Ministry of Justice

Horizon and iHorizon

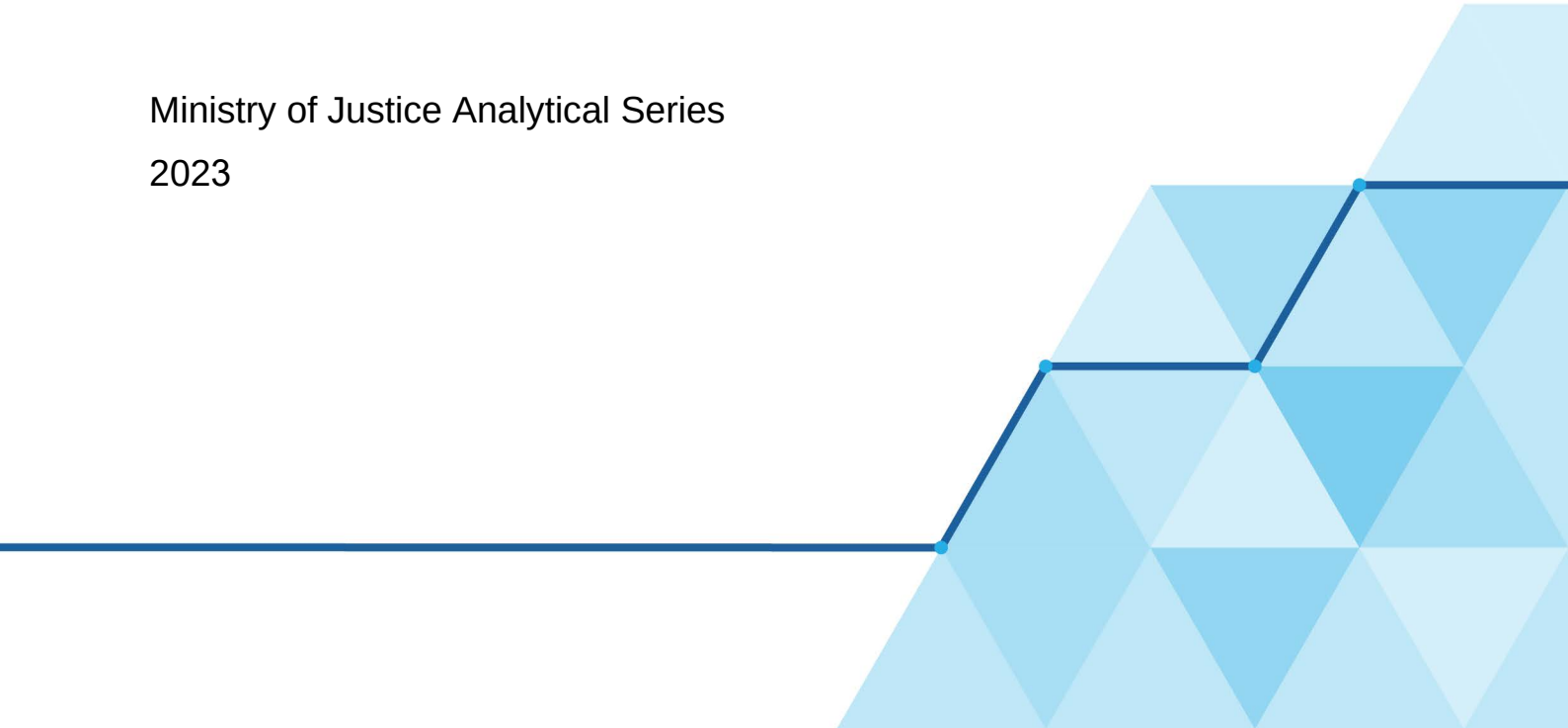
Psychometric analyses of the Success Wheel Measure

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 978-1-84099-987-7

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

Two new assessment measures were developed and administered as part of the clinical outcome study for Horizon and iHorizon. The Success Wheel Measure (SWM) is a 5-item scale designed to measure progress on the five domains of Horizon: (1) Managing life's problems, (2) Healthy relationships, (3) Healthy sexual interests, and (4) Healthy thinking and (5) Sense of purpose (representing desistance). The Horizon Motivational Scale (HMS) is a 4-item scale designed to measure four elements of overall motivation for Horizon participation: enthusiasm, direction, commitment, and holistic attitude.

A third measure was implemented against which the SWM was compared. The Sex Offender Treatment and Interventions Progress Scale (SOTIPS) is a validated measure composed of 16 items shown to have a significant association with sexual recidivism and classified into sexual, criminal, co-operation, self-management, and social stability.

