Ministry of Justice

Kaizen – An Accredited Offending Behaviour Programme:

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

© Crown copyright 2024

This publication is licensed under the terms of the Open Government Licence v3.0 except where otherwise stated. To view this licence, visit nationalarchives.gov.uk/doc/open-government-licence/version/3

Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned.

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 44 0

Acknowledgements

For their subject matter advice and guidance, I would like to thank colleagues from HMPPS Interventions Services. 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

Kaizen is an accredited offending behaviour programme delivered by His Majesty's Prison and Probation Service (HMPPS) for adult men in custody. Kaizen is for people who have been assessed as high or very high risk of reoffending that have been convicted of either general violence, intimate partner violence and/or sexual offence(s). The programme supports participants develop goals and skills to lead a life without reoffending.

The aim of this interim outcome evaluation was to determine whether Kaizen participants were making positive progress against key programme aims, measured by the Success Wheel Measure (SWM). The SWM, designed by HMPPS, assesses participant progress in the following domains: (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 the SWM scores, and whether these changes varied between assessment domains.

Methodological approach

An uncontrolled before-after evaluation was conducted using SWM scores. Routine programme data was collected for 529 men who completed Kaizen between January 2021 and March 2023. After removing incomplete data, the final sample included 289 for this evaluation. SWM scores were summed and standardised to allow direct comparison of those with four target domains to those with five.[1] For example, an individual with four domains present would normally have a maximum SWM sum of 20, whilst an individual with five domains present would have a maximum SWM 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.

[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.

The potential influence of individual and programme delivery factors on improvement against programme targets was investigated using a linear regression model. The factors included were: (1) participant initial SWM score, (2) the programme format (1:1, group, or small group format), (3) motivation levels prior to the programme, (4) risk status (high/other), and whether they had any or a combination of (5) general violence, (6) intimate partner violence, and/or (7) sexual offending history.

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 the Success Wheel domains (i.e., the individual domains that made up the total of the SWM scores, each rated 1-5).

Interpreting results

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

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

Key results

- A large increase in total SWM scores from pre-to-post programme participation was found, indicating positive progress against Kaizen targets.
- Participants with lower initial insight and skills showed greater improvements, with a large effect size.
- The small group format (2-3 participants) led to greater SWM score changes compared to the normal group format (8 participants), though this effect size was small.
- No other factors predicted SWM score changes.

- Positive changes in SWM scores were seen across all rating types, with greater changes noted when rated by facilitators. Pre-programme strengths had a greater impact when rated by participants.
- All SWM domains showed positive score changes, with lower changes in the Sense of Purpose domain compared to Managing Life’s Problems, Healthy Thinking and Positive Relationships domains.
- The influence of individual and programme delivery factors was consistent across domains.

Conclusions

These interim outcome evaluation results suggest that participation in Kaizen is associated with positive changes in key programme targets. Participants with lower levels of insight appear to benefit more. The results also suggest that change was largely equivalent across individuals with different offence histories, reinforcing Kaizen’s broader scope of offending.

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

2. Introduction

2.1 Kaizen

Kaizen is a His Majesty's Prison and Probation Service (HMPPS) accredited[2] offending-behaviour programme designed for adult males[3] who have been convicted of sexual (SO), intimate partner violence (IPV) or general violence (GV) offences. Programme allocation is determined according to the participant's recorded offending history and assessment of their risk, needs, strengths, and responsivity factors using a Programmes Needs Assessment (Ramsay et al., 2019). This means participants should be assessed as high or very high risk of reoffending on either the OASys Sex Predictor (OSP),[4] OASys Violence Predictor (OVP) or the OASys Electronic Spousal Assault Risk Assessment (ESARA), and as presenting with at least one or more strong criminogenic need items in the Self-Management, Offence-Supportive Attitudes and Relationships domains. However, in certain circumstances, Kaizen Treatment Managers can make discretionary decisions to select individuals who fall outside these criteria if a high intensity programme such as Kaizen is judged to be the most appropriate match to their risk and need profile. Kaizen is operationalised in three strands based on the three offence types (i.e., SO, IPV, GV). Participants may have offence histories that would qualify them for multiple strands and strand allocation is at the discretion of the programme Treatment Manager.

Kaizen is 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 and Henfrey, 2017). Kaizen aims to

[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] Kaizen is accessible to adult males and transgender men and transgender women.

[4] Prior to 2021, the actuarial risk screen was used with the Risk Matrix 2000. It is possible a small number of people in the sample who attended the Kaizen sexual offending strand based on their Risk Matrix 2000 score.

help people develop goals and skills that facilitate change in four criminogenic need domains targeted by the programme, namely, ‘Self-Management’, ‘Offence-Supportive Attitudes’, ‘Relationships’, and ‘Sexual Interests’. A fifth domain that includes desistance factors called a ‘Sense of Purpose’ is also targeted.

Kaizen has a flexible delivery model, enabling both group and individual format. Delivered as a group, Kaizen uses a rolling format with up to eight participants per strand each attending two to four two-hour sessions per week for 160-hours. An alternative delivery format (ADF) was introduced during the COVID-19 pandemic to deliver Kaizen to small groups of two to three participants. Regardless of delivery format, participants complete core and optional content based on their assessed strengths and needs.

A tool that supports the strengths-based approach of Kaizen 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)

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). 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.

