Ministry of Justice

The Learning Disabilities and Challenges (LDC) suite of accredited offending behaviour programmes

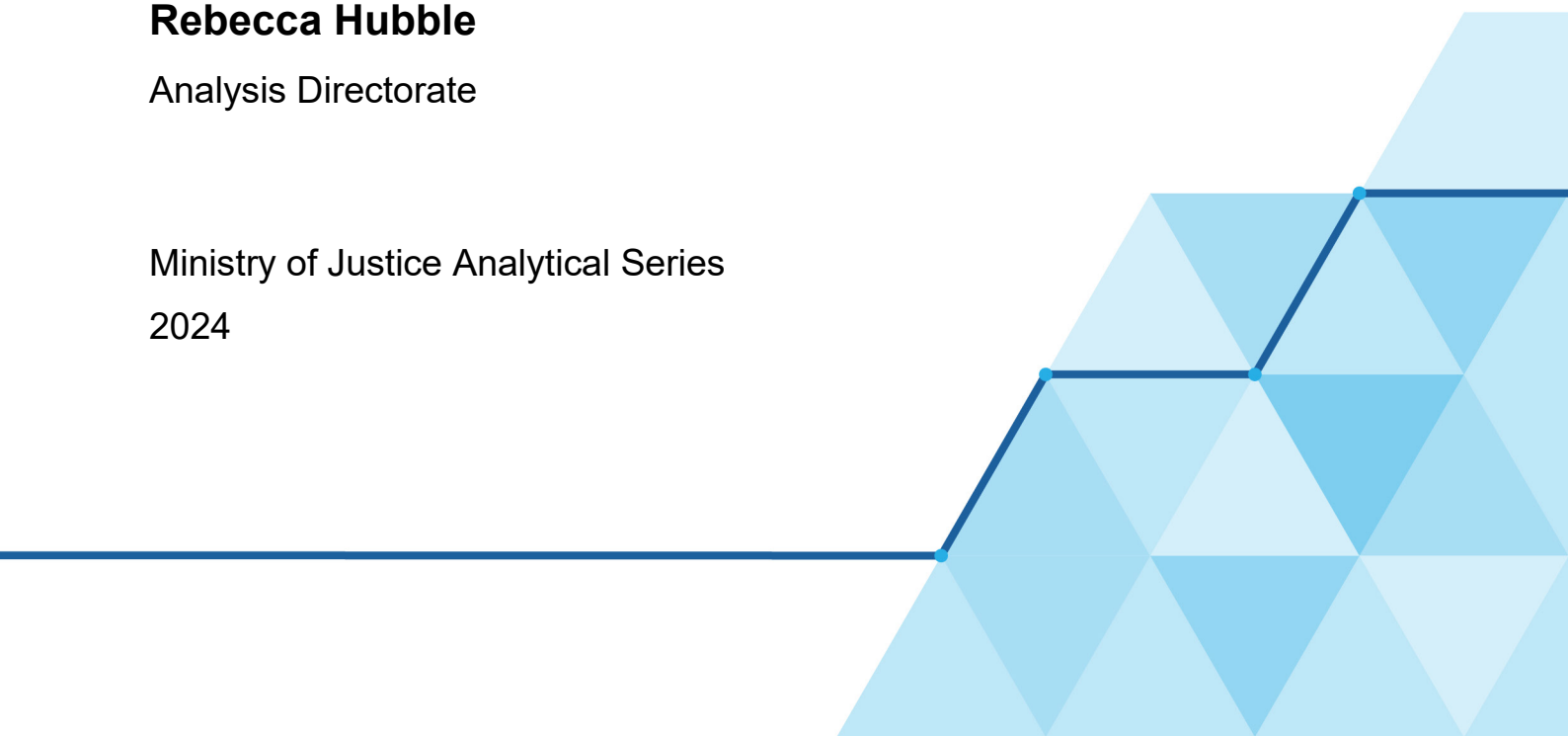
An uncontrolled before-after evaluation of clinical outcomes

Rebecca Hubble

Analysis Directorate

Ministry of Justice Analytical Series

2024



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

Disclaimer

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

First published 2024

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

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

ISBN 978 1 911691 43 3

Acknowledgements

For their subject matter advice and guidance, I would like to thank colleagues from HMPPS. From Ministry of Justice Analysis Directorate, my thanks go to Aleksandra Plochocka for diligent quality assurance checks. I would also like to thank colleagues who reviewed and commented on this report, including Eleftherios Nomikos, Alana Diamond and two anonymous external independent peer reviewers.

Contents

List of tables

List of figures

1. Summary

Introduction and evaluation aims

The Learning Disabilities and Challenge (LDC) suite are accredited offending behaviour programmes delivered by His Majesty's Prison and Probation Service (HMPPS) in custody and the community for adults assessed as having mild impairments in intellectual and adaptive functioning. There are four programmes within the suite: Becoming New Me Plus (BNM+), New Me Strengths (NMS), Living as New Me (LNM), and the Heathy Sex Programme (HSP). Our interim outcome evaluation attends to BNM+ and NMS only. This is because LNM and HSP are secondary programmes, i.e., they are intended to be delivered after completion of a primary programme.

BNM+ is for men with LDC assessed as high or very high risk of reoffending, who present with at least one or more strong criminogenic needs across multiple domains and have a general violence, intimate partner violence and/or sexual offending conviction(s). NMS is designed for people with any offence convictions, who have been assessed as medium or above risk of reoffending, with sufficient levels of criminogenic need addressed by the programme.

The aim of this interim outcome evaluation was to determine whether BNM+ and NMS participants were making positive progress against programme targets reflected in the Success Wheel Measure (SWM). The SWM, designed by HMPPS, is the core metric for measuring participant progress against programme targets for participants of BNM+ and NMS. The assessment domains are: (1) Managing Life's Problems, (2) Healthy Thinking, (3) Positive Relationships, (4) Healthy Sex (for those with a sexual offence conviction only), and (5) Sense of Purpose (desistance from crime). The research also aimed to identify if individual (relating to the person) or programme delivery factors affected changes in SWM scores, and whether these changes varied between assessment domains.

Methodological approach

An uncontrolled before-after evaluation was conducted using SWM scores from 327 men who completed BNM+ or NMS between November 2020 and March 2023, of which 228 (after missing assessment data were removed) comprised the evaluation sample. SWM scores were summed and standardised to allow direct comparison of those with four target intervention domains to those with five.[1] For example, an individual with four domains present would normally have a maximum sum of 20, whilst an individual with five domains present would have a maximum sum of 25. The fifth domain was present for individuals deemed to require a domain related to sexual offending. A paired t-test was used to compare pre-to-post programme SWM scores. A paired t-test was used to compare pre- and post-programme SWM scores.

The impact of individual and programme delivery factors on SWM was analysed using a linear regression model. Factors included were: (1) initial SWM scores, (2) which programme they had participated in (BNM+ or NMS), (3) intervention format (1:1, group, or small group format), and whether they had any combination of a history of (4) general violence, (5) intimate partner violence, and/or (6) sexual offending.

The effect of rating type, (i.e., whether insight and skills were being rated by the programme facilitator, participant, or both), on SWM score changes was examined using a repeated measures ANOVA, and a multivariate model was used to compare predictor variables across rating types. The same methods were then applied to explore differences in SWM score changes between SWM domains.

Interpreting results

The lack of a control group means changes in the SWM scores cannot be directly attributed to BNM+ or NMS participation, as they could be due to unobserved factors such as natural improvement or SWM scorer bias.

[1] The overall scores are reflective of the participants' criminogenic need profiles, and it was deemed suitable to standardise the scorings to allow for direct comparisons between individuals, regardless of the number of domains.

Furthermore, SWMs validation is limited to sexual offending which is an additional limitation. Therefore, results should be viewed as indicative rather than conclusive of the suite's effectiveness.

Key results

- A large increase in total SWM scores from pre-to-post programme participation was found, indicating positive progress against BNM+ and NMS targets.
- Participants with lower initial insights and skills showed greater improvement, with a large effect size.
- Participants of BNM+ had a greater degree of pre-to-post programme participation change compared to those that participated in NMS. This effect size was small.
- Participants with a history of general violence had a smaller pre-to-post programme participation change compared to those without, this effect was very small.
- No other individual or programme delivery factors were found to predict change in total SWM score.
- Facilitator ratings had greater pre-to-post change in total SWM score.
- The impact of individual and programme delivery factors did not differ by rating type.
- All Success Wheel domains had positive pre-to-post programme participation change in SWM score.
- Positive change was lower in the Sense of Purpose domain compared to the Managing Life's Problems domain.
- The difference in pre-to-post programme participation change in SWM score between BNM+ and NMS participants was found to be smaller in the Sense of Purpose domain than the Managing Life's Problems and Healthy Thinking domains.

