Ministry
of Justice

Horizon and iHorizon

An uncontrolled before-after study of clinical outcomes

Ian Elliott and Olivia Hambly

Data and Analysis Directorate

Ministry of Justice Analytical Series

2023

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

Disclaimer

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

First published 2023

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

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

ISBN 978 1 84099 986 0

Acknowledgements

We would like to thank the HMPPS programme developers and facilitators for collecting and reporting the data for this study. Additional data for the psychometric analyses were collected at HMPPS Newbold Revel Training Centre by Ministry of Justice staff Kiran Horne, Sarah Louise Smith, Jessie Brener, Keely Wilkinson, and Lydia Baxter, to whom we owe a debt of gratitude. Our thanks also go to Johannes Huber for diligent quality assurance checks. We would also like to thank the HMPPS Learning and Development team, David Butler-Trump in particular, for their assistance in the collection of normative data for this project. Finally, we would like to thank the various colleagues who reviewed and commented on this report, especially Alana Diamond, members of the Correctional Services Accreditation and Advice Panel, and two anonymous academic peer reviewers.

Contents

List of tables

List of figures

1. Summary

Introduction and study aims

Horizon is an accredited offending behaviour programme delivered by HMPPS in both custody and the community for adult men with a sexual conviction. iHorizon is designed for individuals with convictions relate to indecent images of children only. Both target individuals at medium risk and above according to the Risk Matrix 2000s risk assessment tool and whose primary clinical need is to address sexual offending and are designed to enable participants to build constructive lives that do not involve further offending.

The aim of this evaluation was to establish whether Horizon and iHorizon participants were demonstrating positive progress (i.e., acquiring pro-social insights and skills) across 5 key treatment targets: (1) Managing life's problems, (2) Healthy relationships, (3) Healthy sexual interests, (4) Healthy thinking, and (5) Sense of purpose (desistance from crime).

Methodological approach and interpreting findings

This report presents an uncontrolled before-after study examining progress for participants on Horizon and iHorizon using scores on the Success Wheel Measure (SWM) assessment of progress. Routine and supplementary data was collected for 1,041 adult male Horizon participants and 122 adult male iHorizon participants at 27 delivery sites (19 custodial and 8 community) from groups starting between November 2018 and January 2020. Programme facilitators assessed the extent to which iHorizon/ Horizon participants demonstrated improvements across the five treatment targets, measured by the SWM.

This study used a series of "mixed-design" multilevel model analyses first to examine whether pre-to-post change had occurred and then whether this change was affected by other factors, such as: (1) the extent to which insight and skills already existing at the start of the programme; (2) whether the presence of those insights and skills were being judged by the facilitator or the participant; (3) levels of motivation prior to engaging in the programme; (4) whether participants were maintaining innocence; (5) whether the programme was delivered in custody or the community. Finally, analyses were conducted to ensure the extent of change was consistent across the five different treatment targets.

The absence of a no-treatment control group with whom to compare samples means any changes cannot be directly attributed to participation on Horizon or iHorizon. As observed change could be due to unobserved factors, findings must be considered as indicative.

Key findings

When all other variables were accounted for, large pre-to-post individual change in aggregate SWM scores was observed. Separate psychometric analyses have indicated that the measures used in this study demonstrated acceptable validity and reliability.

For both Horizon and iHorizon pre-to-post effect was moderated by participant's existing strengths, with the pre-to-post effect diminishing for participants with pre-Horizon total SWM scores greater than 15 (60% of the total). The observed pre-to-post effect was also much larger for scores provided by participants than for scores provided by facilitators.

For Horizon the pre-to-post effect was also higher for those with higher pre-Horizon motivation and for those who participated in custody, but the sizes of those differences were very small. For iHorizon, motivation did not affect pre-to-post change. Participants maintaining innocence did not affect their pre-to-post change on either programme.

Positive pre-to-post change was observed in all Success Wheel domains but observed change was lower in the "Healthy thinking" and "Positive relationships" domain. Observed pre-to-post change was smaller for scores provided by facilitators rather than participants on the Managing life's problems, Heathy thinking, and the Sense of purpose domains.

Conclusions

These findings provide promising evidence that participation in Horizon and iHorizon is associated with positive change in programme participants. They also identify groups with fewer existing insights and skills and higher motivation who may be able to benefit more from participation. Change was consistently observed across all treatment targets, particularly for problem solving, sexual interests, and purpose (e.g., structure and routine).

Nevertheless, the lack of a no-treatment comparison group means that these positive changes cannot be directly attributed to Horizon or iHorizon. The next step should be to examine, using more rigorous methods, whether positive change on Horizon and/or iHorizon participants translates into reductions in future reconviction in the community.

2. Introduction

2.1 Aims of this study

The aim of this evaluation was to assess pre-to-post programme progress for participants on key treatment targets for Horizon and iHorizon. To achieve this, total scores on a measure of programme progress called the Success Wheel Measure (SWM) were used to conduct uncontrolled before-after analyses (UBA: also known as a “pre-post study”). More information on the Success Wheel Measure can be found in Section 3.2 and the supplementary psychometric report. The research objectives were to explore:

- Whether or not participants on Horizon/iHorizon demonstrated improvement on relevant programme-related measures, overall and on different treatment targets;
- Whether or not that improvement was affected by individual and operational factors (e.g., the context in which they participated, their motivation to participate, etc.)

The findings of this study are intended to provide exploratory contextual information on progress for Horizon and iHorizon participants that should inform the interpretation of findings from future impact evaluations (i.e., is there evidence that improvements in strengths translated into reductions in proven reconviction rates). This evaluation, however, cannot directly address the question of whether the programme is responsible for pre-to-post changes detected in participants because it does not compare the outcomes of participants to a no-treatment comparison group. This evaluation also does not aim to examine if pre-to-post change is associated with likelihood of future reoffending.

2.2 Horizon and iHorizon

Horizon is an accredited offending behaviour programme (OBP) delivered by Her Majesty’s Prison and Probation Service (HMPPS) in custody and the community for adult men with a sexual conviction. Introduced in 2016, it replaced the Core Sex Offender Treatment Programme (Core SOTP) in prisons and three community programmes: the Community Sex Offender Groupwork Programme (CSOG-P), the Northumberland Sex Offender Groupwork Programme (NSOGP) and the Thames Valley Programme (TVP). At the time of this study, Horizon was delivered at 20 prisons across England and Wales and the 7 National Probation Service (NPS) divisions. Both Horizon and iHorizon are for individuals who are medium risk and above according to the Risk Matrix 2000s risk assessment tool and whose primary clinical need is to address sexual offending.

Horizon is designed to enable participants to build constructive lives that do not involve further offending. Programme activities and delivery for Horizon are designed to reflect current knowledge in the management of individuals with sexual convictions and the wider psychological and criminological literature. Horizon is designed to adhere to organising principles set out by Carter and Mann (2011; Mann & Carter, 2012) and is underpinned by a biopsychosocial model of change (Carter & Mann, 2011; Walton et al., 2017) and desistance from crime research and theory (Farmer et al., 2012; McAlinden et al., 2017). At the time study data was collected, Horizon consisted of 31 sessions over 62 hours.

iHorizon is designed for individuals with convictions only for possessing, downloading, making and/or distributing indecent images of children, not for contact child sexual abuse or a combination of both. iHorizon was introduced in the community in 2018 and replaced the Internet Sex Offender Programme (iSOTP). The content of iHorizon differs from Horizon in that iHorizon is shorter than the standard version of Horizon, there is reduced focus on problem solving and controlling emotions, and there is an additional focus on problematic internet use. iHorizon is delivered only in the community and not in custody. At the time study data was collected, iHorizon consisted of 23 sessions over 46 hours.

A 2019 process study for Horizon found that completion rates were high and feedback from staff and participants was broadly positive. Nevertheless, several issues were identified for improvements to programme content, staff training, and operational delivery (Wilkinson & Powis, 2019). Changes to the programme were made to address concerns raised. Horizon and iHorizon aim to help participants improve in five domains:

- **Managing life’s problems** (MLP: e.g., controlling feelings; solving problems)
- **Healthy thinking** (HT: e.g., fewer pro-offending beliefs; respecting rights of others)
- **Healthy sexual interests** (HSI: e.g., not using sex to cope with negative events)
- **Positive relationships** (PR: e.g., perspective-taking; assertiveness; negotiation)

- **Sense of purpose** (SOP: e.g., protective factors; being a good member of society)

These domains were drawn from the four domains of the dynamic risk domain model (Thornton, 2013): self-management, distorted attitudes, sexual interests, and relational style. These four domains were reconfigured as strengths-based opportunities for growth (e.g., “distorted attitudes” become “healthy thinking”) and are embedded in Horizon and iHorizon as positive outcomes or “approach goals” using the Success Wheel tool (Walton et al., 2017). The Sense of purpose domain was also included to represent development of a desistance identity (self-efficacy and agency) and citizenship (community participation).