2.2 Aims of this evaluation

This evaluation aims to find quantitative evidence of whether Kaizen, an accredited offending behaviour programme, is meeting its programme aims and whether there are any individual or delivery format factors affecting participant progress. This was investigated by answering three evaluation 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 operational (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 from 529 Kaizen participants that completed the programme between January 2021 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 289 Kaizen participants (54.6% of Kaizen cases). A further four participants were removed from the sample due to missing values in predictor variables (n = 285, 53.9%), see Table A1 for details. Table 1 provides demographic information for this evaluation sample.

Table 1: Demographic information of the Kaizen evaluation sample (n = 285)

Variable	Kaizen
Age	
Mean (SD)	35.7 (10.3)
Range	19-66
Ethnicity	
Aggregated Asian ethnicities	15 (5.3%)
Aggregated Black ethnicities	34 (11.9%)
Aggregated Mixed ethnicities	19 (6.7%)
Aggregated White ethnicities	209 (73.3%)
Unknown	8 (2.8%)
Delivery format	
1:1	67 (23.5%)
Group	164 (57.5%)
Alternative delivery format (small group)	54 (18.9%)
Motivation	
Mean (SD)	5.7 (1.6)
Range	1-8
Meets high risk of reoffending criteria	

Variable	Kaizen
Yes	255 (89.5%)
No	30 (10.5%)
History of proven general violence offence(s)	
Yes	233 (81.8%)
No	52 (18.2%)
History of proven intimate partner violence offence(s)	
Yes	151 (53.0%)
No	134 (47.0%)
History of proven sexual offence(s)	
Yes	113 (39.6%)
No	172 (60.4%)

3.2 Measures

In the main analyses three measures were utilised: the SWM, the Horizon Motivation Scale (HMS) and Risk Criteria.

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 on the five domains of the Success Wheel targeted by Kaizen: (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 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 Kaizen 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. This means Kaizen participants present with a broad and diverse criminogenic need profile (also known as 'high-need'). Analysis of individual domains may underestimate clinical change in participants because of this variation in needs, whereas analysis of the overall SWM scores should reflect overall change.

SWM scores are taken twice during Kaizen, first in session 5 of the programme's introductory module called 'Getting Going' when the Success Wheel is introduced, and then again in the final module called 'Future New Me'. Though this means that the pre-programme scores do not precede the start of Kaizen, 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 have been chosen for this purpose as they should have the best inter-rater reliability due to the enhanced 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 sexual offending programmes called Horizon and i-Horizon (Elliott and Hambly, 2023a). Researchers found strong pre-to-post programme improvement on the SWM. Following that evaluation, the aim here was to analyse SWM scores for participants of Kaizen.

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).

Horizon Motivation Scale

The Horizon Motivation Scale (HMS) is a 4-item scale developed by HMPPS and the Ministry of Justice Analysis Directorate to measure motivation towards participating in Horizon, and later used to measure motivation in other programmes such as Kaizen. It is informed by self-determination theory (Deci and Ryan, 2008, Ryan and Deci, 2017) which defines motivation as the drive to engage in a course of action and recognises the different orientations of motivation such as intrinsic and extrinsic motivations. This theory states there are four aspects of activation and intention for motivation: energy, direction, persistence and equifinality (early life experiences).

The four items in the HMS are: (1) enthusiasm, a positive attitude, energy, and a drive to direct that energy positively, (2) direction, an internal desire and willingness to participate, (3) persistence, commitment to completing the programme and acceptance that doing so will take resolve and perseverance, and (4) holistic, recognition that participation will contribute to living an offence-free life. Facilitators score participants as either 0 (no evidence), 1 (some evidence) or 2 (strong evidence) on each of these items, giving an overall range of 0-8. The analyses in this report use participants' summed HMS score.

The HMS is taken three times during the programme: early in the programme at the same time as the SWM is scored, after the first module, and post-programme at the same time as the SWM is scored. In these analyses, the HMS score used is that taken at the same as the pre-programme SWM score.

The HMS score quantifies the level of motivation a participant has at the point where pre-programme strengths is taken. This is used to evaluate the impact early participant motivation has on change in SWM score in the regression.

Risk criteria

Kaizen is designed for men that have been assessed as high or very high risk on at least one of the following risk assessments: OASys Sexual Predictor (OSP),[5] OASys Violence Predictor (OVP) or the electronic Spousal Assault Risk Assessment (ESARA). These risk assessments are used to predict the likelihood of reoffending with regards to sexual, general violence, or intimate partner violence offences.

When establishing suitability of Kaizen, an individual's risk level and their assessed needs and strengths are considered together. Those considered suitable should be both 'high-

[5] 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 Kaizen sexual offending strand access the programme based on their Risk Matrix 2000 score.

risk' and 'high-need'. For this evaluation, 'high-risk' criteria were confirmed as having been met or not. Whether participants also met the 'high-need' criteria or were selected based on a discretionary decision of a Treatment Manager was unknown due to data availability.

3.3 Missing data and imputation

The analysis of individual and programme delivery factors that may affect the change in SWM score pre-to-post programme participation requires complete case analysis. A review of individual and programme delivery variables of interest showed that some contained missing data (see Table A1).