Any assessments used to make inferences about the individuals tested needs adequate validity and reliability to be considered meaningful. This study aimed to test whether the SWM and HMS (a) measure programme-related skills and insight and pre-programme motivation to engage with and complete Horizon, as intended, (b) measure them to an acceptable standard, and (c) measure them in a way that is comparable to other relevant measures of insight and skills with established validity and reliability (e.g., the SOTIPS).

Methodological approach and interpreting findings

Three forms of psychometric analysis were conducted using subsets of Horizon SWM and HMS data. Structural analyses examined whether the five items of the SWM and the four items of the HMS each measure one general overall dimension, which we presume to be pro-social insights and strengths and motivation to complete the programme respectively.

Construct validity analyses examined (a) the extent to which domains reliably produced consistent scores across administrations, (b) whether items duplicate one another and positively correlate with overall totals, and (c) whether items distinguish between those

higher and lower in need. Concurrent criterion validity analyses examined whether the change indicated by the SWM was comparable to that indicated by the SOTIPS.

Key findings

Structural findings indicated that collectively the individual items of the SWM and HMS appear to measure one dominant overall dimension. However, psychometric outcomes cannot qualitatively confirm that those dimensions are insight/skills and motivation.

Reliability analyses indicated that both measures have acceptable levels of reliability. Given a recommended minimum threshold of 0.80 for acceptability, reliability for the SWM was in the range of 0.83–0.87 and reliability for the HMS was in the range of 0.87–0.91. Item-to-total and item-to-item correlations indicated that, for both measures, items were positively associated with total scores and no items appeared to duplicate one-another.

Item response theory metrics indicated that all five items of the SWM successfully discriminated between high and low overall scorers and that thresholds for increases in score are consistent with increasing presence of insights and skills.

No statistically significant difference was found between measured change on the SWM and the SOTIPS, suggesting that the SWM measured change in this context in a manner and to a degree equivalent to a well-established and previously validated measure.

All four items of the HMS successfully discriminated between high and low overall scorers. However, the findings suggested that demonstrating only a relatively small amount of motivation was required by participants to receive higher overall scores from facilitators.

Conclusions

The findings of this exploratory study provide preliminary positive evidence for the validity of the Success Wheel Measure and the Horizon Motivational Screen. This supports the SWM and the HMS as appropriate for use examining outcomes for Horizon and iHorizon.

Nevertheless, methodological limitations should be considered when interpreting these results. From a data perspective, the SWM and HMS are clinical judgement tools and were specifically designed to be used in a study of clinical change. Consequently, the variability

and nature of scores may be affected by natural subjectivity about participants capacity for change or by facilitators overall implicit beliefs about the effectiveness of Horizon.

From a methodological perspective, there continues to be ongoing academic debate around how these psychometric statistics should be interpreted, particularly those related to how well statistical models fit observed data and how to establish the true number of groups or factors that can be identified as existing in observed data.

The findings also indicate potential for improvements. For example, finding that the SWM only just breached the threshold for acceptability suggests some inconsistency in use across cases and across assessors. Further exploration is recommended to establish whether this should involve improvements to the measure, training of assessors, or both.

Additionally, although it is plausible that the frequency of high scores on the HMS is simply indicative of a highly motivated sample, it is worth investigating whether improvements to thresholds, definitions, or scoring guidelines might make the measure more sensitive.

2. Introduction

2.1 Aims of this study

Any assessments used to draw inferences about the individuals tested should have good validity and reliability to be considered useful and meaningful. According to the American Educational Research Association and their collaborators (AERA, APA & NCME, 2014), validity 'refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests' and that the process of validation 'involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations' (p. 9). They argue that a sound argument for validity should integrate various metrics representing multiple strands of evidence that support a coherent account for the interpretation of the assessment. These strands include (but are not limited to):

1. The extent to which items and components reflect the true nature of the construct being measured (structural and construct validity).

2. The extent to which assessment scores predict performance on other variables or compare with other similar or divergent measures (criterion validity).

The aim of this study is to provide psychometric data on the performance of the newly developed measures used to examine change on Horizon and iHorizon. This report accompanies the uncontrolled before-after study of Horizon and iHorizon and examines the performance of the Success Wheel Measure and the Horizon Motivational Scale.

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, 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 reconvictions. Other small-sample studies have found some (albeit mixed) evidence of short-term change 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 et al, 2012; 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, the cause of failure to find an association between psychometric change and future reconvictions for core SOTP remains unclear.