Conclusions

These interim evaluation results suggest that participation in BNM+ and NMS is associated with positive changes in key programme targets. Participants with reported lower levels of insight and skills appear to benefit more.

These findings are promising but should be considered in light of the evaluation's methodological limitations.

2. Introduction

2.1 Learning Disabilities and Challenges suite

A learning disability, also known as an intellectual disability, is characterised by three components: significant impairments in intellectual functioning, significant impairments in adaptive and social functioning, and the onset of these impairments before adulthood. 'Intellectual functioning' refers to a person's capability of "reasoning, problem solving, abstract thinking, judgment, academic learning and learning from experience" (American Psychological Association, 2013). It can be detected by clinical evaluation and Intelligence Quotient (IQ) testing, for example, using the Wechsler Adult Intelligence Scale – Fourth UK Edition (WAIS-IV UK) (Wechsler, 2010). 'Adaptive and social functioning' refers to a person's ability to cope with daily lifestyle tasks which meet sociocultural standards. This could manifest as problems relating to self-care, engaging with others, or focusing on a task.

The criterion for significant impairment in intellectual functioning is an IQ score that is below two standard deviations of the population mean, such that the top of the selected confidence interval is no higher than 69 (British Psychological Society, 2015). Individuals who function above this range but below intellectual average (IQ = 71 - 84), have previously been classified as having 'borderline intellectual functioning'. Whilst not exhibiting a significant impairment, such individuals can still experience adaptive and learning challenges (Venkatesan, 2017, Wieland & Zitman, 2016). In 2018 His Majesty's Prison and Probation Service (HMPPS) adopted the term 'Learning Disability and Challenges' (LDC) to span both impaired and borderline ranges, related broadly to an IQ between 60 and 80 (Wakeling & Ramsay, 2020). A learning screening tool (LST) was also introduced to be used in conjunction with an adaptive functioning assessment and clinical interview to provide a triangulated approach to identifying individuals with LDC for a suite of accredited offending behaviour programmes designed for their specific capacities and learning needs (Wakeling & Ramsay, 2020).

The LDC suite is comprised of four HMPPS accredited[2] offending behaviour programmes: Becoming New Me Plus (BNM+), New Me Strengths (NMS), Living as New Me (LNM), and the Healthy Sex Programme (HSP) (Ramsay, 2020). The focus of this interim outcome evaluation concerns SWM scores from BNM+ and NMS. All four programmes are based on a biopsychosocial model of change (Mann and Carter, 2012) that integrates existing models of rehabilitation, including Risk, Need and Responsivity (RNR) principles (Andrews, Bonta and Hoge, 1990), the Good Lives Model (GLM) (Ward, 2012) and desistance theory to provide a cognitive-behavioural approach that is strength-based, and skills-focused (see Walton, Ramsay, Cunningham & Henfrey, 2017; Ramsay, 2020). Collectively they aim to help people develop goals and skills that facilitate change in four criminogenic need domains, namely, 'Self-Management', 'Offence-Supportive Attitudes', 'Relationships', and where relevant, 'Sexual Interests'. A fifth domain that includes desistance factors called a 'Sense of Purpose' is also targeted.

Becoming New Me Plus (BNM+)

BNM+ is designed for individuals[3] with LDC that have been assessed as high or very high risk of reoffending on either the OASys Sexual Predictor (OSP)[4], the OASys Violence Predictor (OVP) or the OASys Electronic Spousal Assault Risk Assessment (ESARA), and who present with at least one or more strong criminogenic needs in the Self-Management, Offence-Supportive Attitudes and Relationships domains. Those who access BNM+ have been convicted of a general violence (GV), intimate partner violence (IPV) or a sexual offence (SO). The programme is operationalised into three strands based on these three offence-types, and allocation is determined according to the participant's offending history and assessment of their risk, needs, strengths, and responsivity factors using a Programmes Needs Assessment (Ramsay et al., 2019). Participants may have offence

[2] The Correctional Services Advice and Accreditation Panel (CSAAP), a group of independent experts, assesses programmes based on principles of international 'what works' evidence and use criteria, which includes evaluation, to make accreditation recommendations to HMPPS. Programme accreditation is required for programmes to be delivered across prisons and probation.

[3] All programmes in the LDC suite are access to adult males and transgender men and transgender women.

[4] Prior to 2021, the actuarial risk screen used with the Risk Matrix 2000. It is possible a small number of people in the sample who attended the BNM+ and NMS sexual offending strands accessed those programmes based on their Risk Matrix 2000 score.

histories that would qualify them for multiple strands. Strand allocation is at the discretion of the Treatment Manager.

There are three delivery formats for BNM+: group, individual, and alternative delivery format (ADF). Group format comprises of eight participants attending two to four two-hour sessions per week culminating in about 178-hours of intervention. Individual delivery includes two or three weekly one-to-one sessions, usually lasting 60-90 minutes each. ADF consists of small groups of two to three participants and was implemented during the COVID-19 pandemic. Allocation to individual or group format is at the discretion of the Treatment Manager. Regardless of delivery format, participants complete core and optional content based on their assessed strengths and needs.

New Me Strengths (NMS)

NMS is designed for individuals with any offence convictions, who have been assessed as medium or above risk of reoffending, with sufficient levels of criminogenic need addressed by the programme. There are multiple offence-based strands: GV, IPV, SO, acquisitive (inclusive of those with convictions for theft and robbery offences) and mixed cohort. The mixed cohort includes a mixture of conviction histories, although participants with sexual offences against children take part in NMS in single cohorts. Participants may have offence histories that would qualify them for multiple strands. Strand allocation is at the discretion of the Treatment Manager.

As with BNM+, there are three delivery formats: group, individual and ADF. Delivered as a group, NMS comprises of eight participants per strand each attending two to four two-hour sessions per week, totalling about 76-hours of intervention. Individual delivery includes two or three weekly one-to-one sessions, usually lasting 60-90 minutes each. ADF consists of small group delivery to two to three participants, the same as BNM+.

Healthy Sex Programme (HSP)

HSP is designed for people convicted of sexual offences who present with strong offence-related sexual interests. HSP is a secondary programme, designed to be delivered after participation in another sexual offending programme. Our analyses do not explore the effect of the HSP (see Freel & Wakeling, 2023; Elliott & Martin, 2023 for recent studies),

both because it is a secondary programme and due to the difference in monitoring participant progress.

Living as New Me

LNM is a skills maintenance, or "booster", programme for people that have completed BNM+, NMS or HSP. Our analyses do not explore the effect of the LNM, because it is a secondary programme.

A tool that supports the strengths-based approach of the LDC suite is the 'Success Wheel' designed by HMPPS. The Success Wheel conceptualises the four criminogenic need domains targeted by the programme as opportunities for learning skills to lead a constructive, crime-free life. For example, conversations about 'offence-supportive attitudes' can be reframed as an opportunity to develop 'Healthy Thinking' where new insight and skills can be strengthened to support more balanced, flexible perspectives. Participants are encouraged to set goals and monitor their own progress on the Success Wheel. The five Success Wheel domains are outlined below. The 'Healthy Sex' domain is relevant only to those with a history for sexual offending:

- Managing Life's Problems (MLP: e.g., managing emotions, and impulsivity, positive problem-solving)
- Healthy Thinking (HT: e.g., flexible thinking, respecting the rights of others)
- Positive Relationships (PR: e.g., intimacy, perspective-taking, assertiveness, pro-social relationships)
- Healthy Sex (HS: e.g., managing unhealthy sexual arousal)
- Sense of Purpose (P: e.g., developing protective factors for stability, being an active member of society)

The Success Wheel was designed to be accessible for adults with and without LDC, by using simplified language and icons to help aid understanding.

To create a way of evaluating participant progress for HMPPS accredited programmes the Success Wheel was developed into a measure, called the 'Success Wheel Measure' (see section 3.2 below). The SWM was inspired from a programme evaluation by Marques et al. (2005) to use as a structured clinical judgement of whether a participant derives benefit from a programme. As with the Success Wheel, the SWM uses accessible language and

visual aids, and additional tips are provided in the scoring guidance to support delivery for LDC participants.

