HM Prison &
Probation Service

The Healthy Sex Programme

An exploration of pre-to-post psychological test change

Amy Freel & Helen Wakeling
HM Prison and Probation Service

The author

Amy Freel is a registered forensic psychologist currently working in the HMPPS Risk and Capabilities Unit within the Directorate of Security. Dr Helen Wakeling is a research psychologist working in the Evidence-Based Practice Team, a team in HMPPS headquarters that helps to embed evidence-based practice across the Prison, Probation and Youth Custody Service.

Contents

List of tables

1.   Summary

Introduction and aims

The Healthy Sex Programme (HSP) is a His Majesty's Prison and Probation Service (HMPPS) accredited offending behaviour programme designed to help men with convictions for sexual offences effectively manage and reduce their sexual offending. HSP is delivered on an individual basis within a prison setting and aims to help participants understand their sexual interests and arousal patterns. It helps individuals explore triggers for their offence related thoughts, and develop a range of skills to manage their lives and ultimately to desist from sexual offending behaviour. This study aimed to explore short-term pre and post programme participation clinical change for HSP treatment targets using a series of psychological test measures.

Methodological approach and interpreting findings

The sample included 95 adult males who had completed the HSP between 2013 and 2016 and were predominantly serving indeterminate custodial sentences. To measure short-term change against HSP treatment targets, a series of psychological tests were administered to the HSP sample before and after programme completion. Pre to post change analyses were conducted exploring 13 individual psychological measures which were also grouped into four overall dynamic risk domains (i.e., Healthy Sexual Interests, Healthy Thinking, Positive Relationships and Managing Life's Problems) in line with the programme needs assessment used on the intervention. Analyses were also explored by actuarial risk categories defined by a risk assessment tool (Risk Matrix 2000/s), which estimates the likelihood of sexual reoffending.

There were limitations to this research and the pre-post methodology employed. In absolute terms 95 HSP participants is a relatively small sample size but for a study on this particular group it is not small and importantly it is estimated to be half of the population participating in HSP during the study time period. However, the small sample size specifically may leave the results vulnerable to inflated effect sizes and difficulties in replicability. The study design did not include a control group. This makes it difficult to rule out the possibility that other factors (aside from the intervention) could be contributing to any pre to post changes which are observed.

Additionally, the research relied on the use of self-report psychological tests, which are prone to socially desirable responding.

Results

The results found pre-to-post programme statistically significant positive change on nine of the 13 individual psychological test scales, providing support for HSP. The remaining four psychological test scales did not show statistically significant change pre to post programme, though average change was in the desired direction. Pre-to-post statistically significant change, indicating positive change on treatment targets addressed on HSP, was also found on three of the four overall risk domains: namely, Healthy Sexual Interests, Positive Relationships and Managing Life's Problems. Change on the individual psychological test scales and domains was not related to static risk of proven sexual reoffending (defined using the Risk Matrix 2000/s).

When using clinically significant change methodology, a large proportion of the sample scored within a range deemed to be functional post-programme. However, it should be noted that the majority of pre-HSP scores were already within this functional response range. This may reflect the limitations of using psychological tests. Alternatively, it is possible that this reflects the fact that HSP is a secondary programme and that participants could have made this progress from participation on a previous programme aimed at addressing sexual reoffending. Positively, none of the samples' scores showed deterioration in three of the four domains (only two people were deemed to have scored significantly worse post programme in the Healthy Thinking domain).

Conclusions

This study suggests that there are some positive and statistically significant trends emerging from the psychological test data for HSP. Overall, the findings suggest that participation on HSP is associated with a positive effect on participant progress against key treatment targets. However, the findings from this study are not conclusive, and repeating the study using a control group would be helpful. Further analyses of other outcomes, including service user experience of the HSP and the long-term effects of HSP on sexual reoffending should also be examined.

2. Introduction

2.1 The Healthy Sex Programme

Since 2013, HM Prison and Probation Service (HMPPS) has delivered an accredited[1] cognitive-behavioural treatment (CBT) programme[2] called the Healthy Sex Programme (HSP). The programme aims to increase sexual self-regulation skills in order for people to live safer lives. This is done by targeting a number of known criminogenic needs or risk factors that are associated with sexual offending. The HSP is based on a biopsychosocial model of change (Mann & Carter, 2012; Walton, Ramsay, Cunningham & Henfrey, 2017), which outlines a biopsychosocial conceptualisation of criminogenic needs, and a set of organizing principles for programme design. The model builds upon the Good Lives Model (GLM) (Ward & Laws, 2010) and the risk, need and responsivity (RNR) framework (Andrews & Bonta, 2006). Participants who access the programme are either currently experiencing symptoms of a forensically relevant[3] paraphilia or acknowledge a forensically relevant paraphilia[4] that is in remission, and are assessed as medium static risk or above using the Risk Matrix 2000/s (RM2000/s), which is a statistically derived prediction scale for sexual offending (Thornton, Mann, Webster, Blud, Travers, Friendship & Erikson, 2003). At the time of the data collection, low risk men could access HSP in exceptional circumstances (i.e., when the severity of paraphilic interest deemed the programme would be helpful). It is expected that the majority of those accessing HSP will have engaged previously with a group-based accredited programme for people convicted of sexual offences. At the time of this data collection

---

[1] Accreditation is achieved via an independent external committee of international experts called the Correctional Services Accreditation and Advisory Panel (CSAAP). CSAAP review the programme and attest to the fact that it is designed in line with the latest evidence and thinking about what works to reduce re-offending.

[2] Cognitive behavioural treatment aims to help people to manage their problems by encouraging them to recognise how thoughts and cognitions can affect feelings and behaviour.

[3] 'Forensically relevant' means that if that sexual interest were acted upon, it's more likely to be achieved via offending than non-offending means.

[4] The term 'paraphilia' refers to "an intense and persistent sexual interest other than a sexual interest in genital stimulation or preparatory fondling with phenotypically normal, physically mature, consenting human partners" (American Psychiatric Association, 2013, p.685). "Phenotypically normal" in this definition refers to the 'normal' characteristics of adults. Paraphilia are diverse, representing an intense and persistent arousal pattern in distinct stimuli, sets of stimuli or specific acts.

participants likely would have taken part in a Sex Offender Treatment Programme (SOTP), which was replaced by a new suite of programmes in (2017).

The HSP is delivered by registered psychologists and psychologists in training. The programme aims to help participants gain insight into the patterns of their healthy and unhealthy sexual functioning, and practice skills to support management of paraphilic interests, for example by using behavioural therapy, cognitive reappraisal, urge management and mindfulness strategies. Individuals develop a collaborative bespoke HSP plan, which may include a combination of the following topics: exploring healthy sex and misconceptions about sex, sexual roles in relationships, identity and sexuality, the safe use of pornography, skills for intimacy, managing shame and maintaining self-efficacy. The programme lasts between 12 and 30 sessions. Sessions are delivered at a rate of one or two per week, lasting between 30 and 90 minutes. The original programme (HSPv1) was initially accredited by the Correctional Services Advice and Accreditation Panel (CSAAP) from 2013 to 2019. Implementation of the programme was closely monitored during this time and changes were made prior to the accreditation renewal. The evolved version (HSPv2) which was accredited in 2019 (until 2024), took into account theoretical advances and information gained from the implementation of HSPv1. Both versions of HSP are individual programmes for men convicted of sexual offences who have specific treatment needs in areas related to sexual interest. The current research is based on a sample of participants who completed HSPv1.