2.2 Measures

The Horizon and iHorizon clinical outcome study used 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 at limited sites for a fixed time period for the purpose of the study.

The SWM and HMS were developed for implemented on Horizon and iHorizon in 2018 as routine assessments at all sites for all participants. During the Horizon/iHorizon clinical study, the SOTIPS was administered at select sites in addition to and at the same assessment points as the SWM. The SWM was routinely administered during a 1-to-1 session that occurs after Block 2 (pre-programme) and at the final post-Horizon one-to-one session (post-programme). The ratings were completed independently by the facilitator and participant prior to those sessions. This coaching session follows module 2 of the programme, which means that the pre-programme SWM scores do not technically precede the start of Horizon or iHorizon. Modules 1 and 2, however, are designed to address engagement, build rapport, and introduce programme processes and concepts. The modules that follow the 1-to-1 coaching session after module 2 are those that directly address those strengths targeted by the Success Wheel.

The HMS was routinely administered before Block 1 (pre-programme), during the mid-programme one-to-one session, and during the post-Horizon one-to-one session (post-programme). Although the HMS was administered at the same points as the pre- and post-programme SWM, a decision was made to also administer an HMS at the pre-group individual session at the very start of the programme. Since concepts measured by the HMS are introduced in modules 1 and 2 it was necessary to measure motivation before participants addressed those concepts (engagement, constructive participation, etc). This study uses scores from that pre-group 1-to-1 session as the basis for pre-programme motivation and the final administration as the basis for post-programme motivation.

Success Wheel Measure (SWM)

Horizon is an 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. iHorizon is designed for individuals with convictions only for possessing, downloading, making and/or distributing indecent images of children, not for contact sexual abuse or a combination of both. Both 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., intimacy; 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 included to represent development of a desistance identity (self-efficacy and agency) and citizenship (community participation).

The Success Wheel Measure (SWM) is a 5-item scale with scores ranging from 0–25 that was designed specifically to measure progress on Horizon and iHorizon. The Success Wheel Measure items mirror the 5 domains of the Horizon Success Wheel: (1) Managing life's problems, (2) Healthy relationships, (3) Healthy sexual interests, and (4) Healthy thinking and (5) Sense of purpose (representing desistance). The SWM uses a 5-point Likert scale from 1 (has yet to achieve success in this area) through 3 (moderate success in this area) to 5 (very good success in this area). The SWM was intended as an adaptation of a clinician rating developed as part of an evaluation by Marquez et al. (2005) as a structured clinical judgement of if the individual derived benefit from the programme.

Horizon Motivational Scale (HMS)

The Horizon Motivational Scale (HMS) is a 4-item scale that was designed specifically to measure motivation towards participating in Horizon. It is informed by self-determination theory (Deci & Ryan, 2008; Ryan & Deci, 2017), which defines motivation as the internal or external energy that drives someone to engage in a course of action and concerns four aspects of activation and intention: energy, direction, persistence, and equifinality.

The four items of the scale are referred to as: (1) enthusiasm – a positive attitude, personal energy, and a drive to positively direct that energy; (2) direction – an internal desire to take part in Horizon/iHorizon; (3) commitment – a willingness or commitment to completion in full and an acceptance that doing so will take resolve and perseverance, and (4) holistic – a recognition that programmes positively contribute, alongside other pro-social activities, to efforts to live an offence-free life. Items are scored by facilitators on a scale of 0 (no evidence), 1 (some evidence), and 2 (strong evidence). Total scores range from 0–8.

Sex Offender Treatment and Interventions Progress Scale (SOTIPS)

The Sex Offender Treatment and Interventions Progress Scale (SOTIPS) (McGrath et al., 2013) is a facilitator-administered rating scale composed of 16 dynamic risk items shown to have a statistically significant relationship to sexual recidivism (broadly classified into sexual, criminal, co-operation, self-management, and social stability) and is designed to aid the identification and monitoring of treatment needs in adult male sex offenders. The 16 items are scored on a 4-point scale; 0 (minimal-to-no need for improvement), 1 (some need for improvement), 2 (considerable need for improvement), and 3 (very considerable

need for improvement). Total scores range from 0 to 48 and are organized into three risk/need groups: low (0 to 10), moderate (11 to 20) and high (21 to 48). Static SOTIPS scores, but not dynamic changes in SOTIPS scores, have successfully been found to predict all types of recidivism (sexual, violent, and general) (Hanson et al., 2021).

