour between the two groups, given known differences in proven reconvictions rates. Headline results include all TSP participants 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.

Four key sub-analyses were identified as potentially important moderators of TSP effectiveness (see "Explanation of sub-analyses" section for more details):

- **Suitability for TSP** (ideally suitable or not ideally suitable)
- **Completion of TSP** (completed or not completed)
- **Programme integrity 2016-2019** (broadly maintained or compromised, using assessments from the HMPPS 2016-2019 Intervention Integrity Framework)
- **Risk of reoffending prior to TSP** (Offender Group Reconviction Score (OGRS3): low risk, medium risk, or high risk)

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

- **Exclusivity of TSP** (participation in TSP only and participation in one or more other accredited programmes)
- **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+ for any adjudication; 18-25, 26-30, and 31+ for violent adjudications)
- **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[1])

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

Results of the evaluation

Headline results: Male

Results show that over a 6-month period after starting TSP those who had participated were **less likely to receive an adjudication** compared to males who did not participate in TSP and **received an adjudication less frequently**. These results were statistically significant with mostly very small effect sizes.

They also show that those who participated in TSP were **less likely to receive a violent adjudication** compared to males who did not participate in TSP and **received a violent adjudication less frequently**. These results were also statistically significant with mostly very small effect sizes.

Key sub-analysis results: Male

The results showed that males who participated in TSP and met any of the following conditions were **less likely to receive an adjudication**, and **received an adjudication less frequently**:

- **Ideally suitable for TSP**: This was found for both any adjudication and violent adjudications.
- **Completed TSP**: This was found for both any adjudication and violent adjudications.
- **Considered to be at medium and above risk of reconviction after release from prison** (OGRS scores of 50-100): This was true for any adjudications. For violent adjudications this was found for those at high risk of reconviction only (OGRS scores of 75+).

These results were all statistically significant with mostly very small effect sizes.

Quality of delivery: These results were mixed. Over a 6-month period after starting TSP, males who participated **at a prison where standards were broadly met received any type of adjudication less frequently**. This result was statistically significant with mostly very small effect sizes, but there was not a statistically significant difference between groups in the rate of adjudications.

Headline results: Female

Results showed that over a 6-month period after starting TSP those who had participated **received any type of adjudication less frequently** compared to those who did not participate in TSP, the effect sizes were mostly very small. This result was statistically significant. There was not a statistically significant difference between the TSP and matched comparison group in the rate of adjudications. Violent adjudications could not be tested for female participants.

Key sub-analysis results: 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 over

a 6-month period after starting TSP females who completed TSP **received any form of adjudication less frequently** compared to those who did not participate in TSP. This result was statistically significant, and the effect sizes were mostly very small. There was no statistically significant difference between the TSP and matched comparison group in the rate of adjudications.

Conclusion

This study, like the related impact evaluation of TSP on 2-year reconviction rates (Brinn et al., 2023), uses an established and robust matching technique and a large number of matching variables, and has the largest sample size for any study of its kind to date.

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 three key sub-analyses (ideally suitable, completed TSP, and medium and above risk of reoffending) showed that TSP was statistically significantly associated with reductions in adjudications of any type across all outcome measures. Where programme integrity was maintained, TSP was associated with fewer adjudications of any type.

The overall analysis and three key sub-analyses (ideally suitable, completed TSP, and high risk of reoffending) also show that TSP was statistically significantly associated with reductions in violent adjudications across all outcome measures (but not for moderate to high risk of reoffending). For the smaller female cohort, there were some statistically significant associations between TSP and reductions in adjudications: those who participated in TSP received adjudications less frequently, as did TSP completers.

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. It should be noted, however, that this typically applies to studies of reoffending, not prison misconduct. Overall, across the analyses the effect sizes were mostly very small.

Key results for the male cohort

Rates of any proven adjudication within six months: Headline and key sub-analyses

Headline	**30.3%** of the treatment group received one or more proven adjudications of any type. This is a 2 percentage-point difference when compared to the comparison group or a 6% lower proven adjudication rate[2].		This is a statistically significantly[3] lower rate than the comparison group (**32.3%**).
Met the ideally suitable criteria[4]	**35.7%** of the treatment group received one or more proven adjudications of any type. This is a 1.9 percentage-point difference when compared to the comparison group or a 5% lower proven adjudication rate.		This is a statistically significantly lower rate than the comparison group (**37.6%**).
Completed TSP	**28.1%** of the treatment group received one or more proven adjudications of any type. This is a 3.3 percentage-point difference when compared to the comparison group or an 11% lower proven adjudication rate.		This is a statistically significantly lower rate than the comparison group (**31.4%**).

[2] The percentage point change in the proven adjudication rate 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 % minus comparison group %). In this example, 30.3% minus 32.3% equals a -2 percentage-point change. This is different to the percentage change in the proven adjudication rate, which 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 % minus comparison group %) ÷ comparison group %) multiplied by 100. Using the headline figure as an example: ((30.3% - 32.3%) ÷ 32.3%) x 100 = a 6% change.

[3] 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: (1) a lack of an observable effect to be found; (2) lower underlying "baseline" rates of misconduct; (3) inadequate sample sizes; (4) unobservable variables not accounted for in the evaluation design.

[4] As defined by TSP accreditation panel report: The Correctional Services Accreditation Panel Report 2009-2010 (Annex E)

Programme integrity broadly maintained[5] (2016-19)	**32.8%** of the treatment group received one or more proven adjudications of any type. This is a 0.9 percentage-point difference when compared to the comparison group or a 3% lower proven adjudication rate.		This is not a statistically significantly lower rate than the comparison group (**33.7%**).
OGRS3[6] score 50-74 (medium risk)	**30.7%** of the treatment group received one or more proven adjudications of any type. This is a 1.8 percentage-point difference when compared to the comparison group or a 6% lower proven adjudication rate.		This is a statistically significantly lower rate than the comparison group (**32.5%**).
OGRS3 score 75+ (high risk)	**42.8%** of the treatment group received one or more proven adjudications of any type. This is a 3 percentage-point difference when compared to the comparison group or a 7% lower proven adjudication rate.		This is a statistically significantly lower rate than the comparison group (**45.8%**).

(**Green** arrow for statistically significant finding, **grey** arrow for non-statistically significant.)

[5] As quality assured by HMPPS using the Interventions Integrity Framework.

[6] An OGRS3 score is the percentage likelihood of committing any proven offence within two years leading to reconviction. 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 two years.

Frequencies of any proven adjudication within six months: Headline and key sub-analyses

Headline	An average of **0.68** proven adjudications were received by the men in the treatment group.		This is statistically significantly fewer than the comparison group (**0.75** proven adjudications)
Met the ideally suitable criteria	An average of **0.82** proven adjudications were received by the men in the treatment group.		This is statistically significantly fewer than the comparison group (**0.90** proven adjudications)
Completed TSP	An average of **0.60** proven adjudications were received by the men in the treatment group.		This is statistically significantly fewer than the comparison group (**0.71** proven adjudications)
Programme integrity broadly maintained (2016-19)	An average of **0.72** proven adjudications were received by the men in the treatment group.		This is statistically significantly fewer than the comparison group (**0.81** proven adjudications)
OGRS3 score 50-74 (medium risk)	An average of **0.64** proven adjudications were received by the men in the treatment group.		This is statistically significantly fewer than the comparison group (**0.71** proven adjudications)
OGRS3 score 75+ (high risk)	An average of **1.09** proven adjudications were received by the men in the treatment group.		This is statistically significantly fewer than the comparison group (**1.20** proven adjudications)

(**Green** arrow for statistically significant finding, **grey** arrow for non-statistically significant.)

Rates of proven violent adjudications within six months: Headline and key sub-analyses

Headline	**6.3%** of the treatment group received one or more proven violent adjudications. This is a 0.8 percentage-point difference when compared to the comparison group or an 11% lower proven violent adjudication rate.		This is a statistically significantly lower rate than the comparison group (**7.1%**).
Met the ideally suitable criteria	**7.7%** of the treatment group received one or more proven violent adjudications. This is a 0.7 percentage-point difference when compared to the comparison group or an 8% lower proven violent adjudication rate.		This is a statistically significantly lower rate than the comparison group (**8.4%**).
Completed TSP	**5.6%** of the treatment group received one or more proven violent adjudications. This is a 1.2 percentage-point difference when compared to the comparison group or an 18% lower proven violent adjudication rate.		This is a statistically significantly lower rate than the comparison group (**6.8%**).
OGRS3 score 50-74 (medium risk)	**6.2%** of the treatment group received one or more proven violent adjudications. This is a 0.3 percentage-point difference when compared to the comparison group or an 5% lower proven violent adjudication rate.		This is not a statistically significantly lower rate than the comparison group (**6.5%**).
OGRS3 score 75+ (high risk)	**9.6%** of the treatment group received one or more proven violent adjudications. This is a 2 percentage-point difference when compared to the comparison group or a 17% lower proven violent adjudication rate.		This is statistically significantly lower rate than the comparison group (**11.6%**).

(**Green** arrow for statistically significant finding, **grey** arrow for non-statistically significant.)

Note: Programme integrity sub-analyses could not be conducted for proven violent adjudications due to insufficient sample sizes.

Frequencies of proven violent adjudications within six months: Headline and key sub-analyses

Headline	An average of **0.08** proven violent adjudications were received by the men in the treatment group.		This is statistically significantly fewer than the comparison group (**0.09** proven violent adjudications)
Met the ideally suitable criteria	An average of **0.10** proven violent adjudications were received by the men in the treatment group.		This is statistically significantly fewer than the comparison group (**0.11** proven violent adjudications)
Completed TSP	An average of **0.07** proven violent adjudications were received by the men in the treatment group.		This is statistically significantly fewer than the comparison group (**0.09** proven violent adjudications)
OGRS3 score 50-74 (medium risk)	An average of **0.08** proven violent adjudications were received by the men in the treatment group		This is not statistically significantly fewer than the comparison group (**0.08** proven violent adjudications)
OGRS3 score 75+ (high risk)	An average of **0.13** proven violent adjudications were received by the men in the treatment group.		This is statistically significantly fewer than the comparison group (**0.16** proven violent adjudications)

(**Green** arrow for statistically significant finding, **grey** arrow for non-statistically significant.)

Key results for the female cohort

<u>Rates</u> of any proven adjudication within six months: Headline and key sub-analyses[7]

Headline	**40.6%** of the treatment group received one or more proven adjudications of any type. This is a 3.1 percentage-point difference when compared to the comparison group or a 7% lower proven adjudication rate.		This is not a statistically significantly lower rate than the comparison group (**43.7%**)
Completed TSP	**38.1%** of the treatment group received one or more proven adjudications of any type. This is a 3.6 percentage-point difference when compared to the comparison group or a 9% lower proven adjudication rate.		This is not a statistically significantly lower rate than the comparison group (**41.7%**).

(**Green** arrow for statistically significant finding, **grey** arrow for non-statistically significant.)

[7] There was insufficient power to conduct sub-analyses to investigate the moderating effect on prison misconduct of ideal suitability, programme integrity, or OGRS risk score for female participants.

Frequencies of any proven adjudication within six months: Headline and key sub-analyses

Headline	An average of 1.12 proven adjudications were received by the women in the treatment group.		This is statistically significantly fewer than the comparison group (1.37 proven adjudications)
Completed TSP	An average of 0.96 proven adjudications were received by the women in the treatment group.		This is statistically significantly fewer than the comparison group (1.24 proven adjudications)

(**Green** arrow for statistically significant finding, **grey** arrow for non-statistically significant.)

Impact on prison misconduct

Overall estimates

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 receive any proven adjudication within six months could be **lower by between 1 and 3 men**. **This is a statistically significant result.**

The number of proven adjudications of any type received within six months could be **lower by between 4 and 10 adjudications**. **This is a statistically significant result.**

The number of men who receive a proven violent adjudication within six months could be **lower by between 0 and 1 men**. **This is a statistically significant result.**

The number of proven violent adjudications received within six months could be **lower by between 0 and 2 violent adjudications**. **This is a statistically significant result.**

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 receive any proven adjudication within six months could be **higher by 1 or lower by 7 women**. **This is not a statistically significant result.**

The number of proven adjudications of any type received within six months could be **lower by between 5 and 43 adjudications**. **This is a statistically significant result.**

Note: “What you can and can’t say” statements for the male and female results can be found in Annex 1.

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[8].

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.
- 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 four individual sessions), resulting in around 38-hours of contact time (dose). TSP is divided across three

[8] Prison Education and Accredited Programme Statistics 2021 - 2022 - GOV.UK (www.gov.uk)

modules (Self-Control, Problem Solving, and Positive Relationships). It 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 & Lipsey, 2005; Lipsey & Landenberger, 2006; French & Gendreau, 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., 2003), and the community (e.g., Hollin et al., 2004; Palmer et al., 2007; Hollin et al., 2008; McGuire et al., 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). 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.