The effectiveness of rehabilitative programmes for men convicted of sexual offences has not been demonstrated consistently. Some researchers have concluded that the evidence is insufficient to indicate the programmes evaluated are effective (e.g., Dennis, Khan, Ferriter, Huband, Powney & Duggan, 2012; Mews, Di Bella, & Purver, 2017). More recently others have concluded that although firm conclusions await better quality evaluation, the available international evidence indicates that cognitive-behavioural treatment (CBT) programmes can reduce general and sexual recidivism (e.g., Mpofu, Athanasou, Rafe & Belshaw 2016; Schmucker & Losel, 2017; Gannon, Olver, Mallion, & James, 2019). The inconsistency in the findings may depend on the quality of the studies, the quality of the implementation of the interventions as well as

the design of the studies and the criteria for study inclusion. We still need more high quality and robust research to draw firm conclusions.

However, there is good evidence that the principles of RNR have been shown to apply to programmes for sexual offenders (Hanson, Bourgon, Helmus & Hodgson, 2009), and greater effects are found for higher risk men (Schmucker & Losel, 2017). Despite this, owing to the heterogeneous use of the term "CBT", it is not clear which elements are effective (Schmucker & Losel, 2017). CBT-based programmes designed to address paraphilic disorders are especially difficult to evaluate. This is because throughput is low and the therapeutic approach used is holistic, blending both classic behavioural therapy (BT) and cognitive–behavioural methods (McGrath, Cummings, Burchard, Zeoli & Ellerby, 2010), making it difficult to establish which modality of therapy is effective and to what extent.

This study intended to advance knowledge on the effectiveness of HSP. Evaluation of HSP is particularly difficult due to the low yearly throughput of the programme, relatively low proven reoffending rates and the difficulty in identifying a comparison group. However, the completion of a series of studies, each testing different indicators of programme effect, can help to build a picture around effectiveness (see Collaborative Data Outcome Committee, 2007). The present study aimed to look at the short-term change on criminogenic needs measured psychometrically pre-to-post programme (e.g., Wakeling, Freemantle, Beech & Elliott, 2011; Wakeling, Beech & Freemantle, 2013).

2.2   Measuring clinical change

Using psychological tests as an indicator of programme effectiveness has been shown in some research to be associated with reductions in sexual recidivism (e.g., Beech & Ford, 2006; Marques, Wiederanders, Day, Nelson & Van Ommeren, 2005). Furthermore, higher risk individuals can show worse psychological test scores than lower risk individuals, both before and after a programme (e.g., Williams, Wakeling & Webster, 2007), suggesting that psychological tests can be useful in identifying those at greater risk of reoffending. However, these findings are inconsistent. For example, Wakeling and colleagues using the HMPPS Sex Offender Treatment Programme

(SOTP) psychological test battery, found that clinically significant change on scales assessing the criminogenic targets of SOTP added only limited value to actuarial risk prediction, and was not associated with recidivism, with the exception of a few scales measuring relationship functioning (Barnett, Wakeling, Mandeville-Norden & Rakestrow, 2012; 2013; Wakeling et al., 2011; Wakeling et al., 2013). Of interest however was the finding that combined psychological test scores reflecting overall risk domains were more useful in predicting recidivism than observing individual psychological test scales.

There are the well-documented limitations of pre-to-post change methodology using psychological tests, including the fact that such measures may be prone to socially desirable responding, and that they may not relate to longer term outcomes (e.g., reoffending). These limitations are important to consider in understanding these inconsistent findings (e.g., Wakeling & Barnett, 2014), as well as the relatively short follow up periods used by Wakeling and colleagues which, due to the low recidivism base rates of sexual reconviction, may have made statistically significant differences difficult to detect. Despite these limitations, the findings led to the development of a refined psychological test battery, consisting of four short scales measuring four dynamic risk domains consistent with Thornton (2002; 2013); namely, Healthy Sexual Interests, Healthy Thinking, Positive Relationships and Managing Life's Problems (previously defined by Thornton (2002) as: Sexual interests, distorted attitudes, socio-affective functioning and self-management, respectively).

The scales were previously tested with reconviction data to determine their usefulness (Wakeling, 2014, p. 116). In summary, this research produced similar results to the abovementioned outcomes using the full SOTP test battery; that is, psychological test scores marginally enhanced actuarial risk prediction, and the Positive Relationships domain was a better predictor of recidivism overall. In addition, pre-programme domain scores significantly predicted reconviction outcome.

2.3   Aims and objectives of this study

The refined psychological test battery (Wakeling, 2014) was implemented for HSP by HMPPS between 2013 and 2015. The present study aimed to explore short-term pre and post programme participation clinical change for HSP treatment targets using this test battery of psychological tests. Based on the abovementioned findings that higher risk individuals can show worse psychological test scores than lower risk individuals, both before and after a programme, and that greater programme effects are found for higher risk men, the following hypotheses were developed:

1.   (a) there will be a statistically significant difference between pre and post programme psychological test scores for individual measures and domain outcomes, and (b) change will be clinically significant and reliable for domain outcomes;

2.   there will be a statistically significant difference between pre-psychological test scores by level of risk and post-psychological test scores by level of risk. Scores for higher risk men will show a greater level of treatment need than lower risk men; and

3.   there will be a statistically significant interaction between pre-to-post psychological test scores and level of risk of reoffending, in that higher risk men will show a greater level of change than lower risk men.

3. Methodology

3.1 Sample

The sample for this study consisted of 95 adult males with convictions for sexual offences. All participants had completed HSP in HMPS between 2013 and 2016 across 13 prison sites. Dropouts were not included in the sample as only data for those who completed both pre and post programme psychological test data was collected.[5] Participants' mean age was 53 (range 26 -75, SD=12.11). The majority of the sample described themselves as White British (90%) with the largest proportion serving indeterminate sentences (81%). The largest proportion of men were categorised as being very high risk (33%) with the smallest proportion being in the low risk category (11%). Table A1 shows the full demographic characteristics.

3.2 Measures

Risk Matrix 2000/s

The RM2000/s (Thornton et al., 2003) was completed for all men in custody who had convictions for sexual offences, prior to engaging in any accredited programmes. This is an actuarial risk assessment, used with adult males convicted of a sexual offence (see Appendix 2 for further details).

Psychological tests

The refined battery of psychological tests (see Wakeling, 2014, p. 116, for early development of these measures, and Williams, 2007) was administered to each participant by trained staff before and after the HSP. The psychological scales included are described in full in Appendix 2, and table B2 outlines the internal consistency scores (Cronbach's alpha) for each scale. A summary is provided below:

Healthy Sexual Interests

My Private Interests Measure is a 54-item scale measuring sexual paraphilia, originally developed for participants attending adapted SOTPs (Williams, 2007) and was subsequently used more widely for all SOTP participants. It has four subscales:

---

[5] Whilst exact drop-out rates at the time of this research are not known, average drop-out rates from one year prior to the period of data collection, were around 5% for HSP.

sexual preference for children, sexual preoccupation, sexualised violence, and other offence-related sexual interests. Higher scores equate to greater problems in these areas. The measure has excellent internal reliability (α = .92-.94), has been shown to discriminate higher from lower risk individuals, and has shown good convergent validity with other measures of sexual interests (Farren & Barnett, 2014).

Healthy Thinking

Booklet 2 (Interventions Services, 2013) is a 39-item measure containing 2 scales, the Attitudes towards Women Scale, and the Attitudes towards Children Scale, with good and excellent internal consistency respectively (α = .83, α = .95; Wakeling, 2014). Higher scores on both subscales indicate greater problems in these areas. Wakeling (2014) also found that overall pre-treatment scores on these measures combined differed between recidivists and non-recidivists in a large sample of individuals who had undergone an intervention for men convicted of sexual offences in prison.