2.3 Measuring improvement

In a review, Olver and Stockdale (2020) concluded that positive pre-to-post change on sexual offending programmes has been established for risk-relevant constructs such as attitudes to offending, problems maintaining intimate relationships, hostility and anger, and problematic sexual interests. However, they also note mixed results for constructs like well-being and that for constructs like victim empathy, large observed pre-to-post change has typically not translated into reductions in reoffending. Many of their conclusions are drawn from Mann et al. (2010) and relevant meta-analyses (e.g., Helmus et al., 2013).

In a review of methods to evaluate change in men with sexual convictions, Olver and Stockdale (2020) note that although pre- and post-programme administration of self-report psychological tests is common, emerging evidence better supports clinician-rated tools. They argue that self-report measures are narrow in focus and often not designed with risk of reconviction in mind, whereas clinician rated tools are transparent, less prone to socially desirable responding, and assess a range of constructs. They concluded that useful information is provided by routine assessment using psychometrically sound measures. Such measures include the Stable 2007 (Brankley et al., 2017; Hanson et al., 2007), Sex Offender Treatment Intervention Progress Scale (SOTIPS: McGrath et al., 2013), and Violence Risk Scale–Sex Offenders (VRS-SO: Wong et al., 2007), which target meaningful risk factors from theoretically-sound interventions, using suitable statistical procedures.

In the context of HMPPS accredited programmes targeting individuals with sexual convictions, Wakeling et al. (2013) and Wakeling and Barnett (2014) also concluded that limited support existed for the use of relevant self-report psychological tests to assess change and to predict future sexual reconvictions. Other small-sample studies have found some (albeit mixed) evidence of short-term change for treatment targets on programmes for cohorts with sexual convictions using pre-to-post methods and self-report assessments (e.g., Beech et al., 1999; Beech and Ford 2006; Harkins 2008; Keeling et al., 2006; Wakeling et al., 2013). However, a rigorous evaluation of longer-term impact on reconviction of Core SOTP did not find participation to be associated with lower reconviction rates (Mews et al., 2017). As such, reasons why an association was not found between psychometric change and future reconvictions for core SOTP remains unclear.

2.4 Establishing improvement on Horizon/iHorizon

At the time this study was being designed, there were insufficient programme completer numbers to conduct a robust reoffending impact study. Impact evaluations are reliant on large enough sample sizes and suitable follow-up periods to detect statistically significant change between treatment and comparisons groups and therefore to deliver reliable conclusions. This, coupled with relatively low baseline proven reoffending rates for the population of individuals with sexual convictions, means it can take a long time (sometimes upwards of 8 years) to generate the data needed to evaluate proven reoffending.

In any case, Epstein and Klerman (2012) make the case that programmes should be required to “pass their own logic model” and recommend seeking evidence of pre-/post-programme improvement on short-term outcomes prior to designing an impact study. They suggest researchers examine what contributions to change are being made by the constituent parts of the programme’s content and how that fits with our expectations of what should be changing. If we cannot establish that the expected skills and insights are being correctly obtained by the participants, we cannot be confident that those skills and insight are available to change future behaviour and generate long-term impact. Conversely, positive change on relevant treatment targets that is observed may not result in lower rates of reconviction, in which case programme theory should be reconsidered.

3. Methodology

3.1 Sample

In total, routine and supplementary clinical data was collected for 1,163 adult males who participated in Horizon (n = 1,041, 89.5%) or iHorizon (n = 122, 10.5%), including those who completed and those who did not complete. Data were collected from 27 delivery sites (19 of 20 custodial and all 7 community sites) for all groups starting between November 2018 and January 2020. This time period was chosen as it coincided with the implementation of the interim measures at programme sites. Data for an additional measure used to validate the routine measures (SOTIPS: see Appendix 1) were collected at 7 sites (4 custodial, 3 community) for 147 of the 1,041 Horizon participants (14.1%).

Complete cases (no missing data) were selected for analysis, resulting in samples of 886 Horizon participants (85.1% of Horizon cases) and 92 iHorizon participants (75.4% of iHorizon cases). Table 1 provides demographic data for the analytical samples.

Table 1: Demographic data for the Horizon and iHorizon samples

Variable	**Horizon**	**iHorizon**
Age		
Mean (SD)	40.1 (14.7)	38.5 (14.9)
Range	18–82	20–75
Ethnicity		
Aggregated Asian ethnicities (ethnicity codes A1, A2, A3, A9)	53 (6.0%)	2 (2.2%)
Aggregated black ethnicities (B1, B2, B9)	18 (2.0%)	3 (3.2%)
Aggregated mixed ethnicities (M1, M2, M3, M9)	12 (1.35%)	0 (0.0%)
Other ethnicities (O9)	4 (0.45%)	0 (0.0%)
Aggregated white ethnicities (W1, W2, W9)	784 (88.5%)	87 (94.6%)
Unknown	15 (1.7%)	0 (0.0%)

Variable	Horizon	iHorizon
Maintaining innocence		
Yes	145 (16.4%)	2 (2.2%)
Partially	60 (6.7%)	5 (5.4%)
No	681 (76.9%)	85 (92.4%)
Unknown	13 (1.3%)	1 (0.8%)
Delivery context		
Custody	434 (49.0%)	0 (0.0%)
Community	452 (51.0%)	92 (100.0%)

Note: Some categories in these groups were combined for analytical purposes (e.g., maintaining innocence).

3.2 Measures

The main Horizon analyses utilised three measures: The Success Wheel Measure (SWM), the Horizon Motivation Scale (HMS), and the Sex Offender Treatment and Interventions Progress Scale (SOTIPS). The SOTIPS was implemented for Horizon participants only at a selection of sites for a fixed time period, for the purposes of this study. Descriptions and psychometric properties of the measure are in the supplementary psychometric report.

The Success Wheel Measure (SWM) is a 5-item scale with scores ranging from 0–25 that was designed by HMPPS and the Data and Analysis evaluation team to specifically to measure progress in the 5 domains targeted by Horizon and iHorizon: (1) Managing life's problems, (2) Healthy relationships, (3) Healthy sexual interests, (4) Healthy thinking, and (5) Sense of purpose. The Horizon Motivational Scale (HMS) is a 4-item scale that was designed specifically by HMPPS and Data and Analysis to measure motivation towards participating in Horizon: (1) enthusiasm, (2) direction (i.e., internal desire and willingness), (3) commitment, and (4) holistic (i.e., recognition that programmes are one of several pro-social activities to wider efforts to live an offence free life). The Sex Offender Treatment and Interventions Progress Scale (SOTIPS) (McGrath et al., 2012) is composed of 16 dynamic risk items shown to have a statistically significant relationship to sexual recidivism (classified into sexual, criminal, co-operation, self-management, and social stability).

The SWM and HMS were implemented on Horizon and iHorizon in 2018 as routine assessments at all sites for all participants. Pre-programme SWM scores were collected at the first 1-to-1 coaching session, which follows module 2 of the programme. Although this means that the pre-programme SWM scores do not precede the start of Horizon or iHorizon, modules 1 and 2 are related to engagement and rapport and the introduction of programme concepts respectively. The modules that follow the 1-to-1 coaching session are those that directly address those strengths targeted by the Success Wheel.

Although an HMS was administered at the 1-to-1 coaching session after module 2, an HMS was also administered at the end of the pre-group individual session at the very start of the programmes. This administration protocol was implemented as it was considered crucial to measure motivation before participants took part in activities specifically targeting concepts measured by the HMS (engagement, constructive participation, etc). This study uses scores from the 1-to-1 coaching session as the basis for pre-programme motivation and the final administration as the basis for post-programme motivation.

Psychometric findings

Psychometric analyses provide information that allows researchers to judge the quality of psychological or criminological assessments. Since the SWM and HMS were new measures, specifically developed for Horizon/iHorizon, and are central to these analyses we need to know if they measured the things they intended to measure, to a reasonable standard, and that they compare favourably to other similar measures. Psychometric findings indicate that the SWM and HMS performed well (see supplementary report).

Items on both the SWM and HMS appeared to primarily measure one general factor in common (i.e., strengths and motivation) to a degree that met established thresholds. Both the SWM and HMS showed acceptable-to-good levels of reliability. The different items measured the general factors in independent ways while positively contributing to meaningful overall totals, captured a reasonable range of scores, and effectively discriminated between participants who are high and low on the general factor. The SWM also appeared to measure programme progress in a similar manner to the Sex Offender Treatment Intervention Progress Scale (SOTIPS: McGrath et al., 2012). The SOTIPS is well-established as an assessment for this population and successful validation studies have been conducted (e.g., Hanson et al., 2021). However, the SOTIPS requires prohibitively more resource to complete and is not tailored to the aims of Horizon/iHorizon.

3.3 Design

Several analyses were undertaken each utilising a form of multilevel regression modelling to explore (a) overall pre-to-post change on SWM scores, (b) the effect of several key individual and operational variables on pre-to-post change, and (c) whether any pre-to-post differences themselves differ by Success Wheel item. A multilevel modelling approach was chosen as the data collected for Horizon and iHorizon was hierarchical in nature (i.e., the variables of interest are grouped together in a tree-like structure). Statistical analyses were conducted using the R statistical software (R version 3.5.1).