With relevance to the outcome under investigation in this evaluation, there is some evidence to suggest that cognitive-behavioural programmes can reduce prison misconduct. An international meta-analysis conducted by French & Gendreau (2006) found a weak association between behavioural programmes and prison misconduct. The number of criminogenic needs targeted and delivery integrity were found to be important moderators of effect size. A more recent international review assessing institutional misconduct across prison and hospital settings found less clear results (see Papalia et al. 2019).

A randomised controlled trail (RCT), commissioned by the MoJ in 2009, explored the effects of ETS on prison behaviour as a secondary outcome, including adjudications, warnings, minor reports, and security information reports as well as a behavioural checklist rated by prison staff (McDougall et al., 2009). The number of security reports was statistically significantly reduced in the three months from starting ETS. Adjudication reports were also reduced, but this result did not reach statistical significance. The RCT did therefore provide some initial evidence that ETS may have supported pro-social behaviour in prison but further evaluation in this area has been needed.

Summary of methodology

The aim of this evaluation is to assess the impact of TSP on prison misconduct. It is a good quality quasi-experimental evaluation[9] with a large sample. The study includes individuals who participated in TSP between 2011 and 2019[10]. A long study period was chosen to increase the sample size of our 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[11].

Defining prison misconduct

For this study prison misconduct is defined based on proven prison adjudications. Adjudications are the 'procedures by which offences against the Prison or Youth Offender Institution (YOI) Rules allegedly been committed by offenders are dealt with'[12].

Rule breaking incidents are put before an adjudicator (a prison governor or governor-delegated prison officer in internal adjudications) to judge. The adjudicator is presented with evidence of the incident (from the reporting officer, accused offender, any witnesses, and any existing physical evidence) before deciding if the charge has been proven. If a prison misconduct is proven, the adjudicator then decides what punishment(s) to give.

For serious incidents, where the punishment of additional time in custody added is possible, the case will be referred to an Independent Adjudicator (IA). IAs are District Judges or Deputy District Judges approved by the Lord Chancellor to undertake the role of adjudicator in a case. IAs conduct these more serious adjudications following the same process outlined above, but with the unique ability to give additional days as punishment.

The most serious disciplinary offences can be referred to the police and prosecuted in the courts (rather than adjudicated). Crimes in prison that must be referred to the police (e.g., assault, murder, rape) are set out in Annex A of the Crimes in Prison Referral Agreement[13]. Other prison incidents are considered on a case-by-case basis but may be referred to the police depending on aggravating factors or if requested by the victim.

[9] Equivalent to a Level 4 on The Maryland Scientific Methods Scale (SMS) (Farrington et al., 2003). For further discussion see Sherman (1998).

[10] There have been some changes to the delivery of TSP during this time 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 over-ride (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.

[11] See, for example, Weisburd et al. (2001).

[12] HMPPS prisoner discipline procedures (adjudications)

[13] HMPPS NPCC CPS - Crime in Prison Referral Agreement - 2019

Data and analytical methodology

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

- Prison adjudications data to provide prison misconduct outcomes.
- Police National Computer (PNC) to provide criminal histories.
- Offender Assessment System (OASys) to provide risks and needs information.
- Prison population data to provide the self-reported ethnicity variable.
- Measure of Quality of Prison Life (MQPL) survey data to provide offender perceptions of prison safety.
- Safety in Custody statistics to estimate levels of prison violence at each prison.

This suite of data, comprising 94 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 accredited offending behaviour programme evaluation to date. Males and females were analysed in separate PSM models due to the potential for differences in prison misconduct, given known differences in reoffending behaviours between the two groups[14]. To minimise the potential effect of differences in participation year on outcome measures, TSP participation year (or pseudo-TSP participation year for the comparison group)[15] 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 (prison misconduct) to predict a "propensity score" [16]. 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[17] (see standardised differences annex for more detail).

The prison adjudications rates for the treatment and comparison groups were then compared. The rates are calculated using the weighted values[18] for each person after matching. Two prison adjudication outcomes were used to estimate the impact of the intervention over a six-month period and were applied for both the male and female cohorts:

1. A **binary** prison misconduct outcome: the number of people who receive a proven adjudication, expressed as a percentage of the group.

[14] Women in Prison - Justice Committee (parliament.uk)

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

[16] 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.

[17] 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".

[18] 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.

2. A **frequency** prison misconduct outcome: the number of proven adjudications received, expressed per person.

The same general headline measures were applied to eight sub-analyses which examined the effect of TSP on specific sub-groups (see the "Explanation of sub-analyses" section below). Each analysis undergoes a unique PSM process.

Both binary and frequency outcomes were generated for two forms of prison misconduct:

- **Any** type of proven adjudication.
- Proven adjudications categorised as "**violence**" (e.g., any fights and/or assaults).

The outcome measures in this study solely examine the impact of TSP participation on prison adjudications in the 6-months after starting TSP. An additional report examining the impact of TSP participation on proven reconvictions in the two years after release from a prison will be published alongside this report (see Brinn et al., 2023).

The TSP start date was selected as the point at which follow-up would start because (a) we cannot use an index release date like the reconviction study as our sample has not yet been released and (b) we cannot use a TSP "end" date as our study uses an "intent to treat" approach (the TSP group includes anyone who started TSP, regardless of whether or not they completed it) and so not all the sample will have a recorded "end" date. TSP end date data were also less complete than start dates. However, it does mean that some adjudications may have been received by the sample *during* participation in TSP, not after.

A 6-month follow-up period was chosen as data testing indicated that it was (a) the optimum balance between greatest sample size and capturing the greatest number of outcomes (enough records to allow for robust sub-analyses) and (b) the optimum way to ensure a balance between participants on both relatively shorter and longer sentences. We discuss the limitations of using start dates and 6-month follow-up periods in Annex 5.

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 adjudication outcomes between the treatment and comparison groups is compared using statistical significance testing, which returns a "*p*" (probability) value. A *p*-value is the probability of obtaining a result as large as the one observed if, in truth, there is no real difference to be found. A low *p*-value indicates that it is unlikely that the result observed would have been found if an effect or difference did not truly exist.

In this report, *p*-values are interpreted using a 0.05 significance level. If the *p*-value is less than 0.05, the difference between the two groups is considered "statistically significant" (i.e., is small enough to support concluding that there is evidence of a real effect). The direction of the difference indicates whether the treatment effect is positive or negative.

The estimated differences shown in the later tables are the 95% confidence intervals[19] for the differences between the relevant treatment and comparison group outcomes. Larger sample sizes lead to increased power to find a statistically significant effect. In turn, this leads to an increased tendency to detect differences that are statistically significant, even when the clinical significance of those differences is modest[20].

In this analysis we also provide an estimation of "effect size" as an indication of the magnitude of any statistically significant effects found using Cohen's *d*. Cohen's *d* statistic, is a standardized effect size for measuring the difference between two group means, and 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 interpreting 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 effect sizes within specific research areas (e.g., prison misconduct).

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 et al., 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 large real-world 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

[19] A range of values for which you can be 95% confident it contains the "true" mean of the population.

[20] 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 $p < 0.05$.

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 TSP and comparison groups and the magnitude of that change.

For additional insight, odds ratios (OR) for the rates of prison adjudications rate for males and females are included in Annex 12. These show the odds that an outcome will occur (in this case prison misconduct) given exposure to an intervention such as 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 TSP and of prison misconduct 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 prison misconduct 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.

Explanations of sub-analyses

The two headline analyses (male and female) included all individuals in the TSP group and comparison group. Further analyses were undertaken to examine the specific effects of TSP for relevant subgroups. 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 sub-analyses did not meet the required power threshold due to small sample sizes. These can be found in Annex 7, along with the full list of analyses undertaken.

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 significantly different for participants who met the “risk” and “need” criteria for the programme in full (i.e., those who were identified as “ideally suitable”) and those who accessed the programme through a discretionary risk override or did not meet the suitability criteria (i.e., those who 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 risk assessment score greater than or equal to 50 (medium to high risk).

And one of either:

2a. Needs assessment: score greater than or equal to 7 on the seven OASys[21] items (see Table 1), or;
2b. Needs assessment: score equal to 5 on the seven 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 seven items scored from the OASys assessment as part of the ideal suitability criteria are:

Table 1. TSP treatment targets and their associated items on OASys.

TSP targets	Target OASys item
Stop and think	**11.7.** Awareness of consequences
Emotional awareness	**11.4.** Temper control
Problem solving	**11.6.** Problem solving

[21] 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.

TSP targets	Target OASys item
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 **Seeing the whole picture**	**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 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 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 relevant no-treatment (did not participate in TSP) 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+) OGRS 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 of reoffending scores.

OGRS3 is defined as the “percentage likelihood of committing any offence within two years leading to reconviction (proven reoffending)”. This is based on static (or unchangeable) factors such as age at current conviction, age at follow-up (i.e., 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 two years.

Bands of OGRS3 scores were used to create three subgroups of increasing risk for the analysis: 25-49 (low risk), 50-74 (medium to high risk), and 75+ (high to very high risk). Each OGRS3 band was matched to a no-treatment (did not participate in TSP) comparison group. As a reminder, TSP is targeted at those at medium and above risk, so those subgroups are of particular interest.

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

Participation in TSP only (during the same sentence)

This sub-analysis measured the isolated treatment effect of the TSP accredited programme for those who have participated in no other accredited programmes during the same sentence. If offenders have participated in other accredited 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. **Participated in TSP only**: these individuals did not participate in another accredited programme during this sentence **before** they participated in TSP. They may, however, have participated in another accredited programme during the same

custodial sentence **after** participation in TSP, **or** during a different custodial sentence (before or after the current custodial sentence). This group was compared to a matched comparison group who didn't take part in TSP.

2. **Participated in another Accredited Programme**: these individuals participated in another accredited offending behaviour programme before they participated in TSP (or, for the comparison group, their "pseudo-start date") 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 didn't take part 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 TSP), see Annex 9.

Ethnic groups

The effect of TSP for participants may differ depending on an individual's protected characteristics. This sub-analysis sought to investigate the impact of TSP for different ethnic groups. Each ethnic group was compared to a matched "no treatment" comparison group (did not participate in TSP).

This analysis refers to self-reported ethnicity as obtained from the prison population data. Four subgroups are used for this sub-analysis: "Asian and British Asian", "Black, Black British, Caribbean and African", "Mixed and multiple ethnic groups", and "White", as per the Office for National Statistics high-level aggregate categories[22]. Further breakdowns of self-reported ethnicities included within these groupings can be found in Annex 10.

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 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 TSP.

LDC is measured using the HMPPS Learning Screen Tool[23] (LST). 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, authors 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

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

[23] The development of a screen to identify individuals who may need support with their learning (publishing.service.gov.uk)

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 of age on the impact of TSP. Each age band was matched to a “no treatment” comparison group who did not participate in TSP. The age of an individual is measured at the time at which they received TSP.

For our analyses measuring proven adjudications of any type, ages were categorised into four sub-groups: 18-25, 26-30, 31-49, and 50+, inclusive of the minimum and maximum of the range. The groups of 18-25 and 26–30-year-olds were selected based on evidence that some young adults take longer to develop psychosocial maturity, which may impact their likelihood to engage in accredited programmes, and their offending behaviours (Monahan et al., 2013). The age group 50+ was selected to investigate the impact of TSP for relatively older prisoners.

For our analyses measuring proven violent adjudications, the sample size for those over 50 years of age was not large enough to achieve statistically reliable results. For violent adjudications, ages were categorised into three sub-groups, 18-25, 26-30, and 31+. This allowed us to test the effect of TSP on violent adjudications for youngest group (18-25 years) and those for whom psychosocial maturity may continue to be an issue (26-30 years). Although the 31+ group allowed us to run some form of violent adjudication analysis for prisoners for whom psychosocial maturity is less likely to be an issue, it is likely to be a very varied group and may not detect any effects specific to those in the 31-49 and 50+ groups.

Male results in detail

Tables 2 and 5 presents the sample sizes for both the treatment group and the comparison group for male participants, for any proven adjudications and violent adjudications respectively. This includes the unweighted and weighted number of individuals receiving adjudications in the comparison group, with the weighted numbers being used to calculate the adjudication rates[24]. Where sample sizes are relatively small, they will be unlikely to produce a statistically significant result and thus have a lower likelihood of supporting conclusions with an acceptable level of confidence.

Tables 3-4 and 6-7 show the 6-month measures for proven adjudications for both the treatment group and the comparison group. Rates are expressed as percentages and frequencies expressed per person. Effect sizes (expressed as a Cohen's *d* statistic) are also included to estimate the magnitude of the differences 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 characteristics of the male treatment group can be found in Annex 11.