Positive Relationships

The Relationships Scale (Interventions Services, 2013) is an 81-item measure consisting of 3 subscales. Subscale 1: Personality 1 Questionnaire (relating to self-esteem and ruminations), Subscale 2: Personality 2 Questionnaire (relating to perspective taking and locus of control), and Subscale 3: Emotional Loneliness (relating to beliefs about meaningful relationships). On each of the subscales, higher scores equate to greater problems in these areas. Wakeling (2014) found the internal consistency of this overall measure to be good (α = .91).

Managing Life's Problems

Self-Management Scale (Interventions Services, 2013) is an 18-item measure of impulsivity, which has been shown to have good internal consistency (α = .91; Wakeling, 2014). Again higher scores on this measure equate to greater levels of problems in this area.

3.3   Procedures

Analysis 1: Statistical Differences

Raw data was collated for all individual's pre- and post-HSP. Overall domain scores were then summed using the individual scale scores. Examination of differences in pre- and post-programme scores by risk level were examined by a series of ANOVAs[6] to determine whether scores were related to static risk level. Pre-to-post change on the individual scale scores and the overall domain scores were then examined using further ANOVAs.

Analysis 2: Clinical significance (CS) and reliable change index (RCI)

The Jacobson and Truax (1991) method for calculation of CS was employed for domain change only (see Mandeville-Norden, Beech & Hayes, 2008). This method allowed an examination of the extent to which individuals were in the functional range[7] on each domain after the HSP (clinical significance), and whether any pre-to-post change was statistically reliable. To determine whether individuals were in a functional range after the HSP, post-programme psychological test scores were analysed. If a score fell one standard deviation (SD) or more away from the pre-programme mean (in the direction of functional responding) for that population, it was regarded as clinically significant or "functional" on that measure (Wampold, 2001).

A Reliable Change Index (RCI) was also calculated for each individual on domain score outcomes. In the absence of a control group, as suggested by Jacobson and Truax (1991), RCI was calculated using the SD of the pre-programme sample. The RCI was calculated as follows:

$$RCI = \frac{x1 - x2}{SE}$$

---

[6]   ANOVAs or analysis of variance are used to analyse the differences among means.

[7]   Within CS methodology, 'functional' range refers to the scores being one standard deviation or more away from the mean in the direction of normative functioning.

In this equation, $x1$ is the pre-programme score and $x2$ is the post-programme score. SE is equal to:

$$\frac{s^1}{\sqrt{1 - r^{xx}}}$$

In this equation, s^1 is the standard deviation of the sample's pre-programme score; and r^{xx} is the internal consistency (coefficient alpha) statistic of the measure (which is used as a substitution of the test-retest reliability of the measure as the latter was not available for the measures used in this research; see Jacobson & Truax, 1991). Jacobson and Truax (1991) suggest that an RCI of 1.96 or above reflects real change (with a 95% confidence level of $p < 0.05$).

Individuals were then categorised into one of five programme change categories suggested by Jacobson, Roberts, Berns, and McGlinchey (1999) by combining the CS and RCI. This was done for each domain and the categories were:

- **Deteriorated**: individuals who demonstrate reliable change but in the undesired direction;
- **Unchanged**: individuals who demonstrate no reliable change;
- **Improved**: individuals who demonstrate reliable change but are not within the range of normal functioning post-programme;
- **Recovered**: individuals who demonstrate reliable change and clinical significance;
- **Already okay**: individuals who were in the desirable range both pre- and post-programme.

3.4   Limitations and interpreting findings

First, caution must be applied to the findings due to the methodological approach. The sample for this study was relatively small. The small sample size is primarily due to HSP being resource intensive, resulting in a limited annual throughput. But at the time we also relied on psychological test data being sent in centrally to HMPPS. The latest published figures show that for dates which broadly align with the study period around 200 people completed the programme (HMPPS, 2021). As such the present sample, although small, represents around half of all those undertaking this

programme within the years of data collection for this study. The small sample had an effect on the power of the analysis and as such, also the ability to generalise results. This was particularly the case when the sample was broken down into risk groups for analysis. This may mean that it is difficult to detect small effects, which are statistically significant. But it is also possible that effect sizes detected using underpowered analyses may over-estimate the true size of the effect (Button et al., 2013). Another important limitation is the lack of a control group. This reduces confidence in the ability to attribute pre-to-post change to participation in the programme.

Second, this study examined participation on HSPv1 and not HSPv2, which is the current version of the programme being delivered. As such, replication of this study using the current measures used on HSPv2 would be useful to help generalise the applicability of these findings to the most recent version of the programme. Furthermore, the results should not be confused with the effect of HSP on recidivism. A proportion of the sample were categorised as 'recovered' according to their pre-to-post programme scores. Although some studies indicate that individuals differing in their psychological test profiles can have different rates of recidivism (e.g., Beech & Ford, 2006), the trends of self-reported "recovery" in this study should not be inferred to be indicative of the capacity of HSP to reduce recidivism. The association between pre-to-post psychological test change and recidivism is far from clear (see Wakeling and Barnett (2014) for a review). Future studies should continue to explore this relationship.

Third, psychological test scales are prone to socially desirable responding as well as relying on a level of self-insight (Tierney & McCabe, 2001; Beech, Bartels & Dixon, 2013). These problems are well-documented in the general psychological test literature but may be especially pertinent within a prison setting given that individuals often believe that sentence progression is dependent on their progress within an intervention (Wakeling & Barnett, 2014). The HSP environment is designed to be supportive and provide a safe place to test out new skills. It may therefore be possible that individuals completed HSP believing they were better skilled than they actually were, reflecting this in their post-HSP psychological test responses. Socially desirable responding was not explored in the current study in an attempt to keep the

battery of psychological tests to be completed minimal. It is also possible that individuals gained increased insight during HSP participation and therefore there may be a group of individuals whose scores appear to suggest a deterioration in post scores, but which actually reflects increased insight with their problematic areas.

Fourth, reliable offence typology data was not gathered for the study sample. Future research efforts should attempt to explore the effects of HSP by offence typology (e.g., rape, child abuse, etc.) and/or specific paraphilia to determine whether there are differences by type of offence. Additionally, it is worth noting the limitations of RCI and CS techniques. Both RCI and CS create categorical variables which can limit precision and increase noise. Further, the cut-offs used within RCI and CS can be, at least to some degree, arbitrary, and RCI also requires the use of test retest reliability which may not always be available, and wasn't for the current research. Although Chronbach's alpha can be used as a proxy, it is not always an unbiased estimate of test-retest reliability.

And finally, there are limitations with the pre-post design used in this study. These include the fact that change can be a function of natural changes over time and may not necessarily be attributable to the intervention being tested. There is also the issue of regression to the mean, which describes the phenomena that extreme scores pre-intervention are likely to 'regress' down or up to them an even if the intervention has no effect. Test effects may also have an impact; that is, practice effects or test items generating retrospective positive learning independent of the intervention.

4. Results

As part of the data cleaning process skewness and kurtosis[8] were examined. This identified uneven distribution in some of the individual scales and domains. This was expected in line with the sample size and nature of some of the scale information. It was observed most prominently in the HSI (Healthy Sexual Interests) domain. A majority of the sample scored low on these scales, perhaps expectedly in line with socially desirable responding or as a result of the impact of prior accredited programmes attended before HSP. As such, when individuals did report specific interests they appeared as outliers and the data obtained was not normally distributed. However, despite the identification of some uneven distribution, these outliers were not removed in order to minimise the loss of data.

4.1 Pre-to-post programme scores

Table 1 shows mean scores and standard deviations (displayed in brackets) for each of the individual scales and overall domain scores. Sample sizes differ for each measure due to the completion rates of individual scales. If more than 10% of any scale was not completed a score was not calculated. An individual required both pre and post scores to be included in the analysis. Approximately 20% of the sample had missing pre or post scores for each of the measures. The statistical significance of change on each measure is examined in section 4.2 of the reports.