2.2 Aims of this evaluation

This evaluation aims to evaluate whether two accredited offending behaviour programmes from the Learning Disabilities and Challenges (LDC) suite, BNM+ and NMS, are meeting their intervention aims, and whether there are any individual or programme delivery factors affecting participant progress. This was investigated by answering three research aims:

a) explore overall pre-to-post programme participation change in scores on a measure of programme progress called the Success Wheel Measure (SWM),
b) investigate the effect of individual (relating to the person, e.g., motivation) and delivery (relating to the programme, e.g., delivery format) variables of interest on pre-to-post programme participation change in SWM scores, and
c) whether any pre-to-post participation differences differ by rating type or Success Wheel domain.

3. Methodology

3.1 Sample

Administrative data was collected routinely for 327 men who participated in Becoming New Me Plus (n = 170, 52%) or New Me Strengths (n = 157, 48%). Data was collected from men that completed BNM+ and NMS between November 2020 and March 2023. This period was chosen as it coincided with the implementation of the Success Wheel Measure at programme sites.

Cases that had missing Success Wheel Measure data were excluded from this analysis, resulting in a sample of 114 Becoming New Me Plus participants (67.1% of BNM+ cases) and 116 New Me Strengths participants (73.9% of NMS cases). A further two participants were removed from BNM+ due to missing values in other predictor variables (n = 112, 65.9%), see Table A1 for details. Table 1 provides demographic information for these analytical samples.

Table 1: Demographic information of the BNM+ and NMS evaluation sample split by programme (n = 228)

Variable	BNM+	NMS
Age		
Mean (SD)	38.2 (13.8)	39.8 (11.8)
Range	19-68	21-72
Ethnicity		
Aggregated Asian ethnicities	5 (4.5%)	4 (3.4%)
Aggregated Black ethnicities	14 (12.5%)	5 (4.3%)
Aggregated Mixed ethnicities	3 (2.7%)	6 (5.2%)
Aggregated White ethnicities	87 (77.7%)	99 (85.3%)
Unknown	3 (2.7%)	2 (1.7%)
Delivery format		
1:1	36 (32.1%)	29 (25.0%)
Group	61 (54.5%)	78 (67.2%)
Alternative delivery format (small group)	15 (13.2%)	9 (7.8%)

Variable	BNM+	NMS
History of general violence offence(s)		
Yes	83 (74.1%)	63 (54.3%)
No	29 (25.9%)	53 (45.7%)
History of intimate partner violence offence(s)		
Yes	58 (51.8%)	36 (31.0%)
No	54 (48.2%)	80 (69.0%)
History of sexual offence(s)		
Yes	81 (72.3%)	52 (44.8%)
No	31 (27.7%)	64 (55.2%)

3.2 Measures

Success Wheel Measure

The Success Wheel Measure (SWM) is a 5-item scale designed by HMPPS and the Ministry of Justice Analysis Directorate to measure progress in the five domains targeted by Becoming New Me Plus and New Me Strengths: (1) Managing Life's Problems, (2) Healthy Thinking, (3) Positive Relationships, (4) Sense of Purpose and (5) Healthy Sex for those with a history of sexual offending or unhealthy sexual interests. Each domain is scored 1-5 on a Likert scale in each of these domains, from 1 (no or little evidence of success in this area) through 3 (moderate achievement in this area) to 5 (very good success in this area). This means, for participants where the Healthy Sex domain is not applicable, overall SWM scores will range between 4-20, whilst for those where the Healthy Sex domain is applicable, overall SWM scores will range between 5-25. To make the scores between these groups comparable, the summed SWM scores were standardised (using z-scores) for the main analyses.

The overall SWM score is of interest to these analyses because BNM+ has been designed to support individuals who are assessed as having at least one or more strong criminogenic needs in the Managing Life's Problems, Healthy Thinking, and Positive Relationships domains. NMS is designed for individuals who present with a more moderate level of criminogenic need sufficient for the programme. This means BNM+ participants present with a broad and diverse criminogenic need profile (also known as

'high-need'), whilst NMS participants present with a more moderate profile. Analysis of individual domains may underestimate clinical change in participants because of these variations in needs, whereas analysis of the overall SWM scores should reflect overall change.

SWM scores are taken twice during the programmes, first in the mid-programme individual session which takes place after the 'Supporting New Me' block when the Success Wheel is introduced in group sessions and again at the end in the post programme individual session that follows 'Being New Me' (the final block in NMS and BNM+). Though this means that the pre-programme scores do not precede the start of either programme, the initial sessions prioritise engagement, building rapport and introducing the programme concepts. Understanding the Success Wheel is important before the scoring of the SWM.

There are three sets of SWM: scores given by the participant, scores given by the facilitator, and a joint score which must be agreed between participant and facilitator. The analyses of overall change and assessing the influence of individual and programme delivery factors of interest will use facilitator-rated scores only, to maintain an acceptable statistical power. Facilitator-rated scores were chosen for this purpose as they should have the best inter-rater reliability due to the enhanced level of guidance in scoring that facilitators are given and greater amount of experience in using the SWM compared to participants.

SWM scores were first used to evaluate participant progress on two non-LDC sexual offending programmes called Horizon and i-Horizon (Elliott and Hambly, 2023a). Researchers found strong pre-to-post improvement on the SWM. Following that evaluation, the aim was to analyse SWM scores for attendees of BNM+ and NMS.

Psychometric analyses of the SWM scores of people convicted of sexual offending, found that the SWM domains appeared to measure one dominant overall dimension, whilst also being distinct from one-another (Elliott and Hambly, 2023b).

3.3 Missing data

The analysis of individual and programme delivery factors that may affect the change in SWM score pre-to-post programme requires complete case analysis. A review of the

individual and programme delivery variables of interest showed that some contained missing data (see Table A1). Of note are the motivation variable and the meets high-risk criteria variable, both individual factors, which were both excluded from these analyses due to high rates of missingness.

3.4 Evaluation design

The main evaluation aims of this project were to (a) explore overall pre-to-post change in SWM score, (b) investigate the effect of individual and programme delivery variables of interest on pre-to-post change in SWM score, and (c) whether any pre-to-post differences differ by rating type or Success Wheel domain.

Overall change in SWM score

A comparison of means, by paired t-test, was chosen to establish the change in summed SWM score pre-to-post programme.

The effect of programme delivery and individual factors on change in SWM score

Programme delivery and individual factors of interest on pre-to-post change in SWM score were explored using a linear regression model. The regression model quantified the impact of each variable on the change in SWM score. The individual and programme delivery factors of interest in this analysis were:

- Pre-programme strengths (i.e., the participant's first SWM score),
- Programme (i.e., BNM+ or NMS),
- Delivery format (i.e., group, 1:1 or ADF small group),
- Whether they have a history of proven general (all) violence offences,
- Whether they have a history of proven intimate partner violence offences, and/or
- Whether they have a history of proven sexual offending.

Note that offence histories are not mutually exclusive. Due to the small categories and overall sample size, no other control variables were introduced, such as age and ethnicity.

A parsimonious model achieves the desired level of explanatory power (i.e., what percentage of variation in scores is explained by the predictor variables, also known as R^2) with the fewest predictor variables. The parsimonious model was found using forward stepwise regression, which adds each predictor variable to the model in order of statistical

significance and assesses whether the addition of the predictor variable has improved the model's prediction ability.

Comparison of change in SWM score by rating type and by domain

To compare pre-to-post programme participation differences across rating type (i.e., facilitator, participant, or joint score) and Success Wheel domain, two methods were used. A repeated measures ANOVA explored whether there was a difference in overall change in SWM score by rating type and by Success Wheel domain. Whether there was a difference in the influence of individual and programme delivery factors by rating type and Success Wheel domain was explored by fitting the change in SWM score for each rating type, and change in SWM score for each domain, to a multivariate model and the coefficients were compared using simultaneous general linear hypothesis.

The data was subject to screening before conducting the analyses to ensure it met the assumptions for each test, e.g., normality was assessed by examining skewness and kurtosis ahead of conducting any t-tests.

Power analyses were conducted for the proposed analyses, and they indicated that an effect size equivalent to less than a 1-point change in SWM score could be detected in the paired t-test, and effect size equivalent to a 4-point change in SWM score could be detected in the full regression model.

All tests were conducted with $\alpha = 0.05$. Results are indicated as statistically significant where $p < 0.05$. Effect sizes have been provided for statistically significant results, in the form of Cohen's d or Cohen's f^2. To aid interpretation of the effect sizes, Cohen's d statistic is typically categorised as a small effect when 0.2-0.49, a medium effect when 0.5-0.79, and a large effect when 0.8 or greater. Cohen's f^2 statistic is considered a small effect when 0.02-0.14, a medium effect when 0.15-0.34, and a large effect when 0.35 or greater (Cohen, 1988).