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

Table 2. Sample sizes for male cohort after matching, for 6-month **any type of adjudication** analysis (weighted numbers in parenthesis)

Analysis	Treatment group size	Comparison group size	No. with adjudications in treatment group	No. with adjudications in comparison group
Overall (headline)	12,938	82,784	3,921	27,173 (26,729.46)
Met ideal suitability criteria	8,236	32,210	2,937	14,758 (12,268.60)
Did not meet ideal suitability criteria	3,564	35,539	666	7,604 (6,981.91)
Completed TSP	11,853	82,990	3,329	27,208 (26,038.16)
Did not complete TSP	1,077	75,499	588	25,901 (31,477.78)
Programme integrity broadly maintained (2016-19)	1,664	13,394	545	5,016 (4,515.39)
With OGRS3 risk score 25-49	2,516	16,268	500	3,574 (3,419.85)
With OGRS3 risk score 50-74	5.773	25,582	1,770	9,285 (8,314.57)
With OGRS3 risk score 75+	3,626	22,857	1,553	11,444 (10,477.64)
Participated in TSP only	11,668	77,390	3,579	25,721 (25,527.21)
Participated in another Accredited Programme	1,276	5,207	348	1,419 (1,418.20)
Asian and Asian British ethnicity	863	5,323	259	1,538 (1,757.99)
Black, Black British, Caribbean, and African ethnicity	1,766	8,561	625	3,143 (3,329.38)
Mixed and multiple ethnic groups	664	3,181	230	1,239 (1,143.33)
White ethnicity	9,426	63,271	2,728	20,506 (19,459.71)
More likely to present with characteristics of LDC	2,541	13,061	955	5,548 (4,910.64)
Less likely to present with characteristics of LDC	7,669	43,735	2,129	13,353 (13,287.27)
Aged between 18-25	5,255	24,008	2,171	11,382 (10,578.33)
Aged between 26-30	2,652	15,267	762	5,479 (4,760.16)
Aged between 31-49	4,298	33,618	928	8,899 (7,906.81)
Aged 50+	706	7,692	57	671 (629.28)

Male results summary: Any type of adjudication

Tables 4-5 show that there were 25 positive statistically significant results among the 42 analyses. These provide evidence that:

Overall
- Participants are **less likely to receive any type of adjudication** than non-participants.
- Participants **receive fewer adjudications** than non-participants.

Met the ideal suitability criteria
- Participants are **less likely to receive any type of adjudication** than non-participants.
- Participants **receive fewer adjudications** than non-participants.

Completed TSP
- Participants are **less likely to receive any type of adjudication** than non-participants.
- Participants **receive fewer adjudications** than non-participants.

Did not complete TSP
- Participants are **more likely to receive any type of adjudication** than non-participants.
- Participants **receive more adjudications** than non-participants.

Programme integrity broadly maintained
- Participants **receive fewer adjudications** than non-participants.

OGRS3 risk score 50-75 (medium risk)
- Participants are **less likely to receive any type of adjudication** than non-participants.
- Participants **receive fewer adjudications** than non-participants.

OGRS3 risk score 75+ (high risk)
- Participants are **less likely to receive any type of adjudication** than non-participants.
- Participants **receive fewer adjudications** than non-participants.

Participated in TSP only
- Participants are **less likely to receive any type of adjudication** than non-participants.
- Participants **receive fewer adjudications** than non-participants.

Black, Black British, Caribbean, and African ethnic groups

- Participants are **less likely to receive any type of adjudication** than non-participants.
- Participants **receive fewer adjudications** than non-participants.

White ethnic group

- Participants are **less likely to receive any type of adjudication** than non-participants.

Less likely to present with characteristics of LDC

- Participants are **less likely to receive any type of adjudication** than non-participants.
- Participants **receive fewer adjudications** than non-participants.

Ages 18-25

- Participants are **less likely to receive any type of adjudication** than non-participants.
- Participants **receive fewer adjudications** than non-participants.

Ages 26-30

- Participants are **less likely to receive any type of adjudication** than non-participants.
- Participants **receive fewer adjudications** than non-participants.

Ages 31-49

- Participants are **less likely to receive any type of adjudication** than non-participants.
- Participants **receive fewer adjudications** than non-participants.
- Participants **receive fewer adjudications** than non-participants.

Non-completion of TSP

There are several plausible explanations for why we found a negative statistically significant effect for those who started but did not complete TSP. These include, but are not limited to and may be a combination of the following issues:

- Various studies indicate that non-completers are different to completers and may have higher risk of future criminal sanctions (e.g., McMurran & Thedosi, 2007). Non-completion of TSP may be truly associated with an increased likelihood of engaging in prison misconduct in the 6 months after starting TSP.
- For methodological reasons, we use the TSP start date as the beginning of the follow-up period for the TSP group. Therefore, non-completion could be the result of an adjudication, not the other way around, inflating the adjudication rate for the TSP group in the completion analysis (see Annex 5 for a more detailed explanation).
- Because no-one in the comparison sample actually participated in TSP during the sentence selected, and could therefore be classified as having not completed, we compared completers to a matched selection of 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 or fail to complete TSP if they were assigned to the intervention.
- The TSP group had a smaller sample size relative to other analyses. This means that statistical power was lower for this analysis. The negative finding, therefore, could be a “false positive” effect (i.e., a result that is found to be statistically significant when there was no true difference) detected simply by chance.

It is not possible to confirm which of these explanations is true, or which combinations of explanations is true, or if any of these explanations is true. Because there are specific identified methodological reasons why this might not be indicative of a real impact, it is not appropriate to conclude that non-completion of TSP is associated with increased levels of prison misconduct.

Table 3. Proportion of males who received a proven adjudication **of any kind** in a 6-month period after starting TSP compared with a matched comparison group.

Analysis	Treatment group size	Comparison group size	Treatment group rate (%)	Comparison group rate (%)	Estimated difference (% points)	Standardised effect size (Cohen's *d*)	Statistically significant difference?	*p*-value
Overall (headline)	12,938	82,784	30.3	32.3%	-2.8 to -1.1	-0.043	YES	<0.001
Met ideal suitability criteria	8,236	32,210	35.7	37.6	-3.6 to -1.3	-0.050	YES	<0.001
Did not meet ideal suitability criteria	3,564	35,539	18.7	19.6	-2.3 to 0.4	-0.024	NO	0.162
Completed TSP	11,853	82,990	28.1	31.4	-4.2 to -2.4	-0.072	YES	<0.001
Did not complete TSP[25]	1,077	75,499	54.6	41.7	9.9 to 15.9	0.260	YES	<0.001
Programme integrity broadly maintained (2016-19)	1,664	13,394	32.8	33.7	-3.4 to 1.4	-0.020	NO	0.432
With OGRS3 risk score 25-49	2,516	16,268	19.9	21.0	-2.8 to 0.5	-0.028	NO	0.180
With OGRS3 risk score 50-74	5,773	25,582	30.7	32.5	-3.2 to -0.5	-0.040	YES	0.006
With OGRS3 risk score 75+	3,626	22,857	42.8	45.8	-4.7 to -1.3	-0.061	YES	0.001
Participated in TSP only	11,668	77,390	30.6	33.0	-3.2 to -1.4	-0.050	YES	<0.001
Participated in another Accredited Programme	1,276	5,207	27.3	27.2	-2.7 to 2.8	0.001	NO	0.979
Asian and Asian British ethnicity	863	5,323	30.0	33.0	-6.3 to 0.3	-0.065	NO	0.075
Black, Black British, Caribbean, and African ethnicity	1,770	8,960	35.4	39.0	-6.0 to -1.0	-0.075	YES	0.004
Mixed and multiple ethnic groups	664	3,181	34.6	35.9	-5.3 to 2.7	-0.027	NO	0.522
White ethnicity	9.426	63,271	28.9	30.4	-2.8 to -0.8	-0.039	YES	<0.001
More likely to present with characteristics of LDC	2,541	13,061	37.6	37.6	-2.1 to 2.0	0.000	NO	0.989

[25] See the "Non-completion of TSP" section above for more details on this finding.

Analysis	Treatment group size	Comparison group size	Treatment group rate (%)	Comparison group rate (%)	Estimated difference (% points)	Standardised effect size (Cohen's *d*)	Statistically significant difference?	*p*-value
Less likely to present with characteristics of LDC	7,684	44,162	27.7	30.4	-3.8 to -0.2	-0.059	YES	<0.001
Aged between 18-25	5,255	24,008	41.3	44.1	-4.2 to -1.3	-0.055	YES	<0.001
Aged between 26-30	2,652	15,267	28.7	31.2	-4.3 to -0.6	-0.053	YES	0.011
Aged between 31-49	4,298	33,618	21.6	23.5	-3.2 to -0.6	-0.053	YES	0.004
Aged 50+	706	7,692	8.1	8.2	-2.3 to 1.9	-0.006	NO	0.920

(**Green** rows represent positive statistically significant results. **Grey** rows represent statistically non-significant results. **Red** rows indicate negative statistically significant results.)

Table 4. Frequency of males who received a proven adjudication **of any kind** (adjudications per person) in a 6-month period after starting TSP compared with a matched comparison group.

Analysis	Treatment group size	Comparison group size	Treatment group freq.	Comparison group freq.	Estimated difference	Standardised effect size (Cohen's *d*)	Statistically significant difference?	*p*-value
Overall (headline)	12,938	82,784	0.68	0.75	-0.10 to -0.04	-0.031	YES	<0.001
Met ideal suitability criteria	8,236	32,210	0.82	0.90	-0.13 to -0.04	-0.049	YES	<0.001
Did not meet ideal suitability criteria	3,564	35,539	0.37	0.38	-0.05 to 0.02	-0.016	NO	0.364
Completed TSP	11,853	82,990	0.60	0.71	-0.09 to -0.14	-0.078	YES	<0.001
Did not complete TSP[26]	1,077	75,499	1.62	1.15	0.32 to 0.62	0.199	YES	<0.001
Programme integrity broadly maintained (2016-19)	1,664	13,394	0.72	0.81	-0.17 to -0.00	-0.049	YES	0.040
With OGRS3 risk score 25-49	2,516	16,268	0.39	0.41	-0.07 to 0.03	-0.018	NO	0.393
With OGRS3 risk score 50-74	5,773	25,582	0.64	0.71	-0.11 to -0.03	-0.049	YES	0.001
With OGRS3 risk score 75+	3,626	22,857	1.09	1.20	-0.19 to -0.01	-0.056	YES	0.001
Participated in TSP only	11,668	77,390	0.69	0.77	-0.11 to -0.05	-0.049	YES	<0.001
Participated in another Accredited Programme	1,276	5,207	0.59	0.57	-0.06 to 0.01	0.019	NO	0.547
Asian and Asian British ethnicity	863	5,323	0.65	0.74	-0.19 to 0.02	-0.058	NO	0.101
Black, Black British, Caribbean, and African ethnicity	1,770	8,960	0.78	0.87	-0.17 to -0.01	-0.060	YES	0.021
Mixed and multiple ethnic groups	664	3,181	0.80	0.89	-0.23 to 0.05	-0.052	NO	0.211
White ethnicity	9.426	63,271	0.66	0.72	-0.10 to -0.03	-0.042	YES	<0.001
More likely to present with characteristics of LDC	2,541	13,061	0.90	0.96	-0.14 to 0.02	-0.033	NO	0.118

[26] See the "Non-completion of TSP" section above for more details on this finding.

Analysis	Treatment group size	Comparison group size	Treatment group freq.	Comparison group freq.	Estimated difference	Standardised effect size (Cohen's *d*)	Statistically significant difference?	*p*-value
Less likely to present with characteristics of LDC	7,684	44,162	0.60	0.68	-0.12 to -0.05	-0.057	YES	<0.001
Aged between 18-25	5,255	24,008	1.03	1.12	-0.14 to -0.03	-0.043	YES	0.004
Aged between 26-30	2,652	15,267	0.59	0.67	-0.13 to -0.02	-0.055	YES	0.006
Aged between 31-49	4,298	33,618	0.41	0.47	-0.10 to -0.03	-0.059	YES	<0.001
Aged 50+	706	7,692	0.12	0.12	-0.05 to 0.04	-0.013	NO	0.774

(**Green** rows represent positive statistically significant results. **Grey** rows represent statistically non-significant results. **Red** rows indicate negative statistically significant results.)

Table 5. Sample sizes for male cohort after matching, for 6-month **violent adjudication** analysis (weighted numbers in parenthesis)

Analysis	Treatment group size	Comparison group size	No. with violent adjudications in treatment group	No. with violent adjudications in comparison group
Overall (headline)	12,938	82,614	817	6,038 (5,858.63)
Met ideal suitability criteria	8,256	32,391	632	3,569 (2,717.95)
Did not meet ideal suitability criteria	3,563	35,533	131	1,485 (1,440.01)
Completed TSP	11,852	83,046	667	6,048 (5,651.15)
With OGRS3 risk score 25-49	2,516	16,156	98	707 (721.32)
With OGRS3 risk score 50-74	5,778	25,420	357	1,898 (1,646.21)
With OGRS3 risk score 75+	3,625	22,722	348	2,837 (2,643.81)
Participated in TSP only	11,668	77,112	763	5,788 (5,706.61)
More likely to present with characteristics of LDC	2,505	12,035	201	1,205 (1,011.08)
Less likely to present with characteristics of LDC	7,672	43,551	456	2,902 (2,912.30)
Aged between 18-25	5,260	24,680	590	3,621 (3,148.41)
Aged between 26-30	2,663	15,508	109	999 (798.59)
Aged 31+	5,005	41,393	118	1,398 (1075.72)

Note: Because violent adjudications occur less frequently in prisons than non-violent types of adjudications, the number of participants required in a sample to detect differences in violent adjudications between groups is larger (see Annex 6 for a description of “statistical power”). Consequently, not all sub-analyses had large enough samples to be run for violent adjudications as an outcome.