Table 1. Raw scores (scale and domain)

Domain	Scale	*n*	Pre-HSP mean (SD)	Post-HSP mean (SD)
Healthy Sexual Interests	Sexual Preference for Children	75	4.0 (6.2)	2.6 (4.5)
	Sexual Preoccupation	74	3.4 (4.2)	1.8 (2.3)
	Sexualised Violence	75	0.9 (2.3)	0.4 (1.3)

---

[8] Skewness and kurtosis examine how normally distributed the data is. Skewness measures how symmetrical the distribution of scores is. Kurtosis determines the heaviness of the distribution tails.

Domain	Scale	n	Pre-HSP mean (SD)	Post-HSP mean (SD)
	Other Offence-related Sexual Interests	75	0.7 (1.6)	0.3 (1.1)
	Overall	**75**	**8.7 (9.6)**	**4.9 (6.5)**
Healthy Thinking	Attitudes Towards Women	77	5.2 (4.9)	4.1 (3.9)
	Attitudes Towards Children	77	7.3 (12.3)	4.6 (9.3)
	Overall	**77**	**12.5 (15.3)**	**8.7 (11.6)**
Positive Relationships	Ruminations	77	8.0 (7.1)	3.5 (4.3)
	Openness to Men and Women	77	22.2 (8.2)	18.4 (7.1)
	Emotional Congruence with Children	77	13.2 (10.9)	9.1 (8.4)
	Locus of Control	77	8.7 (4.6)	6.2 (4.4)
	Empathic Concern	77	10.4 (4.1)	9.0 (4.21)
	Emotional Loneliness	76	50.4 (10.2)	51.5 (9.7)
	Overall	**76**	**113.3 (27.8)**	**97.7 (22.3)**
Managing Life's Problems	Impulsivity	72	8.4 (9.1)	4.5 (6.3)
	Overall	**72**	**8.4 (9.1)**	**4.5 (6.3)**

Note: Managing Life's Problems contains one psychological test scale and therefore represents the individual scale score and the domain score.

The overall means improved on all domains and all but one individual scale (Loneliness scale pre-HSP mean = 50.43, SD = 10.20; post-HSP mean = 51.50, SD = 9.65) pre-to-post programme. This suggests that overall, individuals' demonstrated improvement in reported level of need following HSP. It is of note, that standard deviations for some of the measures are large, indicating a wide range of scores in the sample. This is discussed further below.

Two one-way ANOVAs were completed to examine differences in the pre-programme and post-programme domain scores for the different levels of risk of reoffending. Overall mean scores are presented in Table 2.

Table 2. Pre and post programme domains scores by risk level

Domain	Risk	Pre-HSP n	Pre-HSP mean (SD)	Post-HSP n	Post-HSP mean (SD)
Healthy Sexual Interests	Low	10	6.8 (7.4)	8	6.3 (6.3)
	Medium	23	8.0 (11.8)	24	6.1 (7.6)
	High	23	10.6 (12.1)	21	4.1 (6.1)
	Very high	29	9.5 (9.0)	29	6.5 (10.2)
	Overall	**85**	**9.1 (10.4)**	**82**	**5.8 (8.1)**
Healthy Thinking	Low	10	9.2 (12.0)	9	8.3 (13.1)
	Medium	24	12.1 (14.2)	24	10.4 (14.3)
	High	23	13.3 (17.2)	21	9.6 (12.7)
	Very high	29	16.6 (18.3)	30	7.0 (9.7)
	Overall	**86**	**13.6 (16.2)**	**84**	**8.8 (12.1)**
Positive Relationships	Low	10	102.5 (30.1)	9	100.8 (22.2)
	Medium	24	113.8 (24.7)	24	99.1 (26.4)
	High	23	118.2 (25.1)	21	100.5 (23.2)
	Very high	28	117.6 (30.3)	30	96.4 (21.0)
	Overall	**85**	**114.9 (27.3)**	**84**	**98.7 (22.9)**
Managing Life's Problems	Low	9	5.1 (5.2)	9	2.4 (3.3)
	Medium	24	10.0 (9.5)	22	6.8 (10.0)

Domain	Risk	Pre-HSP n	Pre-HSP mean (SD)	Post-HSP n	Post-HSP mean (SD)
	High	23	9.1 (10.8)	21	5.8 (8.9)
	Very high	29	8.3 (7.9)	28	4.6 (6.1)
	Overall	**85**	**8.6 (9.0)**	**80**	**5.3 (7.9)**

Pre-programme mean scores for domains (Healthy Sexual Interests, Healthy Thinking and Positive Relationships) were generally higher (indicating negative change) for higher levels of risk, however these differences were not statistically significant: Healthy Sexual Interest ($F(3,81) = 0.407$, $p = 0.748$, $r = 0.08$); Healthy Thinking ($F(3,82) = 0.623$, $p = .602$, $r = 0.10$); and Positive Relationships ($F(3,81) = 0.897$, $p = 0.446$, $r = 0.08$). For the Managing Life's Problems domain, there was less of a pattern in scores by risk, with the medium risk group scoring the highest. However, the differences in pre-programme scores were not statistically significant ($F(3,81) = 0.671$, $p = .572$, $r = 0.10$). This indicates that the relationship with pre-HSP psychological test scores and risk is not clear. All the effect sizes reported here were small (using established guidelines for r; Rice & Harris, 2005).

Overall, there is a less clear pattern for post-programme domain scores by risk. None of the differences in scores by risk level were statistically significant: Healthy Sexual Interests ($F(3,78) = 0.393$, $p = 0.758$), Healthy Thinking ($F(3,80) = 0.395$, $p = 0.757$, $r = 0.08$), Positive Relationships ($F(3,80) = 0.161$, $p = .922$, $r = 0.05$), and Managing Life's Problems ($F(3,59.8) = 1.56$, $p = .215$, $r = 0.15$). For the Managing Life's Problems domain, Welch's F test (for unequal variances) was used to correct for a violation in Levene's test of homogeneity. Again, effect sizes were small. We didn't examine pre to post change on the domains by risk level due to small sample sizes.

4.2   Pre-HSP to post-HSP Effect

A repeated measures ANOVA was completed to examine the pre-to-post effect on psychological test scores during HSP. These are presented in Table 3. Nine individual scales in the Healthy Sexual Interest, Positive Relationships and Managing

Life's Problems domains indicated statistically significant changes in the desired direction: sexual preference for children, sexual preoccupation, other offence-related sexual interest, ruminations, openness to men and women, emotional congruence with children, locus of control, empathic concern, and impulsivity. Four individual scales did not show statistically significant change, though for three of these measures change was observed in the desired direction (sexualised violence, attitudes towards women, and attitudes towards children). For the remaining measure, emotional loneliness, scores got very slightly worse pre to post-HSP. This is perhaps not surprising as the measures asks people about their loneliness at the time of their offending, and the slight increase could well be due to increased insight into their levels of loneliness at this time. The effect sizes ranged from small to large, and were largest for the measures of ruminations, openness to men and women, and impulsivity.