The process of multi-level modelling is to build up from a simple linear regression model to more complicated models to examine the extent to which variables of interest predict our outcomes. We begin with a null model that examines the variance in the data if no conditions are imposed, as if the scores were generated at the same time, by the same person, with the same level of motivation, and so forth. Each of the conditions is individually added one-by one to ask: ‘Does applying this condition to the model improve its ability to predict the outcomes?’ We can then explore the best-fitting (or most “parsimonious”) model to see what constituent parts are generating any effects we find.

Not only do multi-level models allow us to explore complicated relationships between conditions that we have imposed, but also helps to overcome or minimise the effects of statistical matters we would otherwise affect analyses, such as differences between fixed and random effects, issues of homogeneity of regression slopes and assumptions of independence, and missing data, because these issues can be specified in the model.

Analyses

We planned to explore three key research questions:

- Whether or not there are meaningful differences between aggregate total SWM scores measured at pre-programme and post-programme assessments.
- Whether or not any pre-to-post differences in aggregate SWM scores are moderated by individual and operational variables: existing pre-programme strengths, pre-programme motivation, whether ratings are provided by facilitator clinical judgement or participant self-report, delivery context, and maintaining innocence.

- Whether or not any pre-to-post change differed between the five items of the Success Wheel (Managing life’s problems, Healthy thinking, Healthy sexual interests, Positive relationships, and Sense of purpose).

Because this study is exploratory in nature and may be considered "hypothesis-generating" rather than "hypothesis testing" our aim is simply to explore potential relationships between variables. In this approach, we do not have prior assumptions (or we have only intuitive assumptions) about the relationships between individual and operational variables and clinical outcomes for participants on Horizon or iHorizon. Nevertheless, based on insights generated by our data and analyses, we hope to subsequently develop post-hoc hypotheses about how variables might moderate the effect of Horizon and iHorizon on outcomes of interest that we can test more robustly.

3.4 Limitations

Due to the before-after nature of the study, we could only include those who contributed both pre- and post-programme scores. This has the effect of removing all those who did not complete the programme since they do not have post-programme data. As a consequence, the findings of this study only provide indications of changes in perceived strengths over the duration of Horizon/iHorizon for those who complete the programme.

Also due to the methodology, changes in scores provide only indicative evidence for change (positive, negative, or no-change) during the period in which the participants attended Horizon, since the lack of a no-Horizon control group with whom to compare our treated sample means we cannot draw causal inferences. Furthermore, this study aims to provide insight into what variables affect progress over time on the programme but not whether the programmes affect likelihood of reoffending. There are three key limitations, described in detail below (Marsden & Torgerson, 2012; Torgerson & Torgerson, 2008).

The first is temporal change. Many psychosocial problems are self-limiting and temporal improvements on relevant variables can be seen irrespective of any intervention. Furthermore, if the effect takes time to manifest one or more other variables might be introduced or changed during the intervention period. Therefore, for some, change is a function of natural changes over time and not necessarily attributable to the intervention in question. The greater the time between tests the greater the influence of temporal effects may be. Furthermore, there may be an existing temporal change underway that a single pre-test will fail to recognize (e.g., an improvement on socio-affective variables from the time of arrest and the time when the intervention begins).

The second limitation is regression to the mean. When any number of individuals are measured, performance on a single test with an error component will vary and scores will range from the highest to the lowest, but most scores will cluster around the mean. When tested again, individual with pre-intervention scores that represent outliers ("extreme" high or low scores) tend to "regress" down or up to the mean. As Clifton and Clifton (2019) explain, "if an extreme measure is observed at baseline, then its value is likely to be less extreme in the post-intervention measure, even if the intervention has no effect" (p. 2). For some, change is a function of natural corrective processes and not attributable to the intervention. Methods to counteract these effects (and those of temporal change) include the use of aggregated scores from multiple pre-tests (e.g., Shadish et al., 2002).

The third is the potential influence of test effects. The before-and-after design relies on the forms of measurement used. It is possible that improvements can be attributable to factors such as practice effects or test items themselves generating retrospective positive learning independent of the intervention. Methods to counteract these effects include the use of two or more tests for single treatment variables and including participants who only receive the post-test (an adaptation of the Solomon four-group design: Shadish et al., 2002). Similarly, as Horizon uses a novel measure we do not have data on test-retest reliability and consequently cannot rule out the possibility that effects – particularly effects with small effect sizes – are not due to error in the measure and random noise in data it produces.

To improve the rigour of future clinical studies, we recommend two improvements to future study designs, since we cannot randomise participants to a no-Horizon group (see Marshall and Marshall (2007) and Seto et al. (2008) for a review of the strengths and limitations of randomised controlled trials for cohorts with sexual convictions). The first is the use of the difference-in-difference (DID) study design (see Lee (2016) for an overview of DID). In the DID design, a programme is provided to both an eligible group and a non-programme comparison group between the two time periods of a pre-/post-programme design, so that any changes in either known or unknown variables that could influence outcomes measures (e.g., maturation, test effects, etc.) are likely to be experienced equally by both groups. Another approach would be to utilise multiple assessment points, rather than just two (pre and post). As Shadish et al. (2002) note, this provides information on (a) how the groups being compared differ initially and (b) the magnitude of initial group differences, making it easier to identify and account for other sources of change.

The sample sizes, although relatively large in the context of forensic clinical datasets, are still small at an absolute statistical level. We conducted power analyses that indicated that we would require approximately 650 participants to see an overall effect equivalent to a one-full-segment (20%) improvement on the SWM. Given the large effect sizes that we detected the analyses appear adequately powered. As already noted, the specific sample size for the iHorizon analyses, however, is not conducive to confidence in those findings.

Although we could utilise around 90% of the data, approximately 10% of the data were missing. Missingness analyses indicated that some data used in the analysis was likely to be missing not-at-random. Specifically, statistically significantly more facilitator pre-programme total SWM data was missing from the community (94% of scores available) than custody (99.4%) and from those partially maintaining innocence (84.1% of scores available) than those strictly maintaining (98.2%) or not maintaining (97.4%). No concerns about missingness for motivation data was identified. Solving the problem of missingness is a trade-off of biases. Deleting cases with missing data can add bias and/or negatively affect the quality of regression models and their parameters. However, imputing those data points can also add bias if the probability of missing data is related to the observed data (i.e., the pattern of missing data can be predicted by one or more of the variables). For a discussion of these issues see Pepinsky (2008). We decided to remove cases with missing data from our analyses and our outcomes should be interpreted with this in mind.

In addition, the SWM is a clinical judgement tool administered by facilitators to individuals that they hope to see succeed (both for the participant and their own benefit). Thus, the test effect may also include an element of natural confirmation and/or self-fulfilment bias as participants and facilitators focus on the positives (both in terms of progress and in terms of establishing an evidence-base for the programmes). There may also be some questions about how the SWM items are being interpreted by facilitator and participants. The novelty of the SWM, and the lack of psychometric data (e.g., test-retest reliability) prior to the analyses conducted for this study, meant that clinically significant change over the duration of the programme could not be calculated. It is also important to reiterate that the pre-Horizon SWM is not completed before Horizon begins. Scores are generated after modules 1 and 2 (engagement, rapport, and introductions) but before any material related to Success Wheel domains. It is therefore possible that change resulting from the modules covering engagement and rapport are not effectively captured in these data.

We also cannot rule out that participants and facilitators may be incentivised to engage in socially desirable responding – or operationally desirable responding in the case of facilitators. Concerns about “gaming” the SWM by purposefully restricting pre-Horizon scores and/or inflating post-Horizon scores to create a positive effect were considered during its development and strengthened the rationale for a pre-programme independent objective clinical judgement and the use of collaborative scores as extra safeguards.