Male results summary: Violent adjudications

Tables 6 and 7 show that there were 16 positive statistically significant results out of 26 analyses. These provide evidence that:

Overall
- Participants are **less likely to receive a violent adjudication** than non-participants.
- Participants **receive fewer violent adjudications** than non-participants.

Met the ideal suitability criteria
- Participants are **less likely to receive a violent adjudication** than non-participants.
- Participants **receive fewer violent adjudications** than non-participants.

Completed TSP
- Participants are **less likely to receive a violent adjudication** than non-participants.
- Participants **receive fewer violent adjudications** than non-participants.

OGRS3 risk score 75+ (high risk)
- Participants are **less likely to receive a violent adjudication** than non-participants.
- Participants **receive fewer violent adjudications** than non-participants.

Participated in TSP only
- Participants **are less likely to receive a violent adjudication** than non-participants.
- Participants **receive fewer violent adjudications** than non-participants.

Less likely to present with characteristics of LDC
- Participants are **less likely to receive a violent adjudication** than non-participants.
- Participants **receive fewer violent adjudications** than non-participants.

Ages 18-25
- Participants are **less likely to receive a violent adjudication** than non-participants.
- Participants **receive fewer violent adjudications** than non-participants.

Ages 26-30
- Participants are **less likely to receive a violent adjudication** than non-participants.
- Participants **receive fewer violent adjudications** than non-participants.

Table 6. Proportion of males who received a proven **violent** adjudication in a 6-month period after starting TSP compared with a matched comparison group.

Analysis	Treatment group size	Comparison group size	Treatment group rate (%)	Comparison group rate (%)	Estimated difference (% points)	Standardised effect size (Cohen's *d*)	Statistically significant difference?	*p*-value
Overall (headline)	12,938	82,614	6.3	7.1	-1.2 to -0.3	-0.031	YES	0.001
Met ideal suitability criteria	8,197	31,221	7.7	8.4	-1.4 to -0.1	-0.027	YES	0.026
Did not meet ideal suitability criteria	3,563	35,533	3.7	4.1	-1.0 to 0.3	-0.019	NO	0.258
Completed TSP	11,852	83,046	5.6	6.8	-1.6 to -0.7	-0.049	YES	<0.001
With OGRS3 risk score 25-49	2.516	16,156	3.9	4.5	-1.4 to 0.3	-0.028	NO	0.174
With OGRS3 risk score 50-74	5,778	25,420	6.2	6.5	-1.0 to 0.4	-0.012	NO	0.399
With OGRS3 risk score 75+	3,625	22,722	9.6	11.6	-3.1 to -1.0	-0.066	YES	0.001
Participated in TSP only	11,668	77,112	6.5	7.4	-1.3 to -0.4	-0.034	YES	0.001
More likely to present with characteristics of LDC	2,505	12,035	8.0	8.4	-1.6 to 0.8	-0.014	NO	0.529
Less likely to present with characteristics of LDC	7,672	43,551	5.9	6.7	-1.3 to -0.2	-0.031	YES	0.012
Aged between 18-25	5,260	24,680	11.2	12.8	-2.5 to -0.6	-0.047	YES	0.001
Aged between 26-30	2,663	15,508	4.1	5.2	-1.9 to -0.2	-0.050	YES	0.013
Aged 31+	5,005	41,393	2.4	2.6	-0.7 to 0.2	-0.016	NO	0.291

(**Green** rows represent positive statistically significant results. **Grey** rows represent statistically non-significant results. **Red** rows indicate negative statistically significant results.)

Table 7. Frequency of males who received a proven **violent** adjudication (adjudications per person) in a 6-month period after starting TSP compared with a matched comparison group.

Analysis	Treatment group size	Comparison group size	Treatment group freq.	Comparison group freq.	Estimated difference (% points)	Standardised effect size (Cohen's *d*)	Statistically significant difference?	*p*-value
Overall (headline)	12,938	82,614	0.08	0.09	-0.02 to -0.00	-0.024	YES	0.012
Met ideal suitability criteria	8,197	31,221	0.10	0.11	-0.02 to -0.00	-0.025	YES	0.039
Did not meet ideal suitability criteria	3,563	35,533	0.05	0.05	-0.01 to 0.01	-0.008	NO	0.649
Completed TSP	11,852	83,046	0.07	0.09	-0.02 to -0.01	-0.039	YES	<0.001
With OGRS3 risk score 25-49	2.516	16,156	0.05	0.06	-0.02 to 0.01	-0.011	NO	0.630
With OGRS3 risk score 50-74	5,778	25,420	0.08	0.08	-0.01 to 0.01	-0.008	NO	0.602
With OGRS3 risk score 75+	3,625	22,722	0.13	0.16	-0.05 to -0.01	-0.063	YES	<0.001
Participated in TSP only	11,668	77,112	0.09	0.10	-0.02 to -0.00	-0.024	YES	0.018
More likely to present with characteristics of LDC	2,505	12,035	0.11	0.11	-0.02 to 0.01	-0.012	NO	0.575
Less likely to present with characteristics of LDC	7,672	43,551	0.08	0.09	-0.02 to -0.00	-0.027	YES	0.030
Aged between 18-25	5,260	24,680	0.16	0.17	-0.03 to -0.00	-0.030	YES	0.046
Aged between 26-30	2,663	15,508	0.05	0.06	-0.03 to -0.01	-0.059	YES	0.002
Aged 31+	5,005	41,393	0.03	0.03	-0.09 to 0.00	-0.022	NO	0.117

(**Green** rows represent positive statistically significant results. **Grey** rows represent statistically non-significant results. **Red** rows indicate negative statistically significant results.)

Female results in detail

Table 8 presents the sample sizes for both the treatment group and the comparison group for female participants. This includes the unweighted and weighted number of individuals receiving adjudications in the comparison group, with the weighted numbers being used to calculate the adjudication rates. Fewer analyses could be run for female participants due to smaller numbers of female participants relative to male participants. Where sample sizes are relatively small, they will be unlikely to produce a statistically significant result and thus have a lower likelihood of supporting conclusions with an acceptable level of confidence.

Tables 9 and 10 show the 6-month measures for proven adjudications 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 estimate the magnitude of the differences 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 characteristics of the female treatment group can be found in Annex 11.

Female results summary

Tables 9 and 10 show that there were 4 positive statistically significant results in the analyses. These provide evidence that:

Overall

- Participants **receive fewer adjudications** than non-participants.

Completed TSP

- Participants **receive fewer adjudications** than non-participants.

Participated in TSP only

- Participants are **less likely to receive any type of adjudication** than non-participants.
- Participants **receive fewer adjudications** than non-participants.

Table 8. Sample sizes for female cohort after matching, for 6-month **any type of adjudication** analysis (weighted numbers in parenthesis)

Analysis	Treatment group size	Comparison group size	No. with adjudications in treatment group	No. with adjudications in comparison group
Overall (headline)	953	2,743	387	1,010 (1,198.87)
Completed TSP	859	2,670	327	992 (1,114.53)
Participated in TSP only	880	2,583	361	960 (1,170.30)

Table 9. Proportion of females who received a proven adjudication **of any kind** in a 6-month period after starting TSP compared with a matched comparison group.

Analysis	Treatment group size	Comparison group size	Treatment group rate (%)	Comparison group rate (%)	Estimated difference (% points)	Standardised effect size (Cohen's *d*)	Statistically significant difference?	*p*-value
Overall (headline)	953	2,743	40.6	43.7	-6.7 to 0.5	-0.059	NO	0.095
Completed TSP	859	2,670	38.1	41.7	-7.4 to 0.1	-0.075	NO	0.055
Participated in TSP only	880	2,583	41.0	45.3	-8.1 to -0.5	-0.087	YES	0.026

Table 10. Frequency of females who received a proven adjudication **of any kind** (adjudications per person) in a 6-month period after starting TSP compared with a matched comparison group.

Analysis	Treatment group size	Comparison group size	Treatment group freq.	Comparison group freq.	Estimated difference (% points)	Standardised effect size (Cohen's *d*)	Statistically significant difference?	*p*-value
Overall (headline)	953	2,743	1.12	1.37	-0.43 to -0.05	-0.091	YES	0.012
Completed TSP	859	2,670	0.96	1.24	-0.46 to -0.11	-0.115	YES	0.001
Participated in TSP only	880	2,583	1.14	1.45	-0.50 to -0.10	0.106	YES	0.003

(**Green** rows represent positive statistically significant results. **Grey** rows represent statistically non-significant results. **Red** rows indicate negative statistically significant results.)

Acknowledgements

From the Ministry of Justice, the authors would like to thank Lucy Cuppleditch, Alana Diamond, and Panos Zerdevas. We would also like to thank several colleagues in Data and Analysis who have supported the completion of this project: Harry Smart, José Vieira and Manpreet Sidhu for their assistance with sub-analyses; Jon Mitchell, Allan Cox, and Zeeshan Zyed for their support and advice on data relating to prison violence; David Higgins for his expertise regarding the MQPL assessment; Ouarda Ahmad and Daniel Hawksworth for their assistance with adjudications data; Michael Williams, Katie Lovewell, and Helen Williams for assistance with PNC data; Wendy Smith-Yau for a literature review that directly informed this project; Eleftherios Nomikos, Aleksandra Plochocka, and Maxime Brooks for providing additional layers of Quality Assurance.

We would like to acknowledge our additional collaborators from the Justice Data Lab: John Preston, Aimee Brinn, Annie Sorbie, Rosina Costello, Tyler Opoku, Emily Sampson, and James Teasdale. We also appreciate ongoing data and methods advice from Aidan Mews, Philip Howard, Imran Ejaz, and David Dawson.

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. Substantial thanks also go to Sinead Bloomfield in the CSAAP Secretariat, whose prior work on the role of accredited programmes on prison misconduct substantially informed this work.

Please note: The contributions of the first and second authors were approximately equal.

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 and can't say about the results

✓ **What you can say about the 6-month any adjudication findings for males**

"This analysis provides evidence that support from the Thinking Skills Programme may decrease the number of male prisoners receiving proven prison adjudications in custody during a subsequent 6-month period."

"This analysis provides evidence that support from the Thinking Skills Programme may decrease the number of proven prison adjudications received by a male prisoner in custody during a subsequent 6-month period."

✗ **What you can't say about the 6-month any adjudication findings for males**

"This analysis provides evidence that support from the Thinking Skills Programme increases or has no effect on the number of male prisoners receiving proven prison adjudications in custody during a subsequent 6-month period."

"This analysis provides evidence that support from the Thinking Skills Programme increases or has no effect on the number of proven prison adjudications received by a male prisoner in custody during a subsequent 6-month period."

✓ **What you can say about the 6-month violent adjudication findings for males**

"This analysis provides evidence that support from the Thinking Skills Programme may decrease the number of male prisoners receiving proven violent prison adjudications in custody during a subsequent 6-month period."

"This analysis provides evidence that support from the Thinking Skills Programme may decrease the number of proven violent prison adjudications received by a male prisoner in custody during a subsequent 6-month period."

✗ **What you can't say about the 6-month violent adjudication findings for males**

"This analysis provides evidence that support from the Thinking Skills Programme increases or has no effect on the number of male prisoners receiving proven violent prison adjudications in custody during a subsequent 6-month period."

"This analysis provides evidence that support from the Thinking Skills Programme increases or has no effect on the number of proven violent prison adjudications received by a male prisoner in custody during a subsequent 6-month period."

✓ What you can say about the 6-month any adjudication findings for females

“This analysis does not provide clear evidence that support from the Thinking Skills Programme increases, decreases, or has no effect on the number of female prisoners receiving proven prison adjudications in custody during a subsequent 6-month period.”

“This analysis provides evidence that support from the Thinking Skills Programme may decrease the number of proven prison adjudications received by a female prisoner in custody during a subsequent 6-month period.”

✗ What you can’t say about the 6-month any adjudication findings for females

“This analysis provides clear evidence that support from the Thinking Skills Programme increases or decreases the number of female prisoners receiving proven prison adjudications in custody during a subsequent 6-month period.”

“This analysis provides evidence that support from the Thinking Skills Programme may increase or has no effect on the number of proven prison adjudications received by a female prisoner in custody during a subsequent 6-month period.”

Annex 2: Description of CSAAP

The Correctional Services Accreditation and Advice Panel (CSAAP) is comprised of independent international academics and expert practitioners who advise HMPPS on accrediting programmes. CSAAP also provides 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 every 5 years.