Where the dependent variable is the sum of standardised SWM score, the exact score value is not reported. This is because the standardisation made their interpretation less meaningful. Instead, attention should be given to magnitude and equivalent SWM point change where appropriate.

3.5 Limitations and interpreting results

In the absence of an impact evaluation on reoffending, due to an insufficient volume of programme completers to facilitate a robust impact evaluation, an interim outcome evaluation has been chosen. The notion that programmes should be established to 'pass their own logic model' was posited by Epstein and Klerman (2012), who have recommended investigating pre-post programme improvement ahead of conducting an impact evaluation. They suggest that researchers examine the contributors to change and compare this with the expectations which informed the programme design. For BNM+ and NMS, the theory of change posits that by the end of the programme, providing delivery and individualised targeting assumptions have been met, participants will have acquired learning in how to change, having strengthened skills and insight. Without establishing that the participants are obtaining the expected skills and insights, we cannot be confident that those skills and insight can change future behaviour and long-term impact, such as reducing reoffending. Equally, positive change in programme targets may not lead to lower rates of reoffending, in which case the programme theory of change should be reconsidered.

This evaluation does not aim to examine if pre-to-post change in SWM scores are associated with the likelihood of reoffending. Any evidence of progression on the SWM against the programme targets identified by the theory of change for BNM+ and NMS should be interpreted as providing indicative support.

The data used in these analyses had already been collected as part of routine implementation of BNM+ and NMS. Using this retrospective data, though convenient, has several limitations.

Firstly, the SWM data used to assess change was collected as part of programme delivery and therefore there is no SWM data for non-participants. This therefore prohibits comparison to a control group and means we cannot attribute observed change to the programme – any recorded change in SWM score can only be considered indicative of real change resulting from the intervention. This is because of three limitations in within-subject designs. The first is temporal change, which is that change may happen over time regardless of any intervention. The second is regression to the mean. This statistical

phenomenon can make natural variation in repeated data look like meaningful change. This occurs when “extreme” values by the edge of the data are measured again and are found closer to the mean. In a repeated measures design such as this before-after evaluation, this means that not all change can be attributable to the intervention. The third is the potential influence of test effects. Improvement in measures can be attributed to practice effects or the test items themselves generating change independent of the intervention.

Secondly, when using retrospective data there is no control over the consistent recording of data at point of collection. As highlighted in section 3.1 on sample size, there are significant proportions of participants that completed BNM+ (n = 56, 32.9%) and NMS (n = 41, 26.1%) within the defined period, but were excluded from the analytical sample due to missing SWM scores. The removal of participants without SWM scores may have created a sample distinct from the wider BNM+ and NMS population and may produce biased estimates in the results. Missing data in other variables also restricted the analyses. Motivation is recorded on the Horizon Motivation Scale (HMS) three times pre-to-post programme to assess a participant’s motivation. Pre-programme motivation is of theoretical interest on the efficacy of the programme but was not possible for these analyses because a significant proportion of the values were missing, see Table A1. Excluding the variable of interest was of greater importance from an analytical robustness perspective to avoid the removal of further cases.

Another limitation of using historic data is the lack of control over what variables were collected. This restricted the evaluation to available variables rather than all theoretically relevant variables. For example, the LDC suite is suitable for people that maintain their innocence. This decision was taken after evidence showed that denial was not a factor associated with increased likelihood of sexual reoffending (Mann, Hanson and Thornton, 2010) and that taking active responsibility for the future had more value for individuals than passive responsibility for the past (Ware and Mann, 2012). The LDC suite is accessible to anyone who acknowledges problems within their lives that they want to address and work on. Whether maintaining innocence affects the efficacy of the programme is of theoretical interest but when programme data was initially collected the levels were restricted to a binary ‘yes/no’. Due to the stigma associated with offending, especially sexual offences,

people often use some minimisation, meaning ‘yes/no’ lacks important nuance. This was changed in later iterations of data collection to ‘Complete denial/Some minimisation/No minimisation’ but could not be explored in these analyses due to its late implementation and subsequent missingness in earlier cases.

Despite these limitations posed by using retrospective administrative data, this data provides the opportunity to perform an evaluation of interim outcomes efficiently.

It should be noted that the period during which data was collected coincides with regime changes due to the COVID-19 pandemic. The pandemic impacted on all parts of prison life, including the delivery of accredited offending behaviour programmes. Programme delivery was paused temporarily at the start of the pandemic. When it recommenced, it did so in line with exceptional delivery procedures designed to enable small groups to meet in line with social distancing rules and mixing restrictions. Delivery also recommenced in a professional climate of safe systems of working including use of masks, increased hand washing, and social distancing. The effect of these changes likely influenced working relationships and participants’ opportunities to practice skills.

Success Wheel Measure scores were selected to measure change from pre-to-post programme participation. The SWM was developed and originally implemented for Horizon, an accredited offending behaviour programme for adult men with a sexual conviction. Elliott and Hambly (2023b) investigated the validity of SWM as a measure of clinical change against the Sex Offender Treatment and Interventions Progress Scale (SOTIPS), an established scale that has been found to predict sexual recidivism (Hanson et al., 2021), and concluded that SWM could be considered reasonably valid for the Horizon cohort. The BNM+ and NMS cohort is distinct from the Horizon population, according to type of previous convictions (for those without a sexual conviction), risk-level and learning needs. Whilst people with different offending histories share many of the same criminogenic needs, and there is reason to cautiously suggest the same applies to individuals with LDC (Walton et al., 2017), the construct validity of the SWM has only been established for those in the BNM+ and NMS cohort, who like Horizon participants, have a history of proven sexual offending (they represent 72.3% of BNM+ and 44.8% of NMS samples respectively, see Table 1). Construct validity has not been established for those

with histories of intimate partner violence or general violence. This needs to be considered when interpreting results.

The SWM is a clinical judgement tool administered by programme facilitators and participants. The SWM is at risk of socially desirable responding. The SWM scale is a Likert scale which makes it easy to determine what the socially desirable responses are, and participants may choose to rate their strengths greater than they are. This may be exaggerated in cases where performance on the programme informs sentence management. The facilitator and joint scores can be seen as a solution to minimise this effect, but facilitators are still at risk of being bias. Though guidance in how to score SWM is given to facilitators, it is natural for a facilitator to want to see the participants they are working with succeed and increase the evidence base for the success of the programme, which could unconsciously bias them to focus on evidence that points to positive change and minimise contradictory evidence.

4. Results

4.1 Comparison of pre-post SWM scores

A paired t-test was conducted to investigate the overall change in SWM score from pre-to-post programme. There was a statistically significant difference in standardised SWM scores from pre-programme and post-programme; (t(227) = -19.31, p <.001, d = 1.28). This can be considered a large effect size and represents a 3.4 - 4.4-point change in total SWM score (a 31.4% - 31.7% increase).

4.2 Linear regression model

Full model

Multiple linear regression was used to investigate whether the programme delivery and individual factors of interest statistically significantly predicted change in SWM score. See Table B1 for planned contrasts of categorical predictors.

The overall regression, which included pre-programme strengths, programme, delivery format and history of SO, GV and IPV as predictor variables, was statistically significant (R^2 = 0.469, F(7, 220) = 27.75, p < .001). This means that 46.9% of the variation in change in SWM score can be accounted for by these predictor variables. The effects of each predictor variables whilst all others are held constant are detailed in the following paragraphs.

Pre-programme strengths was found to significantly predict change in SWM score (*b* = -0.666, p < .001, f^2 = 0.750). This can be considered a large effect. This means that higher pre-programme strengths predicted smaller change in SWM score.

Programme was found to significantly predict change in SWM score (*b* = -0.327, p = .001, f^2 = 0.045) for NMS compared to BNM+. This can be considered a small effect. This means that NMS predicted smaller change in SWM score compared to BNM+.

Delivery format was not found to significantly predict change in SWM score (1:1, *b* = -0.032, p = .759; ADF, *b* = -0.017, p = .913). This means the delivery format did not affect the change in SWM score.

It was found that having a history of general violence offences significantly predicted a change in SWM score though the effect size was very small (*b* = -0.218, p = .044, f^2 = 0.014), whereas having a history of intimate partner violence offences (*b* = 0.050, p = .620), or sexual offences (*b* = -0.011, p = .914) did not. This means that those with a history of general violence had a smaller change in SWM score compared to those without a history of general violence. Having a history of intimate partner violence or sexual offences did not impact change in SWM score.

Parsimonious model

The parsimonious model was found to include pre-programme strengths, the type of programme (BNM+ or NMS) and history of general violence (R^2 = 0.468, F(3, 224) = 65.73, p < .001). This means 46.8% of the variation in change in SWM score can be accounted for by pre-programme strengths, programme, and history of general violence. Details of the stepwise regression can be found in Table B3.