Of note was the motivation variable, which had approximately 10% of values missing. There were three options to approach this missing data: (1) remove participants with missing motivation values, (2) remove the motivation variable from the analysis, or (3) impute the missing motivation values. Reviewing the missing data showed that the data could be considered as missing-at-random (MAR), meaning that this missingness could be predicted by other variables in the analysis. This means that removing participants with missing motivation values would not be appropriate as it would bias the sample and could impact the outcome of the analysis. As the variable is of theoretical interest to the analysis, option 3 was chosen. Three data imputation methods were considered, and the median imputation method was chosen. This means missing HMS data was replaced by the median HMS scores. Further explanation of this decision process can be found in Annex A, Imputation methods.

3.4 Evaluation design

The main evaluation aims 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 vary 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 scores was explored using a linear regression model. The regression model quantified the impact of each variable on the change in SWM score. The programme delivery and individual factors of interest is this analysis were:

- pre-programme strengths (i.e., the participant's first SWM score),
- delivery format (i.e., group, 1:1, or ADF small group),
- pre-programme motivation (i.e., the participant's first HMS score),
- whether individuals meet the high risk of reoffending criteria,[6]
- 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 offences.

Note that the offence history variables 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 of change in total SWM score explored whether there is a difference in change by rating type and by Success Wheel domain. A multivariate model was used to assess the influence of programme delivery and individual factors, by rating type and Success Wheel domain. The change in SWM score for each rating type, and

[6] Note that meeting the high-risk criteria was used rather than meeting the high-risk and high-need due to data availability.

change in SWM score for each domain, were fitted to the 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 following 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 an 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, these figures are not reported. This is because the standardisation has made the scores meaningless to interpret in isolation. Instead, attention is 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 sample to evaluate proven reoffending, 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 Kaizen, 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 intervention targets identified by Kaizen's theory of change should be interpreted as providing indicative support.

The data used in these analyses had already been collected as part of routine implementation of Kaizen. 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. 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 Kaizen (n = 240, 45.4%) within the defined period, but were excluded from the analytical sample due to missing SWM scores.

Observations of this missingness showed that it would be considered as missing at random (MAR), meaning the pattern of missing data could be predicted by other variables of interest in the data. This means our sample may not represent the wider Kaizen population and may produce biased estimates in the results.

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, Kaizen 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 a future offence-free lifestyle had more value for individuals (Ware and Mann, 2012). Kaizen 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 indicative 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 Kaizen cohort is different from the Horizon population according to type of previous convictions (for those without a sexual conviction) and risk-level. In short, whilst people with different offending histories share many of the same criminogenic needs (Walton et al., 2017), the construct validity of the SWM has only been established for those in the Kaizen cohort, who like Horizon participants, have a history of proven sexual offending (they represent 39.6% of the sample, 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 their 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

The overall change in SWM score from pre-to-post Kaizen participation found a statistically significant difference in standardised SWM scores from pre-programme and post-programme; (t(284) = -24.892, p < .001, d = 1.47) using a paired t-test, see Table B4 for summary statistics. This can be considered a large effect and represents a change of 4.1-5.2-point change in total SWM score (a 36.9% - 40.6% increase in score).

4.2 Linear regression model

Full model

Multiple linear regression was used to investigate whether the individual and programme delivery 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, delivery format, motivation, high-risk criteria, and history of SO, GV and IPV as predictor variables, was statistically significant (R^2 = 0.414, F(8, 276) = 24.35, p < .001), see Table B2 for the regression output. This means that 41.4% of the variation in change in SWM score can be accounted for by these predictor variables. The effects of each predictor variable whilst all others are held constant are detailed in the paragraphs below.

Pre-programme strengths was found to significantly predict change in SWM score (*b* = -0.694, p < .001, f^2 = 0.626). This can be considered a large effect. This means that higher pre-programme strengths predicted smaller change in SWM score.

Delivery format was found to significantly predict change in SWM score, with ADF (groups of two-three participants) being significantly associated with an increase in change in SWM score (*b* = 0.405, p < .001, f^2 = 0.045) compared to group delivery, the reference category. This means that ADF predicted greater change in SWM score compared to group delivery. This can be considered a small effect. There was no significant difference between 1:1

delivery and group delivery, (*b* = 0.082, p = .408). This means 1:1 delivery did not affect change in SWM score compared to group delivery.

Motivation did not significantly predict change in SWM score (*b* = 0.043, p = .102), and neither did meeting the high-risk criteria (*b* = 0.099, p = .463). This means neither motivation nor meeting the high-risk criteria influenced change in SWM score.

It was found that having a history of general violence offences (*b* = -0.088, p = .463), intimate partner violence offences (*b* = 0.033, p = .700), or sexual offences (*b* = -0.011, p = .908) did not significantly predict a change in SWM score. This means that having a history of general violence, intimate partner violence or sexual offending did not impact change in SWM score.

Parsimonious model

The parsimonious model was found to include pre-programme strengths plus delivery format (R^2 = 0.405, F(3, 281) = 63.68, p < .001). This means 40.5% of the variation in change in SWM score can be accounted for by pre-programme strengths and delivery format. Details of the stepwise regression can be found in Table B3.