Annex 3: Details of matching criteria

Below is a table of variables (Table A3.1) 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; Arab and other ethnic groups; Unknown
UK Nationality	Categorical	UK; Non-UK; Unknown
Age at participation in TSP	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
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
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)	-

Variable	Type	Categories
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)	-
Time since conviction	Continuous (integer)	-
Number of previous prison adjudications in 3-months prior to TSP start date	Continuous (integer)	-
Number of previous violent prison adjudications in 3-months prior to TSP start date	Continuous (integer)	-
Accredited Programmes		
Year of participation in TSP (start date)	Categorical	2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019

Variable	Type	Categories
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 categorised 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 categorised 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; Significant
Does the offender have problems with numeracy?	Categorical	Unknown; None; Some; Significant
Does the offender have problems with reading?	Categorical	Unknown; None; Some; Significant
Does the offender have problems with writing?	Categorical	Unknown; None; Some; Significant
Does the offender have learning difficulties?	Categorical	Unknown; None; Some; Significant
Does the offender recognise the impact and consequences of offending on their victim /community/wider society?	Categorical	Unknown; None; Some; Significant
Does the offender currently have a relationship with their close family members?	Categorical	Unknown; None; Some; Significant
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; None; Some; Significant
Is the offender easily influenced by criminal associates?	Categorical	Unknown; None; Some; Significant
Does the offender have a manipulative or predatory lifestyle?	Categorical	Unknown; None; Some; Significant
Does the offender engage in recklessness and risk-taking behaviour?	Categorical	Unknown; No; Yes

Variable	Type	Categories
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; None; Some; Significant
Is the offender motivated to tackle their drug misuse?	Categorical	Unknown; None; Some; Significant
Are drug use or obtaining drugs a major activity or occupation for the offender?	Categorical	Unknown; None; 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; None; Some; Significant
Has the offender engaged in binge drinking or excessive use of alcohol in the last 6 months?	Categorical	Unknown; None; Some; Significant
Has the offender frequently or seriously misused alcohol in the past?	Categorical	Unknown; None; Some; Significant
Does the offender have a history of violent behaviour related to alcohol use at any time?	Categorical	Unknown; None; Some; Significant
Is the offender motivated to tackle their alcohol misuse?	Categorical	Unknown; None; Some; Significant
Does the offender have difficulties coping with everyday life?	Categorical	Unknown; None; Some; Significant
Does the offender currently have psychological problems, including depression?	Categorical	Unknown; None; Some; Significant
Does the offender self-harm, have they attempted suicide, or do they possess suicidal thoughts or feelings?	Categorical	Unknown; None; Some; Significant
Does the offender currently have psychiatric problems?	Categorical	Unknown; None; Some; Significant

Variable	Type	Categories
What level of interpersonal skills does the offender possess?	Categorical	Unknown; None; Some; Significant
Does the offender have issues with impulsivity?	Categorical	Unknown; None; Some; Significant
Does the offender demonstrate aggressive or controlling behaviour?	Categorical	Unknown; None; Some; Significant
Can the offender appropriately control their temper?	Categorical	Unknown; None; Some; Significant
Does the offender possess the ability to recognise problems?	Categorical	Unknown; None; Some; Significant
Does the offender possess appropriate problem-solving skills?	Categorical	Unknown; None; Some; Significant
Is the offender aware of the consequences of their actions?	Categorical	Unknown; None; Some; Significant
Is the offender able to understand other people's point of view?	Categorical	Unknown; None; Some; Significant
Is the offender able to engage in concrete/abstract thinking?	Categorical	Unknown; None; Some; Significant
Does the offender possess pro-criminal or offence-supportive attitudes?	Categorical	Unknown; None; Some; Significant
Does the offender have positive attitudes towards staff?	Categorical	Unknown; None; Some; Significant
Does the offender have positive attitudes towards supervision and/or their licence?	Categorical	Unknown; None; Some; Significant
Does the offender understand their motivation for offending?	Categorical	Unknown; None; Some; Significant
Is the offender motivated to address their offending behaviour?	Categorical	Unknown; None; Some; Significant
Does the offender possess any physical or mental health conditions?	Categorical	Unknown; None; Some; Significant
Does the offender understand the importance of completing programmes?	Categorical	Unknown; None; Some; Significant
On the basis that they could be released imminently back into the community, what risk does the offender currently pose to known adults?	Categorical	Unknown; Low; Medium; High; Very High
On the basis that they could be released imminently back into the community, what risk does the	Categorical	Unknown; Low; Medium; High; Very High

Variable	Type	Categories
offender currently pose to prison staff?		
Measure of the Quality of Prison Life variables		
MQPL Harmony with Staff	Continuous (numerical)	-
MQPL Meeting Individual Needs	Continuous (numerical)	-
MQPL Legitimate Bureaucracy	Continuous (numerical)	-
MQPL Staff Use of Authority and/or Incentives	Continuous (numerical)	-
MQPL Ordered Regime	Continuous (numerical)	-
MQPL Progression & Rehabilitation	Continuous (numerical)	-
MQPL Policing/Supervision	Continuous (numerical)	-
MQPL Prisoner Conflict, Drugs & Exploitation	Continuous (numerical)	-
MQPL Personal Safety	Continuous (numerical)	-
Prison variables		
Number of prison moves in the 3 months prior to TSP participation	Continuous (integer)	-
Simplified prison predominant function[27]	Categorical	Category A; Category B; Category C; Category D; Female Category A; Female Category B; Female Category D; Young Offender Institution
Annualised level of prison violence[28]	Continuous (numerical)	-

[27] This is often referred to as a prison's "category". To account for changes in prison function over time information on function for each year was compiled using a combination of the 2022 Prison Estate Register and relevant reports from the HM Inspectorate of Prisons 2009-2020.

[28] This is an estimated 12-month rate for each prison, calculated for the purpose of this study, using data and calculations consistent with the "Safety in Custody" reports produced by the MOJ Data and Analysis Prison Safety Statistics and Analysis Team and using published prison population statistics.

Annex 4: Methodological approaches

This study evaluates the relative receipt of proven prison adjudications in a cohort of treated and comparison (untreated) offenders within six months after starting TSP to estimate the impact of the intervention on prison misconduct.

The treatment group is comprised of those who started TSP during a prison sentence. This includes participants where there was intention-to-treat (ITT)[29] but who did not necessarily complete the full programme requirements. The comparison group includes those who did not attend (i.e., start) TSP during their sentence[30].

Pseudo-start dates

The date at which an individual in the treatment group starts 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 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.

It is important to note that for this prison misconduct study it is the TSP start date that is used to determine the beginning of the follow-up period. In reconviction impact studies it is typically the date of release that determines the start of the follow-up period. Since our outcome of interest occurs during the prison sentence, we cannot use release date. We also cannot use any form of TSP “end” date as our comparison group, whose start dates we have estimated, do not have an accurate end date. This means our 6-month follow-up period begins at the point at which TSP begins. While this means our follow-up period includes time spent engaged in TSP and is therefore not a traditional “follow-up”, it was chosen as the best compromise between data accuracy and methodological robustness. Given that TSP is shorter, relative to other accredited programmes, our testing indicated that for most the majority of the follow-up period is likely to occur after TSP is completed.

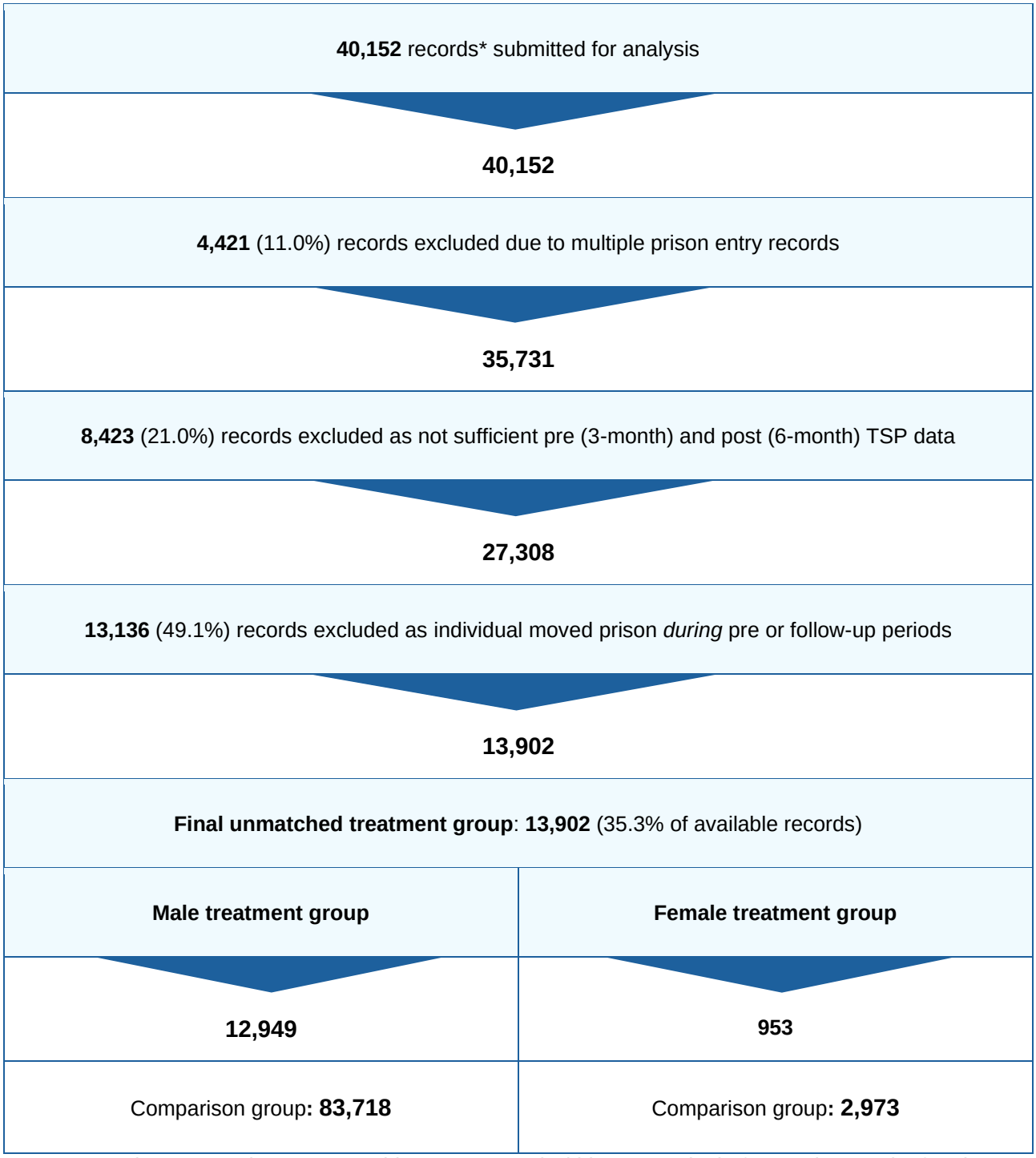
Propensity score matching

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 (prison misconduct) 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.

[29] 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.

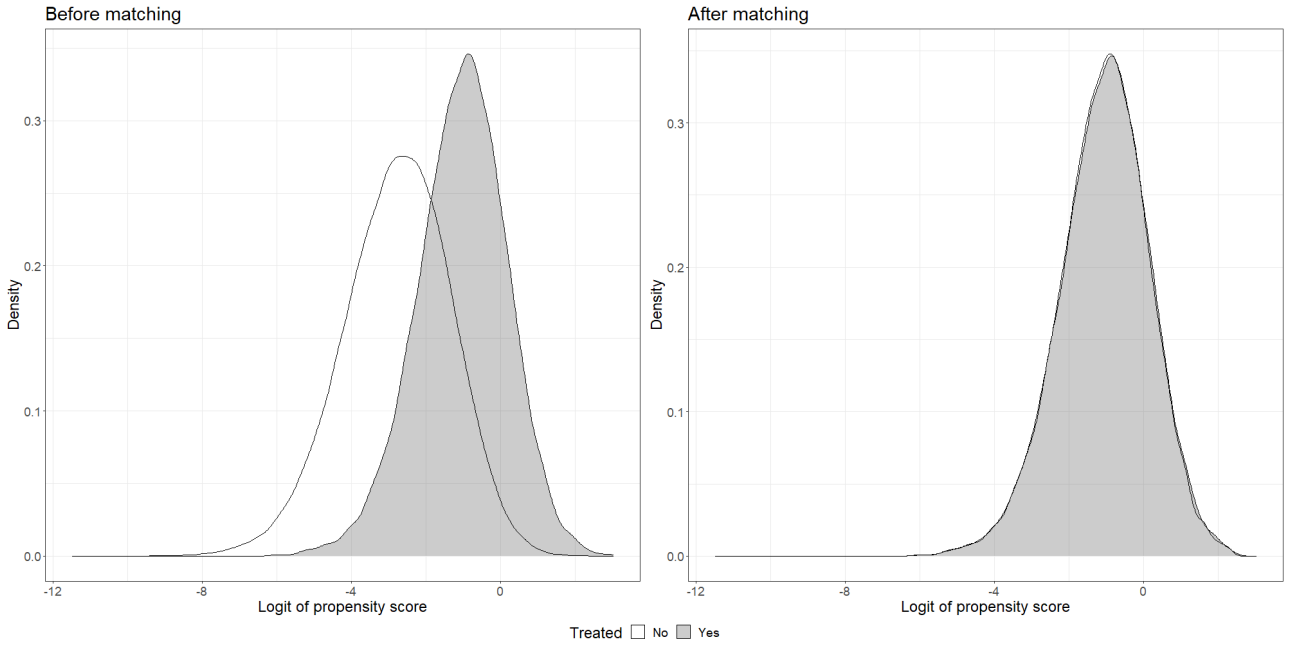
[30] For a comprehensive explanation of how (a) those who participated in TSP twice were dealt with and (b) those in the comparison group who had multiple eligible sentences, see Brinn et al. (2023: Annex 13).