Table 3: Repeated measures ANOVA for individual psychological test scales

Domain	Scale	df (within)	F	p	r
Healthy Sexual Interests	Sexual Preference for Children	71	5.73	0.019	0.27
	Sexual Preoccupation	70	9.67	0.003	0.35
	Sexualised Violence	71	3.22	0.077	0.21
	Other Offence-related Sexual Interests	71	13.20	0.001	0.40
Healthy Thinking	Attitudes Towards Women	73	3.58	0.063	0.22
	Attitudes Towards Children	73	2.49	0.119	0.18
Positive Relationships	Ruminations	73	39.05	<.0001	0.59
	Openness to Men and Women	73	15.88	<.0001	0.42
	Emotional Congruence with Children	73	6.02	0.017	0.28

Domain	Scale	df (within)	F	p	r
	Locus of Control	73	14.12	<.0001	0.40
	Empathic Concern	73	6.30	0.014	0.28
	Emotional Loneliness	72	2.59	0.112	0.19
Managing Life's Problems	Impulsivity	68	21.80	<.0001	0.49

Note: McGrath and Meyer (2006) provide thresholds based on the relationship of r to Cohen's d: $r =$ 0.10 (small, equivalent to $d = 0.20$), $r = 0.24$ (medium; equivalent to $d = 0.50$), and $r = 0.37$ (large; equivalent to $d = 0.80$).

A mixed factorial ANOVA was completed to examine the effect of the HSP on domain scores and explore whether there were differences in effect by risk level. Statistically significant differences were observed in three of the four domains, Healthy sexual interests, Positive relationships, and Managing life's problems.

Mean scores for Healthy Sexual Interests differed significantly before and after HSP ($F(1,70) = 18.76$, $p < 0.001$, $r = 0.47$), with a medium effect size. The mean scores for Healthy Sexual Interests decreased significantly from pre- (mean = 8.66, SD = 9.58) to post-programme (mean = 4.91, SD = 6.49), demonstrating improvements in scores following HSP. There was no statistically significant interaction with risk ($F(3,70) = 1.99$, $p = 0.124$, $r = 0.17$).

For Healthy Thinking, although the scores changed in the desired direction from pre- (mean = 12.52, SD = 15.31) to post-programme (mean = 8.66, SD = 11.64), this change was not statistically significant ($F(1,73) = 3.53$, $p = 0.064$, $r = 0.22$). Whilst we would hope individuals to improve in healthy thinking, this is not a key target of HSP. In addition, there was no statistically significant interaction with risk ($F(3,73) = 1.60$, $p = 0.197$, $r = 0.15$).

For Positive Relationships, mean scores demonstrated statistically significantly change ($F(1,72) = 22.15$, $p < 0.001$, $r = 0.50$), with a medium effect size. Examination of the mean scores indicated change in a positive direction, from pre- (mean = 113.33, SD = 27.83) to post-programme (mean = 97.74, SD = 22.25) demonstrating improvements in scores following HSP. There was no statistically significant interaction with risk ($F(3,72) = 2.29$, $p = 0.085$, $r = 0.18$).

For Managing Life's Problems, the mean scores demonstrated statistically significant change ($F(1,68) = 21.80$, $p < 0.001$, $r = 0.49$), again with a medium effect size. Examination of the mean scores indicated change in a positive direction from pre- (mean = 8.44, SD = 9.06) to post-programme (mean = 4.49, SD = 6.34) demonstrating improvements in scores following HSP. There was no statistically significant interaction with risk ($F(3,68) = 0.09$, $p = 0.966$, $r = 0.04$).

4.3   Clinical and reliable change for domain scores

Clinical significance (comparison with a functional range) and reliable change were examined for domain scores. Table 4 shows the number and percentages of the sample who were regarded as scoring in a clinically significant range post-programme.

Table 4: Clinical Significance (CS) at post-HSP[9]

	Functional range		Non funct. range	
	n	%	n	%
Healthy Sexual Interests	79	83.2	5	5.3
Healthy Thinking	77	81.1	9	9.5
Positive Relationships	83	87.4	3	3.2
Managing Life's Problems	75	78.9	7	7.4

---

[9]   For each domain there a small proportion of the sample for whom clinical significance could not be calculated (there were between 9 and 13 missing cases for each domain).

Large proportions of the population were within a functional range post-HSP. The lowest levels of clinical significance were in Managing Life's Problems ($n = 75$; 78.9%), while the highest were in the Positive Relationships domain ($n = 83$; 87.4%), indicating that overall, a high proportion of individuals had post-programme scores comparable to a functional group (normal range). This suggests a large proportion of the sample, according to clinically significant change (CSC) methodology, appeared "functional" or within normal range post-programme. Table 5 shows the number and percentages of the sample who changed reliably pre-to-post programme within each domain.[10]

Table 5. Reliable change index (RCI)[11]

Domain	No change		Positive change		Negative change	
	n	%	n	%	n	%
Healthy Sexual Interests	54	56.8	18	18.9	3	3.2
Healthy Thinking	59	62.1	15	15.8	4	4.2
Positive Relationships	37	38.9	35	36.8	5	5.3
Managing Life's Problems	48	50.5	24	25.3	1	1.1

The greatest proportion of individuals within each domain did not meet the threshold of "reliable" change. However, the second largest group for all domains, was those showing reliable change for pre-to-post scores in the desirable (positive) direction. Very few individuals demonstrated undesirable (negative) change that was reliable.

---

[10] Jacobson and Truax (1991) suggest that an RCI of 1.96 or above reflects real change (with a 95% confidence level of $p < 0.05$).

[11] For each domain there a small proportion of the sample for whom reliable change index could not be calculated (there were between 17 and 22 missing cases for each domain).

The RCI indicated that the greatest proportion of positive reliable change was observed in Positive Relationships (n = 35; 36.8%), followed by Managing Life's Problems (n = 24; 25.3%), Healthy Sexual Interests (n = 18; 18.9%), and finally Healthy Thinking (n = 15; 15.8%).

4.4　Programme outcome status

The information produced was converted into the five programme outcome groups described above. Table 6 below presents the overall outcomes for each of the four domains. It should be noted that for some categories there are very small numbers.

Table 6. Programme outcome status

	Healthy Sexual Interests		Healthy Thinking		Positive Relationships		Managing Life's Problems	
	n	%	n	%	n	%	n	%
Unchanged	2	2.1	5	5.3	0	0.0	4	4.2
Improved	1	1.1	1	1.1	1	1.1	0	0.0
Recovered	17	17.9	14	14.7	34	35.8	24	25.3
Already okay	52	54.7	54	56.8	37	38.9	44	46.3
Deteriorated	0	0.0	2	2.1	0	0.0	0	0.0
Missing	23	24.2	19	20.0	23	24.2	23	24.2

The findings indicate that the greatest proportion of participants reported scores that were classified as "already okay" and within a functional range prior to HSP (38.9-56.8%). This could be because the participants had all completed other accredited programmes prior to participation in HSP, and therefore could have already made substantial gains in these areas before HSP. The second largest group for all domains was the "recovered" group (14.7-35.8%). Very few individuals fell into the unchanged, improved or deteriorated groups. The Positive Relationships domain indicated the greatest level of individuals categorised as "recovered" following HSP (n = 34; 35.8%), with the Healthy Thinking domain indicating the lowest levels (n =

14; 14.7%). The Healthy Thinking domain was the only domain to show deterioration in scores (albeit for a very small number: $n = 2$; 2.1%) and had the greatest proportion of unchanged scores ($n = 5$; 5.3%).

5.   Discussion

This study aimed to explore the pre-to-post psychological test change results for individuals who had completed HSP, with an additional focus on whether there was an interaction with risk level. Although raw scores indicated some support for positive pre-to-post psychological test change for individual measures and domain scores, further analysis suggested mixed support for the hypotheses.

Hypothesis 1a predicted a statistically significant difference between pre and post psychological test scores for individual scales and overall domains. The analyses demonstrated mixed support for this. Improvements in raw scores, pre-to-post HSP, were statistically significant for nine of the thirteen individual scales. Of the four scales not showing significant improvement, two were in the Healthy Thinking domain. The Emotional Loneliness scale within the Positive Relationships domain, indicated a (non-significant) pre-to-post deterioration. This scale measures an individual's self-reported level of loneliness prior to offending. As such, it is possible that the increase in scores, albeit non-significant, could be a reflection of developed insight rather than an indicator of actual deterioration. Further it is worth noting that positive change on this measure may not be possible due to the fact that it is asking about loneliness at the time of offending.