Pre-programme strengths statistically significantly predicted change in SWM score (*b* = -0.668, p < .001, f^2 = 0.631). This effect size can be considered large. This means that higher pre-programme strengths predicted smaller change in SWM score. Delivery format was found to significantly predict change in SWM score, with alternative delivery format being significantly associated with an increase in change in SWM score (*b* = 0.387, p < .001, f^2 = 0.041) compared to group delivery, the reference category. This effect size can be considered small. This means that ADF predicted larger change in SWM score compared to group delivery. There was no significant difference between 1:1 delivery and group delivery, (*b* = 0.068, p = .483).

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 (i.e., who gave the score) on change in SWM score. There was a statistically significant difference in change in rating type between at least two groups (F(1.65, 467.95) = 15.52, p < .001, using Greenhouse-Geisser sphericity correction).

Pairwise comparisons using paired t-tests with Hommel's correction showed a significant difference between facilitator and joint ratings (t(284) = 4.50, p < .001, d = 0.253), and a significant difference between facilitator and participant ratings (t(284) = 4.37, p < .001, d = 0.304). This effect size can be considered small. This means that the change in SWM score was greater when rated by facilitators compared to participant- or joint-ratings but not by a large amount. 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 by t-test results.

Regression Performance

To determine whether there was a difference in the influence of individual and programme delivery factors by SWM 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 was a statistically significant but very small difference found in the pre-programme strengths coefficient between facilitator-rated SWM scores and participant-rated SWM scores (*b* = 0.124, p = .025, d = 0.153) and between participant-rated SWM scores and joint-rated SWM scores (*b* = -0.058, p = .034, d = 0.143). This means that pre-programme strengths had a greater impact on participant-rated change in SWM score compared to facilitator- or joint-rated change. There was no significant difference between facilitator-rated and joint-rated SWM scores (*b* = 0.066, p = .132). This means that pre-programme strengths had an equivalent impact on facilitator- and joint-rated change in SWM score.

There were no statistically significant differences in the delivery format coefficient between rating types. This means that delivery format had an equivalent impact on change in SWM score across rating types.

4.4 Comparison of domains

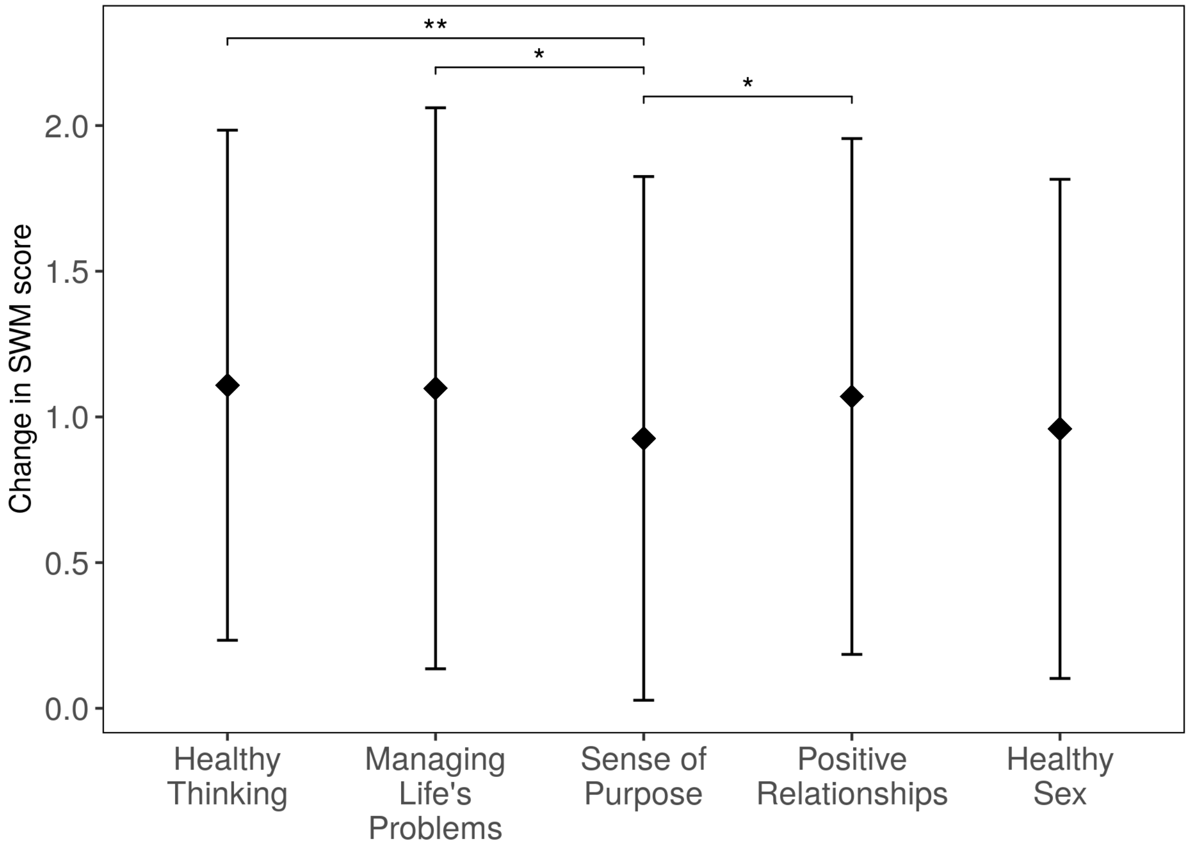
The comparison of domains must consider that only part of the analytical sample has scores in all five domains (n = 122, 42.8%). The following methods were run to compare the common four domains across the whole sample, and then compare the Healthy Sex domain to the others in the subsample where this domain is scored.