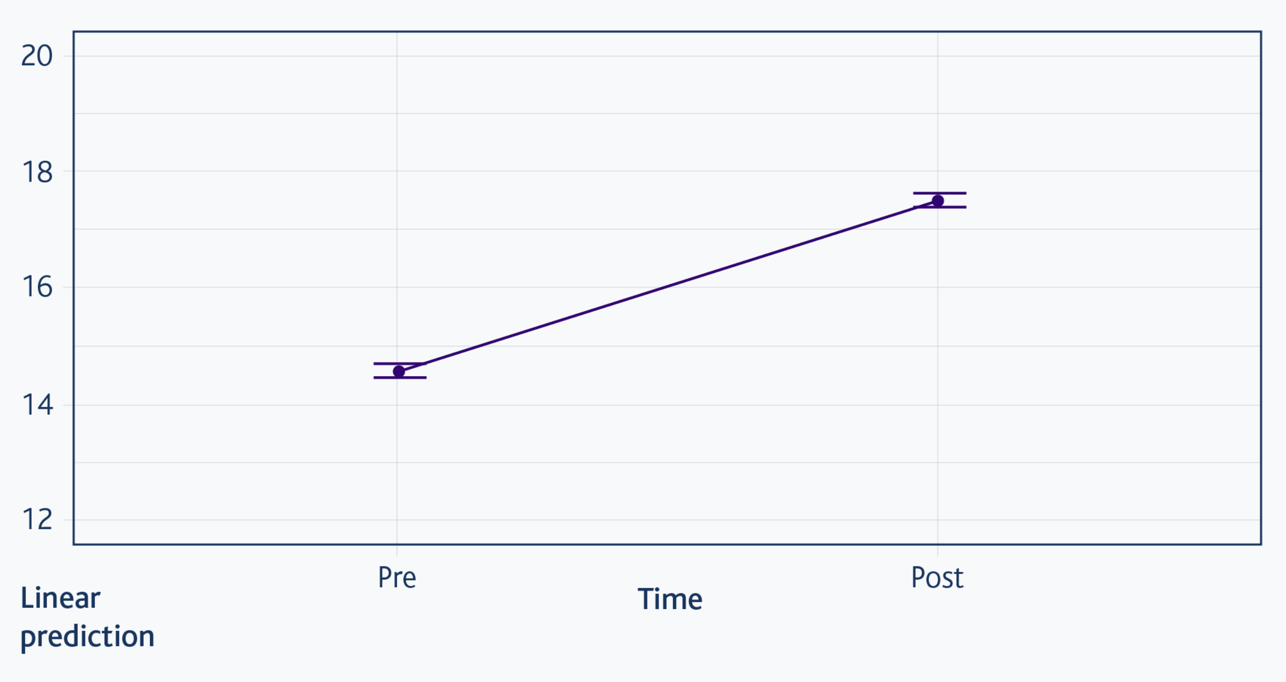
4. Findings

The following sections provide an overview of the findings of each of the three Horizon analyses. More detailed statistical information is available for each analysis in technical appendices at the end of the report, and these are indicated in the text. All effects are reported as statistically significant if they are lower than a threshold of $p < 0.05$. Effect sizes for contrasts are provided as *r* correlational coefficients. According to McGrath and Meyer (2006), we can interpret an *r* value of 0.10–0.23 as a small effect, 0.24–0.36 as a moderate effect, and greater than 0.37 as a large effect. For partial eta squared (η^2_p), Miles and Shevlin (2001) suggest the thresholds of 0.02–0.13 for a small effect, 0.13–0.26 for a moderate effect, and greater than 0.26 as a large effect. See Appendix A for statistical outputs of all analyses, contrasts, and post-hoc tests.

4.1 Horizon

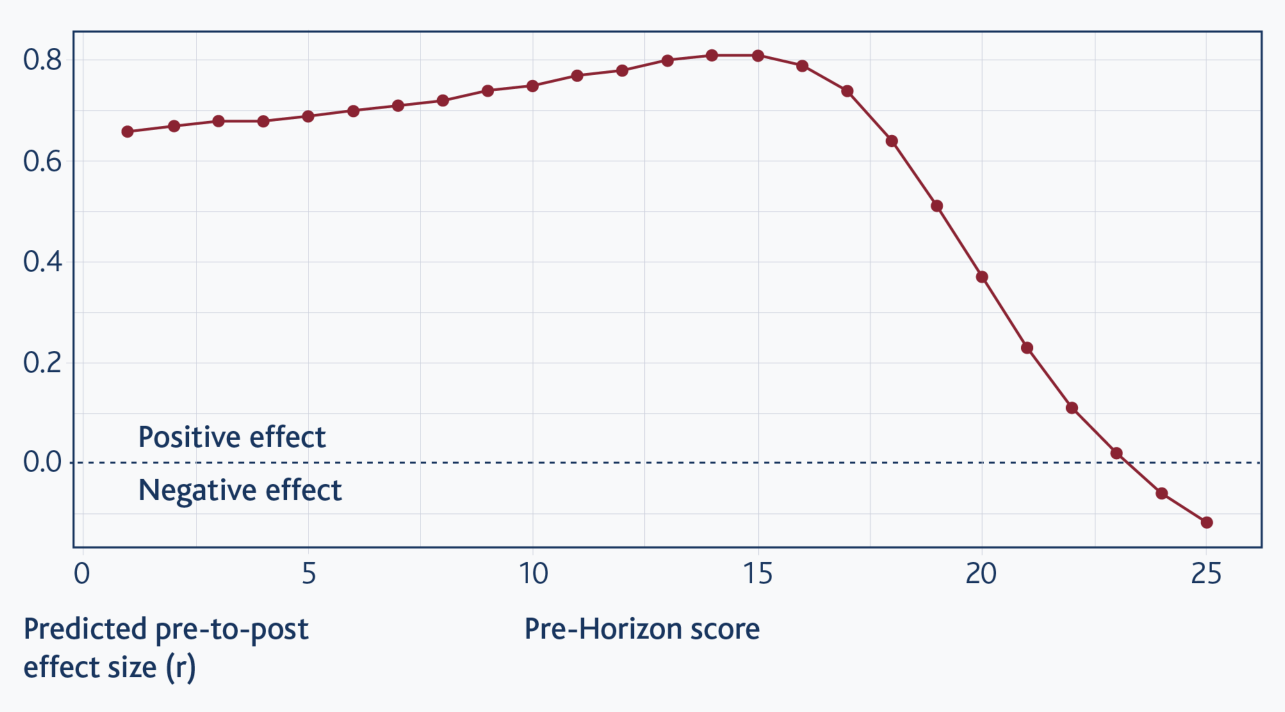
After controlling for pre-Horizon strengths, pre-Horizon motivation, rater, delivery location, and maintaining innocence, overall SWM scores were statistically significantly higher at the post-Horizon assessment point than at the pre-Horizon assessment point (X^2 (13) = 1697.31, $p < 0.0001$, $\eta^2_p = 0.61$). This would be considered a very large effect (see Figure 1) and represented a 4.1 increase in overall total SWM points (a 29.3% change in score).

Figure 1: The pre-to-post difference in scores observed in the Horizon sample



Further analyses indicated that this large pre-to-post effect was statistically significantly different depending on the participant's pre-Horizon baseline strengths (X^2 (14) = 641.22, *p* < 0.0001, η^2_p = 0.29), with the size of the pre-to-post difference diminshing as pre-Horizon strengths increase (see Figure 2). This would be considered a large sized effect.

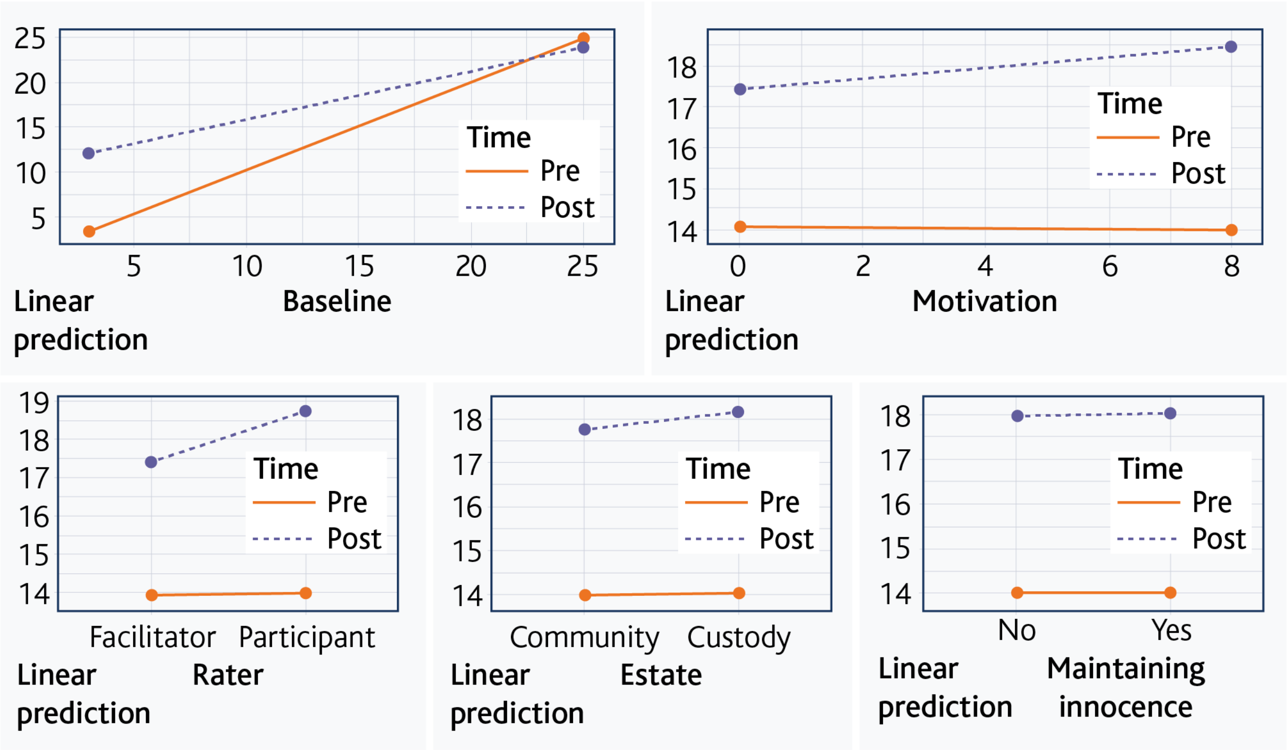
Figure 2: The predicted magnitude of effect sizes for the pre-to-post difference at each pre-Horizon score



The pre-to-post effect was different depending on who provided the rating (X^2 (15) = 119.51, p < 0.0001, η^2_p = 0.06), with the pre-to-post difference being larger for participant-rated strengths than for facilitator-rated strengths. This would be considered a small effect. Additional analyses indicated that facilitator and participant pre-Horizon scores were positively and statistically significantly, but weakly, correlated (see Table A4).

