Recidivism and Static-99R risk assessments in a Swedish cohort of mentally disordered sexual offenders

Christian Baudin

Centre for Ethics, Law and Mental Health

Department of Psychiatry and Neurochemistry

Institute of Neuroscience and Physiology

Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2021

Cover illustration by Pia Moberg

SVANENMÄRKET

Trycksak
3041 0234

To my father, who could not quite make it

ABSTRACT

Mentally disordered sexual offenders (MDSOs) constitute a heterogeneous offender subgroup presenting a wide range of clinical comorbidity. The recidivism rates for such a disparate subgroup vary substantially, which complicates estimations of recidivism risk. The commonly used risk assessment instrument Static-99R has been tested on MDSOs, but it is uncertain how it performs in a Swedish population. The aims of this thesis were to (1) present a Swedish cohort of MDSOs in detail regarding cohort characteristics and recidivism pattern over approximately 20 years of follow-up, (2) examine the predictive qualities of the Static-99R in the cohort, and (3) analyze the psychometric properties of the Static-99R using modern test theory. The cohort comprised 146 convicted adult male sexual offenders who underwent a forensic psychiatric investigation (FPI) as part of their legal process in Sweden between 1993 and 1997. Data were collected from official registries, written court documents, and the FPI reports. Prevalence rates for major mental illnesses were substantial, and most were previously convicted—one third for prior sexual offenses. Half were reconvicted over the follow-up period, though primarily for non-sexual offenses. Recidivism rates for sexual offenses were 9.6% after 5 years, 13.0% after 10 years, 15.1% after 15 years, and 17.1% after 20 years. Recidivism was primarily clustered around the first 5 years after the index offense, with an average of 6.5 years until first new sexual or violent offense. Static-99R performed adequately over 5–20 years of fixed follow-up after release (AUC .73-.79) and the cohort was most similar to the "routine sample" norms when estimating recidivism risk, although with 30% fewer recidivists predicted than expected across all risk categories. The Static-99R demonstrated several psychometric issues, in particular items 1, 3 and 5. This was improved by experimenting with fewer response categories for items 1 and 5, and the exclusion of item 3. In conclusion, the cohort of MDSOs demonstrated internationally comparable recidivism rates as well as high rates of major mental illness. Additionally, the Static-99R may be administered to MDSOs in Sweden, using the "routine sample" norms when estimating recidivism risk, and items 1, 3, and 5 may merit highlighting in future revisions of the instrument concerning this particular subgroup.

SAMMANFATTNING PÅ SVENSKA

Den här avhandlingen undersöker en grupp vuxna sexualbrottsförövare med allvarlig psykiatrisk problematik, huvudsakligen i form av en eller flera psykiatriska diagnoser såsom personlighetssyndrom, psykossjukdom och substansbrukssyndrom. För att kunna minska risken för återfall behövs både en djup kunskap om förövargruppen man arbetar med och ett fungerande riskskattningsinstrument. Ett vanligt förekommande riskskattningsinstrument är Static-99R som består av tio frågor som berör kända riskfaktorer för återfall i sexualbrott. Det är i dagsläget oklart om Static-99R kan användas för sexualbrottsförövare med allvarlig psykiatrisk problematik i Sverige. Syftet med avhandlingen var därför att (1) beskriva förövarkaraktäristika och återfallsmönster hos en kohort sexualbrottsförövare med allvarlig psykiatrisk problematik, (2) utreda hur väl Static-99R presterar i just den här förövargruppen och (3) djupare undersöka de psykometriska egenskaperna hos Static-99R i förhållande till denna specifika grupp.

Förövarna visade sig i hög utsträckning vara belastade med allvarliga psykiatriska diagnoser och försvårande bakgrundsfaktorer och de flesta var tidigare dömda för både sexual- och våldsbrott. Hälften återföll under uppföljningsperioden—huvudsakligen för icke-sexuella brott. Återfalls-frekvenserna för sexualbrott var 10%, 13%, 15% och 17% mätt över 5, 10, 15 och 20 års uppföljning. Återfall var vanligast runt femårsstrecket och genom-snittstid till första återfall i sexual- eller våldsbrott var 6,5 år efter det så kallade indexbrottet. Static-99R kunde förutse nytt återfall i sexualbrott över 5–20 år efter frigivning med acceptabel precision och den normgrupp som passade bäst för att uppskatta återfallsrisk var den som av instrumentets utvecklare kallas "routine sample". Vid en djupare analys så uppvisade Static-99R vissa psykometriska svagheter—huvudsakligen frågorna 1 (ålder vid frigivning), 3 (fälld för någon åtalspunkt innefattande icke-sexuellt våld vid indexbrottet) och 5 (förekomst av tidigare sexualbrott). Resultaten förbättrades genom att slå ihop svarsalternativ för fråga 1 och 5 samt ta bort fråga 3 från instrumentet.

Sammantaget visar avhandlingen att kohorten sexualbrottsförövare med allvarlig psykiatrisk problematik visserligen är heterogen och kliniskt belastad, men att den inte avviker på ett sätt som förhindrar förövarkategorin från att riskbedömas med Static-99R. Givet att fynden i avhandlingen reproduceras i framtida studier med större stickprov så kan kommande revisioner av Static-99R med fördel fokusera särskilt på att förenkla frågorna 1 och 5 samt utvärdera huruvida fråga 3 kan behöva förändras.

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by Roman numerals.

I.   Baudin, C., Nilsson, T., Wallinius, M., Sturup, J., & Andiné, P. (2020). A 24-year follow-up study on recidivism in male mentally disordered sexual offenders with and without psychotic disorders. *Journal of Forensic Sciences*, 65(5), 1610–1618. https://doi.org/10.1111/1556-4029.14327

II.   Baudin, C., Nilsson, T., Sturup, J., Wallinius, M., & Andiné, P. (2021). A Static-99R validation study on individuals with mental disorders: 5 to 20 years of fixed follow-up after sexual offenses. *Frontiers in Psychology*, 12, article 625996. https://doi.org/10.3389/fpsyg.2021.625996

III. Baudin, C., Grimby-Ekman, A., Nilsson, T., Wallinius, M., & Andiné, P. Rasch analysis of Static-99R in a cohort of individuals with mental disorders convicted of a sexual offense (manuscript).

ACKNOWLEDGEMENTS

Fredrik Wagnström for never doubting my capability when I considered giving up during my most stressful times. All of you shared my load on parts, or the entirety, of this journey, and for this I am thankful.

My last, and most personal paragraphs, I dedicate to those closest to me. My family, all of whom contributed in their own ways: My beloved partner **Fanny Kuparinen** whom I met around the time I started my studies. You started out an amazing person and never quite stopped. My dear mother **Beatrice Baudin** who strived to keep me relaxed and comfortable during my visits and tried her best to read my papers. My greatly missed father **Sigurd Baudin** who passed away during my studies. You are still my role model and I wish I had told you when I had the chance. My wiser sister **Alexandra Baudin**, who taught me to read and put up with my awkwardness growing up. My curious nephew **Didrik Baudin** who is much cooler than I was his age. Soon enough you will be too grown up to ask your uncle so many questions, and I will miss the old days.

Finally, for my dear hobbits, some of whom have known me for well over a decade, I present to you the following in-group non sequitur: *catface*. You know who you are, and I hope you know how precious you all are to me. Oh, and to my good old army buddy: Only now do I understand why you would insist on describing your own doctoral studies as the "most fun" you have ever had.

CONTENTS

ABBREVIATIONS

AUC   Area Under the Curve

COMPAS   Correctional Offender Management Profiling for Alternative
               Sanctions

DIF   Differential Item Functioning

DSM   Diagnostic and Statistical Manual of Mental Disorders

DSM-III-R   Diagnostic and Statistical Manual of Mental Disorders, 3rd
               edition, Revised

DSM-IV   Diagnostic and Statistical Manual of Mental Disorders, 4th
               edition

E/O   Expected by Observed

FPI   Forensic Psychiatric Investigation

HART   Harm Assessment Risk Tool

ICC   Intraclass Correlation Coefficient

J   Youden index

M   Mean

Mdn   Median

MDSO   Mentally Disordered Sexual Offender

N   Study cohort

n   Study cohort subgroup

NPV   Negative Predictive Value

OR   Odds Ratio

PCA   Principal Component Analysis

P/E	Predicted by Expected
PPV	Positive Predictive Value
PSI	Person Separation Index
ROC	Receiver Operating Characteristics
RRASOR	Rapid Risk Assessment for Sexual Offense Recidivism
RUMM	Rasch Unidimensional Measurement Models (software)
SACJ-Min	Structured Anchored Clinical Judgment Scale-Minimum
SD	Standard Deviation
SPSS	Statistical Package for the Social Sciences (software)

1 INTRODUCTION

1.1 Sexual offending and mental disorders

During their lifetime, it is estimated that 20% of US women have been raped and twice as many have been the victim of other types of contact sexual violence (1,2). Sixty percent of the victims endure this tragedy in adulthood (1), but almost half of those who are victimized sexually during childhood are revictimized later in life (3). While the traumatic burden of being sexually violated is immeasurable, the socioeconomic lifetime cost for all adult US victims is an estimated 3.1 trillion dollars—a mere third of which is paid by government sources (4). Comparable extensive studies in Sweden are lacking, but the yearly national survey on how criminal offenses affect people living in Sweden paints a similar picture. For women aged 20–24, half report serious concern about being a victim of sexual violence, which is only slightly higher than the 30% of women who report being actually victimized (5). Women of all ages report less concern and less victimization than those aged 20–24, but only marginally so and still roughly ten times that of men (5). In short, sexual violence affects us all, whether by doing harm to our personal, physical, or mental well-being, or by fracturing the social and economic fabric of society (6).

Recidivism rates tend to vary depending on the type of offense and the characteristics of the offender in question. General sexual offenders found in prison populations tend to reoffend sexually at rates of 4–25% over 5 years, 7–35% over 10 years, and 24–45% over 15 years (7–12). Unfortunately, it is difficult to summarize and present a consistent picture of the pattern of recidivism, mainly due to large variations in definitions used, populations examined, and follow-up length (13). Most studies present follow-up periods of 4–5 years, with rare and notable exceptions (8,14).

Mentally disordered sexual offenders (MDSOs) differ from general sexual offenders by being a heterogeneous offender category presenting a wide range of psychiatric diagnoses. As a group they represent most major mental illnesses found in the general population, such as psychotic disorders, mood disorders, substance use disorders, personality disorders, and paraphilic disorders (15–18). Considering their large variance in clinical composition, recidivism studies on MDSOs are markedly affected by these methodological

shortcomings. Consequently, data is even more inconsistent for this offender subgroup, particularly over longer follow-up periods. Two of the longest—but not largest—studies found recidivism rates of 17–47% over 20–25 years, but the subjects' diagnoses were not explicitly specified by the authors (14,17). Shorter studies targeting sexual offenders with psychotic disorders in particular have found recidivism rates of 6–34% over 5–8 years (19,20).

1.2 Legal prerequisites for mentally disordered offenders in Sweden

In contrast to most jurisdictions, an offender cannot be found unaccountable for his or her crime under Swedish law. Nationally, this is a highly debated condition that has engaged clinicians, legal experts, and politicians for several decades. In recent years, two official reports by the Swedish Government have been published, recommending changes with regard to criminal responsibility in Sweden, but concrete legislative changes have yet to materialize (21,22). Instead of criminal unaccountability, the medicolegal term "severe mental disorder" has been in use since 1992 (23). Whether or not an offender has a severe mental disorder has a major impact on the type of sanction available to the court for adjudication, particularly because it may preclude the offender's imprisonment. Instead, these offenders are sentenced to compulsory forensic psychiatric care, although exemptions were put in place in 2008 permitting the court more leeway in making their decision. Additionally, for particularly high-risk offenders with a severe mental disorder, the court may stipulate a special court supervision to protect the public. Essentially, the special court supervision adds an additional step in the process of releasing an offender from compulsory in-patient forensic psychiatric care by requiring the approval of an Administrative Court[1] in addition to the opinion of the medically responsible psychiatrist.

There are no clearly defined diagnoses or mental states that inherently fulfill the requirements for a severe mental disorder, and, since its conception, preparatory works have made suggestions to aid clinicians in adhering to the intent of the law. This has resulted in a range of psychiatric conditions in which psychotic disorders are the primary indicator of a severe mental disorder, but other major mental illnesses such as severe neurodevelopmental disorders and severe personality disorders with psychotic episodes or compulsive elements

---

[1] Förvaltningsrätt (Swedish)

are also to be considered on a case-by-case basis. Ultimately, it is up to the clinicians to present their arguments to the court regarding why a particular condition or diagnosis may or may not fulfill the criteria of a severe mental disorder.

In Sweden, these decisions are based on court-ordered pretrial forensic psychiatric investigations (FPIs) conducted by the National Board of Forensic Medicine (24). This government body was formed in 1991 and comprises several branches that are responsible for a wide range of medicolegal investigations. One of the branches is the forensic psychiatric department, primarily located in Sweden's two largest cities: Stockholm and Gothenburg. These two units handle all FPIs conducted in Sweden, and the investigations are usually carried out by specially trained clinical teams consisting of a forensic psychiatrist, a forensic clinical psychologist, a forensic social worker, and ward staff. Any diagnoses must follow the latest version of the Diagnostic and Statistical Manual of Mental Disorders (DSM) classification system, which is version 5 at the time of writing (25).

1.3   Actuarial risk assessments of offenders

"The best predictor of future behavior is past behavior." This quote has been attributed to a range of psychologists—most commonly Edward Thorndike (1874–1949)—and was expanded upon by clinical psychologist Paul Meehl in 1954 with regard to the use of statistical predictions of behavior such as criminal recidivism (26). While the nature of the quote may invoke a bleak view of mankind and the philosophical elements of free will, it is the cornerstone of any actuarial risk assessment instruments in use today.

Although the first assertive steps towards the use of actuarial risk assessment instruments were taken by urban sociologist Ernest Burgess in the form of a large-scale study in 1928 (27), from the 1960s and onward—after the publication of Meehl's seminal work—actuarial risk assessment instruments started seeing consistent use by clinicians working with offenders (28). These instruments are but one of several risk assessment methods and their goal is to predict recidivism by linking a particular tendency to reoffend to a particular set of so-called predictors. These predictors have essentially been sifted out by conducting a large number of recidivism studies over the decades measuring the effect of a myriad of variables assumed to affect the tendency of an offender to commit a new offense. Some of these variables, for example age at criminal

onset, a general antisocial pattern, or substance use, have repeatedly been linked to recidivism for many types of offenders and offenses (8,17,20,29,30), while results are less conclusive for other variables, such as various types of mental illnesses (8,31). This is generally as expected, as predictors tend to have different effects depending on type of offense and offender subgroup. Some predictors are highly specific for a type of offense, for example sexual deviancy for predicting sexual recidivism (8,30,32)

The premise of actuarial risk assessment encompasses the use of empirically derived predictors in an attempt to quantify the offender's risk of reoffending. A specific value (score) is assigned to one or more predictors (items) depending on their expected impact on the probability of reoffending (recidivism risk). Larger values indicate large effects on the probability of reoffending and smaller values indicate small effects. Next, all offenders in the sample used (the norm group) are followed for a specific length of time (usually a few years) and any instances of recidivism are linked to a specific total score. A higher total score indicates a higher risk for recidivism, often presented as a percentage of the total sample. For example, if 100 persons with a total score of 10 in a sample of 1000 reoffended during the follow-up, the recidivism risk for any other offender with a total score of 10 is predicted to be 10%. If the total score range is too high, say a score of 0 to 100, there may be far too few recidivists per total score to draw any conclusions regarding these particular subjects' recidivism risk. This is especially glaring if there are no recidivists at all for a certain total score. Where there are no recidivists, there is nothing to measure, and while a 0% recidivism risk is mathematically possible, it is highly unlikely in a real-life scenario. To remedy this, all offenders within a certain total score range may be aggregated into larger groups (risk levels or risk categories). A specific risk category can then be linked more accurately to a particular probability of reoffending.