Overall change

Paired t-tests found 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 between at least two groups (F(3, 852) = 4.942, p = .002).

Pairwise comparisons between the four core domains across the whole sample, using paired t-tests with Hommel's correction, showed a small significant difference between Managing Life's Problems and Sense of Purpose (t(284) = 3.01, p = .014, d= 0.185), Healthy Thinking and Sense of Purpose (t(284) = 3.40, p = .005, d = 0.206), and Positive Relationships and Sense of Purpose (t(284) = 2.54, p = .046, d = 0.161), see Figure 1. This means that the change in SWM score was greater for the Managing Life's Problems, Healthy Thinking and Positive Relationships domains compared to the Sense of Purpose domain. All other pairwise comparisons did not show a significant difference, 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 did not show any significant differences. This means change in SWM score was similar across the domains for this subset of the sample. See Table B8 for summary statistics and Table B10 for pairwise comparisons by t-test results.

Figure 1: Pairwise comparison of change in SWM score by domain pre-to-post programme participation[7] [8]



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.

[7] 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 sub-sample. The combination into one figure is for illustrative purposes.

[8] Significant results of the pairwise t-tests are added, p < 0.05 is shown as * and p < 0.01 is shown as **.

There were no statistically significant differences in the pre-programme strengths coefficient or the delivery format coefficient between domains. This means that the predictor variables had an equivalent effect on the change in score across domains.

5. Conclusion

This interim outcome evaluation sought to provide indicative evidence as to whether Kaizen is meeting its 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 programme change, and whether this differed by SWM rating type or domain.

Pre-to-post programme change in SWM scores were found for Kaizen participants, with a large effect size: on average the post-Kaizen scores were greater than pre-Kaizen scores.

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.

How Kaizen was delivered had a small but statistically significant effect on change in SWM score. When all other variables were held constant, those that received Kaizen through alternative delivery format (small group of two to three participants) had a higher rate of improvement compared to those that received group or 1:1 delivery.

Motivation was not found to affect the change in SWM score statistically significantly. This finding suggests that low pre-programme motivation does not pose an obstacle to positive pre-to-post change in SWM scores, as found on average. This is consistent with Kaizen’s ethos for meaningfully engaging participants, since many may not possess internal motivation for change. As such Kaizen, requires that people make an honest effort to learn skills for change, whilst being open and respectful of others in that process, rather than make a commitment to change itself. It should be noted that around 10% of the motivation data was imputed as the median HMS score.

There was no statistically significant difference in change in SWM score between those that met the high-risk criteria and those that did not. This finding indicated that the rate of change from pre-to-post programme for participants who met the high-risk criteria was similar to those who had been selected at the discretion of a Treatment Manager.

Having a history of general violence, intimate partner violence and/or sexual offending convictions did not show a significant difference in change in SWM score compared to those without a history. This finding indicates that participants' offence type history does not impact change in their SWM scores.

The comparison of rating types found positive pre-to-post change for each rating type, and that the magnitude of change was significantly greater for facilitator-rated scores than participant-rated and joint-rated scores. It was found that pre-programme strengths had a less positive impact on change in SWM score when rated by participants compared to when rated by facilitators or jointly. The effect of delivery format did not significantly differ between the rating types.

Domain-specific change was found to be positive for all domains, though the change in the Sense of Purpose domain to be significantly lower than the changes in the Healthy Thinking, Positive Relationships and Managing Life's Problems domains across the analytical sample. There were no significant differences between coefficients for strengths and delivery across domains.

It should be noted that all results presented here are only indicative that participation in Kaizen 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 Kaizen. 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 his 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 programme 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 Kaizen cohort in terms of offence type and risk level. People with different offending histories share many of the same criminogenic needs (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 for those with histories of sexual offending. Further research that aims to validate the SWM for these other cohorts would help broaden its utility as a measure of clinical change.

Overall, in line the programme logic for Kaizen, 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

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.

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

Deci, E. L., & Ryan, R. M. (2008). Self-determination theory: A macrotheory of human motivation, development, and health. Canadian Psychology, 49(3), 182.

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

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.

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., 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

Ryan, R. M., & Deci, E. L. (2017). Self-determination theory: Basic psychological needs in motivation, development, and wellness. London, U.K.: Guilford Publications.

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.

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, 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.

Glossary of terms

Accredited offending-behaviour programme: Offending-behaviour programmes, such as Kaizen, 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 50th 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).

Paired t-test: A statistical test used to compare the means of repeated observations from a sample.

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, motivation, risk, 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	30 (10.4%)	Imputation	Removal of cases would bias sample, as data was found to be missing at random (MAR). Proportion of missing data within acceptable range to impute.
Meets high risk criteria	2 (0.7%)	Remove cases	Binary variable difficult to impute with the sample size. Proportion of missing data very low and unlikely to bias results.
History of general violence	0	-	-
History of intimate partner violence	1 (0.3%)	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	1 (0.3%)	Remove cases	Binary variable difficult to impute with the sample size. Proportion of missing data very low and unlikely to bias results.

Imputation methods

Three imputation methods were considered to impute MAR motivation data: (1) random number imputation, (2) mean/median value imputation, and (3) K-Nearest Neighbour (KNN) imputation.