3. Methodology

3.1 Sample and data

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. As the SWM and HMS are central to our analyses of individual progress on Horizon and iHorizon it is important to know if they are (a) measuring the latent trait they are supposed to be measuring, (b) measuring those traits to an acceptable standard, and (c) comparable to other relevant measures of those traits.

For the purpose of this study, the Horizon data alone was isolated for analysis since it is the larger set. Cases with missing SWM and HMS data were removed from the sample. The Horizon sub-sample was then separated into two subsets, one with participants from selected sites where additional validation measures (SOTIPS) had been implemented ("validation" sample: n = 147) and those from remaining sites ("test" sample: n = 1016). Finally, the test subset was separated into separate subsets containing complete cases of SWM scores (n = 849, 95.0%) and complete cases of HMS scores (n = 855; 95.6%). The test sample was used to produce psychometric and normality data for the SWM and HMS. The validation sample was used to compare the SWM to the SOTIPS.

Representativeness of the validation sample

Chi-square tests of association were used to assess the representativeness of the validation group sample compared to the test sample. No statistically significant effects of age and maintaining innocence were found. A statistically significant association was found between ethnicity and validation group assignment (χ^2 (4) = 9.50, p = 0.0498, FET p = 0.025, ϕ = 0.10) with post-hoc tests indicating that this was the result of underrepresentation of Asian (z = 5.1) and mixed race (z = 2.1) participants in the smaller validation sample. The magnitude of the effect of ethnicity, however, was small.

Normality and score distributions

Prior to any psychometric analyses, statistical checks were carried out on the SWM and HMS datasets used in those analyses to ensure that the data did not excessively deviate from statistical normality. This included specific tests of skew (whether the distribution of scores around the mean is symmetrical or leans heavier to the right or left) and kurtosis (whether the distribution of scores around the mean is narrower or wider), with skew indices between -0.5 and 0.5 and excess kurtosis indices between -1 and 1 used to consider distributions sufficiently “normal” (Lehman, 1991; Westfall, 2014).

As illustrated in Table 1, neither the item scores nor total scores for the SWM critically exceeded the thresholds for skewness or excess kurtosis, indicating that scores were broadly normally distributed. However, for the HMS, the total scores slightly exceed the threshold for skew, where higher overall motivation scores are more frequent than lower (higher motivation is the norm). This is also the case for the Commitment item, where higher commitment scores are more frequent than lower (higher commitment is the norm).

Table 1: Skewness and excess kurtosis scores for the psychometric test subsamples

Measure	Item	Mean	SD	Skewness	Excess kurtosis
SWM	MLP	2.99	0.81	-0.12	-0.35
	HT	3.01	0.84	-0.15	-0.29
	HSI	2.94	0.87	-0.00	-0.19
	PR	2.86	0.79	0.06	-0.30
	SOP	3.28	0.86	-0.26	-0.25
	Total	**15.08**	**3.08**	**-0.15**	**-0.08**
HMS	Enthusiasm	1.31	0.64	-0.39	-0.70
	Direction	1.24	0.68	-0.35	-0.86
	Commitment	1.52	0.56	-0.63	-0.63
	Holistic	1.27	0.64	-0.20	-0.69
	Total	**5.34**	**1.98**	**-0.53**	**-0.27**

3.2 Design

Structural analysis

Structural analyses examine the extent to which data support assumptions about internal structure in terms of (a) the nature of the data and how it can be quantified and (b) the appropriateness of the domains into which items are grouped. The number of dimensions a test can have will always lie somewhere between 1 (all items test one latent variable) and the total number of items (every item tests its own latent variable). Although the SWM and HMS have multiple domains, they are intended to be "essentially unidimensional" in that its items should primarily measure one clearly dominant latent construct in common (e.g., offence-relevant strengths in the SWM) (Slocum-Gori & Zumbo, 2010; Nandakumar & Stout, 1993). The psychometric analyses are also dependent on the measure being at least essentially unidimensional (De Ayala, 2009; Engelhard, 2013; Reckase, 1979).

Unidimensionality was assessed via parallel analysis (Horn, 1965). This procedure compares the observed data to randomly generated data (of the same size) and compares the eigenvalue for each factor (indicating how much of the common variance it accounts for) in the observed data to their corresponding factor that was generated randomly. Our parallel analyses used a generalised least squares method, a polychoric correlational matrix, and 50 resampled datasets to generate comparison eigenvalues.