Since the instruments are designed to be prognostic rather than diagnostic—essentially attempting to predict the future—and use conclusions drawn at group level and apply them to specific individuals, the reasoning behind the clinical use of risk assessment instruments is inductive by nature (33). This is not necessarily bad, but it is an important distinction to make for those involved. There is no guarantee, logically speaking, that an offender will or will not reoffend after undergoing a risk assessment—there is only a probability of reoffending induced from a sample of (hopefully) similar offenders. Whether or not this probability is considered large enough to cause

concern essentially boils down to academically and professionally supported subjectivity, with the decision-making entrusted to the clinicians and the courts. No one "almost reoffends"—one either does or does not. What all parties involved strive to attain is the highest degree of certainty given the circumstances.

1.3.1   Risk assessments of offenders in Sweden

Generally speaking, risk assessments of mentally disordered offenders occur in three distinct forms: a) for offenders currently serving life in prison who *may or may not be* mentally disordered, b) for offenders already sentenced to forensic psychiatric care that *are* mentally disordered, and c) for offenders undergoing an FPI where the court is considering stipulating a special court supervision along with the sentence (23). When conducting these risk assessments, a wide range of tools available to the clinicians are generally used rather than actuarial risk assessment instruments exclusively.

Life imprisonment is arguably the gravest legal sanction afforded under Swedish law. One should note that sexual offenses are generally not a focal point of this offender subgroup. Murder is by far the most common offense resulting in a life sentence, with only two out of a total of 232 prisoners between 1995 and 2013 having received a life sentence for a different offense (34). Historically, the only way for a convict to be released after being served a life sentence was to apply for a pardon. Pardons are granted by the Swedish Government on a case-by-case basis and are considered an exceptionally rare and completely opaque, extrajudicial occurrence. However, another option was established in 2006: conversion of a life sentence to a fixed number of years, which nowadays usually translates to roughly 25 years (34,35). With this option, any prisoner who has served at least 10 years of their life sentence may apply for conversion, and as part of the legal process a risk assessment conducted by the National Board of Forensic Medicine is required. Unlike FPIs, the experts handling these risk assessments commonly work in pairs, headed by either a forensic psychologist or a forensic social worker. While this concerns a small number of prisoners (approximately 3% of the total prison population), the numbers have increased since the 1960s and are expected to continue to grow (34). As a result, if the prisoner's application for conversion is denied, he or she is entitled to reapply every 12 months for as long as they wish, which results in a steadily increasing number of risk assessments being conducted every year.

Approximately 8% of male offenders undergoing compulsory forensic psychiatric care committed a sexual offense as part of their index sentence (36). Violent offenses are far more common, committed by more than half of all those sentenced (36). In Sweden, offenders in forensic psychiatric clinics are required to accept treatment until they are deemed psychiatrically well enough to be released. If the offender refuses to participate or fails to improve for other reasons, it is theoretically possible for them never to be released into the public again. Consequently, forensic psychiatric care has the potential to be a life sentence, and sometimes it effectively is. Under these conditions, a lot of trust is put in the psychiatrist medically responsible for the patient in question to correctly establish when the patient is well enough to be released. The exception is in cases where special court supervision has been stipulated, where the psychiatrist must argue on behalf of the patient to the local administrative court in order for them to approve the release. For this offender group, risk assessments are most common when the extent and degree of a patient's privilege are under consideration. For example, if a patient requests ground privileges or leave, a risk assessment is required. In 2019, approximately 70% of forensic psychiatric patients were assessed through some form of risk assessment instrument, although unaided clinical judgment seems to be common practice (36). Considering the amount of time some of these offenders spend in in-patient care, the risk assessments tend to add up.

When an offender is undergoing an FPI where the court is considering stipulating a special court supervision, the resulting decision depends on how much hypothetical danger the offender poses to the community. This is a complicated question, and in order to reduce the impact of chance and bias, risk assessment instruments are used. As with any FPI, the risk assessments are also conducted by the National Board of Forensic Medicine, either administered by the forensic psychiatrist himself or delegated to the forensic psychologist or forensic social worker. Depending on the outcome of the risk assessment instruments used (usually several different types), the team either will or will not recommend that a special court supervision be stipulated, but the court always has the final word. In their legal conception, special court supervision and the associated risk assessments were meant to be an exceptional addition to the FPI. However, it has become highly prevalent for the courts to specifically ask for this addition, and more than 90% of all FPIs conducted yearly now administer these risk assessments (37).

In sum, risk assessments are integral to several areas of the legal process and affect the sanctions adjudicated, but only for a comparatively small number of offenders—many of whom are mentally disordered. However, for offenders sentenced to prison, no risk assessment is required as part of their release.

1.3.2   Issues affecting offender research and risk assessment utility

There are several issues pertaining to sexual offender research. Many of these issues concern a general lack of agreed upon definitions and criteria for various outcomes and observations. Some issues may have a negative impact on the use of risk assessment instruments, primarily by complicating the scoring procedure or any attempts at validating them using new offender subgroups. The most significant examples are considered below.

Recidivism is not a universally agreed upon event. In Sweden, mainly due to the detailed registers kept by the National Council for Crime Prevention, the word generally refers to an entirely new conviction on one or more charges, whereas it is more commonly defined as a simple arrest in the USA and Canada. As expected, equating recidivism with arrests results in a vastly more inclusive definition that does not, in and of itself, indicate guilt. While this decreases underreporting—which is suspected to be highly prevalent in sexual offender research (38)—it increases the likelihood of targeting actual non-offenders. By using data from arrests to generate research on offender characteristics, we may inadvertently misrepresent actual reoffenders. On the other hand, by limiting ourselves to "hard" convictions we may catch reoffenders with greater accuracy but miss reoffenders who, for whatever reason, avoid being sentenced. There is a clear tradeoff between the two approaches, and depending on the purpose of the research, both are valid choices. This makes it more difficult to compare various studies, however.

There is no clear definition of what constitutes a *sexual offense*. While most studies accurately describe the criminal offenses included in the term, it still hampers comparisons with other studies. Swedish studies almost always refer to Chapter 6 of the Swedish Criminal Code[2] (39), which specifically targets sexual offenses, but even this may be inaccurate. The consumption (not the production) of child pornography, which may or may not be considered a

---

[2] Titled "On sexual offences"

sexual offense, is discussed outside[3] of Chapter 6, for example. The lack of a universal definition of sexual offense is particularly glaring when you consider that some jurisdictions do not consider marital rape a criminal offense, or that the age of sexual consent varies globally from 11 to 21 years of age (or is entirely nonexistent) (40).

Mental disorder is a term describing a broad and fairly abstract notion including a great variety of clinical diagnoses and symptoms (15–18). Of course, one may argue there is a categorical difference between a sexual offender suffering from schizophrenia and one with attention-deficit hyper-activity disorder, but with no clear boundaries as to what does or does not constitute an MDSO, it is yet again difficult to compare studies and aggregate research results properly. Additional confusion ensues when some juris-dictions unintentionally introduce strictly legal terms such as "sexually violent predators" and "mentally disordered offenders" in California or "severe mental disorder" in Sweden, in which clinical and legal discourse essentially blend together.

Lastly, *population* is, comparably, a well-defined term which in this context refers to the specific offender groups and subgroups that a risk assessment is targeted at. Sexual offenders are a heterogeneous population, which has been discussed academically since at least the mid-eighties (41). They may differ vastly with regard to age, personal and criminal background, victim preference, personality, modus operandi, level of sexual arousal, clinical diagnoses, religious belief, ethnicity, recidivism rates, and more (41–52). One risk assess-ment instrument may not perform as well in one population as in another, despite similarities in jurisdiction or culture. Instead, new populations need to be sufficiently calibrated to the norm group for comparisons of recidivism risk to be reasonable.

1.4   The Static-99R risk assessment instrument

Since it was first created in 1999, the Static-99, and its successor Static-99R (53,54), has become the de facto gold standard for risk assessment instruments predicting the recidivism risk for sexual offenders across the globe (55,56). While the original Static-99 performed well, a decade of researching field validity indicated that some changes were desirable, resulting in the release of

---

[3] In Chapter 16, "On offences against public order"

the Static-99R in 2009. In addition, many minor revisions have occurred throughout the years, as seen in Figure 1. The original Static-99 has been cross-validated using a Swedish prison population (57), and a large Swedish sample was included when the norm groups were revised (54,58). However, the current revision of the Static-99R is yet to be validated on a Swedish population, and it is the recommendation of the developers that only the latest revision be used in clinical settings (53).

Interrater reliability for the Static-99R is high for both experienced and unexperienced users, probably since it focuses solely on static risk factors and requires very little formal training of the assessor (53). The instrument consists of 10 separate items that award between -3 and 3 points, with most items simply awarding 0 or 1. The total score varies between -3 and 12 points and reflects a particular nominal risk category (sometimes described as a risk level) out of the five available: I (very low risk), II (below average risk), III (average risk), IVa (above average risk), and IVb (well above average risk)[4].

The Static-99R was originally developed using follow-up periods of 5–10 years (54), so most validation and replication studies focus on this timeframe (53,59–61). However, a recent study by the developers themselves suggests a model for statistically estimating lifetime and residual risk (62). This model further extends the use of the instrument over expansive timeframes, but recidivism studies with fixed follow-up periods longer than 10 years are still rare due to the resources required.

The construct validity of the Static-99R has previously been explored by Brouillette-Alarie, Babchishin, Hanson, and Helmus using factor analysis (63), who found three dimensions linked to sexual recidivism: youthful stranger aggression, persistence/paraphilia, and general criminality. This was later replicated by Brouillette-Alarie, Proulx, and Hanson with similar results (64). These findings build upon previous discussions regarding other actuarial risk assessment instruments where two dimensions are generally linked to sexual recidivism: sexual deviancy/sexual criminality and antisocial behavior/general criminality (63). MDSOs have not been specifically targeted by previous research, and, to our knowledge, no studies have explored underlying

---

[4] See Appendix A for an example of the Static-99R coding sheet

theoretical dimensions of the Static-99R separately from observed recidivism, that is the total score demonstrated by a subject whether they reoffend or not.

Regarding the essential purpose of the Static-99R—predicting recidivism—it performs moderately well (53). The Static-99R tends to produce an AUC of .69–.70, but this can vary depending on how analyses are made (53,65). The AUC—referring to the area under the curve (AUC) of receiver operating characteristics (ROC) analyses—is probably the most highly prevalent measurement of how well a risk assessment instrument discriminates between recidivists and non-recidivists (66,67). The AUC can be described as a statistical index demonstrating the likelihood that the total Static-99R score of a recidivist is higher than that of a non-recidivist (68,69), and a major advantage of the index is that it is uninfluenced by the low base rates endemic in sexual offender research in general (68,70). The AUC may be negatively affected by the size of the total score range, however (71). A value of .50 means there is a 50% probability, which is identical to random chance, whereas 1 indicates a perfect positive prediction and 0 a perfect negative (68). Naturally, a risk assessment instrument must perform substantially better than random chance and most instruments end up with an AUC of approximately .65–.70 at best in clinical practice.

Two norm groups comprising samples from seven countries on two continents are available for comparison when estimating recidivism risk (54), but there are no established methods for deciding which norm group is preferable in a specific population, despite the choice greatly affecting the outcome (72,73). The norm group referred to as "routine sample" (called the "routine norm group" in this thesis for readability) is considered a catch-all for general sexual offenders, whereas the norm group referred to as "high-risk/high-need sample" (called the "high-risk/high-need norm group" in this thesis) was designed for especially violent sexual offenders and MDSOs. Consequently, the Static-99R allows for use with the MDSO subgroup, but it is unclear if this is applicable to a Swedish setting.
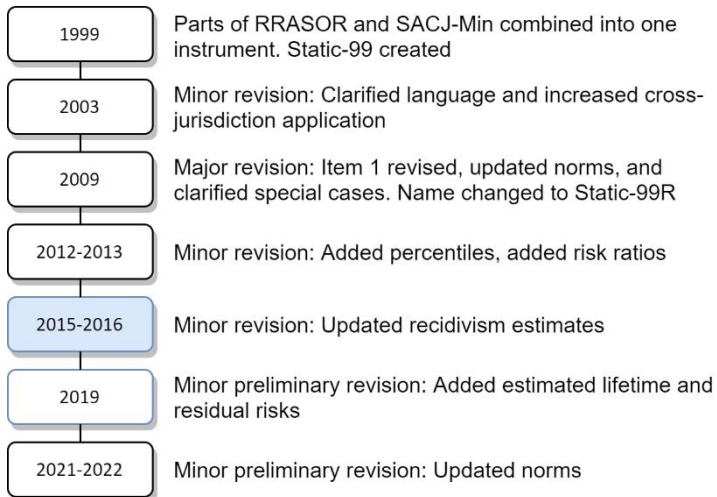
Figure 1. Development history of the Static-99R from the original 1999 release and onward. The blue square denotes when the latest complete version of the coding rules were published.

2 AIMS

The overarching aim of this thesis is to evaluate the utility of the Static-99R risk assessment instrument in a cohort of MDSOs and whether said offenders differ markedly from the Static-99R target group. This is examined through three separate studies, each focusing with increased detail on a specific area of both the cohort and the Static-99R. This is achieved firstly by presenting the cohort in detail with regard to the offender characteristics and recidivism patterns of the cohort demonstrated in Study I; secondly, by examining the predictive qualities of the Static-99R via a validation study using said cohort in Study II; and, thirdly, by examining the psychometric properties of the Static-99R using the coding sheets and risk levels already established for subjects from said cohort in Study III.

2.1 Specific aims

I. Establish and present a) offender characteristics—clinical ones in particular—of MDSOs and b) patterns of recidivism over up to 24 years of follow-up regarding both new sexual offenses and violent offenses of said cohort.

II. Determine the predictive qualities of the Static-99R in MDSOs with regard to a) how well the instrument discriminates between recidivists and non-recidivists, b) how well the instrument is calibrated in relation to the specific cohort, and c) whether the predictive qualities are maintained over longer timeframes than the typical five-year follow-up period.

III. Explore the construct validity of the Static-99R when assessing MDSOs with regard to a) how well the underlying trait measured by the instrument is unidimensional and quantifiable and b) whether there are any areas of the Static-99R that merit highlighting in future revisions of the instrument.

3 SUBJECTS AND METHODS

3.1 The cohort

The same cohort was studied in all three studies and comprised all males in Sweden who between January 1, 1993 and December 31, 1997 were obliged to undergo a court-ordered FPI during their trial for a sexual offense (N = 146). In short, all subjects committed a sexual offense at some point before or during 1997, were charged with the offense, underwent the FPI between the dates stated above, and were subsequently convicted shortly thereafter. The most common index offense sanction was prison (44.5%, $n = 65$), followed by forensic psychiatric care (43.2%, $n = 63$), and probation (12.3%, $n = 18$). Some offenders were technically convicted in 1998 and some committed the offenses as early as the 1980s, but the inclusion criteria focused on the date of the FPI. For a visual presentation of the cohort across the three studies, see Figure 2.

All index sexual offenses were committed against a victim aged 15 years or older, which is the age of criminal responsibility and sexual consent in Sweden, and almost all of them were female (97.3%). While all index offenses consisted of one or more sexual charges, additional charges of different offense types were also possible. This included both non-sexually violent acts and general criminal conduct such as theft, drug offenses, and traffic offenses. However, at least one charge of the index offense sentence was listed under Chapter 6 of the Swedish Criminal Code (39), with some subjects producing two (12.3%, $n = 18$) or even three (1.4%, $n = 2$) sexual offense charges during their index offense. More than half the cohort (53.4%, $n = 78$) were charged and convicted of a non-sexual violent offense in addition to the sexual offense, and 28.0% ($n = 41$) for a violent and a general offense as well.

By far the most common index sexual offense charge was rape (62.3%, $n = 91$), followed by sexual molestation (28.8%, $n = 42$), and sexual coercion (13.0%, $n = 19$). Sexual exploitation was a slightly rarer occurrence (4.8%, $n = 7$). Additionally, one offender (0.7%) was convicted of sexual intercourse with a child under the age of 15 but also targeted adult victims during the same index offense and was therefore included. *Rape*, *sexual molestation*, *sexual coercion*, *sexual exploitation*, and *sexual intercourse with a child*[5] are all

---

[5] Våldtäkt, sexuellt ofredande, sexuellt tvång, sexuellt utnyttjande, sexuellt umgänge med barn (Swedish)

Swedish legal terms translated into English using a glossary published by the Swedish National Courts Administration (74). Note that the definitions may or may not align with how the terms are commonly used by the general population and that they may have changed since 1993.