The pre-to-post effect was statistically significantly larger for those with higher pre-Horizon motivation (X^2 (16) = 23.30, p < 0.0001, η^2_p = 0.01). This would be considered a small effect. Finally, the pre-to-post difference was statistically significantly larger for those who participated in Horizon in custody (X^2 (17) = 9.69, p = 0.002, η^2_p < 0.01). This would be considered a very small effect. Whether or not participants were maintaining innocence did not have a statistically significant effect on the pre-to-post difference (X^2 (18) = 0.47, p = 0.493, η^2_p < 0.01). Figure 3 illustrates these interaction effects.

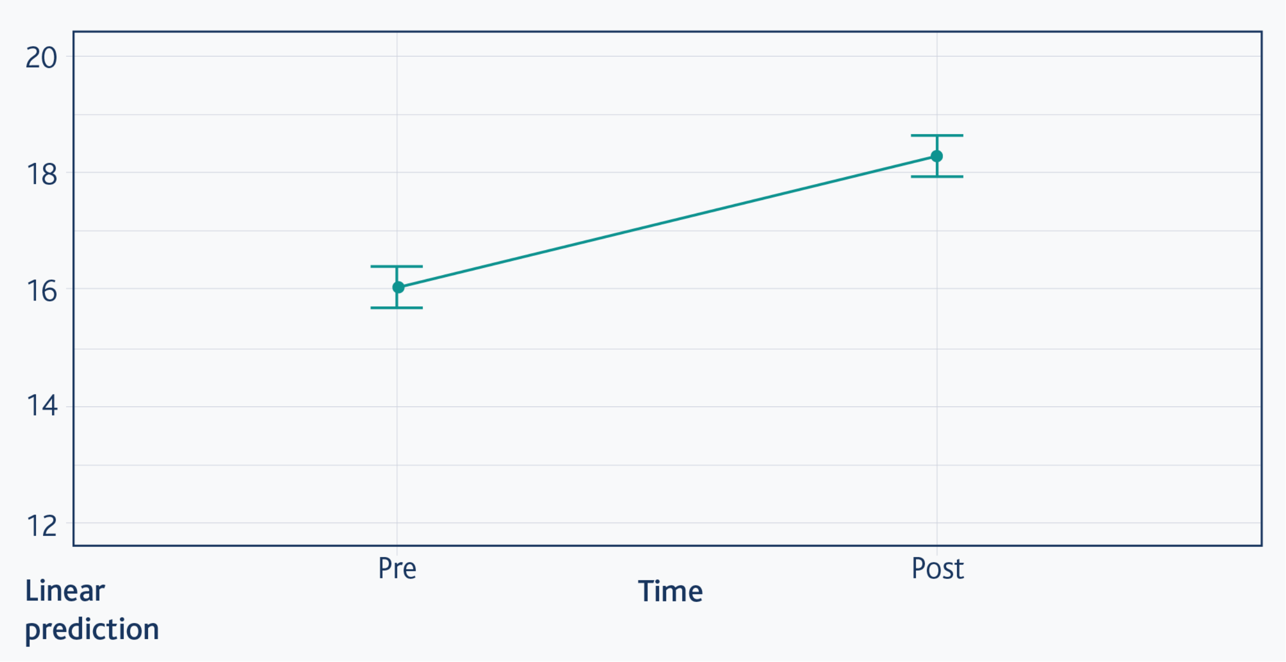
Figure 3: The effects of operational variables on the pre-to-post difference in total Success Wheel Scores observed in the Horizon sample



4.2 iHorizon

For the smaller iHorizon sample, after controlling for baseline, rater, motivation, and maintaining innocence, aggregate post-iHorizon SWM scores were also statistically significantly higher than pre-iHorizon scores (X^2 (12) = 169.20, p < 0.0001, η^2_p = 0.58) (see Figure 4). This would be considered a very large effect and represented a 2.3 increase in overall total SWM points (a 14.1% change in score).

Figure 4: The pre-to-post effect for iHorizon

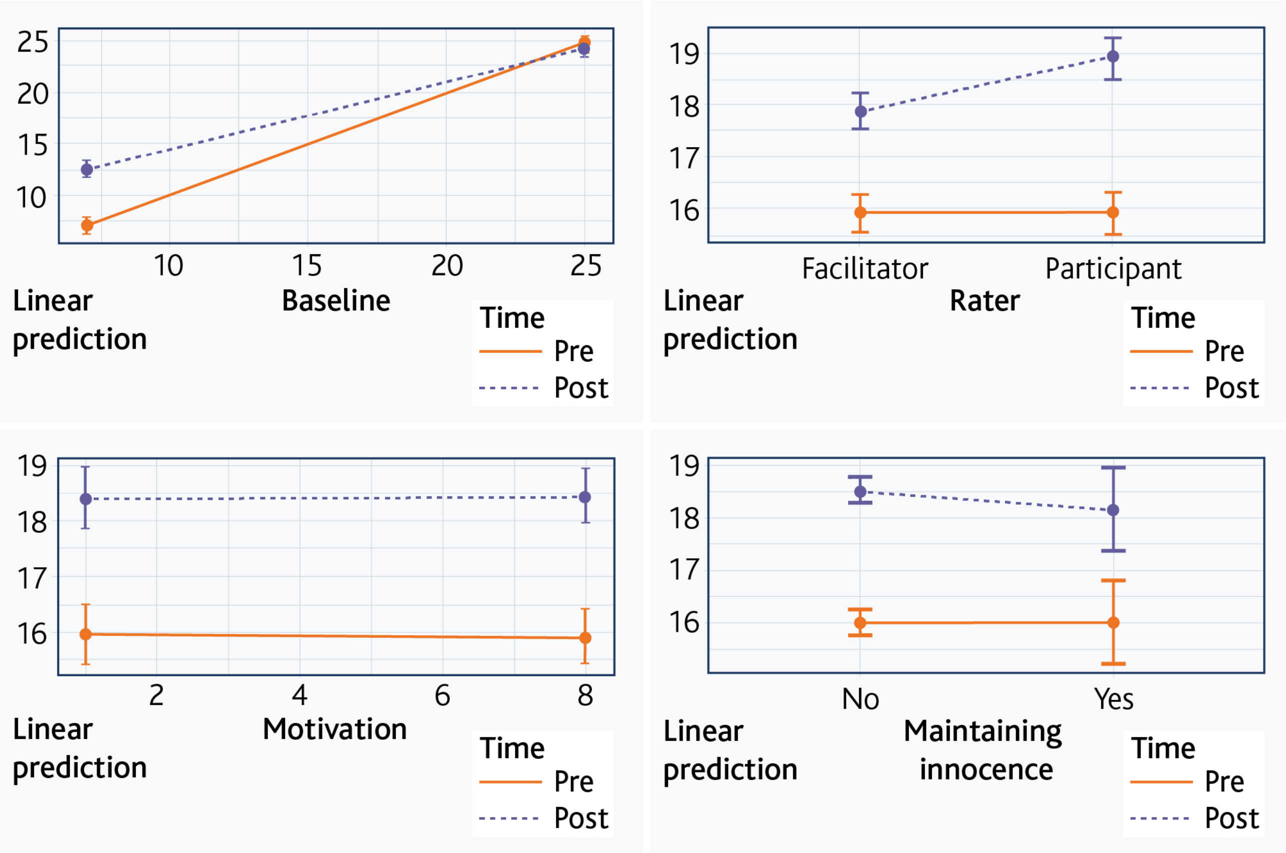


Further analyses indicated that this large pre-to-post effect was statistcally significantly different dending on the participant's pre-iHorizon baseline strengths (X^2 (13) = 51.37, p < 0.0001, η^2_p = 0.24. This would be considered a moderate size effect.

The pre-to-post effect was also different depending on who provided the rating (X^2 (14) = 11.79, p < 0.001, η^2_p = 0.06), with the pre-to-post difference being larger for participant-rated strengths than for facilitator-rated strangths. This would be considered a small effect. Additional analyses indicated that facilitator and participant pre-iHorizon scores were statistically significantly positively correlated, with moderate correlations (Table A4).

Neither pre-Horizon motivation (X^2 (15) = 0.02, p = 0.876, η^2_p < 0.01) or whether or not participants were maintaining innocence (X^2 (16) = 0.44, p = 0.513, η^2_p < 0.01) had a statistically significant effect on the pre-to-post difference. Figure 5 illustrates these interaction effects.