Additionally, indices for the quality of the "fit" of a unidimensional (one factor) model to the data were obtained via confirmatory factor analysis. The thresholds for an acceptable fit were a comparative fit index (CFI) of greater than 0.90, a Tucker-Lewis Index (TLI) of greater than 0.90 and a root mean square error of approximation of less than 0.06 (Engelhard, 2013; Hu & Bentler, 1999; Slocum-Gori & Zumbo, 2010). In a review of common thresholds for statistics, Lance et al (2006) note that discussions about model fit are still being debated in the relevant statistical literature, but that fit metrics (e.g., TLI) above 0.90 are not necessarily evidence of "good fit" for data but conversely that 'models whose TLI and NFI were less than 0.90 could usually be improved substantially' (p. 205). Finally, Velicer's (1976) minimum average partial (MAP) test was conducted. The MAP test removes components one-by-one until proportionately more unsystematic variance remains compared to systematic variance (i.e., the variance is not common variance).

Construct validity

Construct validity is the extent to which assessments generate scores on which inferences can be drawn about individual assessment-subjects or groups of assessment-subjects based on those data. This represents the extent to which higher or lower classifications on items of scales can be used to draw inferences on whether one assessment-subject (or group) has greater clinical need than another with different values. It also represents the extent to which the various items and domains representing different aspects of clinical need collectively contribute to classifications in the expected manner.

Internal consistency was examined using well-established metrics from classical test theory (CTT). CTT is a body of psychometric theory that predicts outcomes of psychological assessment such as the difficulty of items based on the premise that an observed or obtained score on an assessment is the sum of a true score and some amount of error. Reliability, derived from classical test theory, is the extent to which an assessment returns the same results consistently when used in the same context and on repeated occasions. Various standards have been cited for reliability. Lance et al. (2006) concluded that, ‘a more reliable measure is better than a less reliable one’, that “adequate reliability” is dependent on the circumstances, and that a paper by Nunnally (1978) widely cited for a standard of greater than 0.70 for reliability actually recommends 0.80 for measures used in applied research, as did others (e.g., Carmines and Zeller, 1979).

Typically, reliability is examined by assessing individuals at different time intervals (known as “test-rest” reliability). It is possible to estimate that reliability using data from one administration of an assessment using statistical tests. Recent reviews suggest that a range of tests should be used to provide evidence of reliability (Choo, 2022; McNeish, 2018; Revelle & Zinbarg, 2009; Sijtsma, 2009). We calculated five reliability metrics based on recommendations in the literature: Cronbach’s alpha, McDonald’s omega, the algebraic greatest lower bound (GLB), Guttman’s lambda 2, and ten Berge and Zerglers’s mu-2. To aid interpretation of reliability metrics, a Monte Carlo simulation study found lambda 2 and mu-2 to be consistently accurate indicators for unidimensional measures (Cho, 2022).

Nonrelevance and redundancy of items was also examined. Nonrelevance refers to items that are not independently associated with total scores, illustrated by item-to-total correlations below a threshold of 0.20 (Hinkle et al., 2003). Redundancy refers to two or

more items replicating information provided by one-another, illustrated by item-to-item correlations greater than a threshold of 0.80 (Nunnally & Bernstein, 1994).

Whereas CTT assumes all items measure the underlying latent trait to an equal extent, Item response theory (IRT) assumes different items measure the trait to varying extents. An IRT graded response model was used to estimate: (1) “difficulty” – the quantity of latent trait required to move up from one point to the next (e.g., the amount of strengths required to move from a score of 1 to 2); and (2) “discriminant power” – how well items successfully discriminate between participants with different amounts of the latent trait. Discriminant values lower than 0.65 were considered low, 0.65 to 1.34 considered moderate, 1.35 to 1.69 considered high, and greater than 1.7 considered very high (Baker & Kim, 2004).

Criterion validity

Criterion validity examines how a measure is related to relevant outcomes and is typically divided into either concurrent or predictive criterion validity. Concurrent criterion validity represents the association between the measure and an outcome assessed at the same time and is demonstrated when an assessment correlates well with established measures of the same or similar constructs. Predictive criterion validity represents the extent to which a score on a scale or assessment predicts scores on a different measure or likelihood of some other related outcome at a later point in time. It is not possible to evaluate the predictive validity of the SWM and HMS until participants can be followed-up for relevant outcomes (e.g., new reconvictions) and predictive ability is not addressed in this study.