The cohort was demographically diverse, with a majority originating from Sweden (62.3%, $n = 91$) or another Nordic country (8.2%, $n = 12$), and 8.9% ($n = 13$) being born somewhere else in Europe. The remainder of the subjects, comprising about one fifth of the cohort (20.5%, $n = 30$), originated from outside of Europe. The subjects' age also varied greatly, ranging from 17 to 71 years ($Mdn = 34.0$, $M = 35.3$, $SD = 8.42$). Additionally, one in four subjects had immigrated to Sweden in adulthood, resulting in uncertain data regarding their childhood (26.7%, $n = 39$).

A majority of the subjects still participated at the end of the follow-up period on December 31, 2016 (59.6%, $n = 87$, Mdn age = 54.7, $M = 55.6$, $SD = 8.7$), which ranged from 6 months to 24 years ($Mdn = 20.2$, $M = 17.1$, $SD = 6.9$). Seventeen subjects (11.6%) were followed the entire period from the start of inclusion in 1993 up until 2016. Of those who dropped out during the follow-up, roughly twice as many had died (28.1%, $n = 41$) compared to those who had migrated from Sweden (12.3%, $n = 18$).
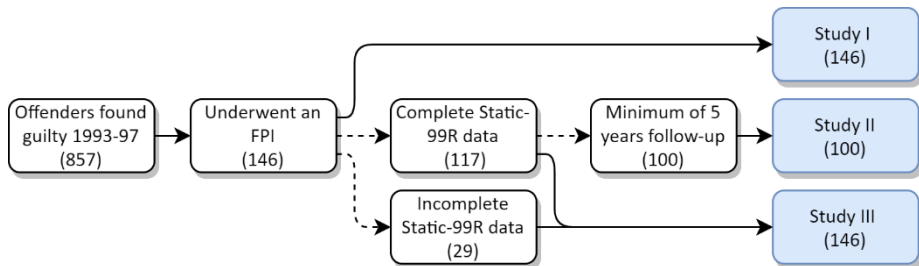
Figure 2. Flowchart visualizing the study cohort included in the three studies. Figure adapted from Figure 1 in both Studies I and II.

3.2 Data processing procedures common for all three studies

Since a single cohort comprising 146 subjects was used in all three studies, large parts of the procedures were interconnected. In brief, each subsequent

study observed and analyzed previously collected data in a different and incrementally more detailed way.

FPI reports were collected from the National Board of Forensic Medicine (including its Central Archive), and the original sentence documents were collected from each individual court, principally by referring to the registers of the National Council for Crime Prevention. This usually concerned District Courts,[6] but in some cases the Courts of Appeal[7] were also involved. In all cases where a subject appealed to the Supreme Court[8]—a quite common occurrence—not a single one was accepted by the court. Dates of either the emigration or the death of a subject were collected from the Swedish Tax Agency's population registry. Release dates from prison and forensic psychiatric care were collected from the Swedish Prison and Probation Service and from each clinic individually.

In Sweden, a unique 12-digit personal identification number is assigned at birth or, in the case of someone who migrates here later in life, as soon as the migration process is complete. With some exceptions, the personal identification number not only is unique to each individual but usually also indicates the person's date of birth, gender, and the municipality in which the person was born. Understandably, the personal identification number offers huge advantages in terms of longitudinal studies using Swedish subjects since the risk of losing or otherwise confusing one subject with another is greatly reduced (75). For all studies in this thesis, personal identification numbers were used to link all data sources to each individual subject before being replaced with a code string in a process commonly referred to as pseudonymization. As long as the code key is saved safely and separately from the data, the subjects' identities stay hidden while the opportunity remains to add additional data later in the process, for example by extending the follow-up period. This was crucial to the thesis as some of the data was collected in 1997 while most of the follow-up data was collected in 2008 and 2016.

Several software products were used in this thesis project, primarily the free and open statistical software jamovi, which is built on top of the R statistical coding language. Study I relied on SPSS 25.0, Study II on jamovi 1.2.9 and Stata/SE 16.1, Study III on jamovi 1.2.22 and RUMM2030Plus, and the thesis

---

[6] Tingsrätter (Swedish)
[7] Hovrätter (Swedish)
[8] Högsta domstolen (Swedish)

framework on jamovi 1.6.15 (76–81). Microsoft Excel was used in all studies for minor descriptive statistics and to offer a simple overview of the data for internal use, as well as in Study II to calculate specific measurements which no specific software offers a solution for (82).

The original collection of data was conducted as part of two previous research projects in 1997 and 2002[9], which was expanded upon in both detail and scope for this thesis. No power calculations were conducted at the time of the original research projects. This was likely disregarded for three reasons. Firstly, power calculations were not as prevalent in the late 1990s, particularly in the field of recidivism research. Secondly, since the purpose of the original research projects were to collect data for later use, no null hypotheses or alternative hypotheses were established as part of the designs. Thirdly, since sexual offenders undergoing FPIs in Sweden are so few, no randomly collected sample was being chosen. Instead, all potential subjects available over several years were included in the original research projects. Power calculations are essentially conducted to limit the risks of type II error, e.g., failing to reject the null hypothesis when it is false (as compared to a type I error – rejecting the null hypothesis when it is true). The risk of a type II error can be reduced by having a large enough sample, which is essential when designing projects using random samples of large research populations. This is a prospective action done in the design phase of a research project. Some researchers argue for the use of post-hoc power calculations, that is power calculations conducted after the experiment is completed using the results to retrospectively evaluate if the sample was large enough. Using power calculations in this manner has been hotly debated, and arguments from statisticians have been put forward that it is a problematic and misleading endeavor (83–86). For example, by assuming that an observed (and often noisy) effect is equal to the true effect being studied, the retrospectively calculated power can vary immensely, causing both reader and researcher to mistakenly overemphasize the results (85). However, post-hoc power calculations can be useful when conducting meta-analytic studier (83), and one should always take observed effect sizes into consideration, for example by evaluating the narrowness of the confidence intervals.

---

[9] Please see subheading 3.4 Ethical statements for more details

3.2.1   Offender data

Most of the offender data was collected from the original FPI reports, which contained detailed descriptions of symptoms as well as formal diagnoses. For some subjects, the original sentence documents also provided insight that was withheld or otherwise missing from the FPI reports, for example, through witness statements in cases where the subject refused to talk to the investigating team during the FPI.

Clinical data was categorized according to the DSM, but more than one version was used during the inclusion period. Both the DSM-III-R and the DSM-IV were used between 1993 and 1997 (87,88), with the former used in 82.2% of the cases ($n = 120$) and the latter in 17.8% of the cases ($n = 26$). For the convenience of the reader, all diagnostic categories have been translated from both versions of the DSM into a simpler format. For example, the term "psychotic disorders" used in the thesis is defined as any diagnosis listed in the "schizophrenia," "delusional disorder," and "psychotic disorders not else-where classified" sections of the DSM-III-R and the "schizophrenia and other psychotic disorders" section of the DSM-IV (Table 1).

Table 1. Diagnostic categories used in the thesis and their DSM-III-R and DSM-IV counterparts. Unless otherwise specified, all disorders in the section or subsection mentioned were included.

Diagnostic category	DSM-III-R (sub)section	DSM-IV (sub)section
Psychotic disorder	"schizophrenia", "delusional disorder", "psychotic disorders not elsewhere classified"	"schizophrenia and other psychotic disorders"
Mood disorder	"mood disorders"	"mood disorders"
Paraphilic disorder	"paraphilias" under "sexual disorders"	"paraphilias" under "sexual and gender identity disorders"
Substance use disorder	"psychoactive substance use disorders" excl. nicotine	"substance-related disorders" excl. nicotine, caffeine, and intoxication/withdrawal
Personality disorder	Axis II personality disorders	Axis II personality disorders

3.2.2   Index offense data

Index offense data was collected primarily from the original sentence documents, with some information being extracted from the FPI reports, particularly the parts written by the forensic social worker, which commonly include a biographic section. While this section may sometimes rely entirely on the offender's self-report during interviews, many parts are usually well-sourced and corroborated by government documents and other external sources. All 146 sentence documents and corresponding FPI reports were read in their entirety, including any documents originating from higher or lower courts involved in the same case.

Sexual offenses are defined as any act listed under Chapter 6 of the Swedish Criminal Code (39), and this is how the term is generally understood in research in the jurisdiction of Sweden. Common examples are rape, sexual molestation, and sexual coercion. Both completed and attempted offenses were included.

3.2.3   Follow-up data

All follow-up data was collected from new sentence documents, starting with the conviction date of the index offense up until December 31, 2016. This mainly consisted of registry data from the National Council for Crime Prevention's register of persons found guilty of offenses, although all written sentence documents (approximately 1,000) were also collected from the individual courts to correctly categorize the type of offense.

Violent offenses were defined as all nonsexual violent acts—whether physical or nonphysical—including coercion as well as arson. These are commonly listed under Chapters 3, 4, and 8[10] of the Swedish Criminal Code (39). General offenses were defined as all other acts not falling under the definitions of sexual or violent offenses and mainly entailed traffic offenses, drug offenses, and theft.

Recidivism was defined as any conviction occurring after the index offense conviction date resulting in a sentence document consisting of one or more

---

[10] Titled "On offences against life and health," "On offences against liberty and peace," and "On theft, robbery and other appropriative offences" respectively

charges. A subject could theoretically recidivate in more than one type of offense, depending on the charges included.

A subject was considered incapacitated when currently serving time in prison or in compulsory forensic psychiatric care. Whether a patient was physically incarcerated at a clinic or not was difficult to ascertain, since this was rarely indicated in the data received from the individual clinics. It is therefore possible that a subject was free to live outside the clinic with strict instructions to take part in whatever treatment the responsible psychiatrist deemed adequate. In contrast to time incapacitated, time at risk was calculated from the date of release from prison or forensic psychiatric care or from the index offense conviction date for those sentenced to probation up until the first new conviction, date of death, date of migration from Sweden, or the end of the follow-up period—whichever occurred first. Note that time incapacitated was handled differently in Study I and Study II, which is described in more detail below.

3.3   Study specific procedures

3.3.1   Study I

The data used for Study I comprised the above-mentioned offender data, index offense data, and follow-up data for all 146 subjects included in the study. No exclusions were required.

In Study I, subjects were never considered incapacitated. Instead, any convictions occurring while they were serving time in prison or in forensic psychiatric care were considered as recidivism. Our rationale was that offenders do not necessarily cease their criminal behavior while imprisoned or undergoing psychiatric treatment, as corroborated by previous national reports (36,89).

The statistical measurements and indexes used were Pearson's chi-square test, Fischer's exact test, and the Mann-Whitney U test for between-group differences. For effect sizes, odds ratios (ORs) were used. To present and explore the time until first reconviction, Kaplan-Meier survival analyses with log-rank tests were used (90).

3.3.2   Study II

The data originally used for Study II consisted of Static-99R risk assessments of all 146 subjects. Unfortunately, 46 subjects needed to be excluded, primarily due to data for item 1 being missing. In the end, complete data on 100 subjects was used in the analyses.

Because of the way the Static-99R coding rules define an offender's capacity to reoffend (53), subjects were considered incapacitated while serving time in prison or in forensic psychiatric care in Study II. This resulted in four subjects who were reconvicted while currently under forensic psychiatric care not being defined as recidivists. However, coding rules are commonly adapted to inconvenient real-life conditions when conducting validation studies (91). The potential impact of these four subjects is explored in Appendix B, however.

As recommended by the developers of the Static-99R, the main author of study II obtained training by a certified Static-99R trainer. In turn, the main author instructed two additional authors of the study on how to assess sexual offenders using the instrument. Considering the extensive experience of all three authors in using several other risk assessment instruments, this training was deemed adequate for scoring the subjects. Together, the three authors scored 54–66 cases each following the latest version of the Static-99R coding rules (53). Ten percent (15 cases) of the total caseload was randomly selected for scoring by all three assessors in order to calculate interrater reliability.

The statistical measurements and indexes used were the intraclass correlation coefficient (ICC) for interrater reliability (92,93), using a two-way mixed-effects absolute agreement single rater model (94). For discrimination, AUC for the ROC and OR derived from a logistic regression for examining Static-99R discrimination were used (66,68,95). While the intercept (B_0) and slope coefficient (B_1) of the regression are sometimes used as an estimated comparison of discrimination, these measurements appear to be less common than AUC and ROC (65,69,71). Additionally, the Youden index (J) (95,96) was used for estimating the Static-99R cut-off score for achieving maximum sensitivity and specificity, i.e., the difference between sensitivity and 1-specificity (95).

In contrast to assessing discrimination, there is no consensus among researchers in the field on what methods are preferable for assessing calibration. For Study II, the E/O index (97), P/E index (96), and Brier score (98) were used

for examining how well the Static-99R performed compared to the norm groups recommended for use.

The E/O index is the ratio of the expected number of recidivists (E), calculated using the regression from the norm group and applying it to the study cohort, to the number of actual recidivists observed in the study cohort (O). As the E/O index approaches 1, perfect calibration is achieved. This is highly unrealistic, as no other sample or cohort would behave identically to the norm group used when developing the instrument. Instead, an E/O index above or below 1 indicates that the instrument either over- or underpredicts recidivism rates in the study cohort (66,97). The P/E index is comparable to the E/O index insofar as it follows the same guidelines, although it is better suited for rare outcomes, e.g., low base rates (96). If no recidivists are observed in the study cohort, dividing E by O requires division by zero. This issue is avoided by instead dividing the predicted number of recidivists in the study cohort (P) by the expected number of recidivists (E)—both the results of regression models. Additionally, by using the P/E index, two predictive regression models are compared instead of a model and absolute numbers linked to a sample. For this thesis, calibration analyses were primarily done using the P/E index, although E, O, and P are all presented.

Lastly, calculating the Brier score is a method for comparing the accuracy of two competing predictive models with dichotomous outcomes (98). When used in the evaluation of risk assessment instruments (96), the probability of recidivism for every single subject as estimated by the regression (a number less than 1) is subtracted from the actual outcome (either 1 or 0). The difference is squared to remove negative values and an average for the entire cohort is calculated. The lower the resulting value is, the better the predictive model performs, with 0 representing perfect prediction and .25 representing random chance.

3.3.3   Study III

The data used for Study III comprised the Static-99R coding sheets of all 146 subjects, including total scores and scores for all ten individual items on 117 subjects and nine individual items for 29 subjects.

The statistical method used was Rasch analysis, which measures an assumed unidimensional trait or underlying construct of an instrument or a scale using a logistic regression model for repeated measures (99–102). As succinctly

described by Bessing et al (103), the "Rasch analysis identifies the relative difficulty of each item on the scale and determines the individual skill or lack of skill that impedes what the scale aims to measure" (p. 3). It is an alternative method supplementing classical test theory (102,104) and can be applied to many different types of scales, almost independently of how the scale design or the number of response categories is presented (100,105–107). Preferably, samples should consist of 150-250 subjects, but larger samples are permissible and may produce more generalizable results (108). Smaller samples are acceptable when conducting pilot studies of explorative nature examining specific subgroups as was the case for study III (108).

The Rasch analysis in Study III was conducted by using the structure of the data from the Static-99R assessments (number of items, number of response categories) to construct what is known as a Guttman scale. A Guttman scale is a hypothetical model indicating how the data should look under perfect circumstances where every single item (and response category) incrementally and evenly adds to the total score of the scale being tested, essentially flawlessly quantifying the underlying trait measured by the scale. This rarely, if ever, occurs in real life, but the model can be used as a theoretical gold standard for what any scale should strive for. By comparing the actual data on the study cohort (individual response category affirmation, individual item score, person total score) and how it departs from the Guttman scale (e.g., what would be assumed of the data were it perfect), the qualities of the scale being tested (and the sample being used for testing the scale) can be measured. This is essentially how a Rasch analysis works.