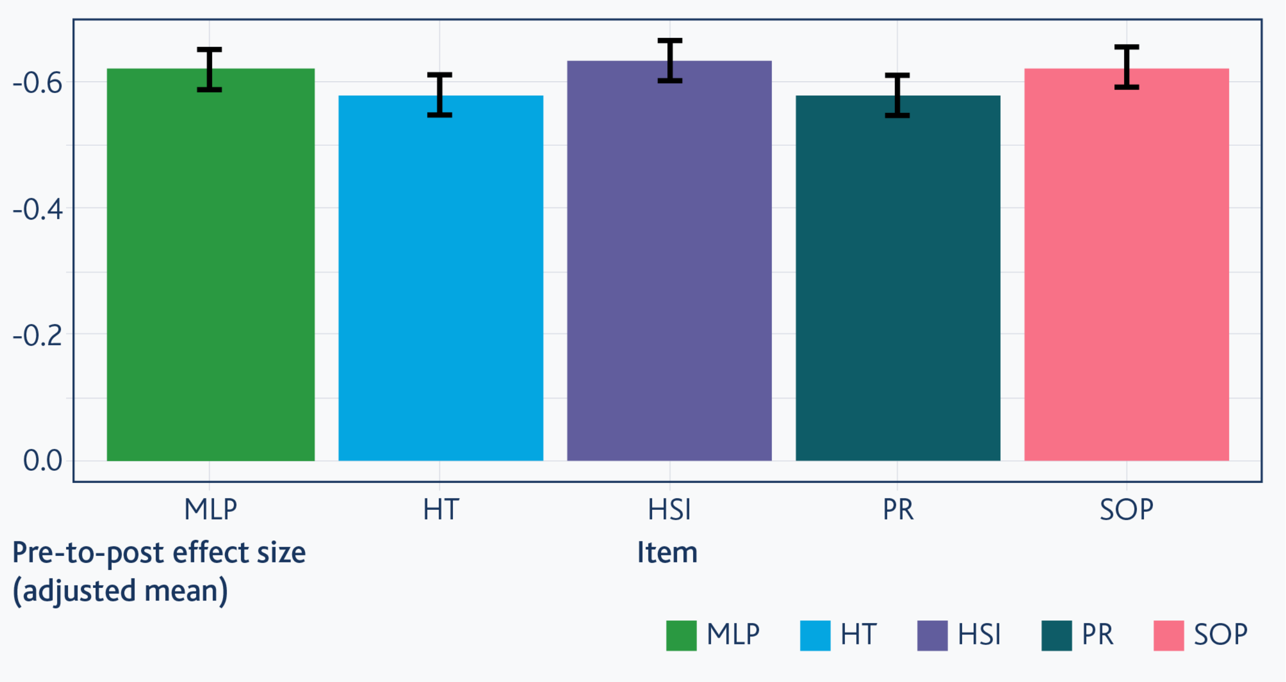
Figure 5: The effects of operational variables on the pre-to-post effect for iHorizon



4.3 Item-level analyses

The item-level analysis for the SWM indicated that, when baseline, motivation, maintaining innocence, and estate were held constant, statistically significant positive pre-to-post effects were found on all Success Wheel Measure items but the magnitude of the pre-to-post effect was different between items. Compared to the “Managing life’s problems” item (the reference category), there were statistically significantly smaller pre-to-post effect on the “Healthy thinking” and “Positive relationships” items were smaller (see Figure 5). However, the effect sizes for the differences between items were very small.

Figure 6: Relative effect sizes for item-level pre-to-post effects



The pre-to-post effect for each SWM item also statistically significantly differed depending on who was providing the rating (participant vs. facilitator: X^2 (22) = 15.55, *p* = 0.004, η^2_p < 0.01). Planned contrasts indicated that there were differences between facilitators and participants on the Managing life's problems item (*b* = 0.07, *t* = 2.03, *p* = 0.042, *r* = 0.02), the Heathy thinking item (*b* = 0.07, *t* = 2.07, *p* = 0.038, *r* = 0.02), and the Sense of purpose item (*b* = 0.09, *t* = 2.88, *p* =.004, *r* = 0.03). However, the effect sizes of both the interaction effect and the contrasts indicated very small differences.

5. Conclusions

We sought to explore whether there had been pre-to-post change in Horizon and iHorizon participants' Success Wheel Measure scores over the time period they attended the programme. This included the combined and independent changes in the 5 domains of the Horizon success wheel: (1) Managing life's problems, (2) Healthy relationships, (3) Healthy sexual interests, and (4) Healthy thinking (representing the domains of the Structured Assessment of Risk and Need (SARN: Mann et al., 2002) and (5) Sense of purpose (representing desistance). Additionally, we sought to explore whether any pre-to-post change we observed was affected by differences between individuals, operational factors, and measurement differences. Finally, we sought to establish the quality of our measures.

Pre-to-post change in SWM scores was observed, with large effect sizes, in both Horizon and iHorizon: on average, post-programme SWM scores were higher than pre-Horizon scores. There was also a large baseline effect on both programmes, whereby the large pre-to-post effect diminishes for those with higher pre-Horizon total SWM scores. The magnitude of pre-to-post improvement rapidly declines for those with pre-Horizon scores greater than 15 out of a total of 25. This is equivalent to scores of 3 or more on all items and 60% of the total possible score. Approximately 45.1% of all Horizon participants received a pre-Horizon facilitator score of greater than 15. We also observe an overall pre-to-post decrease for scores greater than 23. This was expected, as it is more difficult to demonstrate improvement in cases where programme relevant strengths are already high.

Pre-to-post improvement also differed by whose ratings were being observed. Participant SWM scores were, on average, much higher than facilitator scores and the magnitude of pre-to-post change was larger for participants (r = 0.72) than for facilitators (r = 0.54). Since the effects of time and baseline were controlled for in the regression model, this is not simply due to differences between pre-Horizon ratings (i.e., relative pre-Horizon optimism on the part of participants and relative pessimism on the part of facilitators).

There was little-to-no difference in overall SWM scores based on pre-Horizon motivation, maintaining innocence, and delivery context. Although both motivation and delivery context were found to affect pre-to-post differences in total SWM scores in the larger Horizon sample (but not the iHorizon sample), the magnitude of their effects were very small and are unlikely to represent genuine operational differences that require a policy response. A small-but-significant effect of pre-programme motivation on outcomes is to be expected, but these findings do not indicate that a relative lack of motivation is an obstacle to positive change in SWM scores. Similarly, a small-but-significant increase in pre-to-post change for those participating in custody does not represent evidence of a lack of positive change in the community. Alternative explanations might include that, according to its manual, Horizon can be delivered a minimum of twice a week in custody but a minimum of once a week in the community, meaning that the duration between assessments is shorter.

Differences were also observed for the different items on the SWM. Positive pre-to-post change was seen on all items both for facilitators and participants. There are indications that the pre-to-post effect size is slightly smaller for the Healthy thinking and Positive relationships items, but although significant the size of those differences is very small. Also, the findings indicate that although both participants and facilitators observed pre-to-post change, participants self-reported relatively greater change than facilitators on the Managing life’s problems, Healthy thinking, and Sense of purpose items. However, these too were very small statistical differences and do not appear to indicate that facilitators are reporting concerns about participants’ change on those items.

Finally, both the SWM and HMS demonstrated acceptable-to-good levels of reliability and validity. Different items appeared to measure one general factor in independent ways, each measure captured a reasonable range of scores, and both effectively discriminated between participants who are high and low on their respective general factor. The SWM also demonstrated convergent validity with the well-validated SOTIPS measure. Pre-to-post change was also observed on the SWM, with participants demonstrating high levels of enthusiasm and holistic thinking at the pre-programme stage that continued during Horizon, while enthusiasm and commitment showed improvement over time.

These findings provide early, but not definitive, evidence that participation in Horizon and iHorizon is associated with positive change in programme participants. Nevertheless, the lack of a no-treatment control group with whom to compare our samples means that these positive changes cannot be directly attributed to participation on Horizon or iHorizon.

The change that was observed could be attributable to natural individual change over time (e.g., prisoners and probationers learning to live with their circumstances). Test effects may have also contributed, as both facilitators and participants may have been implicitly or explicitly motivated to overstate positive change over the duration of a programme for which they have a stake in its future success. It is also important to emphasise that clinical studies examine “first-order” treatment effects (i.e., are participants gaining skills and insight) and it will also be important for a future robust study to be conducted examining whether any positive change experienced by Horizon and/or iHorizon participants translates into second-order reductions in future reconviction in the community.

There is also an increasing body of evidence to suggest that some amount of change on psychotherapeutic programmes is likely to be due to “extra-therapeutic” variables, such as the quality of the “therapeutic alliance” (the rapport between the facilitator at the participant), and not due to the programme content. In a meta-analytical review of therapeutic alliance effects in the psychotherapeutic literature, Horvath et al. (2011) found a positive relationship between the alliance and outcomes and reported that although therapeutic alliance accounts only for a small proportion of differences in treatment outcomes it is “one of the most robust predictors of treatment success empirical research has been able to document” (p. 15). Other extra-therapeutic variables include delivery conditions and context, facilitators’ therapeutic expertise (and participants’ perceptions of it), and participants’ expectations of success, along with natural self-change, spontaneous improvement, social support, and fortuitous events (Patterson, 2008, Norcross, 2011).

5.1 Clinical implications

The main finding is that positive change is occurring and that should be welcomed. Nevertheless, we list some implications for clinical practice resulting from the findings.