The SOTIPS was used to examine convergent validity for the SWM, or the extent to which the SWM correlates with another well-established and validated measure of treatment progress for a programme that also targets individuals with sexual convictions. If the SWM is effectively measuring progress, there should be no difference in pre-to-post-change on SWM scores and pre-to-post scores on similar measures, represented here by the SOTIPS. Contrary to the SWM, lower scores on the SOTIPS represent positive outcomes, so SOTIPS scores were inversed (positive numbers into negative). The association between the SWM and SOTIPS was tested using a multilevel model approach using time (pre- vs. post-Horizon) and scale (SWM vs. SOTIPS) as predictors. An absence of a statistically significant interaction between measure and time indicates that pre-to-post effects of similar magnitude and in the same direction were observed across both scales.

A bootstrapped Kendall's tau statistic for the association between the SWM and SOTIPS total scores was also conducted to test the basic association between total scores.

3.3 Limitations

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.

It is also important to reiterate that the pre-Horizon SWM is administered after modules 1 and 2 (covering engagement, rapport, and introductions) but before the Success Wheel is introduced. Therefore, any change resulting from modules 1 and 2 is not being captured in these data. We also cannot rule out that participants and facilitators may be incentivised to engage in socially or operationally desirable responding. Concerns about restricting pre scores and/or inflating post scores to create a positive effect were considered during development of the SWM and informed the use of independent facilitator and participant scoring pre-programme and the requirement for collaboration on scores as safeguards.

It is important to reiterate the ongoing debate around the subjective interpretation of the various psychometric indices used in this study. However, interpretation of those metrics continues to be debated. We will argue that the indices for unidimensionality, for example, supported our use of IRT graded response models. These models are highly sensitive to dimensionality, as additional dimensions will introduce confounding variability into our assessment of the relationships between items. Similar questions include: what constitutes "good model fit"? What is an acceptable reliability coefficient? We sought to provide a defence for our decisions, but interpretations should be read with that subjectivity in mind.

4. Findings

4.1 Psychometric analyses

Success Wheel Measure

The parallel analysis indicated that the number of factors in the observed data was 1. The first factor eigenvalue was 2.4 (vs. 0.5 in the resampled data), with a first-to-second eigenvalue ratio of 13.8. The first component eigenvalue was 2.9 (vs. 1.1 in the resampled data), with a first-to-second eigenvalue ratio of 4.2. Confirmatory factor analysis also broadly supported unidimensionality. A one-factor model generated a TLI of 0.98, a CFI of 0.99, and a RMSEA of 0.09 [95% confidence interval: 0.06–0.10]. Finally, the Velicer MAP test for the SWM achieved a minimum average partial square of 0.08 at 1 factor.

Reliability analyses indicated an alpha of 0.83, an algebraic GLB of 0.87, an omega total of 0.83, a lambda 2 of 0.83, and a mu-2 of 0.83. All coefficients exceeded the acceptable threshold of 0.80. Item-to-total correlations (Table 2) ranged from 0.54 to 0.60 indicating that all items contribute positively to the SWM total score (no non-relevant items). Item-to-item correlations ranged from 0.34 to 0.52 indicating that items were not unduly correlated.

Table 2: Psychometric findings for SWM items

Item	CFA factor loading	Item-to-total	IRT difficulty parameters				IRT DA
			***b*1**	***b*2**	***b*3**	***b*4**	
MLP	0.72	0.57	-2.80	-0.89	0.80	3.02	1.86
HT	0.73	0.60	-2.72	-0.88	0.70	2.67	2.05
HSI	0.67	0.54	-2.55	-0.80	0.92	2.68	1.69
PR	0.68	0.55	-2.95	-0.65	1.19	3.39	1.61
SOP	0.67	0.54	-3.27	-1.37	0.27	2.39	1.60

Note: DA = discriminatory ability.

We judged these findings to provide enough evidence for an essentially unidimensional structure and to provide confidence that item response theory findings can be interpreted

confidently. An IRT graded response model indicated that all SWM items had high or very high discriminatory ability. Difficulty parameters ranged from -2.55 to -3.27 for the first threshold, -0.65 to -1.37 for the second, 0.27 to -1.19 for the third, and 2.39 to 3.39 for the fourth, indicating limited variation between items in terms of difficulty (See Table 2).