Rasch analysis is an iterative process consisting of many required subtests applied independently and sequentially, each affording insight in order to adequately gauge how well the data fits the Rasch model and what conclusions can be drawn from the results (99,109–113). If the data fits the Rasch model perfectly (in that the items work invariantly and responses are properly categorized (114)), the scale essentially mirrors an interval scale where floor and ceiling effects are eliminated (107) and one single unit indicates a uniform increase or decrease in the strength of the underlying trait or construct.

There are many alternatives of measuring whether the data fits the Rasch model, and most interact in manners that are difficult to sort out. For Study III, several subtests and various ways of measuring person and item fit were used,

but some were of particular interest to this thesis. For this reason, the following terms and qualities are described:

- Person ability (or skill): The level of the underlying trait demonstrated by a particular subject. Highly skilled subjects tend to affirm most of the items in a scale, and *only* highly skilled subjects are expected to affirm the most difficult items.

- Item difficulty: The point where the probability of a person choosing any of two response categories in a dichotomous item is identical. The item difficulty indicates how much skill is required from a subject in order to affirm it.

- Item–trait interaction: A measurement of the hierarchical ordering of all items along the continuum of the underlying trait (115). A statistically significant result indicates that item difficulty is not stable in relation to subject ability levels, meaning that the difficulty of some items is not the same for two similarly skilled subjects.

- Under-discrimination (or poor discrimination): When a particular item in the scale poorly discriminates highly skilled subjects from low-skilled subjects for an item of that particular difficulty level.

- Item fit residuals: A measurement of the amount left over when comparing the actual performance of an item to that of the perfect theoretical Rasch model (107). Preferably, this should be as close to 0 as possible.

- Disordered thresholds: When an item has more than two response categories to choose from, every additional category in order should add incrementally to the item difficulty.

- Inverse responses: When response categories for an item oppose the general order of the other items in the scale. This occurs when every additional score is supposed to indicate an increase in ability except for a specific item where this is the opposite.

3.4 Ethical statements

All studies in this thesis were approved by the regional ethical review board in Gothenburg in one main and one supplemental application (registration numbers 377-17 and T1056-17). The two applications expanded upon three separate previous research projects: *A retrospective survey of sexual crimes against minors in Sweden* from 1997 and its follow-up project from 2002, as well as *Registry based follow-up of research groups "Gothenburg Forensic Neuropsychiatry Project" and "Gothenburg Sexual Child Abuse Studies"*, also from 2002.

Subjects were considered "not directly affected by the research" and informed consent was not deemed to be required by the ethical review board at the time of the original data collection in 1997. Additionally, it was considered inappropriate to approach all the subjects almost 20 years after the original data collection to engage in any type of informed consent. Instead, all data was pseudonymized using the code key kept separately in a safety box. Data could not be irrevocably anonymized since this would rule out any attempts to properly link follow-up data to the correct subject. All results were presented in group format, partly to make it more difficult to identify any one subject by their individual offender characteristics or recidivism pattern. Since no contact was made with the subjects, no financial compensation for participation was possible.

No female subjects were included in this project, but in total there was only one female available who fit the inclusion criteria. During the early design phases of the thesis, the question of whether this subject should be included was thoroughly discussed. Considering the low prevalence of female offenders in the cohort and that the use of Static-99R may be inappropriate for this offender subgroup (53,116), we ultimately decided to focus on an all-male offender cohort. With a larger sample over a larger timeframe, it might have been possible to include a female subgroup, but considering approximately 1% of convicted sexual offenders in Sweden are female (117), this was deemed unfeasible.

4 RESULTS

4.1 Cohort details and recidivism patterns

The first aim of this thesis was to establish and present a) offender characteristics—clinical ones in particular—of MDSOs and b) patterns of recidivism over up to 24 years of follow-up regarding both new sexual offenses and violent offenses of said cohort. This was examined primarily in Study I.

4.1.1 Offender characteristics

Table 2 presents all the offender characteristics explored in Study I according to separate categories: index offense factors, criminal history factors, social factors, and clinical factors. While most of the variables are self-explanatory, a few require further detail. In respect of index offense factors, "contact offenses" indicate that there was some degree of sexually motivated physical contact, whereas use of physical violence was non-sexual and commonly resulted in the subject being charged with a violent offense in addition to the sexual offense. With regard to developmental factors, clear indications were required of criterion A for conduct disorder and of criterion A.2 for attention-deficit/hyperactivity disorder according to the current DSM version (25). This was articulated plainly in the biographic section of the FPI, usually by describing the child as regularly starting fights, harming small animals, fervently disobeying teachers, frequently running out of the classroom, and so on. In the case of childhood institutionalization, any instances of having lived in or spent time in a family home, foster care, or a juvenile detention center counted.

All but one subject were given a clinical diagnosis during the FPI (99.3%, $n = 145$), the average being two diagnoses per subject, and one fourth of the cohort were on psychotropic medication (24.0%, $n = 35$). Personality disorders were highly prevalent in the cohort with cluster B being the most common (29.5%, $n = 43$) and 15 subjects (10.3%) displaying two or more personality disorders simultaneously. Out of the 38 with an Axis I psychotic disorder, most were categorized as suffering from schizophrenia ($n = 18$), nonspecific psychotic disorder ($n = 12$), or delusional disorder ($n = 5$). An additional nine subjects were described as having personality disorders with psychotic features although with no formal diagnosis of psychotic disorder rendered. The group of subjects with mood disorders almost exclusively presented various types of

depression, with only one subject being described as displaying "mania." The most common types of paraphilic disorder diagnoses were paraphilia not otherwise specified, exhibitionism and pedophilia comprising a total of two thirds of the 15 subjects with a paraphilic disorder. Fewer than one subject in ten were considered intellectually disabled, predominantly presenting mild or marginal symptoms ($n = 11$). Half of the cohort were considered "severely mentally disordered" according to Swedish law (45.9%, $n = 67$), which included almost every single subject with an Axis I psychotic disorder (94.7%, $n = 36/38$).

Statistically significant differences between subjects with a psychotic disorder and subjects without were found only in respect of five variables tested, three of which were clinical, and all with large confidence intervals. Subjects without psychotic disorders were more likely to present comorbidity for personality disorder (OR = 3.5, CI 95% = 1.6–7.5, $p < .001$), substance use disorder (OR = 2.4, CI 95% = 1.0–5.5, $p = .039$), or paraphilic disorder (OR = 12.8,[11] $p = .012$)—the latter of which comprised no subjects with a psychotic disorder at all. For the non-clinical variable, subjects with a psychotic disorder were less likely to have been institutionalized at some period during their childhood (OR = 3.1, CI 95% = 1.0–9.5, $p = .039$).

---

[11] Haldane–Ascombe correction applied (divided by empty cell)

Table 2. Offender characteristics for all 146 mentally disordered sexual offenders (MDSOs). The table is adapted from Table 1 in both Studies I and II.

Factor	Variable	MDSOs (N =146)
Index offense factors	Age at index offense	$Mdn = 34.5$, $M = 35.7$ (17.7-71.6, $SD = 10.5$)
	Number of index offense sexual charges	$Mdn = 1$, $M = 1.2$ (1-3, $SD = 0.4$)
	Contact index offense	123 (84.2%)
	Physically violent index offense	120 (82.2%)
	Index offense substance intoxication	74 (50.7%)
	Male victims	3 (2.1%)
Criminal history factors	Age at first conviction	$Mdn = 21.7$, $M = 24.8$ (15.2-66.6, $SD = 10.1$)
	Previously convicted	100 (68.5%)
	Any previous sexual offense	49 (33.6%)
Social factors	Childhood institutionalization	33 (22.6%)
	Steady partner	51 (34.9%)
	Secondary school diploma	43 (29.5%)
	Employment or studies	46 (31.5%)
Clinical factors	Paraphilic disorder	15 (10.3%)
	Mood disorder	17 (11.6%)
	Substance use disorder	55 (37.7%)
	Personality disorder	90 (61.6%)
	Antisocial personality disorder	21 (14.4%)
	Psychotic disorder	38 (26.0%)
	Intellectual disability	13 (8.9%)
	Childhood hyperactive traits	25/132 (18.9%)
	Childhood conduct disorder traits	35/133 (26.3%)

4.1.2   Recidivism during follow-up

Recidivism rates for new sexual offenses throughout the follow-up period were 9.6% (n = 14) after 5 years, 13.0% (n = 19) after 10 years, 15.1% (n = 22) after 15 years, and 17.1% (n = 25) after 20 years. The survival analyses presented in Figure 3 present recidivism rates visually over up to 24 years of follow-up.
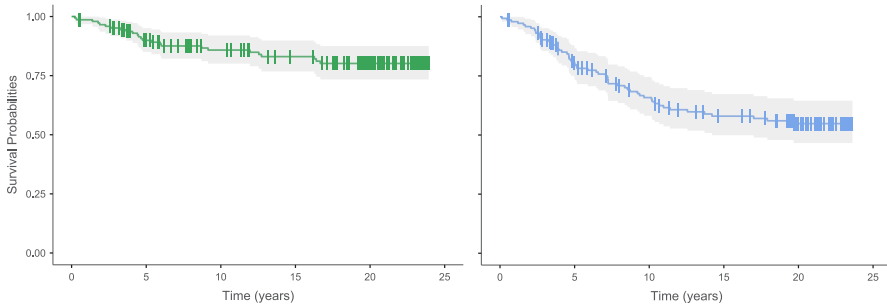
Figure 3. Survival analyses plotting sexual recidivism (left) and violent recidivism (right) over the entire follow-up period. Intersecting lines mark the end of a follow-up for a single subject, also known as a right-censored event.

As seen in Figure 4, recidivism was mainly clustered around the five-year mark, indicating that most reoffending subjects were reconvicted within the first 5 to 10 years after the index offense. For those committing new offenses, the average number of years until first conviction for either a new sexual offense or a violent offense was 6.5 years in both cases (*Mdn* = 4.6, *range* = 0.2–16.7 for a sexual offense and *Mdn* = 5.1, *range* = 0.1–19.7 for a violent offense).
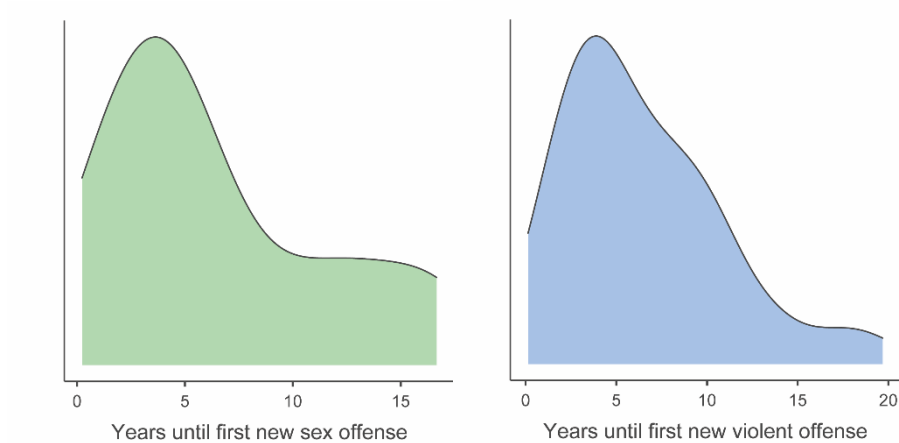
Figure 4. Two graphs visualizing recidivism over the entire follow-up period, essentially reflecting the slope of the survival analyses seen in Figure 3. Peaks on the y-axis denote higher levels of recidivism incidents and valleys lower levels.

Recidivists were commonly reconvicted for one or two new offenses, with a higher propensity for violent offenses than sexual offenses (*Mdn* = 2.0, *M* = 1.9, *range* = 1–5 compared to *Mdn* = 1.0, *M* = 2.0, *range* = 1–13). Half of the subjects were reconvicted for some type of new offense (*n* = 77), although 34% of the recidivists were convicted for strictly non-sexual, non-violent offenses (*n* = 26).

Measured as a proportion (the number of new sentences containing a specific type of charge divided by all new sentences), violent charges were more frequent among the subjects. On average for the cohort, half of all new sentences contained at least one violent charge (*M* = 47%), whereas one fifth contained at least one sexual charge (*M* = 19.7%). It was also more common for subjects to be consistently violent in their recidivism pattern (e.g., demonstrating a proportion of 100%), with more than twice as many subjects having charges of violence in every single new reconviction as compared to sexual charges (26%, *n* = 20/77, versus 11.7%, *n* = 9/77).

4.2 Static-99R validation

The second aim of the thesis was to determine the predictive qualities of the Static-99R in MDSOs with regard to a) how well the instrument discriminates between recidivists and non-recidivists, b) how well the instrument is

calibrated in relation to the specific cohort, and c) whether the predictive qualities are maintained over longer timeframes than the typical five-year follow-up period. This was explored in Study II.

For recidivism studies, a follow-up period can be either fixed or ragged (i.e., variable). A fixed follow-up period differs from a ragged one insofar as it lasts a predefined number of years and applies to all subjects used in the analysis. While a ragged follow-up was possible in Study I, a fixed follow-up was required to validate the Static-99R in Study II. Seventeen of the 146 subjects had less than 5 years of follow-up, predominantly because they were deported or voluntarily emigrated shortly after their release. Furthermore, no item except for item 2 ("ever lived with a lover") may be left unanswered, meaning that any other item demonstrating missing data invalidated the use of a particular subject altogether, leading to their complete exclusion from analyses. This was unfortunately the case for an additional 29 subjects where time of release could not be collected from their respective forensic psychiatric clinics. In sum, 100 subjects were included in the validation of the Static-99R. Only minor differences were found between the 46 excluded subjects and those included, with the exception of prevalence rates of psychotic disorders and whether the index offense resulted in compulsory forensic psychiatric care—both of which were more common among subjects excluded.

Static-99R interrater reliability was considered better than "strong" (118), or between "good" and "excellent" (93) (ICC = .89, CI 95% = .76–.94, $p < .001$), which is consistent with findings from the 11 studies presented in the coding rules (.84–.95) (53).

4.2.1 Discrimination

As seen in Table 3, the Static-99R predicted recidivism over all fixed follow-up periods with moderate accuracy, and the odds of committing a new sexual offense increased by 31–45% with every additional point in the Static-99R total score.

The cut-off score with the largest Youden index was 6, indicating optimal sensitivity and specificity for the first 5 years of fixed follow-up, although a cut-off score of 4 performed only slightly worse ($J = 0.4607$ versus 0.4688). This means that by classifying subjects with a score of 5 and less as non-recidivists and those with 6 or more as recidivists, the Static-99R correctly categorized three out of four reconvicted subjects as recidivists and non-

reconvicted subjects as non-recidivists (72.7% sensitivity, 74.2% specificity). Additionally, one fourth of recidivists were correctly assumed to reoffend sexually (25.8% PPV) and almost all non-recidivists were correctly assumed not to (95.7% NPV).

Table 3. Static-99R score and effect sizes for 5, 10, 15, and 20 years of fixed follow-up after release. All p-values are linked to the logistic regressions (not presented in the table). The table is adapted from Table 4 in Study II.

Fixed follow-up	Static-99R score			Receiver Operating Characteristics	
	M (*SD*, min-max)	*n*/N	p	OR (CI 95%)	AUC (CI 95%)
5 years	4.19 (2.85, -2-11)	100%	< .001	1.45 (1.14-1.84)	.790 (.700-.865)
10 years	4.13 (2.86, -2-11)	91%	.015	1.31 (1.05-1.63)	.734 (.633-.823)
15 years	4.42 (2.77, -1-11)	79%	.005	1.36 (1.10-1.69)	.741 (.636-.838)
20 years	4.31 (2.67, 1-11)	36%	.030	1.40 (1.03-1.91)	.739 (.578-.879)

4.2.2 Calibration

By extrapolating the intercept and slope of the regression from the development of the two norm groups, we were able to calculate expected recidivism rates for every possible total score of the Static-99R (119). This was necessary a) since the published recidivism tables did not include the upper and lower limits of the total score that was demonstrated by the study cohort (54) and b) to ensure that the expected recidivism rates could be measured with greater accuracy. Table 4 presents the recidivism rates for the study cohort and the two norm groups for every Static-99R total score, where the predicted rates were consistently lower than what was expected from both the routine norm group (E_1) and the high-risk/high-need norm group (E_2).