Figure A4.1 Attrition from treatment group to create final cohort



*Age greater than or equal to 18 years old. OASys record within 12 months before and 1 month after the TSP start date. Prison population record within 1 week window either side of the TSP start date.

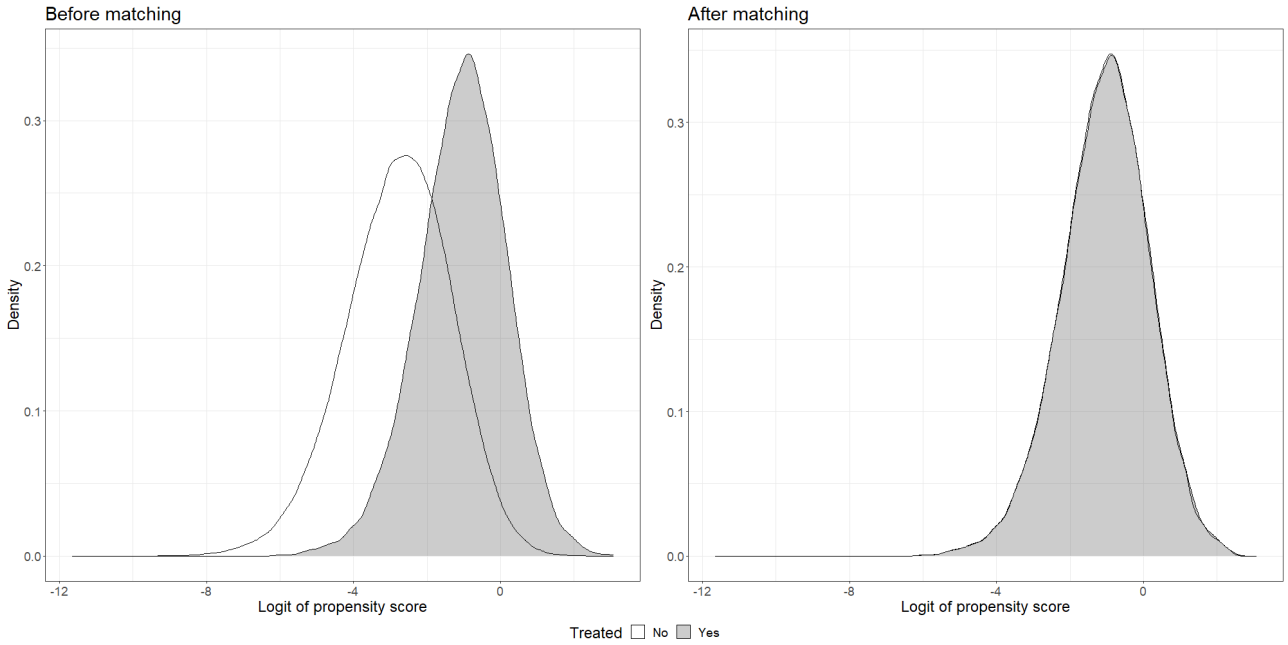
Figure 4.1. Density plot of the logit of propensity scores before and after weighting and matching for the male headline violent adjudications analysis.



Treatment group members were matched to similar untreated offenders, where their propensity scores were within a certain tolerance level (i.e., comparison group scores were within a specified distance from the treatment group score, such as within 0.1 in either direction). 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, weighted proven adjudication rates (any adjudications or violent adjudications) for the treatment and comparison groups were compared. Figures 4.1 and 4.2 show the density of propensity scores before and after matching, to illustrate how the propensity score weighting process creates TSP and non-TSP groups that are comparable on their likelihood to receive TSP. 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 or unmeasured factors could influence the results that emerge from analysis of the observed data.

Figure 4.2. Density plot of the logit of propensity scores before and after weighting and matching for the male headline violent adjudications analysis.



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 be logically inferred where appropriate:

- Section 6, Question 7: Evidence of domestic violence/partner abuse
- 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, we logically imputed entries based on corresponding OASys variables. For example, 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 using the male headline analysis, to measure the possible effect of some methodological decisions on the results (see Table A4.1). 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 6-month male headline analysis for any proven adjudications.

Sentence selection

TSP differs from prior Accredited Programmes that have been evaluated using Propensity Score Matching, due to the presence of multiple participation.

This means that there are some individuals in the treatment group who have participated in TSP during multiple distinct prison stays. In a randomised controlled trial (RCT) design, it would be possible to account for multiple participation by simply excluding individuals with prior participation from either the treatment or comparison groups of the trial. However, because the evaluation of TSP is retrospective and multiple participations in TSP is relatively more frequent than for other accredited programmes, this presents challenges for quasi-experimental evaluation.

If individuals with multiple participations are excluded, it reduces sample sizes. It also selects a specific subset of the prison population that may not be representative of the overall population of people in prison who received TSP. These issues could limit the usefulness and generalisability of any results obtained from subsequent analysis.

Alternatively, if individuals with multiple participations are included, then it is necessary to decide when an individual should appear in the treatment or the comparison group, to produce a reliable estimate of the causal effect of programme participation on the outcomes of interest. This is of particular concern in propensity score matching, because it is necessary to provide selection criteria which classifies observations into treatment and comparison groups in a manner most like an RCT as possible.

Due to the complexity of this evaluation and these different trade-offs, the Justice Data Lab 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 their respective studies. Following exploratory analysis conducted for a related TSP reconviction study (See Annex 13 in Brinn et al., 2023), the method “all participations and all non-participations” (APAN) best balanced methodological robustness and sufficient sample size.

Given the analytical overlap between the TSP reconviction study and this one, the APAN sentence selection methodology was also selected for the prison misconduct impact study.

Table A4.1. A list of sensitivity analyses and their comparison with the headline male all adjudications analysis.

Sensitivity	Explanation	Findings
Standard approach	The chosen approach is displayed here for comparison against other sensitivities. Radius matching (with replacement) using a uniform kernel was applied.	179 variables included in the final model. Mean absolute standardised difference of 0.52%. 12,938 TSP and 82,784 comparisons matched.
Parsimonious	To explore the effect of having fewer variables in the model (greater constraints imposed when determining model variables).	98 variables included in the final model. Mean absolute standardised difference of 1.73%. 12,948 TSP and 83,071 comparisons matched. Results very similar[31] to the standard approach.
Non-parsimonious	To explore the effect of having more variables in the model (fewer constraints imposed when determining model variables)	206 variables included in the final model. Mean absolute standardised difference of 0.50%. 12,944 TSP and 82,787 comparisons matched. Results very similar to the standard approach.
Common support	To explore the effect of having a restriction on "common support". Treatment group members automatically excluded where propensity scores are outside overall range of propensity scores of the comparison group. Comparison group members automatically excluded where propensity scores are outside overall range of propensity scores of the treatment group.	180 variables included in the final model. Mean absolute standardised difference of 0.52%. 12,928 TSP and 82,386 comparisons matched. Results very similar to the standard approach.
Epanechnikov kernel	This explores using an alternative type of kernel matching sometimes used for PSM models.	180 variables included in the final model. Mean absolute standardised difference of 0.52%. 12,942 TSP and 82,945 comparisons matched. Results were very similar to the standard approach.
Matching on propensity scores	To explore the effect of matching on propensity scores rather than the logit of propensity scores.	180 variables included in the final model. Mean absolute standardised difference of 0.54%. 12,938 TSP and 83,718 comparisons matched. Results 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.	72 variables included in the final model (but only 84 available without OASys variables included). Mean absolute standardised difference of 0.54%. 12,946 TSP and 83,452 comparisons matched. TSP results similar to the standard approach, but comparison group rate lower by 0.4%.

[31] Where results are described as "very similar", absolute differences between the TSP and comparison groups are less than or equal to 0.1 percentage point for rates of any proven adjudications and less than or equal to 0.01 for the frequency of any proven adjudications.

Annex 5: Limitations and caveats

Whilst this study uses a recognised evaluation methodology (propensity score matching), which is considered level 4 on the SMS (Scientific Methods Scale) 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 et al. (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 TSP and of the likelihood of receiving a prison adjudication, the effects of variables that are not recorded cannot be known. As such, it is possible that unobserved factors could influence these results (i.e., results may change if important unobserved are included).

- 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 TSP if they were assigned to the intervention.

- These analyses only concern prison misconduct outcome data. There may be other important outcomes to consider for rehabilitation interventions. Examples of this might include reductions in reoffending rates, increased employability, improved mental health, healthier relationships, or positive attitudes. A study of the impact of TSP on rates of proven reconvictions has been published alongside this study (see Brinn et al., 2023).

- As noted in the methodology section, we used TSP start dates as the time point at which follow-up begins as we couldn't use index release dates or TSP end dates. As a result, some adjudications may have occurred during TSP, not after TSP. This means that the number of adjudications may be inflated in the TSP group. Since TSP is a relatively short programme, however, it was agreed that a small number of participants would be likely to have received an adjudication without receiving exposure to any TSP modules.

- We also chose a 6-month follow-up period. We conducted scoping work prior to analysis to determine the effects of using a 6-, 9-, or 12-month follow-up. Two key implications led us to choose a 6-month follow-up. First, any follow-up longer than 6-months would render several important sub-analyses unviable due to the reduction in sample size. Second, enforcing longer follow-ups requires participants with longer prison sentences. This biases our sample towards those with more serious offences. A

smaller proportion of our sample spent more than 6 months in prison after starting TSP. TSP is also not a programme designed for the most serious offenders.

- A sub-analysis was performed to isolate the effect of 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 TSP. The effect of 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 comparison individuals (those who did not participate in TSP during the index sentence) instead undertook a non-accredited offending behaviour programme during the index sentence or some other non-business-as-usual intervention, which might have had an impact on prison misconduct rates.

- This evaluation measures a treatment effect using proven adjudication outcomes (and proven violent adjudication outcomes) in accordance with the standard HM Prison and Probation Service definition as set out in the Prisoner Discipline Procedures (Adjudications) and other relevant policies. As such, the study only accounts for proven adjudications and does not measure treatment effects on prison misconduct that is committed but is either not recorded by prison staff or does not lead to an adjudication hearing. It is also with noting that the recording of adjudications may vary across prisons.

- 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 prison misconduct rates 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.

- Sub-analyses with small sample sizes increases the chance of some levels of a matching variable to be zero. For example, in our sub-analysis of any type of adjudication for the "Asian and Asian British" sub-sample, no post-matching members of the TSP group were on a short sentence ("Less than 6 months"), while some in the comparison group were. It is possible for treatment and comparison groups to differ on

individual categories of one or more variables in a randomised controlled trial or for a group to contain no individuals with a certain characteristic assumed to be shared equally by proportions of both groups. However, it emphasises the need to review the standardised differences tables when interpreting any of our analyses, particularly those with small sample sizes.

- Exploratory work for the “Participated in TSP only” sub-analysis indicated that a small number of comparison group individuals were enrolled on other accredited 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.

- The Measure of the Quality of Prison Life (MQPL) is self-reported data related to prison life from the perspective of the prisoners and is not an objective measure of the observed prison environment. Relatively more objective measures exist of some of the concepts measured by MQPL – albeit there are no perfectly objective measures. For example, both HMPPS and His Majesty’s Inspectorate of Prisons (HMIP) report more granular data on "purposeful activity". MQPL data is also collected at each prison approximately every two years, which is less frequent than other comparable sources of prison data. While this means that MQPL data might not account for all the variation in prison environments, it was considered sufficient for our purposes for two reasons. Firstly, it provided a range of relevant variables from a single source and in a consistent format over the whole time period required. Secondly, its focus on experience means it may account for individual perceptions of the prison environment that are likely to influence behaviour.

- There were several methodological approaches to selecting which participations in TSP to use as the “experimental” one in this study (See Brinn et al., 2023: Annex 13). Since the two studies use the same method and broadly similar data, it was assumed that the same approach could be applied to the prison misconduct study. However, this has two key limitations.

  Firstly, it means that there are some individuals in the treatment group who have participated in TSP more than once. It is possible that participating in TSP multiple times has a different or compounding effect compared to participating in TSP just once. Despite this, all participations in TSP are treated as equal.

  Secondly, randomised controlled trials (RCTs) are widely considered to be the “gold standard” methodology for examining the effectiveness of an intervention (Hariton & Locascio, 2018). An “RCT-like” methodology tested for the reconviction study was the most comparable to a real RCT. However, the ‘all participations all non-participations’ (APAN) methodology was chosen because (a) the “RCT-like” method reduced the sample size and (b) the dataset generated by each method was found to be closely matched.