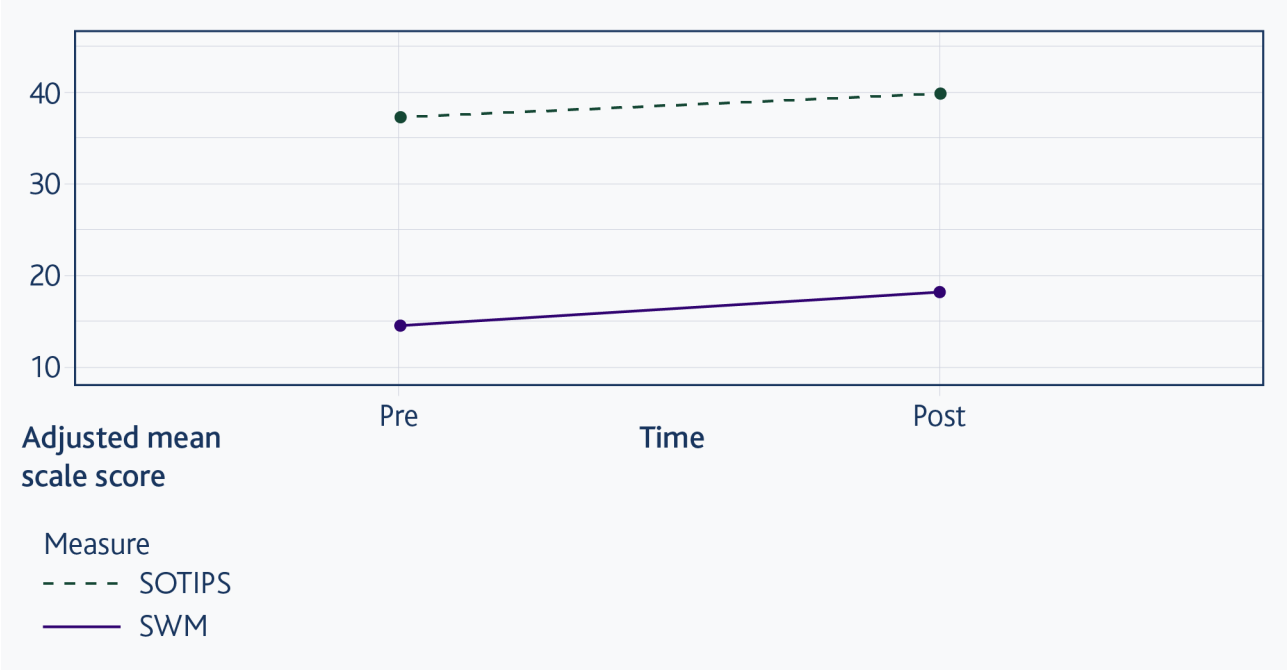
Table 3: Results of an Analysis of Variance (ANOVA) for convergence between measures

Model	df	AIC	BIC	logLik	X^2	*p*
Null	5	4769.1	4790.9	-2379.5		
Time	6	4763.8	4790.1	-2375.9	7.23	0.007
Measure	8	3407.8	3438.4	-1696.9	1358.06	<.0001
Time x Measure	10	3407.7	3442.7	-1695.8	2.09	0.148

R^2 = .88. AIC = Akaike's Information Criteria; BIC = Bayesian Information Criteria; logLik = Log-likelihood.

Independent effects of time (χ^2 (6) = 7.23, p = .007) and scale (χ^2 (7) = 1358.07, p < .0001) were observed (see Table 3). These models indicating that, all else held constant, post-Horizon scores were, on average, higher than post-Horizon scores (r = 0.64) and that SWM scores were, on average, higher than SOTIPS scores, since they are on different scales after SOTIPS scores were inverted (r = 0.97). However, the interaction between time and measure was not statistically significant, (χ^2 (8) = 2.09, p = 0.148): the size of the positive pre-to-post effect did not differ between the measures (see Figure 1).

Figure 1: Interaction effect of Time and Measure in the validation sample



The Kendall's rank correlation tau indicated a statistically significant association between SWM and SOTIPS total scores ($\tau = 0.26$, $z = 4.36$, $p < 0.0001$). This association was supported by a bootstrapped tau ($\tau_{boot} = 0.26$, standard error = 0.06, 95% confidence intervals [0.15, 0.37]). This indicates a moderate association between total scores.

Horizon Motivational Screen

Structural analysis for the HMS was conducted using the same tests as for the SWM. The parallel analysis indicated that the number of factors in the observed data was 1. The first factor eigenvalue was 2.42 (vs. 0.10 in the resampled data), with a first-to-second eigenvalue ratio of 23.4. The first component eigenvalue was 2.79 (vs. 1.11 in the resampled data), with a first-to-second eigenvalue ratio of 4.9. Confirmatory factor analysis also broadly supported essential unidimensionality, where a one-factor model generated a TLI of 0.99, a CFI of 0.99, and an RMSEA of 0.07 [95% confidence interval: 0.03–0.12]. The Velicer MAP test achieved a minimum average partial square of 0.15 at 1 factor.