Pre-programme strengths statistically significantly predicted change in SWM score (*b* = -0.667, p < .001, f^2 = 0.769). This effect size can be considered large. This means that higher pre-programme strengths predicted smaller change in SWM score. Type of programme was found to significantly predict change in SWM score, with NMS being significantly associated with an increase in change in SWM score (*b* = -0.313, p < .001, f^2 = 0.049) compared to BNM+. This effect size can be considered small. This means that NMS predicted smaller change in SWM score compared to BNM+. A history of general violence also significantly predicted change in SWM score (*b* = -0.231, p = .015, f^2 = 0.022). This effect size can be considered small. This means that those with a history of general violence had a smaller change in SWM score compared to those without a history of general violence.

4.3 Comparison of rating types

Overall change

Paired t-tests found statistically significant positive pre-to-post change in SWM score for each rating type (see Table B6). A repeated measures ANOVA was performed to compare the effect of rating type on change in SWM score. There was a statistically significant difference in change in rating type between at least two groups ($F(1.65, 374.51) = 4.408$, $p = 0.018$, using Greenhouse-Geisser sphericity correction).

Pairwise comparisons using paired t-tests with Hommel's correction showed a significant but very small difference between facilitator and joint ratings ($t(227) = 2.48$, $p = .032$, $d = 0.155$), and a statistically significant but very small difference between facilitator and participant ratings ($t(227) = 2.32$, $p = .043$, $d = 0.187$). This effect size can be considered very small. This means that the change in SWM score was greater when rated by facilitators compared to participant- or joint-ratings. There was no significant difference between participant and joint ratings. This means that change in SWM score was similar between participant- and joint-rated SWM scores. See Table B5 for summary statistics and Table B7 for pairwise comparison t-test results.

Regression performance

To determine whether there was a difference in the influence of individual and programme delivery factors by rating type, the change in SWM score for each rating type was fitted to a multivariate model and the coefficients were compared using simultaneous general linear hypothesis. This means that we can compare the influence of each predictor variable on the change in SWM score by rating type, e.g., whether pre-programme strengths affect facilitator- and participant-rated scores equally.

There were no statistically significant differences in the pre-programme strengths coefficient, the programme coefficient, or the history of general violence coefficient between rating type. This means that pre-programme strengths, type of programme and history of general violence had an equivalent impact on change in SWM score across rating types. See Table B11, Table B12, and Table B13 for details of the comparisons.

4.4 Comparison of domains

The comparison of domains must consider that only part of the analytical sample has scores in all five domains (n = 166, 72.8%). The following methods were run to compare the common four domains across the whole sample, and then compare the Healthy Sex domain to the others in the subsample where this domain is scored.

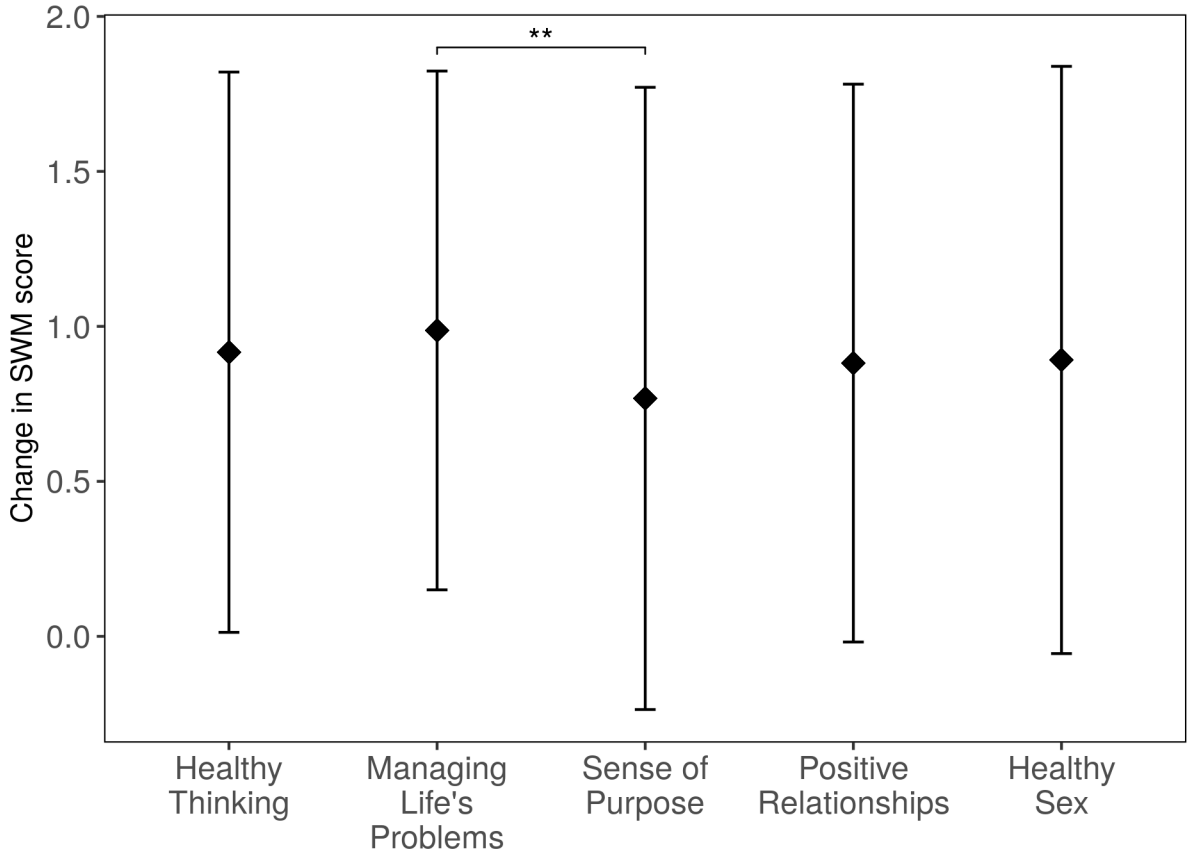
Overall change

Paired t-tests found statistically significant positive pre-to-post change in SWM score for each domain (see Table B9). A repeated measures ANOVA was performed to compare the effect of domain on change in SWM score for the four common domains across the whole sample. There was a statistically significant difference in change in domain score between at least two groups ($F(2.86, 648.09) = 4.025$, $p = .008$, using Greenhouse-Geisser sphericity correction).

Pairwise comparisons between the four core domains across the whole sample, using paired t-tests with Hommel's correction, showed a small statistically significant difference between Managing Life's Problems and Sense of Purpose ($t(227) = 3.24$, $p = .008$, $d = 0.237$). This means that the change in SWM score was greater for the Managing Life's Problem's domain compared to the Sense of Purpose domain. All other comparisons were non-significant, which means change in SWM score was similar between the other domains.

Pairwise comparisons between the five domains across the subsample with a Healthy Sex score, using paired t-test with Hommel's correction, also showed a small statistically significant difference between Managing Life's Problems and Sense of Purpose ($t(165) = 2.95$, $p = .037$, $d = 0.259$). This means the change in SWM score was greater for the Managing Life's Problems domain compared to the Sense of Purpose domain for this subset of the sample. All other comparisons did not show any significant differences. This means change in SWM score was similar between the Healthy Sex domain and the other domains for this subset of the sample. See Table B10 for details of these comparisons.

Figure 1: Pairwise comparison of change in SWM score by domain[5] [6]



Regression performance

To determine whether there was a difference in the influence of individual and programme delivery factors by Success Wheel domain, the change in SWM score for each domain was fitted to a multivariate model and the coefficients were compared using simultaneous general linear hypothesis. This means that we can compare the influence of each predictor variable on the change in domain score, e.g., whether pre-programme strengths have an equivalent impact on change in score across domains.

There were no statistically significant differences in the pre-programme strengths coefficient between domains. See Table B14 for details. There was a significant difference found in the programme coefficient between the Managing Life's Problems domain and the Sense of Purpose domain (b = -0.304, p = .011, d = 0.207) and the Healthy Thinking

[5] Note this figure combines the mean and standard deviation and pairwise comparisons for the four common domains across the whole sample and then additionally the Healthy Sex domain for the subsample. The combination into one figure is for illustrative purposes.

[6] Significant results of the pairwise t-tests are added and p < 0.01 is shown as **.

domain and the Sense of Purpose domain (*b* = -0.228, p = .018, d = 0.193). This means that BNM+ predicted a greater change in SWM score in the Managing Life's Problems and Healthy Thinking domain compared to the Sense of Purpose domain. There were no significant differences in the type of programme coefficient for the remaining domains. This means that BNM+ and NMS had an equivalent effect on the change in score for the other domains. See Table B15 for further detail.