- Sexual offenders are known to have relatively low recidivism rates (Hanson, 2018). It is therefore reasonable to propose that this propensity for relatively less criminal

behaviour after release from prison than those with other conviction types may also be reflected in less prison misconduct before release. For example, a study conducted by MOJ analysts indicated that specialist Category B and Category C prisons for those with sex offenders had the lowest prison assault rate – only Open prisons had lower rates (Dent et al., 2015) Its manual states that TSP is suitable for sexual offenders, so it was decided 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.

- Finally, it was assumed that moving from one prison to another could have an effect on the likelihood that someone would engage in prison misconduct. For example, Cook and Kim (2023) found an association between "code support" (a belief in the use of violence to gain status and protect against victimisation during incarceration) and misconduct in US jail (see also Klatt et al., 2016; Mears et al., 2013). This belief system is described as being imported into custodial settings, and therefore may be "of particular salience when individuals enter prison" (Mears et al., 2013: p. 695), particularly before any counter-measures to those beliefs can be put in place (e.g., programmes, family support, etc.)

  To account for this possibility, we created a matching variable that represents the number of prisons at which an individual had a record of being received in the three months prior to participation in TSP (or for the comparison group, the TSP start date assigned to them). However, this approach does not consider the *nature* of any prison move, such as whether that move is to and from prisons with different functions (e.g., Category A to Category B or vice versa). If it is the specific nature or type of prison move that affects the likelihood of engaging in misconduct, not simply moving prison in and of itself, it may not be accounted for completely in this study.

Annex 6: 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[32] from 0.65 to 0.80) that broadly represent reductions in prison misconduct of between approximately 5 and 10 percentage points in the treatment groups compared to the comparison groups. Statistical power also depends on the baseline rate of prison misconduct in the wider prison population. Using the full adjudications dataset available we estimated baseline rates[33] of 29% for any proven adjudication and 7% for any proven violent adjudication. For any analyses that focused on violent outcomes, we used the estimated violent baseline rate rather than the general estimated baseline rate. 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 (Table A6.1) which were excluded due to their statistical power being less than sufficient (as described above). It was agreed that if there was only one "green" rating for a subgroup, these wouldn't be analysed.

Table A6.1. Power analysis RAG rating of excluded sub-analyses.

Sub-analysis (excluded)	RAG rating
Female Violent Headline	RED
Male Violent Non-completers	RED
Female All Non-completers	RED

[32] An odds ratio (OR) is the odds that an outcome will occur (like a prison adjudication) given exposure to an intervention (like TSP), compared to the odds of the outcome occurring if not exposed to that intervention. An OR greater than 1 indicates the intervention increases the odds of the outcome (more prison adjudications after TSP). An OR of less than 1 indicates the decreased occurrence of the outcome (fewer prison adjudications after TSP).

[33] The "baseline rate" is the rate at which an outcome (e.g., general or violent prison adjudications) naturally occurs in the wider population absent of any intervention.

Sub-analysis (excluded)	RAG rating
Female Violent Completers	RED
Female Violent Non-completers	RED
Female All Ideally suitable	RED
Female All Not ideally suitable	RED
Female Violent Ideally suitable	RED
Female Violent Not ideally suitable	RED
Male All 2014-16 QOD Broadly met	GREEN
Male All 2014-16 QOD Compromised	AMBER
Male All 2016-19 QOD Compromised	RED
Male Violent 2014-16 QOD Broadly met	AMBER
Male Violent 2014-16 QOD Compromised	RED
Male Violent 2016-19 QOD Broadly met	RED
Male Violent 2016-19 QOD Compromised	RED
Female All 2014-16 QOD Broadly met	RED
Female All 2014-16 QOD Compromised	RED
Female All 2016-19 QOD Broadly met	RED
Female All 2016-19 QOD Compromised	RED
Female Violent 2014-16 QOD Broadly met	RED
Female Violent 2014-16 Compromised	RED
Female Violent 2016-19 QOD Broadly met	RED
Female Violent 2016-19 Compromised	RED
Female All Age 18-25	RED
Female All Age 26-30	RED
Female All Age 31-49	RED
Female All Age 50+	RED
Female Violent Age 18-25	RED
Female Violent Age 26-30	RED
Female Violent Age 31+	RED
Male All OGRS 0-24	GREEN
Male Violent OGRS 0-24	RED
Female All OGRS 0-24	RED
Female All OGRS 25-49	RED
Female All OGRS 50-74	RED
Female All OGRS 75+	RED
Female Violent OGRS 0-24	RED

Sub-analysis (excluded)	RAG rating
Female Violent OGRS 25-49	RED
Female Violent OGRS 50-74	RED
Female Violent OGRS 75+	RED
Male Violent Participated in another AP	RED
Female All Participated in another AP	RED
Female Violent Participated in TSP only	RED
Female Violent Participated in another AP	RED
Female All More likely to present with characteristics of LDC	RED
Female All Less likely to present with characteristics of LDC	RED
Female Violent More likely to present with characteristics of LDC	RED
Female Violent Less likely to present with characteristics of LDC	RED
Male All Other ethnicities	RED
Male Violent Asian and Asian British ethnicity	RED
Male Violent Black, Black British, Caribbean, and African ethnicity	RED
Male Violent Mixed and multiple ethnic groups	RED
Male Violent Other ethnicities	RED
Male Violent White ethnicity	GREEN
Female All Asian and Asian British ethnicity	RED
Female All Black, Black British, Caribbean, and African ethnicity	RED
Female All Mixed and multiple ethnic groups	RED
Female All Other ethnicities	RED
Female All White ethnicity	AMBER
Female Violent Asian and Asian British ethnicity	RED
Female Violent Black, Black British, Caribbean, and African ethnicity	RED
Female Violent Mixed and multiple ethnic groups	RED
Female Violent Other ethnicities	RED
Female Violent White ethnicity	RED

Annex 7: Full list of analyses undertaken

The full list of analyses undertaken are listed below in Table A7.1. Male and female sub-analyses were run separately. Sub-analyses that were considered but did not reach thresholds sufficient statistical power are listed in Annex 6. Each analysis was conducted for either rates or frequencies of 6-month “any” or “violent” adjudication outcomes. For each of the following sub-analyses, the treatment group was matched to offenders in England and Wales using demographics, criminal history, individual offending-related risks and needs, and select prison characteristics.

Table A7.1. A list of the analyses undertaken by gender and outcome measured.

Analysis	Male		Female	
	Any	Violent	Any	Violent
Overall (headline)	✓	✓	✓	x
Met ideal suitability criteria	✓	✓	x	x
Did not meet ideal suitability criteria	✓	✓	x	x
Completed TSP	✓	✓	✓	x
Did not complete TSP	✓	x	x	x
Programme integrity broadly maintained (2016-19)	✓	x	x	x
With OGRS3 risk score 25-49 (low risk)	✓	✓	x	x
With OGRS3 risk score 50-74 (medium risk)	✓	✓	x	x
With OGRS3 risk score 75+ (high risk)	✓	✓	x	x
Participated in TSP only	✓	✓	✓	x
Participated in another accredited programme before TSP	✓	x	x	x
Asian and Asian British ethnicity	✓	x	x	x
Black, Black British, Caribbean, and African ethnicity	✓	x	x	x
Mixed and multiple ethnic groups	✓	x	x	x
White ethnicity	✓	x	x	x
More likely to present with characteristics of LDC	✓	✓	x	x
Less likely to present with characteristics of LDC	✓	✓	x	x
Aged between 18-25	✓	✓	x	x
Aged between 26-30	✓	✓	x	x
Aged between 31-49	✓	x	x	x
Aged 31+	x	✓	x	x
Aged 50+	✓	x	x	x

Annex 8: Further information on ideal suitability and programme integrity

Ideal Suitability

Selection onto TSP as part of the programme manager "risk over-ride" group

Not all of the male 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 OGRS3 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 that make TSP a potentially suitable offer for support. These additional eligibility criteria are referred to as "risk over-rides" and allow Treatment Managers to use their experience 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 or above using the Risk Matrix 2000/s[34].
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.

[34] See Risk Matrix Scoring guide (2007) for details. The RM2000/s has been replaced in routine risk assessment by the OASys Sexual (Reoffending) Predictor (OSP) (see Implementation and use of OASys Sexual reoffending Predictor (OSP) Policy Framework).

Table A8.1. Classification of the not ideally suitable into clinical and risk over-ride groups.

Over-ride group	Frequency	Proportion of the not ideally suitable sample (%)
Group 1	443	12.4%
Group 2	18	0.5%
Group 3	404	11.3%
Group 4	1444	40.5%
Group 5	1,781	49.9%

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.

Overall, 1,785 (50.1%) of those participants included in the "not ideally suitable" sub-analysis either qualified for a risk over-ride or met other eligibility criteria. The 3,566 participants included in the male "not ideally suitable" sub-analysis could be placed into one of five groups (see Table A8.1):

- **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 and above on one or more relevant items assessed within OASys (e.g., risk to prison staff, risk to other prisoners, 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.
- **Group 5**: Those who were not eligible for a risk over-ride because:
  1. They did not meet the TSP need criteria in full.
  2. They did not meet the exception to the TSP risk criteria as per group 1-4 above.
  3. 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 practise to ensure effective 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 dated "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 KLOEs.

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 practise 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/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 programmes

A total of 1,634 records of participation in another Accredited Programme prior to TSP exist for our pre-matched TSP treatment group of 13,902.

Table A9.1 lists the programmes attended, the number of TSP participants who attended that programme, and the proportion of the whole treatment group that frequency represents. Some of the programmes listed are delivered by external providers and not HM Prison and Probation Service. Some of these programmes may have been discontinued and are no longer available, and some may no longer be recommended for accreditation by the Correctional Services Accreditation and Advice Panel.

Table A9.1. List of previous participations in APs prior to participation on TSP.

Accredited Programme	**Number of TSP participants**	**Percentage of TSP treatment group**
Total participation in any other accredited programme	1,634	11.8%
Prisons Addressing Substance Related Offending (P-ASRO)	301	2.2%
Building Steps to Recovery (BSR)	201	1.4%
Controlling Anger and Learning to Manage it (CALM)	175	1.3%
Sex Offender Treatment Programme (SOTP) Core Programme (CP)[35]	196	1.4%
RESOLVE[36]	94	0.7%
Short Duration Programme (SDP)	70	0.5%
RAPt 12-Step Programme	62	0.4%
Alcohol Related Violence (ARV)	56	0.4%
Control of Violence for Angry Impulsive Drinkers (COVAID)	78	0.6%
Healthy Relationships Programme (HRP)	21	0.2%
Therapeutic Community (TC)	43	0.3%
Enhanced Thinking Skills (ETS)	44	0.3%
SOTP Extended Programme (EP)	14	0.1%
Self-Change Programme (SCP)	2	0.01%

[35] "CP", "EP", and "RP" are common initialisations of the SOTP core, extended, and rolling programmes. In the absence of evidence that there were other accredited programmes that could be shortened in this way and given that all three initialisations are in this list, it is assumed that these are the programmes being referred to. Our records also indicate that approximately one-third of the participants who have participated in CP, EP and RP have one or more previous or index sexual convictions.

[36] This is not an acronym; it is just typically capitalised.

Accredited Programme	Number of TSP participants	Percentage of TSP treatment group
Democratic Therapeutic Community (DTC)	52	0.4%
Alcohol Dependency Treatment Programme (ADTP)	16	0.1%
Challenge to Change (Kainos CTC)	29	0.2%
Focus on Resettlement (FOR)	2	0.01%
Better Lives Booster (BLB)	12	0.09%
Building Better Relationships (BBR)	15	0.1%
Becoming New Me (BNM)	16	0.1%
Healthy Sex Programme (HSP)	1	0.007%
Prison Partnership 12-Step Programme (PPTSP)	15	0.1%
Bridges[37]	11	0.08%
Sex Offender Treatment Programme (SOTP)[38]	9	0.06%
FOCUS[39]	44	0.3%
Cognitive Skills Booster (CSB)	14	0.1%
Choices Actions Relationships Emotions (CARE)	4	0.03%
SOTP Rolling Programme (RP)	5	0.04%
Juvenile Enhanced Thinking Skills (JETS)	11	0.08%
Timewise[40]	4	0.03%
Adapted Better Lives Booster (ABLB)	3	0.02%
Horizon	1	0.007%
Cognitive Self Change Programme (CSCP)	2	0.01%
High Intensity Pilot (SOTP HI)[41]	1	0.007%
Chromis	4	0.03%

[37] May also be referred to as “The Bridge”.

[38] This was originally referred to in the database as “STOP”. In the absence of evidence there is or was an accredited programme called STOP, this has been assumed to be a clerical error and should be SOTP.

[39] This is not an acronym; it is just typically capitalised.