Reliability analyses indicated an alpha of 0.87, an algebraic GLB of 0.91, an omega total of 0.87, a lambda 2 of 0.87, and a mu-2 of 0.87. All coefficients exceeded the acceptable threshold of 0.80. Item-to-total correlations ranged from 0.50 to 0.65, above the threshold of 0.20, indicating all items contribute positively to total scores. Item-to-item correlations ranged from 0.37 to 0.62, all lower than an acceptable threshold of 0.80 (Table 4).

Table 4: Psychometric findings for HMS items

Item	CFA factor loading	Item-to-total	IRT difficulty parameters		IRT DA
			b1	*b2*	
Enthusiasm	0.85	0.63	-1.54	0.26	3.10
Direction	0.85	0.65	-1.23	0.32	3.02
Commitment	0.75	0.55	-2.66	-0.20	1.90
Holistic	0.64	0.50	-1.98	0.49	1.42

Note: DA = discriminatory ability.

A graded response model indicated that all SWM items had high or very high discriminatory ability (i.e., were able to successfully discriminate between those with different levels of the latent dimension). Difficulty parameters ranged from -2.66 to -1.23 for the first threshold, and -0.20 to 0.49 for the fourth indicating little variation between items in

terms of difficulty, but also that the commitment item was “easier” (i.e., it did not take as much of the latent trait to receive a score of “2” on that item) (see Table A4).

5. Conclusions

The findings of this exploratory study provide early positive evidence for the quality of both the Success Wheel Measure and the Horizon Motivational Screen. However, caveats to this raised in the limitations section apply and should be considered. Specifically:

- Both the SWM and the HMS appear to measure one dominant dimension, albeit psychometric indices do not identify or confirm the qualitative nature of that dimension (i.e., whether it is indeed “strengths” or “motivation”, as expected).
- The reliability of both measures appears to reach acceptable standards, in terms of the observed ratings across different administrations being consistent.
- No SWM or HMS items appear to be replicating one another and all appear to contribute positively to total scores. These findings provide confidence that items on both scales are independent but similarly relevant to their respective latent trait.
- All items on both measures appear to be able to correctly and strongly distinguish between individuals who are higher and lower in their respective latent traits.
- However, the difficulty parameters for the HMS suggest that the test may be too “easy”: it appears to take very little motivation to get higher scores. This is supported by excessive skew towards higher scores. In particular, a participant with overall moderate motivation will likely have received a maximum score for commitment. While it is plausible that this is because programme participants are a highly motivated sample, this is worthy of further investigation as it may be that definitions or scoring guidelines could be improved to improve accurate judgement.

In conclusion, the evidence supports the assertion that the SWM and the HMS are reasonably valid measures but identifies potential improvements that can be investigated.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). The standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Baker F.B., & Kim S. H. (2004). Item response theory: Parameter estimation techniques (2nd ed). Boca Raton, FL: CRC Press.

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.

Carmines, E. G., & Zeller, R. A. (1979). Reliability and validity assessment. Newbury Park, CA: Sage.

Cho, E. (2022). The accuracy of reliability coefficients: A reanalysis of existing simulations. Psychological Methods. Advance online publication. Retrieved June 8, 2022 from https://psycnet.apa.org/doiLanding?doi=10.1037%2Fmet0000475

De Ayala, R. J. (2009). Methodology in the social sciences. The theory and practice of item response theory. New York, NY, US.

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

Engelhard, G., Jr. (2013). Invariant measurement. New York, N.Y.: Routledge.

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.

Hinkle, D. E., Wiersma, W. & Jurs, S. G. (2003). Applied statistics for the behavioral sciences (5th Ed.). Boston, MA: Houghton Mifflin.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179–185.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1–55.

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.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? Organizational research methods, 9(2), 202–220.

Lehman, R. S. (1991). Statistics and research design in the behavioural sciences. Belmont, CA: Wadsworth.

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

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.

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.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. Psychological Methods, 23(3), 412–433.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. Journal of educational statistics, 18(1), 41–68.

Nunnally, J. C. & Berstein, I. H. (1994). Psychometric theory (3rd Ed.). New York, N.Y.: McGraw-Hill.

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.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207–230.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. Psychometrika, 74(1), 145–154.

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

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74(1), 107.

Slocum-Gori, S. L., & Zumbo, B. D (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. Social Indicators Research, 102, 443–461.

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.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. Psychometrika, 41(3), 321–327.

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.

Westfall, Peter H. (2014), Kurtosis as peakedness: 1905 – 2014. R.I.P. The American Statistician, 68(3), 191–195,

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.