These imputation methods were evaluated using a train-test split of the complete case data available and assessing accuracy and error rates. This involved splitting the known data into a training dataset on which each imputation was trained (where relevant) and then run on the test dataset. The imputed values in the test dataset were compared to their actual values to calculate the accuracy and error rate. The imputation was simulated over 1000 partitions of train-test split data to test for the stability of each imputation method.

Random number imputation

Random number imputation involves randomly assigning a valid value in place of the missing value. In the case of HMS score, the theoretical range of values is 0-8, but as none of the sample had a total HMS score of less than 1, the range 1-8 was used.

Mean or median value imputation

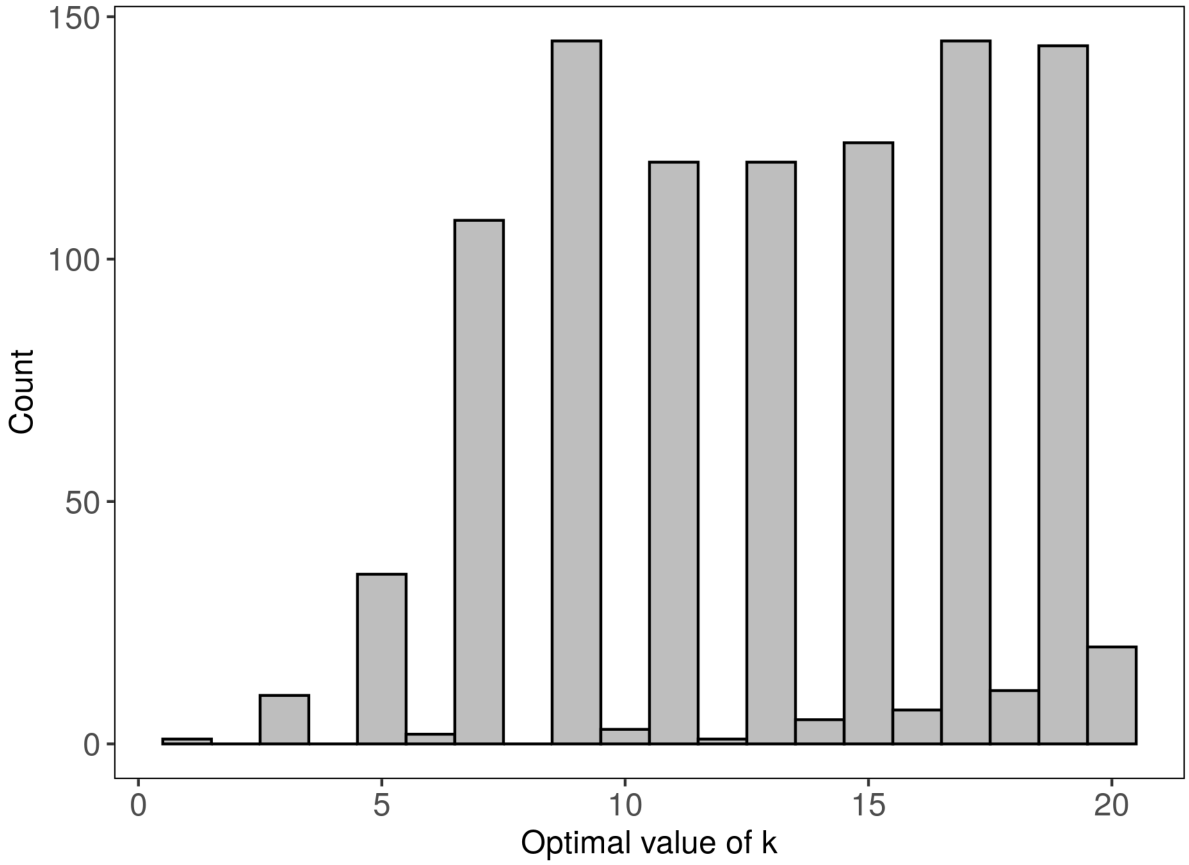
Mean and median value imputation involve assigning the mean or median of the known sample value to all cases of missing values. The total HMS score distribution does not violate normality measures in skewness or kurtosis, meaning the mean and median values are quite similar. The median value was chosen for two reasons: firstly, because the median is an integer, like all other HMS scores, whereas the mean value was decimal, and secondly the data has a small negative skew which means the mean is affected by a small number of low HMS scores.

K-nearest neighbour (KNN) imputation

K-nearest neighbour imputation involves finding a predefined number (k) of cases closest in distance to the missing case to predict its value.

KNN requires a value for k. To find the optimal value of k, a simulation was run where k was equal to the values from 2 to 20, and this was repeated over 1000 partitions of train-test data splits to find the optimal value of k. From this, the optimal value of k was found to be 19 based on highest accuracy and lowest error rate, but also that the optimal value of k varied was dependent on the train-test split and therefore several values of k performed similarly well (see figure A1).