- The baseline effect we observed may be a measurement issue: those with high pre-scores on a measure simply do not have the potential to progress as far on the SWM scale as those with lower pre-programme scores. However, regression to the mean and natural improvement cannot be ruled out. In either case, clinicians may consider how positive progress is communicated to those with higher levels of pre-programme strengths to ensure that they (a) understand why their efforts on Horizon might not be reflected in increases in SWM score and (b) are able to receive positive feedback.

- These baseline findings also highlight the potential benefits of introducing a high-quality needs assessment prior to Horizon and iHorizon, to ensure that those receiving the programme have a need for the programme and can benefit from it. It could be speculated that those assessed as having adequate strengths in areas targeted by the success wheel do not need a strengths-based programme. Conversely, it could equally be speculated that mere rehearsal of skills and insight already possessed may still have a positive impact on pro-social behaviour and the likelihood of reoffending. Although the design of the study does not allow us to conclude that the programmes were the cause of any change in SWM scores, a review of eligibility is recommended.
- Motivation only appeared to have a small effect on the outcomes on the SWM, which provides some early evidence that Horizon and iHorizon may benefit participants even in circumstances where motivation to engage is less (or lower) than ideal. Conversely, maintaining innocence does not appear to affect positive progress. Although we cannot conclude that attendance on the programme is responsible for any changes in SWM scores, these are positive early indications that the inclusion of individuals who are low in motivation to engage with the programme or who are maintaining their innocence does not appear to be a barrier to positive progress during either programme.
- The item-level analyses also indicated that, for Horizon at least, positive progress was observed in all domains of the Success Wheel, albeit with slightly smaller effects for the Heathy thinking and Positive relationships domains. This suggests that the overall positive progress observed at the aggregate level was not predominantly the result of disproportionately large effects in any specific domain or domains or that the overall positive change observed was masking poor outcomes in any specific domains.
- Finally, the psychometric findings provide some early evidence for the quality of the Success Wheel Measure and the Horizon Motivational Scale. They appear to be performing as intended and these findings provide confidence that these

measures, or measures based on the same principles, will have utility for other programmes.

- Finally, from an analytical perspective, these analyses were conducted on 85% of the Horizon sample and 75% of the iHorizon sample. Although these were adequate to achieve our analytical goals, further efforts should be taken, and incentives provided, to ensure that sites are returning complete and accurate data for all participants.

References

Beech, A., Beckett, R. C., & Fisher, D. (1998). STEP 3: An evaluation of the prison sex offender treatment programme. London, U.K.: Home Office.

Beech, A., & Ford, H. (2006). The relationship between risk, deviance, treatment outcome and sexual reconviction in a sample of child sexual abusers completing residential treatment for their offending. Psychology, Crime & Law, 12(6), 685–701.

Brankley, A.E., Helmus, L.M., Hanson, R.K. (2017). STABLE-2007 evaluator workbook: Updated recidivism rates. Unpublished report. Ottawa, ON: Public Safety Canada.

Carter, A., & Mann, R. E. (2011). Organizing principles for an integrated model of change for the treatment of sexual offending. In D. Boer (Ed.), The Wiley handbook on the theories, assessment and treatment of sexual offending (pp. 359–381). London, UK: Wiley-Blackwell.

Clifton, L., & Clifton, D. A. (2019). The correlation between baseline score and post-intervention score, and its implications for statistical analysis. Trials, 20(1), 1–6.

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.

Farmer, M., McAlinden, A. M., & Maruna, S. (2016). Sex offending and situational motivation: Findings from a qualitative analysis of desistance from sexual offending. International journal of offender therapy and comparative criminology, 60(15), 1756–1775.

Hanson, R. K., Harris, A. J. R., Scott, T-L, Helmus, M. L. (2007). Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project. Ottawa, ON: Public Safety Canada.

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.

Harkins, L., Flak, V. E., Beech, A. R., & Woodhams, J. (2012). Evaluation of a community-based sex offender treatment program using a good lives model approach. Sexual Abuse, 24(6), 519–543.

Helmus, L, Hanson, R.K., Babchishin, K.M., Mann, R.E. (2013) Attitudes supportive of sexual offending predict recidivism: a meta-analysis. Trauma Violence and Abuse, 14, 34–53.

Horvath, A. O., Re, A. C. D., Flückiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. In J. C. Norcross (Ed.), Psychotherapy relationships that work: Evidence-based responsiveness (pp. 25–69). Oxford, U.K.: Oxford University Press.

Keeling, J. A., Rose, J. L., & Beech, A. R. (2006). An investigation into the effectiveness of a custody-based cognitive-behavioural treatment for special needs sexual offenders. The Journal of Forensic Psychiatry & Psychology, 17(3), 372–392.

Lee, M. (2016). Matching, Regression Discontinuity, Difference in Differences, and Beyond. London, UK: Oxford University Press.

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.

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

Mann, R. E., O’Brien, M., Thornton, D., Rallings, M., & Webster, S. (2002). Structured assessment of risk and need. London, U.K.: HM Prison and Probation Service.

Marsden, E., & Torgerson, C. J. (2012). Single group, pre-post test research designs: Some methodological concerns. Oxford Review of Education, 38(5), 583–616.

Marshall, W. L., & Marshall, L. E. (2012). Treatment of sexual offenders: Effective elements and appropriate outcome evaluations. In E. Bowen, & S. Brown (Eds.), Perspectives on evaluating criminal justice and corrections. Advances in programme evaluation (pp. 71–94) (Vol. 13). Bingley, U.K.: Emerald Group.

McAlinden, A. M., Farmer, M., & Maruna, S. (2017). Desistance from sexual offending: Do the mainstream theories apply? Criminology & Criminal Justice, 17(3), 266–283.

McGrath, R. J., Cumming, G. F., & Lasher, M. P. (2013). Sex Offender Treatment Intervention and Progress Scale (SOTIPS). https://nicic.gov/sotips-sex-offender-treatment-intervention-and-progress-scale.

McGrath, R. E., & Meyer, G. J. (2006). When effects sizes disagree: The case of *r* and *d*. Psychological Methods, 22(4), 386–401.

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

Miles, J. & Shevlin, M. (2001). Applying Regression and Correlation: A Guide for Students and Researchers. London, U.K.: Sage.

Norcross, J. C., & Wampold, B. E. (2011). Evidence-based therapy relationships: research conclusions and clinical practices. Psychotherapy, 48(1), 98.

Olver, M. E., & Stockdale, K. C. (2020). Evaluating change in men who have sexually offended: Linkages to risk assessment and management. Current Psychiatry Reports, 22, 1–10.

Patterson, C. L., Uhlin, B., & Anderson, T. (2008). Clients' pretreatment counseling expectations as predictors of the working alliance. Journal of Counseling Psychology, 55(4), 528.

Pepinsky, T. B. (2018). A note on listwise deletion versus multiple imputation. Political Analysis, 26(4), 480–488.

Seto, M. C., Marques, J. K., Harris, G. T., Chaffin, M., Lalumière, M. L., Miner, M. H., ... & Quinsey, V. L. (2008). Good science and progress in sex offender treatment are intertwined: A response to Marshall and Marshall (2007). Sexual Abuse, 20(3), 247–255.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Thornton, D. (2013). Implications of our developing understanding of risk and protective factors in the treatment of adult male sexual offenders. International Journal of Behavioral Consultation and Therapy, 8(3–4), 62–65.

Torgerson, D. J., & Torgerson., C. J. (2008). Designing randomised trials in health, education, and the social sciences: An introduction. Basingstoke, U.K.: Palgrave Macmillan.

Wakeling, H., Beech, A. R., & Freemantle, N. (2013). Investigating treatment change and its relationship to recidivism in a sample of 3773 sex offenders in the UK. Psychology, Crime & Law, 19(3), 233–252.

Wakeling, H. C., & Barnett, G. D. (2014). The relationship between psychometric test scores and reconviction in sexual offenders undertaking treatment. Aggression and Violent Behavior, 19(2), 138–145.

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.

Wilkinson, K., & Powis, B. (2019). A process study of the horizon programme. London, U.K.: Ministry of Justice.

Wong, S. C. P., Gordon, A., & Gu, D. (2007). Assessment and treatment of violence-prone forensic clients: An integrated approach. British Journal of Psychiatry, 190 (Suppl.), s66–s74.

Appendix A
Findings and statistics

Reference categories

In each of the multilevel models, not every combination of comparisons between the levels of the category are tested. For categorical variables, planned contrasts in multilevel models use “reference categories” against which other levels of the predictor are compared. This becomes important when interpreting the interactions between predictors in models. The reference category for each predictor are listed in Table A1 below.

Table A1: Reference categories for planned contrasts

Variable	Reference category	Planned contrasts	Analyses to be included
Time	Pre	1. Post vs. Pre	Horizon & iHorizon
Rater	Facilitator	1. Participant vs. Facilitator	Horizon & iHorizon
Maintaining innocence	No	1. Yes vs. No	Horizon & iHorizon
Delivery context	Custody	1. Community vs. Custody	Horizon only
Item	MLP	1. HT vs. MLP 2. HSI vs. MLP 3. PR vs. MLP 4. SOP vs. MLP	Horizon & iHorizon

Horizon analyses

Table A2 presents the statistics for the goodness-of-fit tests between the various models being compared in the main Horizon analysis.