There were no significant differences in the history of general violence coefficient between domains. This means a history of general violence had an equivalent impact on change in SWM score across domains. See Table B16 for further detail.

5. Conclusion

This interim outcome evaluation sought to give indicative evidence to whether Becoming New Me Plus (BNM+) and New Me Strengths (NMS) from the LDC suite of accredited offending behaviour programmes are meeting their programme aims, through exploring pre-to-post programme participation change as measured by the SWM. It also sought to explore whether any programme delivery or individual factors impacted the magnitude of pre-to-post change, and whether this pre-to-post change differed by SWM rating type or domain.

Pre-to-post programme change in SWM scores were found for BNM+ and NMS participants, with large effect sizes: on average the post-programme scores were greater than pre-programme scores by 3.4 - 4.4-points (a 31.4% - 31.7% increase).

Pre-programme strengths was the most influential factor on change in SWM score, with increases in pre-programme strengths predicting reductions in the magnitude of SWM change. Participants with very strong initial pre-programme strengths showed a negative change in total SWM. This is expected, as higher pre-programme SWM scores leave less opportunity to demonstrate improvement.

Whether someone had participated in BNM+ or NMS had a small but statistically significant effect on change in SWM score. When all other variables were held constant, those participating in BNM+ had a higher rate of improvement compared to those that participated in NMS. This may be explained based on risk and need logic, since attendees of BNM+ typically exhibit a more pronounced risk and need profile than their counterparts who access NMS. BNM+ is a lengthier programme because it attends to the higher pre-existing levels of risk and need of its attendees and may provide more opportunities for skills and insight development. This added opportunity may contribute to the larger magnitude of change captured on the SWM.

It was found that those with a history of general violence offending had a statistically significant lower rate of improvement compared to those without a history of general violence offending. The effect size however was very small. Neither a history of intimate

partner violence nor a history of sexual offending significantly affected the change in SWM score pre-to-post programme. This finding indicated that the LDC suite did not have a meaningful differential impact on change dependent on participant offence type histories pre-post programme.

Delivery format was not found to impact change in SWM score. This indicated that participants that engaged with the LDC suite via alternative delivery format were not at a disadvantage from this newly developed delivery method of small groups.

The magnitude of change was significantly higher for facilitator-rated SWM score compared to participant- and joint-rated, however the effect size was very small. When comparing coefficients between rating types, no significant differences were found.

We explored whether there were any differences in change in SWM score by SWM domain and found a small significant difference between the Managing Life's Problems and Sense of Purpose domains, in both the whole sample and the subpopulation with Healthy Sex scores (i.e., those with sexual offence convictions). When comparing coefficients for the parsimonious model, there was a significant difference in the programme coefficient between the Managing Life's Problems domain and Sense of Purpose domain, and the Healthy Thinking domain and the Sense of Purpose domain, with BNM+ participants having a greater pre-to-post change in the Managing Life's Problems and Healthy Thinking domains compared to the Sense of Purpose domain. There were no significant differences in the pre-programme strengths coefficient or history of general violence coefficient between domains.

It should be noted that all results presented here are only indicative that participation in the LDC suite is associated with positive change: causality cannot be inferred from the programme on the observed change due to the lack of control group (i.e., that provides a counterfactual scenario). Without a control group it is not possible to know for sure if the change in SWM scores found in these analyses were a product of natural individual change over time or regression to the mean rather than participation in BNM+ or NMS. Test effects may have also contributed, as both facilitators and participants may be biased to overstate positive change.

It is important to emphasise that this evaluation investigated clinical outcomes, i.e., participants building strengths, developing insight and gaining skills, but this does not necessarily translate to reductions in future reoffending. Evidence that improvements in psychometric scores predict decreases in reoffending is mixed. Pre-programme scores and changes on only a few measures associated with the Relationship domain have so far been associated with reoffending (e.g., Barnett, Wakeling, Mandeville-Norden & Rakestrow, 2012; Wakeling, Freemantle, Beech & Elliott, 2011). The SWM has not been validated for predicting reoffending.

Additionally, this evaluation used the SWM to measure progress towards intervention targets in participants. The SWM has been validated as measuring clinical change using a well-established scale for the Horizon cohort, which has some differences to the BNM+ and NMS cohorts in terms of offence type, risk level and learning needs. People with different offending histories share many of the same criminogenic needs, and there is reason to cautiously suggest this is also the case for those with LDC (Walton et al., 2017). However, the SWM has not been validated for those with proven histories of intimate partner violence and general violence like it has been for those with histories of sexual offending. Further research that aims to validate the SWM for these other cohorts, as well as those with LDC, would help broaden its utility as a measure of clinical change.

Overall, in line the programme logic for BNM+ and NMS, the pre-to-post change in SWM scores provides indicative evidence that participants made progress in learning how to change.

References

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. Criminal Justice and Behavior, 17, 19–52

APA (American Psychiatric Association). Diagnostic and statistical manual of mental disorders. fifth ed. Washington, DC: APA; 2013.

Barnett, G. D., Wakeling, H. C., Mandeville-Norden, R., & Rakestrow, J. (2012). How useful are psychometric scores in predicting recidivism for treated sex offenders? International Journal of Offender Therapy and Comparative Criminology, 56, 420–446.

British Psychological Society (2015) Guidance on the Assessment and Diagnosis of Intellectual Disabilities in Adulthood: A document compiled by a Working Group of the British Psychological Society's Division of Clinical Psychology, Faculty for People with Intellectual Disabilities. https://www.bps.org.uk/guideline/guidance-assessment-and-diagnosis-intellectual-disabilities-adulthood

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

Elliott, I., & Hambly O. (2023a) Horizon and iHorizon: An uncontrolled before-after study of clinical outcomes. (Ministry of Justice Analytical Series unnumbered). London: Ministry of Justice. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1137650/horizon-iHorizon-uncontrolled-study.pdf

Elliott, I., & Hambly O. (2023b) Horizon and iHorizon Psychometric analyses of the Success Wheel Measure. (Ministry of Justice Analytical Series unnumbered). London: Ministry of Justice. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1137660/horizon-iorizon-psychometric-analyses.pdf

Elliott, I., & Martin, E. (2023) Post-release reoffending outcomes for individuals with offence-related sexual paraphilias: An exploratory risk-band analysis. (Ministry of Justice Analytical Series unnumbered). London: Ministry of Justice. https://assets.publishing.service.gov.uk/media/63f4954de90e077bb6c6d19a/sexual-paraphilias.pdf

Epstein, D., & Klerman, J. A. (2012). When is a program ready for rigorous impact evaluation? The role of a falsifiable logic model. Evaluation Review, 36(5), 375–401.

Freel, A., & Wakeling, H. (2023) The Healthy Sex Programme: An exploration of pre-to-post psychological test change. (Ministry of Justice Analytical Series unnumbered). London: His Majesty's Prison and Probation Service. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1137873/healthy-sex-programme.pdf

Hanson, R. K., Newstrom, N., Brouillette-Alarie, S., Thornton, D., Robinson, B. B. E., & Miner, M. H. (2021). Does reassessment improve prediction? A prospective study of the Sexual Offender Treatment Intervention and Progress Scale (SOTIPS). International journal of offender therapy and comparative criminology, 65(16), 1775–1803.

Mann, R.E., Hanson, K.R. & Thornton, D. (2010). Assessing risk for sexual recidivism: Some proposals on the nature of psychologically meaningful risk factors. Sexual Abuse: A Journal of Research and Treatment, 22 (2), 191–217.

Mann, R. E., & Carter, A. J. (2012). Organising principles for the treatment of sexual offending. In B. Wischka, W. Pecher, & H. van der Boogaart (Eds.), Behandlung von straftätern: Sozialtherapie, maßregelvollzug, sicherungsverwahrung [Offender treatment: Social therapy, special forensic hospitals, and indeterminate imprisonment]. Freiburg, Germany: Centaurus.

Marques, J. K., Wiederanders, M., Day, D. M., Nelson, C., & Van Ommeren, A. (2005). Effects of a relapse prevention program on sexual recidivism: Final results from California's Sex Offender Treatment and Evaluation Project (SOTEP). Sexual Abuse: A Journal of Research and Treatment, 17(1), 79–107.