[40] Timewise is not an accredited programme, it is a “shorter cognitive skills programme which specifically focuses on prison violence” (see Prison Safety and Reform White Paper)

[41] This was originally referred to in the database as “HI”. HMPPS documentation suggest that a High Intensity SOTP programme was piloted (see Glossary of Programmes). Another possibility is the Healthy Identity Intervention, delivered by HMPPS for the violent extremist cohort, but that programme was not available at the time this single participation occurred.

Annex 10: Ethnic groups

Self-reported ethnicity groupings were created using the Self Defined Ethnicity – 18+1 Standard as per the Office for National Statistics categories[42].

Table A10.1. Self-reported ethnicity groupings.

Ethnic group	18+1 Self-reported ethnicity code	18+1 Self-reported 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

[42] Self-defined ethnicity - 18plus1

Annex 11: Profile of the treatment groups

The following descriptive statistics aim to provide an overview of the characteristics of 12,938 male and 953 female treatment groups available for the headline analyses. The treatment group included males whose ages ranged from 18 to 91 years and females whose ages ranged from 18 to 65 years. The table below (Table A11.1) contains information on demographics, offence history, offending-related risks/needs, and participation in other accredited programmes.

Table A11.1. Descriptive statistics for the male and female TSP participant samples. Variables are presented as numbers and proportions (%) except where stated.

Variable	Male sample	Female sample
Age at TSP participation (in years)	Mean = 30.4 IQR = 23-36	Mean = 31.9 IQR = 25-37
Ethnic group		
Asian and Asian British	875 (6.6%)	11 (1.2%)
Black, Black British, Caribbean, and African	1,802 (13.6%)	56 (5.9%)
Mixed and multiple ethnic groups	692 (5.2%)	44 (4.6%)
Arab and other ethnic groups	64 (0.5%)	2 (0.2%)
White	9,642 (72.9%)	834 (87.5%)
Unknown/Not stated	144 (1.1%)	6 (0.6%)
Nationality		
UK national	12,516 (94.7%)	922 (96.7%)
Non-UK national	549 (4.2%)	18 (1.9%)
Unknown	154 (1.2%)	13 (1.4%)
OASys learning difficulties and challenges (LDC)		
No problems	10,118 (76.5%)	763 (80.1%)
Some problems	1,319 (10.0%)	61 (6.4%)
Significant problems	511 (3.9%)	21 (2.2%)
Unknown	1,271 (9.6%)	108 (11.3%)
OASys Difficulties with reading, writing, or numeracy		
No problems	9,331 (70.6%)	700 (73.5%)
Some problems	2,904 (22.0%)	197 (20.7%)
Significant problems	789 (6.0%)	42 (4.4%)
Unknown	195 (1.5%)	14 (1.5%)
Participation on other accredited programmes		
Participated in TSP only	11,900 (90.0%)	881 (92.4%)
Participated in another AP before TSP	1,319 (10.0%)	72 (7.6%)

Variable	Male sample	Female sample
Index offence sentence length		
Less than or equal to 6 months	21 (0.2%)	3 (0.3%)
Between 6 and 12 months	107 (0.8%)	4 (0.4%)
12 months to less than 4 years	4,447 (33.6%)	510 (53.5%)
4 to 10 years	5,999 (45.4%)	359 (37.7%)
More than 10 years	1,100 (8.3%)	14 (1.5%)
Indeterminate sentence	690 (5.2%)	13 (1.4%)
Life sentence	855 (6.5%)	50 (5.3%)
Risk assessment		
Offender Group Reconviction Scale Predictor (OGRS3) score	Mean = 60.3 IQR = 49-76	Mean = 63.4 IQR = 51-79
Offender Violent Predictor (OVP) score	Mean = 41.5 IQR = 27-56	Mean = 36.0 IQR = 21-49
Index offence group		
Violence against the person	4,129 (31.2%)	368 (38.6%)
Sexual offences	1,730 (13.1%)	13 (1.4%)
Robbery	1,536 (11.6%)	198 (20.8%)
Theft offences	1,901 (14.4%)	139 (14.6%)
Possession of weapons	333 (2.5%)	52 (5.5%)
Drug offences	2,203 (16.7%)	122 (12.8%)
Summary offences excluding motoring	661 (5.0%)	22 (2.3%)
Fraud offences	211 (1.6%)	9 (0.9%)
Public order offences	356 (2.7%)	14 (1.5%)
Criminal damage and arson	47 (0.4%)	7 (0.7%)
Miscellaneous crimes against society	74 (0.6%)	9 (0.9%)
Summary motoring offences	37 (0.3%)	0 (0.0%)
Unknown	1 (0.0%)	0 (0.0%)
Prior offences		
Previous offences	Mean = 30.0 IQR = 10-40	Mean = 37.7 IQR = 11-52
Previous violent offences	Mean = 3.6 IQR = 1-5	Mean = 4.1 IQR = 1-5
Previous convictions	Mean = 12.8 IQR = 5-18	Mean = 16.7 IQR = 6-23
Previous custodial sentences	Mean = 4.0 IQR = 1-6	Mean = 4.1 IQR = 0-6
Prior prison adjudications		

Variable	Male sample	Female sample
Number of previous proven adjudications in 3 months prior to TSP participation	Mean = 0.3 IQR = 0-0	Mean = 0.6 IQR = 0-1
Number of previous proven violent adjudications in 3 months prior to TSP participation	Mean = 0.04 IQR = 0-0	Mean = 0.1 IQR = 0-0
Time since conviction (in days)	Mean = 518 IQR = 172-518	Mean = 317 IQR = 112-321
Prison moves 3-months prior to TSP	Mean = 0.2 IQR = 0-0	Mean = 0.17 IQR = 0-0

Note: IQR = interquartile range (50% of all the values are between these two numbers).

Annex 12: Odds ratios for binary measures

Table A12.1 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.

An odds ratio (OR) is the odds that an outcome will occur (like a prison adjudication) after exposure to an intervention (e.g., participating TSP), compared to the odds of that outcome occurring if not exposed to an intervention (e.g., not participating in TSP). It is important to note that ORs are *not* probabilities.

- An OR **greater than 1** indicates that the outcome is more likely to occur (e.g., you are more likely to receive an adjudication after participating in TSP).
- An OR of **less than 1** indicates that the outcome is less likely to occur (e.g., you are less likely to receive an adjudication after participating in TSP).
- An **OR of 0.50** indicates that the outcome is half as likely to occur after exposure to an intervention (like participating in TSP) than in the absence of exposure to that intervention.

Table A12.1. Odds ratios for the binary measures by analysis

Analysis	Lower CI	OR	Upper CI
Male: Any adjudications			
Overall (headline)	0.891	0.912	0.933
Met ideal suitability criteria	0.881	0.901	0.921
Did not meet ideal suitability criteria	0.885	0.940	0.994
Completed TSP	0.833	0.854	0.876
Did not complete TSP	1.514	1.682	1.871
Programme integrity broadly maintained (2016-19)	0.894	0.958	1.022
With OGRS3 risk score 25-49 (low risk)	0.875	0.932	0.987
With OGRS3 risk score 50-74 (medium risk)	0.891	0.918	0.946
With OGRS3 risk score 75+ (high risk)	0.850	0.885	0.921
Participated in TSP only	0.876	0.897	0.919
Participated in another accredited programme before TSP	0.939	1.002	1.064
Asian and Asian British ethnicity	0.793	0.870	0.947
Black, Black British, Caribbean, and African ethnicity	0.809	0.856	0.903
Mixed and multiple ethnic groups	0.862	0.945	1.028
White ethnicity	0.892	0.917	0.942

Analysis	Lower CI	OR	Upper CI
More likely to present with characteristics of LDC	0.955	0.999	1.045
Less likely to present with characteristics of LDC	0.855	0.881	0.907
Aged between 18-25	0.868	0.894	0.920
Aged between 26-30	0.845	0.890	0.934
Aged between 31-49	0.853	0.895	0.938
Aged 50+	0.779	0.977	1.115
Male: Violent adjudications			
Overall (headline)	0.843	0.883	0.921
Met ideal suitability criteria	0.886	0.905	0.942
Did not meet ideal suitability criteria	0.788	0.904	1.009
Completed TSP	0.774	0.817	0.858
With OGRS3 risk score 25-49 (low risk)	0.749	0.867	0.971
With OGRS3 risk score 50-74 (medium risk)	0.894	0.951	1.003
With OGRS3 risk score 75+ (high risk)	0,748	0.806	0.861
Aged between 18-25	0.821	0.864	0.905
Aged between 26-30	0.686	0.786	0.875
Aged 31+	0.788	0.905	1.010
Participated in TSP only	0.834	0.875	0.915
More likely to present with characteristics of LDC	0.871	0.951	1.024
Less likely to present with characteristics of LDC	0.830	0.882	0.931
Female: Any adjudication			
Overall (headline)	0.833	0.881	0.929
Completed TSP	0.805	0.858	0.910
Participated in TSP only	0.792	0.840	0.888

Glossary of terms

Adjudication: Adjudications are part of the prison disciplinary system. Any rule breaking activity or accusation of rule breaking that occurs within the prison system can be tried and punished through the internal prison disciplinary system. The hearings are called adjudications.

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.

Confidence levels: 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) will be a key output for analyses as the prison misconduct rates for the treatment and control groups are from samples of larger populations.

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

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.

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.

Meta-Analysis: A meta-analysis consists of two or more independent from each other studies that examine the same measure across different timelines and/or circumstances. The goal of a meta-analysis is to produce a single estimate sourced from several others by statistically analysing and weighting previous findings.

MQPL: The Measuring the Quality of Prison Life (MQPL) survey assesses prisoners' perceptions of the quality of decency and safety at the prison in which they are serving a custodial sentence. The MQPL survey is undertaken at each prison as a part of a rolling three-year programme.

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 Version 3 (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.

Outcome rate: The proportion of individuals in the total sample (and between groups) who receive a proven adjudication.

Outcome frequency: The number of proven adjudications, expressed per person in the total sample (and between groups).

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. 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 (i.e., there is not a true difference to be found between the groups).

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'.

Violent adjudication: An adjudication that is classified in the prison adjudications data as being of the type "Violence" (i.e., fights and/or assaults).

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.

Brinn, A., Preston, J., Costello, R., Opoku, T., Sampson, E., Elliott, I. A., & Sorbie, A. (2023). Reoffending impact evaluation of the prison-based Thinking Skills Programme (TSP). London, U.K.: Ministry of Justice.

Cann, J., Falshaw, L., Nugent, F. & Friendship, C. (2003). Understanding what works: Accredited cognitive skills programmes for adult men and young offenders (Research, Development & Statistics Directorate Research Findings: No. 226). London, U.K.: Home Office.

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.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. London, U.K.: Routledge.

Cook, C. L., & Kim, M. (2023). Assessing the validity of criminological theories of misconduct in jail. Crime and Delinquency. Online first: https://doi.org/10.1177/00111287231170113

Dent, P, Dorrell, D., & Howard, P. (2015) Understanding prison violent trends and correlates. Prison Service Journal, 221, 4-9.

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, U.K.: Home Office.

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). London, U.K.: Routledge.

French, S. A., & Gendreau, P. (2006). Reducing prison misconducts: What works. Criminal Justice and Behavior, 33(2), 185-218.

Friendship, C., Blud, L., Erikson, M., Travers, R., & Thornton, D. (2003). Cognitive-behavioural treatment for imprisoned offenders: An evaluation of HM Prison Service's cognitive skills programmes. Legal and Criminological Psychology, 8(1), 103-114.

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, U.K.: 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.

Klatt, T., Hagl, S., Bergmann, M. C., & Baier, D. (2016). Violence in youth custody: Risk factors of violent misconduct among inmates of German young offender institutions. European Journal of Criminology, 13(6), 727–743.

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 (pp. 57-71). Dordrecht, The Netherlands: Springer.

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.

McMurran, M., & Theodosi, E. (2007). Is treatment non-completion associated with increased reconviction over no treatment? Psychology, Crime and Law, 13(4), 333-343.

Mears, D. P., Stewart, E. A., Siennick, S. E., & Simons, R. L. (2013). The code of the street and inmate violence: Investigating the salience of imported belief systems. Criminology, 51(3), 695-728.

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

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

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

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, U.K.: Ministry of Justice.

Schafer, 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. (2016). Why what works works. [Doctoral thesis, Leicester University]. https://figshare.le.ac.uk/articles/thesis/Why_What_Works_Works/10213172

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.

Tong, L. S. J., & Farrington, D. P. (2006). How effective is the “Reasoning and Rehabilitation” programme in reducing re-offending? A meta-analysis of evaluations in four countries. Psychology, Crime & Law, 12, 3–24.

Wakeling, H. (2018), The development of a screen to identify individuals who may need support with their learning. London, U.K. Ministry of Justice.

Wakeling, H., & Ramsay, L. (2020). Learning disability and challenges in male prisons: programme screening evaluation. Journal of Intellectual Disabilities and Offending Behaviour, 11(1), 49-59.

Weisburd, D., Lum, C. M., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? Annals of the American Academy of Political and Social, Science, 578, 50-70.