Table 4. Recidivism rates for each Static-99R total score observed in (O) and predicted for (P) the study cohort (based on the regression model using data from the study cohort), as well as expected for the routine (E_1) and high-risk/high-need (E_2) norm groups (by extrapolating the intercept and slope from the original norm group regression models (119)).

Static-99R total score	Subjects in cohort	O rates	P rates	E_1 rates	E_2 rates
-2	1	0.00%	0.83%	1.34%	4.46%
-1	2	0.00%	1.19%	1.92%	5.66%
0	4	0.00%	1.71%	2.76%	7.15%
1	11	0.00%	2.46%	3.93%	9.00%
2	15	0.00%	3.52%	5.59%	11.27%
3	8	0.00%	5.01%	7.88%	14.02%
4	16	18.75%	7.08%	11.01%	17.32%
5	12	0.00%	9.93%	15.17%	21.19%
6	12	25.00%	13.75%	20.54%	25.67%
7	7	14.29%	18.74%	27.21%	30.72%
8	4	25.00%	25.01%	35.08%	36.28%
9	2	50.00%	32.54%	43.85%	42.24%
10	3	33.33%	41.10%	53.03%	48.43%
11	3	33.33%	50.22%	62.00%	54.67%

With a 70% probability[12] on average of the regression model correctly predicting recidivism, the study cohort Brier score (0.088) was closer to that of the routine norm group (0.092) than of the high-risk/high-need norm group (0.096). This indicated that the routine norm group was preferable compared to the high-risk/high-need norm group.

Table 5 shows that over the five-year fixed follow-up period, the number of predicted recidivists was 30% lower than expected in the routine norm group across all risk categories (P/E_1 = 0.70) and 46% lower than expected in the high-risk/high-need norm group (P/E_2 = 0.54). Additionally, statistically significant differences were found between the study cohort and the high-

---

[12] Brier score = (Probability-1)^2

risk/high-need norm group, indicating that the high-risk/high-need norm group may be unfit for use.

Table 5. Number of recidivists across Static-99R risk categories for a five-year fixed follow-up period, where the observed number of recidivists (O) is based on the outcome for the study cohort, the predicted number of recidivists (P) on the regression model for the study cohort, and the expected number of recidivists on the regression model for the Static-99R routine norm group (E₁) and high-risk/high-need norm group (E₂). The table is adapted from Table 3 in Study II.

Category	n	Number of recidivists				P/E_1 (CI 95%)	P/E_2 (CI 95%)
		O	P	E_1	E_2		
I	1	0	0.01	0.01	0.04	0.62	0.19
II	6	0	0.09	0.15	0.40	0.62	0.23
III	34	0	1.20	1.90	3.80	0.63 (0.11-3.78)	0.32 (0.05-1.89)
IVa	28	3	2.32	3.58	5.31	0.65 (0.18-2.35)	0.44 (0.12-1.58)
IVb	31	8	7.35	10.10	10.62	0.73 (0.35-1.50)	0.69 (0.34-1.43)
Total	100	11	10.98	15.75	20.18	0.70 (0.39-1.26)	0.54 (0.30-0.98)

Note. A CI of 95% that includes 1 indicates no statistically significant difference between the predicted and the expected number of recidivists ($p > .05$). Due to the low number of expected recidivists, CI was uncalculatable for risk categories I and II.

4.3 Static-99R construct validity

The third aim of the thesis was to explore the construct validity of the Static-99R when assessing MDSOs with regard to a) how well the underlying trait measured by the instrument is unidimensional and quantifiable and b) whether there are any areas of the Static-99R that merit highlighting in future revisions of the instrument. This was accomplished in Study III using a modern test theory approach commonly used for developing and evaluating psychometric instruments such as scales and questionnaires. Since missing data is permissible when using this method, all 146 subjects were included in the study.

4.3.1 Evaluation of the original ten items

The Rasch analysis resulted in a low overall model fit for the MDSO cohort data with a reported scale reliability measured as person separation index (PSI) of approximately 0.50–0.60, and a mean person location slightly below 0, indicating a reasonable targeting of the items. Item difficulty appeared to inconsistently quantify the underlying trait measured, which was indicated by the analysis demonstrating a significant item–trait interaction chi-square and item fit residuals roughly 50% higher than expected.

In particular, three of the individual Static-99R items presented psychometric issues: items 1, 3, and 5. Item 3 was found to be particularly problematic, demonstrating under-discrimination (Figure 5) with fit residuals well beyond the critical cut-off value of +/- 2.5 (114). Both items 1 and 5—the only two polytomous items in the Static-99R—demonstrated problems connected to their response categories. Neither of the two items formed monotonic, slightly overlapping waves across the x-axis as expected from items performing satisfactorily. Instead, one response category was eclipsed by two of the response categories for item 1, and two response categories were overshadowed by the other two response categories for item 5. There were no indications of inverse response categories in any of the ten items.
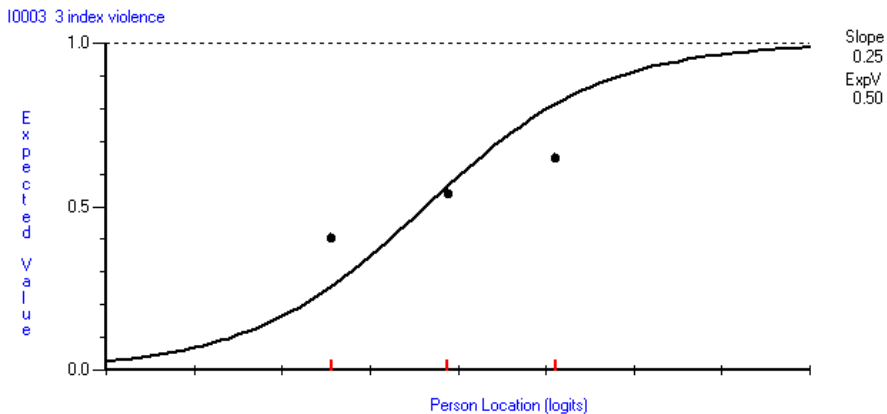

Figure 5. Item characteristic curve for item 3 demonstrating under-discrimination by the class interval averages (black dots) forming a "flatter" slope than the model assumes.

Additionally, although the principal component analysis (PCA) found the scale to be unidimensional, the residuals correlation matrix indicated response

dependency between two item pairs, and uniform differential item functioning (DIF) was indicated for three subgroups of subjects: those with substance use disorders, those with an intellectual disability, and those with a secondary school diploma or a higher level of education attainment.

4.3.2   Suggested alterations for psychometric improvement

By removing item 3 and collapsing the underperforming response categories found in items 1 and 5, several psychometric improvements were made to the Static-99R when assessing MDSOs. As any alterations to one part of the data affect the model frame of reference, it is generally recommended to make any changes one at a time, each change being followed by a new Rasch analysis (114,120). Consequently, the changes to items 1, 3 and 5 were made in three separate steps, by adjusting item 1 first, item 5 second, and item 3 last.

For item 1, response categories 1 and 2 were collapsed into one, resulting in the ordered overlap from one category to the next seen in Figure 6. Every item requires its response categories to reflect a steady, consistent increase in difficulty, meaning that a subject affirming the lowest response category (blue) should possess a lower level of the underlying trait (i.e., demonstrating "less skill" in Rasch terms) than a subject affirming the response category next in line (red), and so on. Figure 6 depicts such a progression in item difficulty. Naturally, there is still a measure of chance involved, which is indicated by the overlap between response categories. Such an overlap is unavoidable in clinical settings.
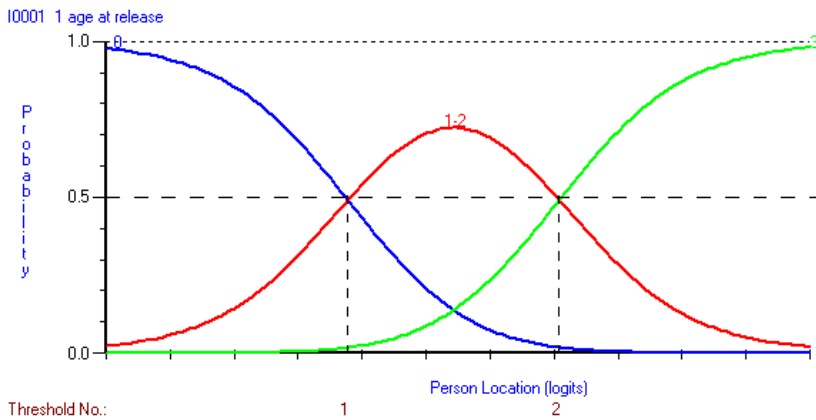
Figure 6. The four original response categories of item 1 collapsed into three to remedy the disordered thresholds not depicted. New, ordered thresholds are denoted 1 and 2 below the x-axis. Note the monotonic overlapping waves.

For item 5, all four response categories were collapsed into two distinct waves, effectively dichotomizing the polytomous item. The changes to items 1 and 5 decreased the complexity of both items while retaining or improving the psychometric properties of the scale. Scale reliability (PSI) was slightly improved.

The changes to items 1 and 5 did not affect the fit residuals of item 3 sufficiently, resulting in its removal still being necessary. Removing item 3 led to no degradation of scale reliability (PSI), targeting, item–trait interaction, or any of the item fit residuals.

Displayed in Figure 7 is the relationship between all subjects' trait levels and the nine items left in the scale. High-skilled subjects tended to affirm item 10, while low-skilled subjects tended to affirm the first of the three response categories now available for item 1. Several of the items (3, 4, 6, and 9) were affirmed by subjects demonstrating an average level of the underlying trait. Gaps in the distribution of items were noticeable primarily in the upper and lower parts of the trait spectrum. These visual gaps combined with a low person/item separation index indicated that more items may be needed for the scale to properly distinguish between subjects of high and low levels of the underlying trait in the study cohort.
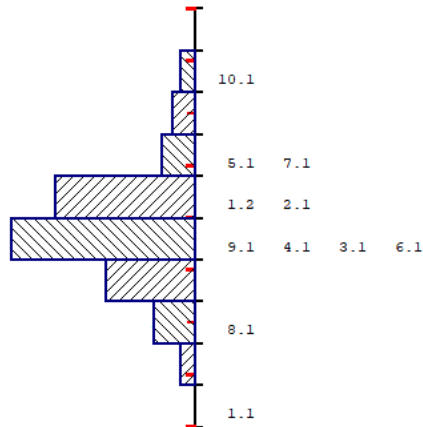
Figure 7. Item map depicting the relationship between number of subjects (left) demonstrating a specific level of the trait (y-axis) in relation to the item thresholds of items 1, 2, and 4–10 (right). Trait level varies across the y-axis, with positive values near the top of the axis and negative values near the bottom. Note that most items are dichotomous and therefore demonstrate one threshold (X.1). Item 1 demonstrates two thresholds (1.1 and 1.2) by having three separate response categories. Consequently, threshold 1.1 separates the first response category from the second, and threshold 1.2 the second response category from the third. See Figure 6 for a graphical depiction.

In sum, as a result of the changes to items 1, 3, and 5, the scale was still found to be unidimensional and no longer demonstrated any response dependency. Signs of uniform DIF were still indicated, which related to subjects diagnosed with substance use disorders, but the other two DIF issues had been resolved automatically by the changes made to items 1, 3, and 5. Lastly, overall scale reliability was improved as compared to the original ten items, but there was still a statistically significant item–trait interaction despite the alterations.

5   DISCUSSION

5.1   Summary of main findings

Aim I: The MDSO cohort was heterogeneous and burdened by extensive psychiatric comorbidity and criminal history. Almost all subjects had a psychiatric diagnosis, with two diagnoses per subject on average. Personality disorders were common, but paraphilic disorders less so. Most subjects had been previously convicted, half of these for sexual offenses. Most recidivists committed one or two new offenses, more often violent than sexual. Charges of violence were frequent, even during incidents of sexual recidivism, and one fourth of all recidivists were convicted for strictly violent offenses compared to one in ten for strictly sexual offenses. Recidivism rates for new sexual offenses were 9.6% after 5 years, 13.0% after 10 years, 15.1% after 15 years, and 17.1% after 20 years. More than half of the sexual recidivists were reconvicted during the first 5 years after the index offense, indicating that longer follow-up periods may not be as pertinent.

Aim II: The predictive qualities of the Static-99R in the mentally disordered cohort were satisfactory for up to 20 years after the offenders were released from incarceration for the index offense. The best discrimination was achieved for the five-year follow-up period, which is the follow-up period that the Static-99R was primarily developed for. Additionally, the odds of committing a new offense increased appropriately with every additional point on the Static-99R in line with previous studies. The norm group statistically best suited for use when estimating the recidivism risk of a subject was the routine norm group.

Aim III: Findings indicated that the underlying trait measured by the Static-99R was unidimensional but inconsistently quantified, with a lower overall scale reliability than preferrable. Interpretations of the results should be tempered accordingly. The only two polytomous items performed better with fewer response categories than the original four, and one additional item was found to be ineffective, warranting its removal.

5.2 Comments on main findings

5.2.1 Aim I

The clinical burden demonstrated by the study cohort in its entirety was expected in view of previous findings on offenders undergoing FPIs in Sweden (121). Psychotic disorders were more frequent in the study cohort than in general sexual offender samples, which tend to demonstrate psychotic disorder rates of 5–10% (18,122). However, there is substantial variation in studies and forensic samples such as the study cohort tend to present higher rates than non–forensic samples (15,18). When comparing the clinical characteristics of the study cohort with international studies of psychosis-focused samples, our subjects with psychotic disorders demonstrated higher rates of comorbid personality disorders but lower rates of paraphilic disorders (19). For those without psychotic disorders, the prevalence rates of substance use disorder and personality disorders were consistent with previous research on MDSOs (15,123). Overall, diagnoses of paraphilic disorders were infrequent considering that the cohort consisted of sexual offenders, but given that only one of the index offense victims were underaged (and the subject concurrently targeted additional, older victims) and that paraphilic disorders often comprise instances of pedophilia, this was not unexpected. Instead, many diagnoses of paraphilic disorder were either non-specific or reflected various explicit paraphilias (e.g., exhibitionism or voyeurism). In merely three cases were the paraphilic disorders paired with specific descriptions of and underlying issues with inhibition such as compulsive sexual behavior. Subjects diagnosed with bipolar disorder were fewer than in international findings (15,124,125), although the description of that specific diagnosis was changed along with the introduction of DSM-IV that coincided with the period of data collection. It is reasonable to suspect that this affected how the subjects were described and clinically classified at the time.

The sexual recidivism rates demonstrated by the study cohort were comparable to those of general sexual offenders with adult victims, although these rates are severely impacted by factors such as follow-up length and recidivism definitions (7–13). The results were also similar to those of previous studies on MDSOs with adult victims, which have found recidivism rates of around 6% for follow-up periods of up to 8 years (20) and 17% over 20 years (17), although they are a far cry from rates as high as 50% over 25 years (14). With a few individual exceptions, there were no indications of specialized sexual

offenders in the study cohort. Instead, non-sexual recidivism was the norm, which is consistent with previous research specifically exploring offender versatility (8,121,126,127).

Understandably, studies following sexual offenders over particularly long follow-up periods are rare, especially those on MDSOs. However, considering that more than half of all sexually reoffending subjects in the study cohort were reconvicted over the first 5 years and two thirds over the first ten years, it is possible that the risk–reward ratio of spending additional resources on recidivism studies spanning a quarter of a century may be overrated. This is particularly true of prospective studies, which may suffer from having research projects inherited from one research group by another over the years, with only marginal rewards from new data that is collected. In addition, as the subjects grow older, their risk of reoffending tends to decline (128).

5.2.2   Aim II

Static-99R discrimination over 5 years was found to be acceptable and comparable with two of the larger meta-studies conducted in recent years using the routine norm group (12,119). This also held true for longer follow-up periods up to and including 20 years after release as seen in Table 3, with the current results being in line with the few studies using fixed follow-up periods of 10 years (12,59–61). To our knowledge, there are no other studies presenting data on the basis of follow-up periods of 20 or even 15 years. The optimal Static-99R total score cut-off of 6 points for separating predicted recidivists from non-recidivists was greater than is commonly found in other studies (96), but the difference between cut-off scores 6 and 4 was minimal when compared using the Youden index. Consequently, there is little basis for concluding that the cohort greatly differs from international studies with regard to the optimal cut-off scores for maximizing sensitivity and specificity.