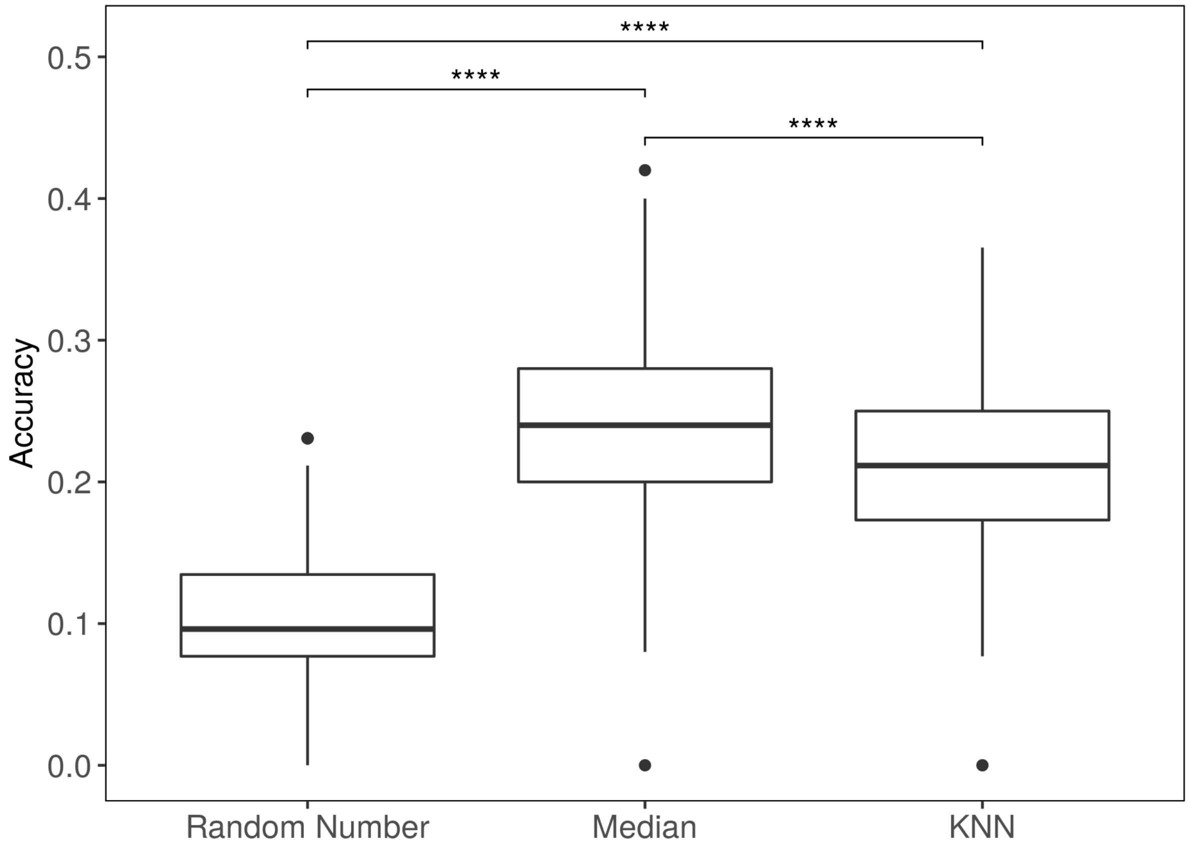
Figure A1: Calculating the optimal value of k through simulation



A comparison of the methods, by simulation over 1000 partitions of train-test data computing accuracy and error, shows that there is a significance difference between the accuracy of all three methods (see figure A2). Random number imputation gives the lowest accuracy. Median imputation performs marginally, but statistically significantly, better than KNN.

The data is presented as a box and whiskers plot to visualise the frequency of simulation outputs. The middle two quartiles (25th-75th percentiles) as a box, with the 50th percentile (also known as median) denoted by a line within the box. The tails represent up to 1.5 * the interquartile range (IQR) plus or minus from the 25th and 75th percentiles, and any dots should be considered outliers.

Figure A2: Comparison of imputation methods by accuracy in train-test data simulations[9]



[9] Significant results of the pairwise comparisons are added, p < 0.0001 is shown as ****.

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. 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
Delivery format	Group	1. 1:1 vs. Group 2. ADF vs. Group
Meets high-risk criteria	No	1. Yes vs. No
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 model

The regression outputs from the full model.

Table B2: Regression outputs for the full model

Variable	Estimate	SE	95% CI (LL and UL)		p
(Intercept)	0.460	0.242	-0.017	0.936	.059
Pre-programme strengths	-0.694	0.053	-0.798	-0.591	<.001
Delivery format (1:1)	0.082	0.099	-0.113	0.277	.408
Delivery format (ADF)	0.405	0.107	0.195	0.616	<.001
Motivation	0.043	0.026	-0.009	0.095	.102
Meets high-risk criteria (Y)	0.099	0.134	-0.166	0.363	.463
History of general violence (Y)	-0.088	0.119	-0.323	0.147	.463
History of intimate partner violence (Y)	0.033	0.086	-0.137	0.203	.700
History of sexual offending (Y)	-0.011	0.094	-0.197	0.175	.908

Table B3 presents the statistics for goodness-of-fit between models to determine the parsimonious model by forward stepwise regression. The null model represents the intercept only and each row after adds the stated variable to the model.

Table B3: Stepwise regression to find parsimonious model

Model	df	RSS	AIC	BIC	F	p
Null		213.51	730.48	736.79		
Pre-programme strengths	1	133.23	598.08	609.04	177.01	<.001
Delivery format	3	127.10	588.65	606.91	6.76	.001
Motivation	4	125.83	587.79	609.71	2.80	.096
Risk	5	125.46	588.94	614.52	0.82	.365
History of general violence (Y)	6	125.25	590.46	619.68	0.47	.494
History of intimate partner violence (Y)	7	125.17	592.30	625.17	0.16	.691
History of sexual offending (Y)	8	125.17	594.29	630.81	0.01	.908

The most complex statistically significant model is change in SWM score predicted by pre-programme strengths and delivery format.