Three of the four domain scores significantly improved pre-to-post programme (Healthy Sexual Interests, Positive Relationships and Managing Life's Problems). No statistically significant pre-to-post change was observed in the Healthy Thinking domain. This finding may be explained to by the fact that HSP targets Healthy Thinking less than Healthy Sexual Interests and Positive Relationships.

Hypothesis 1b predicted that pre-to-post change would be clinically significant and statistically reliable. The analyses also demonstrated mixed support for this part of the hypothesis, which perhaps is more realistic than expecting every individual to demonstrate pre to post clinically significant change. Post-programme, the majority of the sample were found to be in the functional range for all four domains (see Table 4). However, the majority of the sample for each domain were categorised as "already okay" (Table 6). Therefore, the RCI indicated that the highest proportion of

the sample reported no overall reliable change. This could be due to the fact that HSP is a secondary programme and that treatment gains may have already been made in these areas prior to participating in HSP. Pre- and post-programme scores on many of the scales were also indicative of few problems in these areas, and this could possibly be due to socially desirable responding. Individuals whose score demonstrate more problems could then be identified as outliers and this difference in scores has the effect of increasing the standard deviations for the change criteria. The overall small sample size in combination with the large standard deviations results in very strict criteria for individuals to meet in order to show reliable change. Additionally, reliable change (with large standard deviation) is difficult to demonstrate when the capacity for pre-to-post change is limited by the pre-programme score already being low, that is indicating few problems in these areas (particularly when mean scores are low). Demonstrating reliable change may also be the result of methodological issues in the dataset, including missing data. It would be useful to explore this further in future studies using larger samples.

The second highest proportion of the sample were categorised as "recovered". This suggests that aside from those who already reported functional scores, the next largest group indicated "recovery" from a non-functional to a functional range. This is an encouraging finding and indicates that, at least for some of the sample, psychological test domain scores changed reliably in the desired direction to within a functional range. Positively, very few individuals were categorised as deteriorated or unchanged.

In line with previous findings (e.g., Williams et al., 2007), hypothesis 2 predicted a statistically significant difference between pre-psychological test scores by level of risk and post-psychological test scores by level of risk, where higher risk men would show worse scores. Although the pre-programme raw scores were generally worse for higher risk men, neither these nor the post-programme scores significantly discriminated between actuarially determined risk levels. It is likely that this is due to the small size of each risk group, and so future research should aim to obtain a larger HSP sample. The post-programme raw scores demonstrated a less consistent pattern. This is consistent with other research which has shown that post-treatment

psychological test scores may be less reliable than pre-treatment scores, perhaps as a result of socially desirable responding (e.g., Barnett et al., 2013).

Hypothesis 3 predicted that there will be a statistically significant interaction between pre to post psychological test scores and level of risk. This hypothesis was not supported. When examining pre-HSP scores there were no significant differences between domain scores and risk level. Generally, we would expect higher risk men to have higher levels of criminogenic need, but it is possible that the lack of statistical difference at pre-HSP is due to positive changes already having been made on previous accredited programmes prior to engaging with HSP, thereby reducing the difference between the men across risk categories. Post-HSP domain scores didn't differ by risk level significantly either, which could be explained in the same way. It is also possible that an increased insight due to participation in HSP has had an impact on the post-HSP psychological test scores differently for different risk groups. Further, it appears that the amount of change pre-to-post HSP, as explored in hypothesis 1a and 1b, is not associated with level of risk. That is, higher risk men did not report psychological test change to a greater degree than the lower risk men. It is possible that the small sample size again effected on the degree to which significant patterns in the data were detectable. This is an interesting finding that is worthy of further exploration.

6. Conclusion

The current study suggests that, overall, there are some positive and statistically significant trends emerging from the psychological test data for HSP. The findings generally suggest that HSP is having a positive effect on participants, however, further research is required to determine whether this has any relationship with longer-term outcomes (such as reoffending) as this was not investigated in the present study. Further research would also be warranted to determine whether the findings are applicable to the latest version of HSP, and it may also be warranted to explore other methods to examine differences in change scores (including using methods which control for baseline scores and overcome some of the difficulties with RCI and CS as set out in the limitations section).

The present research has also highlighted some of the cautions that need to be applied and the limitations of psychological tests and pre-post methodology. The low pre-programme scores on some of the measures, for example, make it difficult to detect reliable change pre-to-post programme. It is recommended that there should be a focus on data-rich and longer-term HSP outcomes utilising a variety of methodologies, including qualitative investigations and analysis of reconviction data where possible. It is through these combined methodologies, together with the indicative, preliminary findings reported here, that a holistic picture of the performance of HSP may begin to emerge.

References

American Psychiatric Association. (2013). The Diagnostic and Statistical Manual of Mental Disorders. Fifth Edition.

Andrews, D. A., & Bonta, J. (2006). *The psychology of criminal conduct* (4th ed.) Newark, NJ: LexisNexis/ Matthew Bender.

Barnett, G. D., Wakeling, H., & Howard, P. (2010). An examination of the predictive validity of the Risk Matrix 2000 in England and Wales. *Sexual Abuse: A Journal of Research and Programme*, *22*, 443-470.

Barnett, G. D., Wakeling, H., Mandeville-Norden, R., & Rakestrow, J. (2012). Does change in psychometric test scores tell us anything about risk of reconviction in sexual offenders. *Psychology Crime and Law*. *19:1*, 85-110.

Beckett, R.C. (1987). Children and Sex Questionnaire. HM Prison Service, UK. Unpublished.

Beech, A. R., & Ford, H. (2006). The relationship between risk, deviance, programme outcome and sexual recidivism in a sample of child sexual abusers completing residential programme for their offending. *Psychology, Crime & Law, 12*, 685-701.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure. Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365-376.

Capara, G.V. (1986). Indications of aggression: The Dissipation-Ruminations Scale. *Personality and Individual Differences*, *7*, 763-769.

Collaborative Outcome Data Committee. (2007). Sexual offender programme outcome research: CODC guidelines for evaluation Part 1: Introduction and overview (Corrections Research User Report No. 2007-02). Ottawa, Ontario: Public Safety Canada.

Davis, M. H. (1980). A multi-dimensional approach to individual differences in empathy. JSAS *Catalogue of Selected Documents in Psychology*, *10*, 85-100.

Dennis, J.A., Khan, O., Ferriter, M., Huband, N., Powney, M.J., & Duggan, C. (2012). Psychological interventions for adults who have sexually offended or are at risk of offending. (CD007507; Cochrane Database of Systematic Reviews Issue 12). Chichester, UK: John Wiley & Sons.

Eyesenk, S B. G., & Eyesenk, H. J. (1978). Impulsivity and venturesomeness: their place in a dimensional system of personality description. *Psychological Reports, 43,* 1247-1255.

Farren. L., & Barnett, G. (2014). Measuring offence-related sexual interests in convicted sexual offenders: An initial test of the reliability and validity of a self-report scale. *Her Majesty's Prison and Probation Service Analytical Summary.* Found at: Analytical Summary (publishing.service.gov.uk).

Gannon, T. A., Olver, M. E., Mallion, J. S., & James, M. (2019). Does specialized psychological treatment for offending reduce recidivism? A meta-analysis examining staff and program variables as predictors of treatment effectiveness. *Clinical Psychology Review, 73,* 101752.

Grubin, D. (2008). Validation of Risk Matrix 2000 for use in Scotland. Report prepared for the Risk Management Authority.