When using the reference guide (the tables presenting recidivism risk per Static-99R total score) in the evaluators' workbook to compare a subject's risk category with that of the norm groups, the entire range of possible total scores is not presented (54). This is generally not an issue, except when assessing offenders with particularly high or low total scores. In other words, the reference guide does not offer a recidivism risk for a large enough total score span to be reliably used in a clinical environment. This can be remedied by using the intercept and slope of the original regression model for the norm groups to calculate complete estimations of recidivism risk for all total scores,

as was done in Study II. Besides requiring effort, this also demands some degree of know-how that can hardly be expected from Static-99R users in general, especially considering the risk of producing incorrect estimates and those errors potentially having a major effect on the risk assessment. As the number of offenders demonstrating extremely high and low scores is rare, so are precise recidivism rates. It is possible that presenting these numbers in the reference guide instills a false sense of precision and a belief that any offender demonstrating that level of risk is an outlier. It would be particularly challenging to estimate the true recidivism risk for such an individual and the clinician should handle the case with extra care. Even though we are technically able to calculate a probability of reoffending, it may give clinicians a false sense of confidence in their assessments. This is a highly undesirable outcome for risk assessments which is generally recommended to avoid, particularly when conveying risk in a forensic setting, for example in a risk report to the court, where nuances in language are of the utmost importance (129). Consequently, not presenting the tables in full may be the lesser of two evils.

As regards calibration, all the measurements indicated that the routine norm group was preferable to the high-risk/high-need norm group. Subjects in the cohort reoffended in a lesser extent than their Static-99R total score predicted across all risk categories, meaning that the number of recidivists predicted was lower than expected. So, while the routine norm group was more similar to the cohort than the high-risk/high-need norm group, offenders still tended to reoffend in lower numbers than predicted by their Static-99R total score. Apart from drawing the conclusion that the study cohort simply demonstrated a less serious risk profile, two reasons for this discrepancy comes to mind. The first reason pertains to definitions. In this thesis, recidivism was defined as a new conviction, but the Static-99R coding rules are less strict. In the coding rules, charges that do not lead to a conviction count as recidivism as well as sexually motivated offenses in general (53,54), and two samples used when developing the coming revision of the norm groups include arrests (12). Consequently, it is possible that the strict definition used in the thesis resulted in fewer recidivists. The second reason pertains to the study cohort composition. The similarities between the study cohort and the routine norm group may have been amplified by the unavoidable exclusion of 29 subjects where the date of release could not be established, either due to a lack of data or because the subject had never been released from their index offense sentence. This affected item 1 of the Static-99R, which in turn invalidated all total scores for the subjects in question. In view of the fact that all of these subjects were

undergoing forensic psychiatric treatment, it is possible that some of the offenders most impacted by their mental disorders were not present when the calibration was being assessed, which may have rendered the cohort more similar to one selected from a general prison population. Furthermore, as demonstrated in Study I, mental disorders were prevalent throughout the cohort, and a Swedish study using a sample of almost 8,500 first-time sexual offenders found that they were three to five times more likely to be diagnosed with a major mental illness and six times more likely to have undergone psychiatric in-patient care prior to their offense as compared to the general population (122). In sum, the sexual offender population in general seems to carry a substantial clinical burden whether or not they are compelled to undergo an FPI and subsequently may be sentenced to forensic psychiatric care.

While the Static-99R appeared to be better calibrated with regard to higher risk categories than lower risk categories, the low number of recidivists complicates the situation. In fact, no subjects at all were reconvicted in risk categories I, II, and III, which invalidated all E/O indexes for these risk categories. While the P/E index limits the negative effect of these circumstances by using the probabilities from the regression instead of absolute numbers, ratios of two extremely low values are still fragile. A single recidivist in risk category III, for example, would have had a major effect on the outcome. Consequently, any conclusions drawn from any such index must be carefully weighed using the data available.

As shown in Table 5, there was a considerable range for the confidence intervals—from 0.39 to 1.26. While this is not uncommon for a validation study of this type (96), it still suggests a degree of uncertainty in the statement that the model predicted 30% fewer recidivists than expected in the norm group. In a validation study such as Study II, non-significant confidence intervals for the E/O indexes (or, in this case, the P/E indexes) are desirable outcomes. A non-significant result indicates that the norm group with which the expected recidivism rates are compared is similar enough to that of the study cohort for clinical use (66,96,97). For the sake of validation, non-significant results are arguably more important than the degree to which recidivists may be under- or overpredicted.

5.2.3 Aim III

When employing statistical methods uncommon for the field one should generally focus on large, broad samples rather than a highly specialized, smaller cohort as was done in study III. For this reason, it would have been methodologically preferable to use a sample comprising general sexual offenders rather than MDSOs. As regards this thesis, this was an impractical option that would hinder the overall thesis aim of specifically targeting offenders with mental disorders—a naturally small but important offender sub-group. Consequently, the exploratory results of the thesis are of a tentative nature that indicate the psychometric properties of the Static-99R when assessing MDSOs but cannot reliably establish them with certainty. This includes the finding that the Static-99R was unidimensional, which contrasts two previous studies (63,64). However, the findings of those studies were specifically linked to observed recidivism of general sexual offenders rather than the inherent properties of the Static-99R and how it functions as a scale when assessing MDSOs as presented by the thesis.

While the Static-99R did not fulfill the strict assumptions of the Rasch model, this is not an uncommon occurrence. Scale reliability was not as strong as desirable, but the PSI is in itself not enough to evaluate how well a scale performs (114). Scales with many items are more likely to demonstrate a higher PSI than scales that do not have many items (103,114,130), and the Static-99R comprises only ten items. However, scales with an equal or similar number of items have been found to perform as reliably or even better (114,131).

Both item 1 and item 5 performed better with fewer response categories than their original four. This is not an uncommon occurrence for polytomous items, but it does add additional complexity to a scale without an equal gain in information gleaned from the scale. Item 1 is of particular interest, considering that it was revised in 2009 and expanded from two categories to four after research was conducted indicating that age had a substantial impact on reducing the risk of recidivism in sexual offenders. One consideration is the nature of item 1, which relies on the age of the subject when released from incapacitation. What constitutes release is specified in detail in the coding rules, but it may still be difficult for a clinician to specify this date with any certainty, especially if the jurisdiction differs from that in the USA and Canada where the Static-99R was developed. Considering that this is the item with the greatest potential to affect the Static-99R total score (between -3 and 1), this is

not to be taken lightly. When this item was scored in Study II, release dates were unavailable for 29 subjects and were generally less certain for subjects in forensic psychiatric care than for subjects serving time in prison. On the other hand, having subjects self-report their age at release is acceptable according to the coding rules, which in itself indicates some inherent uncertainty for the item (53).

Item 3 demonstrated psychometric issues to such a degree that the removal of the item was deemed necessary in order for the Static-99R MDSO cohort data to achieve a better model fit. While a large enough sample may magnify small differences between observed and expected values, causing them to demonstrate a misfit, the 146 subjects included in Study III were not enough to have this type of an effect (114). The excessive (positive) fit residuals of item 3 indicated that the item under-discriminated (109), as shown in Figure 5, meaning that the item discriminated between subjects demonstrating high levels and low levels of the trait less than expected for an item of this particular difficulty. There have been issues with item 3 in previous Static-99R recidivism research, particularly with regard to whether or not it adds incrementally to the prediction of sexual recidivism at all (132). The Static-99R coding rules state that item 3 appears to be predictive for rape specifically but not for other types of new sexual offenses and only in North American samples (54). In fact, the developers state in the coding rules that "this item may not be a strong predictor and caution in assessment may be warranted" when using the Static-99R in other countries (p. 52). With this in mind, the psychometric issues found in Study III regarding item 3 were less surprising.

An important finding in Study III was the item–trait chi-square, indicating that the items may inconsistently quantify the underlying trait. This was only marginally improved by revising items 1, 3, and 5, but was improved further in Study III by experimenting with even greater revisions to the instrument. When the items in a scale do not demonstrate hierarchical ordering along the continuum of the underlying trait, this indicates that the difficulty of some items may differ for two subjects demonstrating similar levels of the trait. This means that some of the items are unreliable, but none of the subjects demonstrated a sufficiently large misfit to cause concern. Nevertheless, the interaction between subject abilities and item difficulty was problematic and is a clear indication of general model misfit, which in this case suggests that more alterations to the scale were needed. This was eventually remedied in Study III through additional revisions to the item structure which had significant effects

on the overall design of the scale. In view of the gap in the item map depicted in Figure 7, it is possible that additional, well-fitted items may have had a similar remedial effect on both general scale reliability (PSI) and item–trait interaction.

From a clinical standpoint, item 10 was of particular interest. As shown in Figure 7, the item map indicated that subjects affirming item 10 also demonstrated a higher level of the underlying trait by being located in the furthest part of the spectrum. If during the index offense or any of their previous sexual offenses, an MDSO in the cohort had targeted a male victim, they tended to receive a particularly high total score on the Static-99R. Inversely, only subjects with a high total score affirmed this item. This pertained to nine subjects, which is more than expected in view of the fact that a large majority of index offense victims as well as victims of new sexual offenses were female. So, while the cohort demonstrated a clear preference for female victims, those that did not share this preference tended to score higher on the Static-99R. This is consistent with findings that offenders with a flexible victim preference (or a preference for males) tend to reoffend to a greater degree (8).

As with any method, Rasch analysis requires a large enough sample in order to be reliable. The minimum sample size depends on several factors, especially the number of scale items and how well the sample fits the Rasch model. While samples comprising 150-250 subjects is a respectable ambition, 50 subjects is generally sufficient, although as few as 30 is possible in smaller pilot studies that tackle new or otherwise untested, dichotomous scales (108). Scales with polytomous items consisting of many responses are more complicated to work with, considering that it is preferred to have at least 10 observations (events) per category (133), but this was generally not an issue in Study III. Consequently, the 146 subjects included were sufficient for the analyses, but a larger sample or cohort is generally preferred over a smaller one in all types of statistical analyses, Rasch analysis included (108). That said, a large enough sample also amplifies small discrepancies between the data and the model, which may result in more complex interactions between the two (114). It does not disqualify the method from use, but it does put greater pressure on the analyst to unravel the errors that truly reflect the data and those that are simply artifacts of the sample size.

Missing data is permissible when using Rasch analysis, but it is preferrable if such data is missing randomly from the dataset and generally affecting all items

and response categories (109). Unfortunately, this was not the case. Less than 2% of all responses were missing for the 146 subjects, but these exclusively affected item 1, resulting in a 20% loss of data for that single item. Since most parts of the Rasch analysis affect all other parts, there is a chance of localized issues, with item 1 having a cascading effect on other parts of the analysis. It is very likely that better model fit would have been achieved with a larger cohort, a complete dataset, or by having randomly missing data. For these reasons, while the results are no less interesting, they must not be overstated.

5.3   Ethical aspects of risk assessments

As with any clinical practice, risk assessment instruments are not without flaws or critique. Results, including total scores, may be misinterpreted, manipulated, misapplied, or misrepresented by clinicians, intermediaries, administrators, or the client for whom the assessment results are intended, commonly a court (134,135). Considering the weight an actuarial risk assessment result may have in a legal proceeding, it is the duty of all parties involved to conduct themselves ethically and with their utmost discipline, in research as well as clinical practice. An anecdotal example from Sweden is the expansion of the use of risk assessments by the National Board of Forensic Medicine in 2006. The ethical implications of subjecting incapacitated individuals to these instruments were thoroughly discussed, and some argued for being exempted from undertaking the task. The discussions abated over the years and the task eventually became a natural and mandatory part of the job encompassing all occupational groups. Of course, the ethical dilemma was not solved—it seemingly cannot be. Instead, it faded through habituation. Nevertheless, we must be aware that by not furthering our collective knowledge of risk assessments and the recidivism risk of mentally disordered offenders, we reduce our chances of properly taking care of those entitled to it—the patients as well as the members of the communities that they are supposed to reenter.

For mentally disordered offenders, especially those undergoing compulsory forensic psychiatric treatment, the ethical considerations are particularly delicate (136). They are not free to make certain decisions and some interventions are involuntary—including risk assessments. It is difficult to conduct research under such circumstances as it may be argued that subjects are solely consenting in order to appear compliant to the clinical personnel. Depending on the type and degree of mental disorders involved it may even be unclear whether consent can be given at all. That said, having subjects refuse to partake

in scientific research is not uncommon in these contexts, so fear of repercussions appears to have a minor impact. When we exclude a subgroup of patients on the premise that they are too vulnerable from an ethics viewpoint, are we truly minimizing harm to the patients or simply protecting ourselves from making uncomfortable ethical decisions? A recent report by the Swedish Agency for Health Technology Assessment and Assessment of Social Services found that the scientific foundation for whether the treatment of forensic psychiatric patients is effective or not is still uncertain (137). By not conducting research, how can we ever be certain?

In the USA, it has been argued that established predictors may act as a proxy for minority group affiliation, and that a focus on prior offenses may result in specific societal subgroups being unfairly considered high risk (138). Naturally, this could be exacerbated by using ethnicity or sexuality as an actual predictor in and of itself. This too is a sensitive issue that cannot easily be solved, especially considering that computer assisted modelling may entrench current concerns by using data that may or may not be biased. For example, using machine learning, HART (Harm Assessment Risk Tool) found 34 predictors based on over 100,000 police custody events to predict recidivism risk (139). Of these predictors, two relate to postcodes (although one is slated for removal in a future revision). Postcodes may arguably be a proxy for socioeconomic status, minority affiliation, or both, and the use of such a risk factor may lead to positive feedback loop effects. In short, police resources would naturally be funneled to the highest-risk postcode area, likely resulting in more arrests, which would further increase the predictor weight, leading to more police focus, et cetera. For those who consider machine learning an issue for a distant future, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a software similar to HART, has been in use in the USA for 20 years on more than one million offenders—and the algorithm used is considered a trade secret (140). Although general risk assessment tools are the most affected by this, there is no reason for practitioners who use more specialized instruments such as the Static-99R to rest easy. Fortunately, some researchers have started addressing potential issues with machine learning models and risk assessment instruments (134).

6 CONCLUSIONS

In sum, this thesis demonstrated that MDSOs that underwent an FPI as part of their legal process demonstrate high rates of major mental illness and recidivism rates comparable to those of international studies. Despite being followed for up to 24 years after the index offense, many recidivists were reconvicted early on during the follow-up period, indicating that shorter studies with larger samples may be able to take the place of long, resource-heavy follow-up studies.

Furthermore, the actuarial risk assessment instrument Static-99R can be applied to MDSOs in Sweden by using the routine norm group estimates for predicting and conveying recidivism risk for new sexual offenses. Yet, by using the tables published in the evaluators' workbook as described in the coding rules, the recidivism risk for an offender may be overestimated. The magnitude of overestimation is not statistically large enough to invalidate the use of the Static-99R, however. It is also plausible to use the Static-99R to predict recidivism risk over extended periods into the future far beyond what is recommended by the coding rules and the evaluators' workbook—up to and including 20 years after release.

Tentatively, the Static-99R performs adequately on MDSOs in Sweden from a psychometric standpoint guided by modern test theory. However, several issues arose that, if reproduced by additional studies, may require addressing in future revisions of the instrument. In particular, the results from the Rasch analysis using the MDSO cohort showed that it may be possible to reduce the response categories for items 1 and 5 without losing any psychometric value. Given that this indication is replicated in future research projects, preferably using larger samples as well as broad samples of general sexual offenders, such a change to the items in future revisions of the Static-99R may simplify the scoring procedure for clinicians specifically assessing offenders in the MDSO subgroup.