Table B4 presents the mean and standard deviation for facilitator-rated summed standardised SWM scores at pre- and post-programme.

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

Variable	Mean	SD
Before	-0.643	0.800
After	0.635	0.736

Table B5 gives the mean and standard deviation for facilitator-rated, participant-rated and joint-rated change in summed standardised SWM scores.

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

Variable	Mean	SD
Facilitator	1.28	0.867
Participant	0.995	0.994
Joint	1.05	0.934

Table B6 gives the results of pre-to-post change as measured by paired t-tests for each rating type.

Table B6: Pre-to-post paired t-tests by rating type

Variable	df	T	p	Cohen's d
Facilitator	284	-24.892	<.001	1.47
Participant	284	-16.899	<.001	1.00
Joint	284	-18.993	<.001	1.13

Table B7 gives the output of the pairwise comparisons between rating type, conducted using paired t-tests with adjusted p-values using Hommel's correction.

Table B7: Pairwise comparison of change in SWM score by rating type, using paired t-test

Comparison	df	T	p	p.adj	Cohen's d
Facilitator - Participant	284	4.50	<.001	<.001	0.304
Facilitator - Joint	284	4.37	<.001	<.001	0.253
Participant - Joint	284	1.26	.208	.208	0.058

Table B8 combines the mean, standard deviation and count 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	1.10	0.963	285
Healthy Thinking	1.11	0.875	285
Positive Relationships	1.07	0.885	285
Sense of Purpose	0.93	0.899	285
Healthy Sex	0.96	0.857	122

Table B9 gives the results of pre-to-post change as measured by paired t-tests for each domain.

Table B9: Pre-to-post paired t-tests by domain

Variable	df	T	p	Cohen's d
Managing Life's Problems	284	-19.257	<.001	1.14
Healthy Thinking	284	-21.383	<.001	1.17
Positive Relationships	284	-20.407	<.001	1.21
Sense of Purpose	284	-17.398	<.001	1.03
Healthy Sex	121	-12.365	<.001	1.12

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, using paired t-test

Comparison	df	T	p	p.adj	Cohen's d
MLP-HT	284	0.202	.840	.840	0.070
MLP-PR	284	0.531	.596	.840	0.120
MLP-P	284	3.010	.003	.014	0.231
HT-PR	284	0.764	.446	.840	0.048
HT-P	284	3.400	<.001	.005	0.160
PR-P	284	2.540	.012	.046	0.114
HS-HT	121	-1.52	.132	.660	-0.147
HS-MLP	121	-1.78	.077	.396	-0.186
HS-PR	121	-2.24	.027	.205	-0.242
HS-P	121	0.43	.666	.666	0.051

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	0.124	0.048	2.564	.025	0.153
Facilitator-Joint	0.066	0.044	1.506	.132	0.089
Participant-Joint	-0.058	0.024	-2.392	.034	0.143

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

Table B12: Comparison of delivery format coefficient by rating type

Comparison	Estimate	SE	Z	p.adj	Cohen's d
Facilitator-Participant	0.223	0.115	1.927	.108	0.115
Facilitator-Joint	0.084	0.097	0.867	.386	0.051
Participant-Joint	-0.138	0.064	-2.173	.081	0.128

Table B13 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 B13: Comparison of pre-programme strengths coefficient by domain

Comparison	n	Estimate	SE	Z	p.adj	Cohen's d
MLP-HT	285	-0.031	0.047	-0.653	.681	0.039
MLP-P	285	-0.074	0.048	-1.549	.681	0.091
MLP-PR	285	-0.054	0.047	-1.172	.681	0.068
HT-P	285	-0.043	0.049	-0.897	.681	0.052
HT-PR	285	-0.024	0.047	-0.509	.681	0.030
P-PR	285	0.020	0.048	0.412	.681	0.025
MLP-HS	122	-0.174	0.080	-2.182	.095	0.197
HT-HS	122	-0.153	0.077	-1.982	.142	0.180
P-HS	122	0.005	0.085	0.057	.954	0.005
PR-HS	122	-0.103	0.079	-1.301	.387	0.118

Table B14 presents the output from the simultaneous tests for general linear hypotheses when comparing the delivery format (ADF) 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 delivery format coefficient by domain

Comparison	n	Estimate	SE	Z	p.adj	Cohen's d
MLP-HT	285	-0.121	0.096	-1.253	.841	0.075
MLP-P	285	-0.132	0.102	-1.294	.783	0.077
MLP-PR	285	-0.154	0.098	-1.568	.526	0.093
HT-P	285	-0.011	0.101	-0.112	.911	0.006
HT-PR	285	-0.033	0.098	-0.338	.911	0.020
P-PR	285	-0.022	0.103	-0.211	.911	0.013
MLP-HS	122	-0.029	0.204	-0.142	>.999	0.013
HT-HS	122	-0.114	0.199	-0.573	.927	0.052
P-HS	122	-0.034	0.236	-0.142	>.999	0.013
PR-HS	122	0.019	0.216	0.089	>.999	0.008

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- 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.

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.