Ramsay, L. (2020). Strengths-Based Programmes for Men with Sexual Convictions, Who Have Learning Disability and Learning Challenges. In: Hocken, K., Lievesley, R., Winder, B., Swaby, H., Blagden, N., Banyard, P. (eds) Sexual Crime and Intellectual Functioning. Sexual Crime. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-52328-2_3

Ramsay, L., Walton, J.S., Frost, G., Rewaj, C., Westley, G., Tucker, H., Millington, S., Dhar, A., Martin, G. and Gill, C. (2019), "Evaluation of offending behaviour programme selection: the PNA", The Journal of Forensic Practice, Vol. 21 No. 4, pp.264–277. https://doi.org/10.1108/JFP-04-2019-0015

Venkatesan, S. (2017). Demographic, cognitive and psycho-social profile of adults with borderline intellectual functioning. Journal of Contemporary Psychological Research. 4(1), 1–12.

Wakeling, H. C., Freemantle, N., Beech, A. R., & Elliott, I. (2011). Identifying predictors of recidivism in a large sample of UK sexual offenders: A prognostic model. Psychological Services, 8, 307–318.

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.

Walton, J. S., Ramsay, L., Cunningham, C., & Henfrey, S. (2017). New directions: Integrating a biopsychosocial approach in the design and delivery of programs for high risk services users in Her Majesty's Prison and Probation Service. Advancing Corrections: Journal of the International Corrections and Prison Association, 3(1), 21-47.

Ward, T. (2012). The good lives model of offender rehabilitation: Basic assumptions, aetiological commitments, and practice implications. In Offender Supervision (pp. 41–64). Willan.

Ware, J. & Mann, R. E. (2012). How should "acceptance of responsibility" be addressed in sexual offending treatment programs? Aggression and Violent Behavior, 17, 279–288.

Wechsler (2010) Wechsler Adult Intelligence Scale - Fourth UK edition (WAIS-IV UK) (WAIS-IV).

Wieland, J & Zitman, F.G. (2016) It is time to bring borderline intellectual functioning back into the main fold of classification systems. British Journal of Psychology Bulletin, 40(4), 204–206.

Glossary of terms

Accredited offending-behaviour programme: Offending-behaviour programmes, such as BNM+, can be awarded accreditation by the Correctional Services Advice and Accreditation Panel.

Construct validity: How well a scale measures the concept it was designed to measure.

Control variable: A control variable is held constant throughout an analysis to assess the relationship between the dependent variable and the predictor variable(s).

Dependent variable: Also known as the outcome variable, this is the variable measured. In these analyses the dependent variable is SWM score or change in SWM score.

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

HMPPS: His Majesty's Prison and Probation Service.

Linear regression model: Linear regression is a statistical model which estimates the linear relationship between a dependent variable and one or more predictor variables.

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.

Median: This is a measure of central tendency calculated by finding the $\frac{1}{2}\,n^{th}$ value of a dataset containing n values. This is also known as the 50^{th} percentile, such that there is an equal probability of a value falling above or below it.

No significant difference: This means that 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.

p-adjustment: p-adjustment is used in studies with multiple outcomes to account for the multiple comparisons issue, to combat the increased risk of finding a false positive result.

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).

Parsimonious model: The most statistically complex model with the fewest predictor variables.

Predictive validity: How well a scale can predict a concrete outcome, such as recidivism.

Predictor variable: Also known as an independent variable, this is the variable that is being investigated as to whether it influences the dependent variable. In these analyses, pre-programme strengths, delivery format, programme, history of GV, history of IPV and history of SO are predictor variables.

Regression coefficient: A parameter estimate which describes the relationship between a predictor variable and the dependent variable.

Repeated measures ANOVA: A repeated measures ANOVA compares the means across one or more variables that are based on repeated observations.

Sphericity: This refers to the condition where the variances of the differences between all combinations of within-subject conditions are considered equal. This is a key assumption in conducting a repeated measures ANOVA.

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.

Standardisation: The process of transforming data into a consistent and uniform format that can be easily compared. See Appendix C – Z-scores for the method used to standardise SWM scores in these analyses.

Stepwise regression: This is a method of calculating the parsimonious model by adding predictor variables one-by-one. These analyses used forward stepwise regression.

Appendix A
Missing data in predictor variables

Of cases that had complete SWM scores, the following missingness within predictor variables was found alongside the decision taken on how to process that missingness and the rationale for why.

Table A1: Missingness within predictor variables for cases with complete SWMs

Variable	Number and proportion of values missing	Decision taken	Rationale
Delivery format	0	-	-
Motivation	35 (15.2%)	Remove variable from analysis	Removal of cases would bias sample, as data was found to be missing at random (MAR). Proportion of missing data is out of the acceptable range to impute.
Meets high-risk criteria	48 (20.9%)	Remove variable from analysis	Removal of cases would bias sample, as data was found to be missing at random (MAR). Proportion of missing data is out of the acceptable range to impute.
History of general violence	2 (0.9%)	Remove cases	Binary variable difficult to impute with the sample size. Proportion of missing data very low and unlikely to bias results.
History of intimate partner violence	2 (0.9%)	Remove cases	Binary variable difficult to impute with the sample size. Proportion of missing data very low and unlikely to bias results.
History of sexual offending	2 (0.9%)	Remove cases	Binary variable difficult to impute with the sample size. Proportion of missing data very low and unlikely to bias results.

Appendix B
Results and Statistics

Reference categories

In the linear regression models, not every comparison between the levels of each category is tested. For categorical variables, one category is used as the reference category which the other categories are compared to, which is known as planned contrasts. This is important in interpreting the results of the linear regression model. The reference category for each categorical predictor is in Table A1 below.

Table B1: Reference categories for planned contrasts

Variable	Reference category	Planned contrasts
Programme	BNM+	1. NMS vs. BNM+
Delivery format	Group	1. 1:1 vs. Group 2. ADF vs. Group
History of general violence	No	1. Yes vs. No
History of intimate partner violence	No	1. Yes vs. No
History of sexual offending	No	1. Yes vs. No

Regression models

Table B2: Regression table for the full model

Variable	Estimate	SE	95% CI (LL and UL)		p
(Intercept)	1.123	0.147	0.824	1.141	<.001
Pre-programme strengths	-0.666	0.052	-0.767	-0.564	<.001
Programme	-0.327	0.099	-0.522	-0.132	.001
Delivery format (1:1)	-0.032	0.104	-0.236	0.172	.759
Delivery format (ADF)	-0.017	0.156	-0.325	0.291	.913
History of general violence (Y)	-0.218	0.107	-0.429	-0.006	.044
History of intimate partner violence (Y)	-0.050	0.102	-0.251	0.150	.620
History of sexual offending (Y)	-0.011	0.103	-0.215	0.192	.914

Table B3 presents the outputs of forward stepwise regression in developing the parsimonious model. The null model considers only the intercept, and each row adds the stated variable into the model.

Table B3: Stepwise regression in determining the parsimonious model

Model	df	RSS	AIC	BIC	F	p
Null		188.38	607.5	614.4		
Pre-programme strengths	1	106.90	480.3	490.6	179.19	<.001
Programme	2	102.85	473.5	487.2	8.91	.003
History of general violence (Y)	3	100.18	469.5	486.7	5.86	.016
History of intimate partner violence (Y)	4	100.09	471.3	491.9	0.21	.645
Delivery format	6	100.04	475.2	502.7	0.05	.951
History of sexual offending (Y)	7	100.03	477.2	508.1	0.01	.914

The most complex statistically significant model is change in SWM score predicted by pre-programme strengths, programme and history of general violence convictions.

Comparisons

Table B4 presents the mean and standard deviation of facilitator-rated summed standardised SWM scores before and after the intervention.

Table B4: Summary statistics of pre- and post-programme SWM scores

Time	Mean	SD
Before	-0.581	0.880
After	0.584	0.742

Table B5 shows the mean and standard deviation of change in summed standardised SWM scores by rating type.

Table B5: Summary statistics of change in SWM score by rating type

Variable	Mean	SD
Facilitator	1.16	0.911
Participant	0.994	0.924
Joint	1.03	0.881

Table B6 presents the outputs of pre-to-post change in total SWM score by paired t-tests for each rating type.

Table B6: Pre-to-post change in total SWM score by rating type

Variable	df	T	p	Cohen's d
Facilitator	227	-19.31	<.001	1.28
Participant	227	-16.235	<.001	1.14
Joint	227	-17.581	<.001	1.19

Table B7 presents the outputs of pairwise comparison of change in summed standardised SWM scores by rating type, with effect sizes described in Cohen's d. Comparisons were conducted using a paired t-test. P-adjustment was executed using Hommel's correction.