Hanson, R. K., Gizzarelli, R., & Scott, H. (1994). The attitudes of Incest offenders: Sexual entitlement and acceptance of children. *Criminal Justice and Behavior*, *21 (2),* 187-202.

Hanson, R. K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). The Principles of Effective Correctional Programme Also Apply to Sexual Offenders: A Meta-Analysis. *Criminal Justice and Behavior, 36*, 865.

HMPPS Annual Digest (2021). Official Statistics Bulletin. Found at: HMPPS Annual Digest 2020/21 (publishing.service.gov.uk)

HMPPS Interventions Services (2013). *Assessment and Evaluation Manual for Sex Offender Treatment Programmes*. National Offender Management Service, UK. Unpublished.

Jacobson, N. S., Roberts, L.J., Berns, S.B., & McGlinchey, J.B. (1999). Methods for determining the clinical significance of programme effects: Description, application and alternatives. *Journal of Consulting and Clinical Psychology, 67*, 300-307.

Jacobson, N. S., & Truax, P. (1991) Clinical Significance. A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19.

Långström, N., Enebrink, P., Laurén, E.M., Lindblom, J., Werkö S., & Hanson, R.K. (2013). Preventing sexual abusers of children from reoffending: systematic review of medical and psychological interventions. *British Medical Journal* (Clinical research ed.), 347: f4630.

Levenson, H. (1974). Multidimensional locus of control in prison inmates. *Personality and Social Psychology Bulletin*, *1*, 354-356.

Mandeville-Norden, R., Beech, A.R., & Hayes, E. (2008). Examining the effectiveness of a U.K. community based sexual offender programme program for child molesters. *Psychology Crime and Law, 14,* 493-512.

Mann, R. E., & Carter, A. J. (2012). Organising principles for the programme of sexual offending. In B. Wischka, W. Pecher & H. van der Boogaart (Eds)., Behanlung von Straftätern: Sozialtherapie, Maβregelvollzug, Sicherungsverwahrung [Offender programme: Social Therapy, Special Forensic Hospitals, and Indeterminate Imprisonment] Centaurus.

Marshall, W. L. (1995). Sex with Children is Justifiable Questionnaire. HM Prison Service UK. Unpublished

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

McGrath, R., Cummings, G., Burchard, B., Zeoli, S., & Ellerby, L. (2010). Current practices and emerging trends in sexual abuser management: The safer Society 2009 North American Survey. Brandon, VT: Safer Society Press.

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

Mpofu, E., Athanasou, J. A., Rafe, C., & Belshaw, S. H. (2016). Cognitive Behavioral Therapy Efficacy for Reducing Recidivism Rates of Moderate- and High-Risk Sexual Offenders: A Scoping Systematic Literature review. *Interpersonal Journal of Offender Therapy and Comparative Criminology,* 1-17.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. Law and Human Behavior, *29,* 615–620.

Roger, D., & Najarian, B. (1989). The Construction and Validation of a New Scale for Measuring Emotion Control. *Personality and Individual Differences, 10(8),* 845-853.

Russell, D., Peplan, C. A., & Cutrona, C. A. (1980). The revised UCLA Loneliness Scale Concurrent and discriminant validity evidence. *Journal of Personality and Social Psychology*, *39*, 472-480.

Schmucker, M. & Lösel, F. (2017). Sexual offender programme for reducing recidivism among convicted sex offenders: a systematic review and meta-analysis. Campbell Systematic Reviews.

Thornton, D. (2002). Constructing and testing a framework for dynamic risk assessment. *Sexual Abuse: A Journal of Research and Treatment, 14*, 139-153.

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

Thornton, D., Beech, A., & Marshall, W.L. (2004). Pre-programme self-esteem and post-programme sexual recidivism. *International Journal of Offender Therapy and Comparison Criminology, 48*, 587-599.

Thornton, D., Mann, R., Webster, S., Blud, L., Travers, R., Friendship, C., & Erikson, M. (2003). Distinguishing and combining risks for sexual and violent recidivism. *Annals of New York Academy of Sciences*, *989*, 225-235.

Tierney, D. W., & McCabe, M. P. (2001). An evaluation of self-report measures of cognitive distortions and empathy among Australian sex offenders. *Archives of Sexual Behavior. 30*, 495 – 519.

Ward, T. & Laws, D. R. (2010). Desistance from sex offending: Motivating change, enriching practice. *International Journal of Forensic Mental Health*, 9, 11-23.

Wakeling, H. C. (2014). *Examination of the use of self-report psychometrics within sexual offender treatment and in prediction of reoffending.* Doctoral dissertation, University of Birmingham. ProQuest Dissertations and Theses Global.

Wakeling, H. C., & Barnett, G. (2010). Measuring grievance thinking in sexual offenders: The revised Dissipation-Rumination Scale. *Journal of Sexual Aggression*. *17*, 273-289.

Wakeling H. & Barnett, G. (2014). The relationship between psychometric test scores and reconviction in sexual offenders undertaking programme. *Aggression and Violent Behavior*, *19*, 138-145.

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

Wakeling, H., Freemantle, N., Beech, A.R., & Elliott, I. A. (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.

Wampold, B. E. (2001). *The great psychotherapy debate: Models, methods and findings*. Mahwah, N.J: Lawrence Erlbaum Associates.

Webster, S. D., Mann, R. E., Thornton, D., & Wakeling, H. C. (2007). Further validation of the short self-esteem scale with sexual offenders. *Legal and Criminological Psychology, 12,* 207-216.

Williams, F. (2007). My Private Interests Measure for Sexual Offenders with cognitive and/or social deficits. National Offender Management Services. Unpublished.

Williams, F., Wakeling H. & Webster S. (2007). A psychometric study of six self-report measures for use with sexual offenders with cognitive and social functioning deficits. *Psychology Crime and Law, 13(5),* 505-522.

Appendix A
Full demographic breakdown

Table A1. Demographic characteristics

Demographic	n	%
Ethnicity		
White British	85	90
White other	3	3
Black British	2	2
Asian	1	1
Asian / mixed	1	1
Missing	3	3
Sentence type		
Determinate	15	16
Indeterminate	77	81
Missing	3	3
Risk level		
Low	10	11
Medium	27	28
High	25	26
Very high	31	33
Missing	2	2

Appendix B
Clinical and risk measures

Risk Matrix 2000 (RM2000/s)

Risk Matrix 2000 (RM2000/s: Thornton, Mann, Webster, Blud, Travers, Friendship & Erikson, 2003) is a static risk assessment tool which is used with adult males who have been convicted of a sexual offence. At least one of the sexual offences must have been committed when the individual was over 16 years of age. The RM2000/s predicts risk of sexual recidivism and is made up of seven items divided into two scoring stages.

Stage one comprises 3 items: age of the individual on release, number of sentencing occasions for a sexual offence and number of sentencing occasions for any criminal offence. The scores assigned to each of these items are summed and translated into one of four preliminary risk categories: low, medium, high and very high.

The second scoring stage has four further items: whether or not the victim of any of their sexual offences have been male, strangers, whether or not they have ever been in a stable live-in relationship for over two years, and whether or not any of their offences have been non-contact. These items are scored on a dichotomous scale as either present or not. If two or three of these items are present the initial risk category is raised one level (e.g. from low to medium). If all four aggravating factors are present, the initial risk category is raised by two risk levels (e.g. from low to high).

The RM2000/s is completed for all men convicted of sexual offences for whom the tool is applicable prior to engaging in programme as part of the assessment procedure. The RM2000/s has been found to have a good level of predictive validity (Barnett, Wakeling & Howard, 2010; Beech & Brown, 2007; Grubin, 2008; Thornton et al., 2003).