Table A2: ANOVA of model fit for predictors and interactions for Horizon models

Model	*df*	AIC	BIC	logLik	X^2	*p*	η^2_p
Null model	3	19094.6	19113.1	-9544.3			
Independent effects							
Baseline	5	16297.8	16328.7	-8143.9	2800.77	<.0001	0.93
Rater	7	16241.8	16285.1	-8113.9	59.95	<.0001	0.49
Motivation	9	16234.3	16289.8	-8108.1	11.57	0.003	0.03

Model	***df***	**AIC**	**BIC**	**logLik**	***X^2***	***p***	***η^2_p***
Delivery context	10	16231.5	16293.2	-8105.7	4.80	0.028	<0.01
Maintaining innocence	11	16233.2	16301.1	-8105.6	0.23	0.629	<0.01
Time	13	14539.9	14620.2	-7257.0	1697.31	<.0001	0.43
Interaction effects (with time)							
Baseline x Time	14	13900.7	13987.1	-6936.4	641.22	<.0001	0.29
Rater x Time	15	13783.2	13875.8	-6876.6	119.51	<.0001	0.06
Motivation x Time	16	13761.9	13860.7	-6864.9	23.30	<.0001	0.01
Delivery context x Time	17	13754.2	13859.1	-6860.1	9.69	0.002	<0.01
Maintaining innocence x Time	18	13755.7	13866.8	-6859.9	0.47	0.493	<0.01

Note: η^2_p = partial eta squared (approximate effect size). logLik = log-likelihood.

Table A3 presents the fixed effects for the most parsimonious model (the most complex, statistically significant model): the interaction between delivery context and time.

Table A3: Fixed effects and effect sizes for the Horizon Delivery context x Time interaction

Contrast	***b***	**SE**	***df***	***t***	***p***	***r***
(Intercept)	0.00	0.22	1767	-0.01	0.989	
Baseline	1.00	0.01	758	81.46	<.0001	0.90
Rater	0.00	0.09	126	0.02	0.982	0.00
Motivation	0.00	0.02	882	0.05	0.957	0.00
Delivery location	-0.01	0.08	882	-0.07	0.943	0.00
Maintaining innocence	0.05	0.07	882	0.68	0.494	0.00
Time	9.65	0.32	1767	30.51	<.0001	0.35
Baseline x Time	-0.50	0.02	1767	-28.94	<.0001	0.32
Rater x Time	1.34	0.12	1767	10.96	<.0001	0.06
Motivation x Time	0.12	0.03	1767	3.90	<.0001	0.01
Delivery context x Time	0.36	0.12	1767	3.11	0.002	0.01

Table A4 presents the findings from a series of Pearson’s correlations of the associations between facilitator and participant pre-programme SWM scores.

Table A4: Pearsons correlation between facilitator and participant pre-Horizon scores for each SWM item

SWM item	Correlation coefficient	*df*	*t*	*p*
Horizon				
Managing life's problems	0.41	977	14.14	<.0001
Healthy thinking	0.36	970	12.09	<.0001
Healthy sexual interests	0.38	971	12.69	<.0001
Positive relationships	0.39	978	13.14	<.0001
Sense of purpose	0.47	977	16.47	<.0001
iHorizon				
Managing life's problems	0.68	108	9.75	<.0001
Healthy thinking	0.52	108	6.41	<.0001
Healthy sexual interests	0.62	108	8.28	<.0001
Positive relationships	0.51	108	6.12	<.0001
Sense of purpose	0.57	108	7.18	<.0001

iHorizon analyses

Table A5 presents the statistics for the goodness-of-fit tests between the various models being compared in the main iHorizon analysis. Delivery context is not included as iHorizon is only delivered in the community.

Table A5: ANOVA of model fit for predictors and interactions for iHorizon models

Model	*df*	AIC	BIC	logLik	X^2	*p*	η^2_p
Null model	3	1838.2	1849.9	-916.1			
Independent effects							
Baseline	5	1516.7	1536.2	-753.3	325.54	<.0001	0.95
Rater	7	1514.3	1541.7	-750.2	6.31	0.043	0.39
Motivation	9	1518.3	1553.5	-750.2	0.01	0.993	<0.01
Maintaining innocence	10	1520.1	1559.2	-750.0	0.23	0.634	<0.01
Time	12	1354.9	1401.8	-665.4	169.20	<.0001	0.58
Interaction effects (with time)							
Baseline x Time	13	1305.5	1356.3	-639.8	51.37	<.0001	0.24
Rater x Time	14	1295.7	1350.5	-633.9	11.79	0.001	0.06

Model	*df*	AIC	BIC	logLik	X^2	*p*	η^2_p
Motivation x Time	15	1297.7	1356.3	-633.9	0.02	0.876	<0.01
Maintaining innocence x Time	16	1299.3	1361.8	-633.6	0.43	0.513	<0.01

Note: η^2_p = partial eta squared (approximate effect size). logLik = log-likelihood.

Table A6 presents the fixed effects for the most parsimonious model (the most complex, statistically significant model): the interaction between rater and time.

Table A6: Fixed effects and effect sizes for the iHorizon Delivery context x Time interaction

Contrast	*b*	SE	*df*	*t*	*p*	*r*
(Intercept)	-0.02	0.58	181	-0.04	0.969	
Baseline	1.00	0.03	72	29.71	<.0001	0.92
Rater	0.00	0.21	18	-0.02	0.983	0.00
Motivation	0.00	0.04	89	-0.02	0.985	0.00
Maintaining innocence	-0.18	0.28	89	-0.65	0.520	0.00
Time	8.14	0.77	181	10.54	<.0001	0.38
Baseline x Time	-0.38	0.05	181	-8.12	<.0001	0.27
Rater x Time	1.02	0.30	181	3.42	0.001	0.06

Item-level analyses

Table A7 presents the statistics for the goodness-of-fit tests between the various models being compared in the Horizon item-level analysis.

Table A7: ANOVA of model fit for item-level predictors and interactions for Horizon

	df	AIC	BIC	logLik	X^2	p	η^2_p
Null model	13	24020.5	24121.5	-11997.2			
Item	18	23984.7	24124.6	-11974.3	45.80	<.0001	<0.01
Time x Item	22	23977.1	24148.1	-11966.6	15.55	0.004	<0.01
Time x Rater x Item	31	23975.5	24216.5	-11956.8	19.59	0.021	<0.01

Note: η^2_p = partial eta squared (approximate effect size). logLik = log-likelihood. The baseline model controlled for the independent effects of time, pre-Horizon baseline strengths, pre-Horizon motivation, rater, delivery location and maintaining innocence.

Table A8 presents the fixed effects for the most parsimonious model (the most complex, statistically significant model): the interaction between time, rater, and item.

Table A8: Fixed effects and effect sizes for the Time x Rater x Item interaction model

	b	SE	*df*	*t*	*p*	*r*
(Intercept)	0.90	0.03	9353	27.96	<.0001	
Baseline	0.67	0.01	1780	108.49	<.0001	0.87
Rater	0.01	0.00	873	3.59	<.0001	0.01
Motivation	0.13	0.02	1426	5.82	<.0001	0.02
Delivery location	0.03	0.01	873	1.82	0.069	0.00
Maintaining innocence	0.03	0.02	873	1.54	0.123	0.00
Time	0.59	0.02	4084	25.62	<.0001	0.14
HT (vs. MLP)	0.00	0.02	9353	0.25	0.802	0.00
HSI (vs. MLP)	-0.01	0.02	9353	-0.36	0.719	0.00
PR (vs. MLP)	-0.02	0.02	9353	-0.83	0.405	0.00
SOP (vs. MLP)	0.04	0.02	9353	2.17	0.030	0.00
Time x HT (vs. MLP)	-0.04	0.03	9353	-1.55	0.122	0.00
Time x HSI (vs. MLP)	0.03	0.03	9353	1.24	0.217	0.00
Time x PR (vs. MLP)	-0.02	0.03	9353	-0.76	0.450	0.00
Time x SOP (vs. MLP)	-0.01	0.03	9353	-0.37	0.713	0.00
Rater x HT (vs. MLP)	0.03	0.03	9353	1.10	0.270	0.00
Rater x HSI (vs. MLP)	0.06	0.03	9353	2.12	0.034	0.00
Rater x PR (vs. MLP)	0.03	0.03	9353	1.01	0.314	0.00
Rater x SOP (vs. MLP)	-0.02	0.03	9353	-0.83	0.405	0.00
Rater x Time x MLP	0.07	0.03	9353	2.03	0.042	0.00
Rater x Time x HT	0.07	0.03	9353	2.07	0.038	0.00
Rater x Time x HSI	0.03	0.03	9353	0.86	0.388	0.00
Rater x Time x PR	0.02	0.03	9353	0.75	0.455	0.00
Rater x Time x SOP	0.09	0.03	9353	2.88	0.004	0.00