Table B7: Pairwise comparison of change in SWM score by rating type

Comparison	df	T	p	p.adj	Cohen's d
Facilitator - Participant	227	2.32	.021	.043	0.152
Facilitator - Joint	227	2.48	.014	.032	0.185
Participant - Joint	227	0.63	.531	.531	0.038

Table B8 combines the mean, standard deviation and count of facilitator-rated SWM scores for the four common domains for the whole sample, with the summary statistics for the Healthy Sex domain for the subpopulation with scores in this domain.

Table B8: Summary statistics of change in SWM score by domain

Variable	Mean	SD	n
Managing Life's Problems	0.987	0.837	228
Healthy Thinking	0.917	0.904	228
Positive Relationships	0.882	0.900	228
Sense of Purpose	0.768	1.000	228
Healthy Sex	0.892	0.947	166

Table B9 presents the outputs of pre-to-post change in SWM score by paired t-tests for each domain.

Table B9: Pre-to-post change in SWM score by domain

Variable	df	T	p	Cohen's d
Managing Life's Problems	227	-17.807	<.001	1.18
Healthy Thinking	227	-15.315	<.001	1.01
Positive Relationships	227	-14.793	<.001	0.98
Sense of Purpose	227	-11.574	<.001	0.77
Healthy Sex	165	-12.127	<.001	0.94

Table B10 presents the output of the pairwise comparisons between the four common domains across the whole sample and also the pairwise comparisons between the Healthy Sex domain and the other domains within the subsample that have Healthy Sex domains. Pairwise comparisons were conducted using paired t-tests with p-adjustment by Hommel's correction.

Table B10: Pairwise comparison of change in SWM score by domain

Comparison	df	T	p	p.adj	Cohen's d
MLP-HT	227	1.16	.249	.498	0.081
MLP-PR	227	1.80	.074	.221	0.121
MLP-P	227	3.24	.001	.008	0.120
HT-PR	227	0.60	.552	.552	0.039
HT-P	227	2.09	.038	.165	0.156
PR-P	227	1.66	.099	.296	0.120
HS-MLP	165	-1.97	.051	.357	-0.178
HS-HT	165	-1.44	.153	.592	-0.116
HS-PR	165	-1.22	.224	.666	-0.105
HS-P	165	0.816	.416	.832	0.076

Table B11 presents the output from the simultaneous tests for general linear hypotheses when comparing the pre-programme strengths coefficient between rating types for the parsimonious model. P-adjustment has been executed using Hommel's correction.

Table B11: Comparison of pre-programme strengths coefficient by rating type

Comparison	Estimate	SE	Z	p.adj	Cohen's d
Facilitator-Participant	-0.066	0.053	-1.255	.441	0.075
Facilitator-Joint	-0.047	0.046	-1.049	.588	0.068
Participant-Joint	0.017	0.037	0.465	.642	0.030

Table B12 presents the output from the simultaneous tests for general linear hypotheses when comparing the programme coefficient between rating types. P-adjustment has been executed using Hommel's correction.

Table B12: Comparison of programme coefficient by rating type

Comparison	Estimate	SE	Z	p.adj	Cohen's d
Facilitator-Participant	-0.223	0.097	-2.311	.063	0.152
Facilitator-Joint	-0.091	0.083	-1.090	.276	0.073
Participant-Joint	0.132	0.068	1.937	.106	0.129

Table B13 presents the output from the simultaneous tests for general linear hypotheses when comparing the history of general violence coefficient between rating types. P-adjustment has been executed using Hommel's correction.

Table B13: Comparison of history of general violence coefficient by rating type

Comparison	Estimate	SE	Z	p.adj	Cohen's d
Facilitator-Participant	-0.069	0.101	-0.689	.820	0.045
Facilitator-Joint	0.020	0.087	0.228	.820	0.013
Participant-Joint	0.089	0.071	1.252	.631	0.083

Table B14 presents the output from the simultaneous tests for general linear hypotheses when comparing the pre-programme strengths coefficient between the four common domains across the whole sample, and the comparisons of the Healthy Sex domain against the other domains for the subsample with HS scores. P-adjustment has been executed using Hommel's correction.

Table B14: Comparison of pre-programme strengths coefficient by domain

Comparison	n	Estimate	SE	Z	p.adj	Cohen's d
MLP-HT	228	-0.088	0.052	-1.709	.525	0.112
MLP-P	228	-0.026	0.057	-0.453	.900	0.030
MLP-PR	228	-0.033	0.052	-0.631	.900	0.042
HT-P	228	0.062	0.058	1.079	.867	0.071
HT-PR	228	0.055	0.054	1.023	.900	0.067
P-PR	228	-0.007	0.057	-0.125	.900	0.001
MLP-HS	166	-0.126	0.064	-1.982	.146	0.153
HT-HS	166	-0.024	0.064	-0.386	.985	0.029
P-HS	166	-0.084	0.070	-1.206	.552	0.093
PR-HS	166	-0.136	0.063	-2.161	.098	0.168

Table B15 presents the output from the simultaneous tests for general linear hypotheses when comparing the programme coefficient between the four common domains across the whole sample, and the comparisons of the Healthy Sex domain against the other domains for the subsample with HS scores. P-adjustment has been executed using Hommel's correction.

Table B15: Comparison of programme coefficient by domain

Comparison	n	Estimate	SE	Z	p.adj	Cohen's d
MLP-HT	228	-0.015	0.080	-0.192	.848	0.012
MLP-P	228	-0.303	0.097	-3.130	.011	0.207
MLP-PR	228	-0.132	0.084	-1.572	.282	0.104
HT-P	228	-0.288	0.099	-2.918	.018	0.193
HT-PR	228	-0.116	0.088	-1.317	.376	0.087
P-PR	228	0.172	0.100	1.724	.254	0.114
MLP-HS	166	-0.004	0.119	-0.031	.975	0.003
HT-HS	166	0.038	0.112	0.342	.975	0.026
P-HS	166	0.205	0.136	1.504	.531	0.117
PR-HS	166	0.056	0.116	0.478	.975	0.037

Table B16 presents the output from the simultaneous tests for general linear hypotheses when comparing the history of general violence coefficient between the four common

domains across the whole sample, and the comparisons of the Healthy Sex domain against the other domains for the subsample with HS scores. P-adjustment has been executed using Hommel's correction.

Table B16: Comparison of general violence coefficient by domain

Comparison	n	Estimate	SE	Z	p.adj	Cohen's d
MLP-HT	228	-0.082	0.083	-0.988	.806	0.065
MLP-P	228	-0.126	0.101	-1.255	.806	0.083
MLP-PR	228	-0.152	0.087	-1.755	.476	0.116
HT-P	228	-0.045	0.103	-0.434	.806	0.029
HT-PR	228	-0.070	0.092	-0.762	.806	0.050
P-PR	228	-0.025	0.104	-0.245	.806	0.016
MLP-HS	166	-0.213	0.118	-1.808	.282	0.140
HT-HS	166	-0.091	0.112	-0.818	.639	0.063
P-HS	166	-0.109	0.136	-0.810	.639	0.062
PR-HS	166	-0.054	0.116	-0.469	.639	0.036

Appendix C
Formulae

Z-scores

Assessment of the SWM found it measured one underlying dominant dimension, which means it is suitable to use summed SWM scores to assess overall change. However, participants are only scored for domains in which they have needs and only a subset of the sample are scored on the Healthy Sex domain.

To make scores between participants with scores in four domains (range 4-20) and participants with scores in five domains (range 5-25) we converted them to Z-scores. The formula to convert to Z-scores is:

$$Z = \frac{x - \mu}{\sigma}$$

Where Z denotes the standardised score, x represents the observed value, μ is the sample mean and σ is the standard deviation of the sample.

Z-scores were calculated using pre- and post-programme summed SWM scores for the group with only four domain scores and the group with five domain scores. Z-scores can be converted back to raw scores, which has been done for the headline figures to aid interpretation of the results.

Cohen's d

Cohen's d was used to report effect sizes for the comparisons conducted by t-tests and simultaneous general linear hypotheses. The formula for Cohen's d is:

$$d = \frac{M_1 - M_2}{SD_{pooled}}$$

Where M_1 and M_2 denote the mean of group one and two respectively, and SD_{pooled} is the pooled standard deviation of the two groups.

Cohen's f^2

Cohen's f^2 was used to report effect sizes for the coefficients in the regression models. The formula which was used to calculate Cohen's f^2 for the local effect size, where B is the variable of interest is:

$$f^2 = \frac{R^2_{AB} - R^2_A}{1 - R^2_{AB}}$$

R^2_{AB} is the proportion of variance explained for variables A and B together, and R^2_A is the proportion of variance accounted for by A alone. This gives the proportion of variance accounted solely by B.