Although the RM2000/s risk categories can be interpreted in an actuarial way as indicating long term risk of reconviction, it should be remembered that (a) the reconviction rates shown were derived from specific samples, and are therefore

subject to sampling error; (b) reconviction rates may vary from one jurisdiction to another, and over time, depending on the behaviour of the police and the courts; and (c) reconviction is at best a lower-bound estimate of rates of re-offending.

Psychological Test Measures

Healthy Sexual Interests

My Private Interests Measure for Sexual Offenders with cognitive and/or social deficits (Farren & Barnett, 2014; Williams. F, 2007)

This is a 54-item scale measuring sexual interests. The scale covers a variety of different 'interests' and intends to provide an overall picture of participants' sexual interests. The scale has four subscales: Sexual preoccupation, sexual preference for children, preferring sex to include violence or humiliation, and other offence related sexual interests. Participants rate items on a dichotomous true / false scale depending on whether they agree with the statement or not. Items are scored (yes = 2, no = 0, missing = 1). If more than one item is missing from each subscale the score is not computed. Higher scores indicate stronger levels of sexual interests. This measure has excellent internal reliability (α = .92-.94), has been shown to discriminate higher from lower risk individuals, and has shown good convergent validity with other measures of sexual interests (Farren & Barnett, 2014).

Healthy Thinking

Booklet 2 (Interventions Services, 2013).

This is a 39-item measure containing 2 scales, both with good internal consistency (Wakeling, 2014; see Table B2).

- **Scale 1**: Attitudes towards Women Scale, consists of 9 items relating to attitudes and beliefs that support the sexual abuse of women. Item responses are on a 5-point likert scale. Response anchors are (0 = strongly disagree, 1 = disagree, 2 = undecided, 3 = agree and 4 = strongly agree). Higher scores indicate greater agreement with scale items (more strongly held attitudes). Missing items are given a score of 2.
- **Scale 2:** Attitudes towards Children Scale. This consists of 30 items relating to attitudes and beliefs that support the sexual abuse of children. Item responses are on a 5-point Likert scale. Response anchors are (0 = strongly disagree, 1 = disagree, 2 = undecided, 3 = agree, and 4 = strongly agree).

Higher scores indicate greater agreement with scale items (more strongly held attitudes). Missing items are given a score of 2. Total scale: Offence supportive beliefs. Scores from both scales are summed to produce an overall domain score. Higher scores indicate more strongly held attitudes relating to the sexual abuse of women and children.

This scale is unpublished but was devised from four different scales: the Sex with Children is Justifiable Scale (Marshall, 1995); the Children and Sex Questionnaire (Beckett, 1987); the Entitlement to Sex Scale (Hanson, Gizzarelli, & Scot, 1994); and the Women are Deceitful Scale (Interventions Services, 2013). Scale 2 is made up of all 18 items of the original Sex with Children is Justifiable Scale plus 12 items from the original 15 item Cognitive Distortions subscale of the Children and Sex Scale.

Healthy Relationships

Relationships Scale (Interventions Services, 2013).

This is an 81-item measure consisting of 3 main subscales each with good internal consistency (Wakeling, 2014; see Table B2).

- **Subscale 1:** Personality 1 Questionnaire. This consists of 14 items relating to self-esteem and ruminations. Item responses are dichotomous (1 = true, 2 = false). For scoring purposes, items are recorded (0 = false, 2 = true). Missing items are assigned a value of 1. If more than 1 item is missing a total score is not computed. High scores equate to a low self-esteem and high ruminations. Greater scores indicate dysfunction in this area.

- **Subscale 2:** Personality 2 Questionnaire. This consists of 49 items measuring the extent to which respondents are unable to see others' points of view or relate to others and have an external locus of control. Item responses are on a Likert scale (0 = strongly disagree, 1 = disagree, 2 = undecided, 3 = agree, and 4 = strongly agree). Missing items are given a score of 2. If more than 5 items are missing a score is not computed. Greater scores indicate greater difficulties in this area.

- **Subscale 3:** Emotional Loneliness. This consists of 18 items which indicate the extent to which respondents believe they had meaningful relationships, had people close to them, or were lonely in the 6 months preceding their sexual offence. Item responses are on a 4-point Likert scale (1 = completely

false, 2 = mainly false, 3 = mainly true, 4 = completely true). Missing items are assigned a value of 1. If more than 2 items are missing the score is not computed. Greater scores indicate greater levels of loneliness and fewer close and meaningful relationships.

Total Scale: Relationships. Scores from subscales are summed to produce an overall Relationships domain score. Higher scores indicate more difficulties in an individual's relationships with self and others.

The origins of this scale are derived from five different scales; the UCLA Emotional Loneliness Scale (Russell, Peplan & Cutrona, 1980); the Self Esteem Scale (Thornton, Beech & Marshall, 2004; Webster, Mann, Thornton & Wakeling, 2007); the Ruminations Scale (Capara, 1986; Wakeling & Barnett, 2010); the Interpersonal Reactivity Index (Davis, 1980); and the Locus of Control Scale (Levenson, 1974). Subscale 1 is made up of 7 of the original 8 self-esteem items, and 7 of the original 15 rumination items. Subscale 2 is made up of 16 of the original 18 openness to men and women items, all 15 items of the emotional congruence with children subscale of the Children and Sex scale, 9 of the Locus of Control scale items, 5 items of the perspective taking subscale of the Interpersonal Reactivity Index and 4 items of the Empathic Concern subscale of the Interpersonal Reactivity index. Subscale 3 is made up of 18 items of the original UCLA Emotional Loneliness Scale.

Managing Life's problems

Self-Management Scale (Interventions Services, 2013). This is an 18-item measure of impulsivity. These items relate to an individuals' tendency to act without thinking about long-term consequences. Item responses are dichotomous (1 = true, 2 = false). For scoring purposes items are recoded (0 = false, 2 = true). Missing items are assigned a value of 1. Greater scores indicate a greater tendency to act without thinking about long-term consequences. As there is just one scale within this domain, it is of note that when referring to the total scale score, this was repeated and reflected the single impulsivity scale score.

The items for this scale originate from two different scales: the Impulsivity Scale (Eysenck & Eysenck, 1978); and the Aggression Control and Benign Control

subscale of the Emotional Control Questionnaire (ECQ: Roger & Najarian, 1989). Some items have been reworded slightly from the original version to better correspond with a true/false format. The scale comprises 8 items from the Benign Control subscale of the ECQ and 2 items from the Aggression Control subscale of the ECQ. The scale has been shown to have good internal consistency (Wakeling, 2014; see Table B2).

Psychological Test characteristics:

Internal consistency scores (α) for the individual scales and domains can be found in Table B1 below. Although desirable, test-retest reliability scores were difficult to obtain due to the intense resource provision this data collection would require.

Table B1. Internal Consistency scores for individual scales and domains

Domain	Scale	Internal consistency (Cronbach's α)
Healthy Sexual Interests	Sexual Preference for children scale	0.93
	Sexual Preoccupation scale	0.93
	Sexualised violence scale	0.94
	Other sexual interests scale	0.92
	Overall domain 1 scale	0.90
Healthy Thinking	Attitudes towards women scale	0.83
	Attitudes towards children scale	0.95
	Overall domain 2 scale	0.94
Positive Relationships	Ruminations scale	0.82
	Openness to men and women scale	0.89
	Emotional congruence with children scale	0.88
	Locus of control scale	0.82
	Empathic concern scale	0.83

Domain	Scale	Internal consistency (Cronbach's α)
	Loneliness scale	0.95
	Overall domain 3 scale	0.91
Managing Life's problems	Impulsivity scale	0.91
	Overall domain 4 scale	0.91