7 STRENGTHS AND LIMITATIONS

One of the most substantial limitations is the low number of subjects included in all three of the studies, as this affects the results of the thesis and, consequently, the conclusions drawn from these results. The cohort consisted of 146 MDSOs, with the specific number of subjects varying from analysis to analysis: from as few as 100 up to all 146. In addition to the size of the cohort, and in part as a result thereof, the number of recidivists were also few in number. Although the fixed five-year base rate in Study II matched similar validation studies (65,96), the Static-99R risk categories I, II, and III produced zero recidivists. While offenders sorted into categories I and II generally desist from reoffending, the number of recidivists was expected to be approximately 2 and 4 individuals in the routine and high-risk/high-need norm groups respectively. The negative impact is somewhat mitigated by the use of ratios from regression models for calibration (P/E index) rather than the observed number of recidivists (E/O index), but a model is only as good as the data it is built on. The statistical power of the E/O and P/E indexes depends exclusively on the number of recidivists—not on the number of subjects included in the study (97). This is particularly problematic for studies researching recidivism with low base rates such as sexual offenses. The only solution is to use larger samples or extremely recidivistic offenders, and while the former is technically difficult to obtain, the latter does not necessarily reflect the general offender that the risk assessment instrument is targeting. As such, the issue of base rates is largely beyond the control of researchers (38,141–143), but it can be alleviated somewhat by massively expanding the number of subjects included in a study, which will result in less fragile regression models.

Why choose 146 subjects? The research project utilized data from previous projects established in 1997 and 2002, which included all males undergoing an FPI for a sexual offense between 1993 and 1997. When studying MDSOs, two inclusion criteria are vital— mental disorder and sexual offense—but there are no universally agreed upon definitions for either of those terms. Roughly 1% of convicted offenders submit to an FPI (24), amounting to 550 FPIs every year (37), and FPIs do not merely target sexual offenders. However, by cross-referencing data from the National Council for Crime Prevention, this cohort represented 17% of all sexual offenders with adult victims convicted between 1993 and 1997. The cohort thus included 94% of sexual offenders sentenced to forensic psychiatric care and 10% of those sentenced to prison or probation during that period. Population wise, Sweden is a small nation compared to the

USA or Canada, but we still need to examine how well we can predict recidivism in our local populations using the data at hand instead of simply expecting international results to be applicable.

A major strength of the thesis is the level of clinical detail afforded by using the FPIs of all 146 subjects. This generally difficult-to-acquire data is extensive, spanning the entire life of the offender and their recent clinical state in particular. By using diagnoses made during intensive four-week assessments by teams of specialists, the quality of the data is higher than that available from other measurements such as the number of hospitalizations within a specific timeframe prior to the index offense. Additionally, the biographic section of the FPIs occasionally gives insight into parts of prior offenses already culled from other registers, or into parts of the index offense sentence otherwise redacted by the courts. In short, the FPIs are uniquely effective when used appropriately in clinical research and should be included in research on forensic psychiatric populations when available.

8 FUTURE PERSPECTIVES

8.1 Clinical application

By confirming the predictive strength of the Static-99R and the similarities between the cohort of MDSOs and the routine norm group, the use of the Static-99R on MDSOs in a Swedish clinical setting has been found to be appropriate. Considering that it is already being used in everyday practice, this is a reassuring result, despite the flaws primarily related to cohort size. While the results of this thesis affirm the use of the Static-99R on MDSOs in common forensic psychiatric settings, its use in other settings where MDSOs are prevalent is also supported. As sexual offenders undergoing risk assessments as part of converting their life sentence to a fixed number of years were not represented in the study cohort, the use of Static-99R on this specific offender subgroup is not adequately validated. While this does not preclude clinicians from using the instrument during those particular risk assessments, neither does the results from this thesis specifically support any such decision.

8.2 Directions for future research

In order to confirm the results of this thesis, additional, larger validation studies on Swedish MDSOs are required. Considering the high quality of data on offender history that is available to researchers in Sweden, this can be achieved without extensive resources for general sexual offenders, but less so for MDSOs. The argument has been made that samples used should be large enough to result in at least 100 recidivists for stable regression estimates (119,144). As this thesis found a general base rate of approximately 0.1, a sample of no less than 1,000 subjects would be desired for a more extensive validation study. This is large number of subjects for such a specific offender subgroup, particularly for a country the size of Sweden, and it would be even more difficult to assemble enough subjects to consider potential clinical subgroups within the MDSO population or the particularly small group of sexual offenders sentenced to life imprisonment. While the routine sample is generally recommended for use when there are no locally developed norm groups available (67) (which is usually the case), there is much to gain from attempting to develop specific norms based on large Swedish samples. As a tentative first step, cooperative efforts by the Nordic countries in pooling their data may be a way forward to eventually develop separate national norms.

After all, there has been previous research conducted in both Denmark and Sweden using non-revised Static-99 (57,145) and more recently in Norway using the revised Static-99R (146).

In cases where high-quality data on specific MDSOs is unavailable (even by using more lenient inclusion criteria than in this thesis), general prison populations would be adequate for establishing a routine norm group that is likely to be a better fit than those available internationally. Naturally, this requires the aggregation of data in large national databases that are available for researchers and clinicians, which would require extensive ethical discussion, but it is a reasonable aspiration. Such a database is available in Texas, USA, where preliminary steps have been taken to develop local norms for the prison population in that very state (147,148). By establishing a similar database in a Swedish setting, up-to-date Swedish norms would be possible for broad populations of sexual offenders, further improving the scientific and clinical foundation on which the treatment of sexual offenders depends (149).

APPENDIX

A.   Static-99R coding sheet

Static-99R – TALLY SHEET

Assessment date: _____   Date of release from index sex offence: _____

Item #	Risk Factor	Codes			Score
1	Age at release from index sex offence	Aged 18 to 34.9 Aged 35 to 39.9 Aged 40 to 59.9 Aged 60 or older			1 0 -1 -3
2	Ever lived with a lover	Ever lived with lover for at least two years? Yes No			 0 1
3	Index non-sexual violence - Any convictions	No Yes			0 1
4	Prior non-sexual violence - Any convictions	No Yes			0 1
5	Prior sex offences	Charges 0 1,2 3-5 6+	Convictions 0 1 2,3 4+		0 1 2 3
6	Four or more prior sentencing dates (excluding index)	3 or less 4 or more			0 1
7	Any convictions for non-contact sex offences	No Yes			0 1
8	Any unrelated victims	No Yes			0 1
9	Any stranger victims	No Yes			0 1
10	Any male victims	No Yes			0 1
	Total Score	Add up scores from individual risk factors			

Nominal Risk Levels (2016 version)	Total	Risk Level
	-3, -2,	I - Very Low Risk
	-1, 0,	II - Below Average Risk
	1, 2, 3	III - Average Risk
	4, 5	IVa - Above Average Risk
	6 and higher	IVb -Well Above Average Risk

There [was, was not] sufficient information available to complete the Static-99R score following the coding manual (2016 version). I believe that this score [fairly represents, does not fairly represent] the risk presented by Mr. XXXX at this time. Comments/Explanation: _____

_____   _____   _____
(Evaluator name)   (Evaluator signature)   (Date)

Figure A.1. Static-99R coding sheet (also known as the tally sheet), as presented in the coding rules (53)*. Printed in keeping with the intent of Canada's knowledge dissemination policy (copyright held by Public Safety Canada).*

B.   Exploring the effects of pre-release recidivism on calibration

Four subjects were reconvicted for a sexual offense while undergoing forensic psychiatric care, three of whom were not reconvicted during the fixed five-year follow-up and one who was. This gave the impression that three subjects did not commit new sexual offenses, which, while technically correct, is untrue in a pragmatic sense. Considering the low base rate for sexual recidivism, these uncounted reconvictions may have influenced the P/E indexes of one or more risk categories, which is explored further in this appendix.

The three subjects in question scored 1, 5, and 7 points on the Static-99R, corresponding to risk categories III, IVa, and IVb respectively. This raised the total number of recidivists in each category, in order, from 0 to 1, 3 to 4, and 8 to 9, from a total of 11 to 14 out of 100 subjects (a 30% increase in observed recidivists).

When performing the calibration analyses based on the number of expected and predicted recidivists, the routine norm group was still preferable compared to the high-risk/high-need norm group, although the differences were no longer statistically significant, as shown in Table B.1. Predicted recidivists were now 11% lower than expected in the routine norm group across all categories, compared to 31% in the high-risk/high need norm group. While this was an improvement, there was no longer a consistent pattern of over- or under-prediction for all risk categories as compared to any of the norm groups. Instead, more recidivists than expected were predicted in the two lowest risk categories (I and II), with this figure stabilizing in the three higher risk categories (III–IVb). The risk category best reflecting the expected number of recidivists in the routine norm group was IVa, followed closely by III.

Table B.1. Alternative number of recidivists across Static-99R risk categories for a five-year fixed follow-up period which includes three additional subjects who reoffended before release. The observed number of recidivists (O) is based on the alternative outcome for the study cohort, the predicted number of recidivists (P) on the alternative regression model for the study cohort, and the expected number of recidivists on the regression model for the Static-99R routine norm group (E_1) and high-risk/high-need norm group (E_2). The table is adapted from Table 5.

| Category | n | Number of recidivists | | | | P/E_1 (CI 95%) | P/E_2 (CI 95%) |
		O	P	E_1	E_2		
I	1	0	0.02	0.01	0.04	1.42	0.43
II	6	0	0.19	0.15	0.40	1.27	0.47
III	34	1	2.07	1.90	3.80	1.09 (0.28-4.25)	0.54 (0.14-2.13)
IVa	28	4	3.37	3.58	5.31	0.94 (0.32-2.74)	0.64 (0.22-1.85)
IVb	31	9	8.37	10.10	10.62	0.83 (0.42-1.63)	0.79 (0.40-1.55)
Total	100	14	14.02	15.75	20.18	0.89 (0.53-1.50)	0.69 (0.41-1.17)

Note. A CI of 95% that includes 1 indicates no statistically significant difference between the predicted and the expected number of recidivists ($p > .05$). Due to the low number of expected recidivists, a CI was uncalculatable for risk categories I and II.

C. Estimations using the preliminary 2021 norms

Three weeks before this thesis was going to print Lee and Hanson released new preliminary norms for the routine norm group (12). Their overall conclusion was that while the new five-year norms demonstrated slightly lower recidivism risk across all risk categories, the difference was not considered large enough to cause any major concern.

The fact that the P/E index demonstrated a propensity for the study cohort to reoffend at lesser rates than expected according to the norm group warranted new estimations derived from the 2021 norms. Consequently, new recidivism rates for every Static-99R total score were calculated using the intercept and slope from the 2021 regression model. The new comparisons resulted in an improved Brier score of 0.090 and a 14% higher total P/E index of 0.80— closer to a perfect 1. Additionally, when correcting for the three pre-release offending subjects discussed in Appendix B, a further improved total P/E index of 1.02 was achieved.

While it is the intention of Lee and Hanson that the new norms supersede the prior ones from 2016, it is not yet clear when a new official evaluators' workbook will be released. Until then, the new norms may still freely be used by clinicians, and as demonstrated above using the MDSO study cohort, they appear to be preferable regarding calibration.

REFERENCES

1.   Breiding MJ, Smith SG, Basile KC, Walters ML, Chen J, Merrick MT. Prevalence and characteristics of sexual violence, stalking, and intimate partner violence victimization—national intimate partner and sexual violence survey, United States, 2011. MMWR Surveill Summ [Internet]. 2014 Sep;63(8):1–18. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25188037

2.   Smith SG, Zhang X, Basile KC, Merrick MT, Wang J, Kresnow M, et al. National Intimate Partner and Sexual Violence Survey (NISVS): 2015 data brief - Update Release [Internet]. National Center for Injury Prevention and Control, Centers for Disease Control and Prevention. Atlanta, GA; 2018. Available from: https://www.cdc.gov/violenceprevention/pdf/2015data-brief508.pdf

3.   Walker HE, Freud JS, Ellis RA, Fraine SM, Wilson LC. The prevalence of sexual revictimization: a meta-analytic review. Trauma, Violence, Abus [Internet]. 2019 Jan;20(1):67–80. Available from: https://doi.org/10.1177/1524838017692364

4.   Peterson C, DeGue S, Florence C, Lokey CN. Lifetime economic burden of rape among U.S. adults. Am J Prev Med [Internet]. 2017 Jun;52(6):691–701. Available from: https://dx.doi.org/10.1016/j.amepre.2016.11.014

5.   The Swedish National Council for Crime Prevention. Swedish crime survey 2020: English summary of Brå report 2020:8 [Internet]. Stockholm (SE); 2020. Available from: https://www.bra.se/bra-in-english/home/publications/archive/publications/2020-10-15-swedish-crime-survey-2020.html

6.   Craig LA, Browne KD, Beech AR. Assessing risk in sex offenders: a practitioner's guide. 1st ed. Chichester, West Sussex: John Wiley & Sons, Ltd; 2008. 270 p.

7.   Harris AJR, Hanson RK. Sex offender recidivism: a simple question [Internet]. Public Safety and Emergency Preparedness Canada. Ottawa, ON (CA): Solicitor General of Canada; 2004. Available from: https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/sx-ffndr-rcdvsm

8.   Hanson RK, Bussière MT. Predicting relapse: a meta-analysis of sexual offender recidivism studies. J Consult Clin Psychol [Internet]. 1998;66(2):348–62. Available from: https://doi.org/10.1037/0022-006X.66.2.348

9.   Langan PA, Schmitt EL, Durose MR. Recidivism of sex offenders released

from prison in 1994 (NCJ 198381) [Internet]. Washington, DC: The U.S. Deparment of Justice, Office of Justice Programs, Bureau of Justice Statistics; 2003. Available from: https://www.ojp.gov/ncjrs/virtual-library/abstracts/recidivism-sex-offenders-released-prison-1994

10. Vess J, Skelton A. Sexual and violent recidivism by offender type and actuarial risk: reoffending rates for rapists, child molesters and mixed-victim offenders. Psychol Crime Law [Internet]. 2010;16(7):541–54. Available from: https://doi.org/10.1080/10683160802612908

11. Rettenberger M, Briken P, Turner D, Eher R. Sexual offender recidivism among a population-based prison sample. Int J Offender Ther Comp Criminol [Internet]. 2015;59(4):424–44. Available from: https://doi.org/10.1177/0306624X13516732

12. Lee SC, Hanson RK. Updated 5-year and new 10-year sexual recidivism rate norms for Static-99R with routine/complete samples. Law Hum Behav [Internet]. 2021 Feb;45(1):24–38. Available from: https://doi.org/10.1037/lhb0000436

13. Greenberg DM. Sexual recidivism in sex offenders. Can J Psychiatry [Internet]. 1998;43(5):459–65. Available from: https://doi.org/10.1177/070674379804300502

14. Langevin R, Curnoe S, Fedoroff P, Bennett R, Langevin M, Peever C, et al. Lifetime sex offender recidivism: a 25-year follow-up study. Can J Criminol Crim Justice [Internet]. 2004 Oct;46(5):531–52. Available from: https://doi.org/10.3138/cjccj.46.5.531

15. Stinson JD, Becker J V. Sexual offenders with serious mental illness: prevention, risk, and clinical concerns. Int J Law Psychiatry [Internet]. 2011;34(3):239–45. Available from: https://doi.org/10.1016/j.ijlp.2011.04.011

16. Harris AJ, Fisher W, Veysey BM, Ragusa LM, Lurigio AJ. Sex offending and serious mental illness: directions for policy and research. Crim Justice Behav [Internet]. 2010;37(5):596–612. Available from: https://doi.org/10.1177/0093854810363773

17. Kingston DA, Olver ME, Harris M, Wong SCP, Bradford JM. The relationship between mental disorder and recidivism in sexual offenders. Int J Forensic Ment Health [Internet]. 2015;14(1):10–22. Available from: https://doi.org/10.1080/14999013.2014.974088

18. Lewis ET, Dwyer RG. Psychosis and sexual offending: a review of current literature. Int J Offender Ther Comp Criminol [Internet]. 2018;62(11):3372–84. Available from: https://doi.org/10.1177/0306624X17740016

19. Alish Y, Birger M, Manor N, Kertzman S, Zerzion M, Kotler M, et al. Schizophrenia sex offenders: a clinical and epidemiological comparison

study. Int J Law Psychiatry [Internet]. 2007;30(6):459–66. Available from: https://doi.org/10.1016/j.ijlp.2007.09.006

20. Långström N, Sjöstedt G, Grann M. Psychiatric disorders and recidivism in sexual offenders. Sex Abus A J Res Treat [Internet]. 2004;16(2):139–50. Available from: https://doi.org/10.1023/B:SEBU.0000023062.56389.ed

21. SOU 2002:3. Psykisk störning, brott och ansvar [Internet]. Stockholm (SE): Ministry of Justice; 2002. Available from: https://www.regeringen.se/rattsliga-dokument/statens-offentliga-utredningar/2002/01/sou-20023/

22. SOU 2012:17. Psykiatrin och lagen–tvångsvård, straffansvar och samhällsskydd [Internet]. Stockholm (SE): Ministry of Health and Social Affairs; 2012. Available from: https://www.regeringen.se/rattsliga-dokument/statens-offentliga-utredningar/2012/04/sou-201217/

23. Svennerlind C, Nilsson T, Kerekes N, Andiné P, Lagerkvist M, Forsman A, et al. Mentally disordered criminal offenders in the Swedish criminal system. Int J Law Psychiatry [Internet]. 2010;33(4):220–6. Available from: https://doi.org/10.1016/j.ijlp.2010.06.003

24. Hallquist T. Rättspsykiatrisk undersökning. In: Granhag PA, Christianson SÅ, editors. Handbok i rättspsykologi. 1st ed. Stockholm: Liber; 2008. p. 453–66.

25. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5) [Internet]. 5th ed. Washington, DC: American Psychiatric Association; 2013. 991 p. Available from: https://doi.org/10.1176/appi.books.9780890425596

26. Meehl PE. Clinical versus statistical prediction: a theoretical analysis and a review of the evidence [Internet]. Minneapolis: University of Minnesota Press; 1954. 149 p. Available from: https://doi.org/10.1037/11281-000

27. Burgess EW. Factors determining success or failure on parole. In: Bruce AA, editor. The Workings of the Indeterminate Sentence Law and Parole in Illinois. Springfield, IL: Illinois State Parole Board; 1928. p. 205–49.

28. Craig LA, Rettenberger M. A brief history of sexual offender risk assessment. In: Laws DR, O'Donohue W, editors. Treatment of sex offenders [Internet]. 1st ed. Springer International Publishing; 2016. p. 19–44. Available from: https://doi.org/10.1007/978-3-319-25868-3_2

29. Hanson RK, Morton-Bourgon K. Predictors of sexual recidivism: an updated meta-analysis [Internet]. Public Works and Government Services Canada and the Department of Public Safety and Emergency Preparedness. 2004.

Available from: https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/2004-02-prdctrs-sxl-rcdvsm-pdtd/

30. Hanson RK, Morton-Bourgon KE. The characteristics of persistent sexual offenders: a meta-analysis of recidivism studies. J Consult Clin Psychol [Internet]. 2005 Dec;73(6):1154–63. Available from: https://doi.org/10.1037/0022-006X.73.6.1154

31. Mann RE, Hanson RK, Thornton D. Assessing risk for sexual recidivism: some proposals on the nature of psychologically meaningful risk factors. Sex Abus A J Res Treat [Internet]. 2010 Jun;22(2):191–217. Available from: https://doi.org/10.1177/1079063210366039

32. Babchishin KM, Karl Hanson R, Hermann CA. The characteristics of online sex offenders: a meta-analysis. Sex Abus A J Res Treat [Internet]. 2011;23(1):92–123. Available from: https://doi.org/10.1177/1079063210370708

33. Hart SD, Michie C, Cooke DJ. Precision of actuarial risk assessment instruments. Br J Psychiatry [Internet]. 2007 May;190(S49):s60–5. Available from: https://doi.org/10.1192/bjp.190.5.s60

34. The Swedish National Council for Crime Prevention. Livstidsdomar–utveckling och faktisk strafftid [Internet]. Stockholm (SE): Ministry of Justice; 2015. Available from: https://www.bra.se/publikationer/arkiv/publikationer/2015-03-24-livstidsdomar---utveckling-och-faktisk-strafftid.html

35. SFS 2006:45. Lag (2006:45) om omvandling av fängelse på livstid [Internet]. Sweden: Stockholm: Ministry of Justice; 2006. Available from: https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/lag-200645-om-omvandling-av-fangelse-pa_sfs-2006-45

36. Swedish National Forensic Psychiatric Register RättspsyK. Annual 2019 (Årsrapport 2019) [Internet]. Gothenburg (SE); 2019. Available from: https://registercentrum.blob.core.windows.net/rattspsyk/r/Arsrapport-2019-rygqgg6dUU.pdf

37. The National Board of Forensic Medicine. Årsredovisning 2020 [Internet]. Stockholm (SE); 2020. Available from: https://www.rmv.se/aktuellt/las-rattsmedicinalverkets-arsredovisning-for-2020/

38. Scurich N, John RS. The dark figure of sexual recidivism. Behav Sci Law [Internet]. 2019 Mar;37(2):158–75. Available from: https://doi.org/10.1002/bsl.2400

39. SFS 1962:700. The Swedish Criminal Code (Brottsbalken) [Internet]. Sweden: Stockholm: Ministry of Justice; 1962. Available from: https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/brottsbalk-1962700_sfs-1962-700

40.   AgeOfConsent.net. Highest and Lowest Ages of Consent [Internet]. AgeOfConsent.net. 2021 [cited 2021 Feb 10]. Available from: https://www.ageofconsent.net/highest-and-lowest

41.   Knight RA, Rosenberg R, Schneider B. Classification of sexual offenders: Perspectives, methods, and validation. In: Burgess AW, editor. Rape and sexual assault: a research handbook. New York: Garland Publishing, Inc.; 1985. p. 222–93.

42.   Guay JP, Proulx J, Cusson M, Ouimet M. Victim-choice polymorphia among serious sex offenders. Arch Sex Behav [Internet]. 2001;30(5):521–33. Available from: https://doi.org/10.1023/A:1010291201588

43.   Sjöstedt G, Långström N, Sturidsson K, Grann M. Stability of modus operandi in sexual offending. Crim Justice Behav [Internet]. 2004;31(5):609–23. Available from: https://doi.org/10.1177/0093854804267094

44.   Planty M, Langton L, Krebs C, Berzofsky M, Smiley-McDonald H. Female victims of sexual violence, 1994-2010 (NCJ 240655) [Internet]. Washington, DC: The U.S. Deparment of Justice, Office of Justice Programs, Bureau of Justice Statistics; 2013. Available from: https://www.ojp.gov/ncjrs/virtual-library/abstracts/female-victims-sexual-violence-1994-2010

45.   Robertiello G, Terry KJ. Can we profile sex offenders? A review of sex offender typologies. Aggress Violent Behav [Internet]. 2007 Sep;12(5):508–18. Available from: https://doi.org/10.1016/j.avb.2007.02.010

46.   Levenson JS, Becker J, Morin JW. The relationship between victim age and gender crossover among sex offenders. Sex Abuse [Internet]. 2008;20(1):43–60. Available from: https://doi.org/10.1177/1079063208314819

47.   Cann J, Friendship C, Gozna L. Assessing crossover in a sample of sexual offenders with multiple victims. Leg Criminol Psychol [Internet]. 2007;12(1):149–63. Available from: https://doi.org/10.1348/135532506X112439

48.   Harris DA, Knight RA, Smallbone S, Dennison S. Postrelease specialization and versatility in sexual offenders referred for civil commitment. Sex Abus A J Res Treat [Internet]. 2011 Jun;23(2):243–59. Available from: https://doi.org/10.1177/1079063210384276

49.   Sim DJ, Proeve M. Crossover and stability of victim type in child molesters. Leg Criminol Psychol [Internet]. 2010;15(2):401–13. Available from: https://doi.org/10.1348/135532509X473869

50.   Marshall WLL. Diagnostic issues, multiple paraphilias, and comorbid disorders in sexual offenders: their incidence and treatment. Aggress Violent

Behav [Internet]. 2007 Jan;12(1):16–35. Available from: https://doi.org/10.1016/j.avb.2006.03.001

51. McPhail I V., Hermann CA, Fernane S, Fernandez YM, Nunes KL, Cantor JM. Validity in phallometric testing for sexual interests in children: a meta-analytic review. Assessment [Internet]. 2019 Apr;26(3):535–51. Available from: https://doi.org/10.1177/1073191117706139

52. Harris DA, Mazerolle P, Knight RA. Understanding male sexual offending: a comparison of general and specialist theories. Crim Justice Behav [Internet]. 2009;36(10):1051–69. Available from: https://doi.org/10.1177/0093854809342242

53. Phenix A, Fernandez Y, Harris AJR, Helmus M, Hanson RK, Thornton D. Static-99R coding rules: revised–2016 [Internet]. Ottawa, ON (CA); 2016. Available from: http://static99.org/pdfdocs/Coding_manual_2016_v2.pdf

54. Phenix A, Helmus L-M, Hanson RK. Static-99R & Static-2002R evaluators' workbook [Internet]. Ottawa, ON (CA); 2016. Available from: http://static99.org/pdfdocs/Evaluators_Workbook_2016-10-19.pdf

55. Kelley SM, Ambroziak G, Thornton D, Barahal RM. How do professionals assess sexual recidivism risk? An updated survey of practices. Sex Abus [Internet]. 2020 Feb;32(1):3–29. Available from: https://doi.org/10.1177/1079063218800474

56. Rice AK, Boccaccini MT, Harris PB, Hawes SW. Does field reliability for Static-99 scores decrease as scores increase? Psychol Assess [Internet]. 2014 Dec;26(4):1085–94. Available from: https://doi.org/10.1037/pas0000009

57. Sjöstedt G, Långström N. Actuarial assessment of sex offender recidivism risk: a cross-validation of the RRASOR and the Static-99 in Sweden. Law Hum Behav [Internet]. 2001;25(6):629–45. Available from: https://doi.org/10.1023/A:1012758307983

58. Långström N. Accuracy of actuarial procedures for assessment of sexual offender recidivism risk may vary across ethnicity. Sex Abus A J Res Treat [Internet]. 2004 Apr;16(2):107–20. Available from: https://doi.org/10.1023/B:SEBU.0000023060.61402.07

59. Lee SC, Hanson RK, Fullmer N, Neeley J, Ramos K. The predictive validity of Static-99R over 10 years for sexual offenders in California: 2018 update [Internet]. Ottawa, ON (CA): Carleton University and California Department of Justice; 2018. Available from: https://saratso.org/pdf/Lee_Hanson_Fullmer_Neeley_Ramos_2018_The_Predictive_Validity_of_S_.pdf

60. Smid WJ, Kamphuis JH, Wever EC, Van Beek DJ. A comparison of the predictive properties of nine sex offender risk assessment instruments. Psychol Assess [Internet]. 2014 Sep;26(3):691–703. Available from:

https://doi.org/10.1037/a0036616

61.   Olver ME, Mundt JC, Thornton D, Beggs Christofferson SM, Kingston DA, Sowden JN, et al. Using the Violence Risk Scale-Sexual Offense version in sexual violence risk assessments: updated risk categories and recidivism estimates from a multisite sample of treated sexual offenders. Psychol Assess [Internet]. 2018 Jul;30(7):941–55. Available from: https://doi.org/10.1037/pas0000538

62.   Thornton D, Hanson RK, Kelley SM, Mundt JC. Estimating lifetime and residual risk for individuals who remain sexual offense free in the community: practical applications. Sex Abus A J Res Treat [Internet]. 2021 Sep;33(1):3–33. Available from: https://doi.org/10.1177/1079063219871573

63.   Brouillette-Alarie S, Babchishin KM, Hanson RK, Helmus L-M. Latent constructs of the Static-99R and Static-2002R. Assessment [Internet]. 2016 Feb;23(1):96–111. Available from: https://doi.org/10.1177/1073191114568114

64.   Brouillette-Alarie S, Proulx J, Hanson RK. Three central dimensions of sexual recidivism risk. Sex Abus A J Res Treat [Internet]. 2018 Sep;30(6):676–704. Available from: https://doi.org/10.1177/1079063217691965

65.   Helmus L, Hanson RK, Thornton D, Babchishin KM, Harris AJR. Absolute recidivism rates predicted by Static-99R and Static-2002R sex offender risk assessment tools vary across samples. Crim Justice Behav [Internet]. 2012 Sep;39(9):1148–71. Available from: https://doi.org/10.1177/0093854812443648

66.   Helmus LM, Babchishin KM. Primer on risk assessment and the statistics used to evaluate its accuracy. Crim Justice Behav [Internet]. 2017 Jan;44(1):8–25. Available from: https://doi.org/10.1177/0093854816678898

67.   Hanson RK, Babchishin KM, Helmus L, Thornton D. Quantifying the relative risk of sex offenders. Sex Abus A J Res Treat [Internet]. 2013 Oct;25(5):482–515. Available from: https://doi.org/10.1177/1079063212469060

68.   Reeves SG, Ogloff JRP, Simmons M. The predictive validity of the Static-99, Static-99R, and Static-2002/R: which one to use? Sex Abus [Internet]. 2018 Dec;30(8):887–907. Available from: https://doi.org/10.1177/1079063217712216

69.   Hanson RK, Lunetta A, Phenix A, Neeley J, Epperson D. The field validity of Static-99/R sex offender risk assessment tool in California. J Threat Assess Manag [Internet]. 2014 Jun;1(2):102–17. Available from:

https://doi.org/10.1037/tam0000014

70.   Hanson RK, Thornton D. Static 99: improving actuarial risk assessments for sex offenders [Internet]. The Department of the Solicitor General and Her Majesty's Prison Service. Ottawa, ON (CA); 1999. Available from: https://www.securitepublique.gc.ca/cnt/rsrcs/pblctns/sttc-mprvng-actrl/sttc-mprvng-actrl-eng.pdf

71.   Helmus L, Thornton D, Hanson RK, Babchishin KM. Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders. Sex Abus A J Res Treat [Internet]. 2012 Feb;24(1):64–101. Available from: https://doi.org/10.1177/1079063211409951

72.   Elwood RW, Kelley SM, Mundt JC. The 2015 Static-99R: alternative recidivism tables for high-risk offenders. Int J Offender Ther Comp Criminol [Internet]. 2017 Oct;61(14):1593–605. Available from: https://doi.org/10.1177/0306624X15623803

73.   DeClue G, Zavodny DL. Forensic use of the Static-99R: choosing a comparison group. Open Access J Forensic Psychol [Internet]. 2013;5:151–82. Available from: https://www.researchgate.net/publication/262911023_Forensic_Use_of_the_Static-99R_Part_3_Choosing_a_Comparison_Group

74.   Swedish National Courts Administration. Swedish-English Glossary [Internet]. 5th ed. Stockholm (SE): The communication department, Swedish National Courts Administration; 2019. 286 p. Available from: www.domstol.se/ordlista

75.   Ludvigsson JF, Otterblad-Olausson P, Pettersson BU, Ekbom A. The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. Eur J Epidemiol [Internet]. 2009;24(11):659–67. Available from: https://doi.org/10.1007/s10654-009-9350-y

76.   The jamovi project. jamovi 1.2.22 [Software]. 2020. Available from: https://www.jamovi.org

77.   The jamovi project. jamovi 1.2.9 [Software]. 2020. Available from: https://www.jamovi.org

78.   RUMM Laboratory Pty Ltd. RUMM2030Plus [Software]. 2020. Available from: http://www.rummlab.com.au

79.   StataCorp. Stata statistical software 16.1 [Software]. College Station, TX: StataCorp LLC; 2019. Available from: https://www.stata.com

80.   IBM Corp. IBM SPSS Statistics for Windows 25.0 [Software]. Armonk, NY: IBM Corp; 2017. Available from: https://www.ibm.com/analytics/spss-statistics-software

81.   The jamovi project. jamovi 1.6.15 [Software]. 2021. Available from:

https://www.jamovi.org

82.   Microsoft Corporation. Microsoft Excel [Software]. Microsoft Corporation; 2018. Available from: https://office.microsoft.com/excel

83.   Hoenig JM, Heisey DM. The abuse of power. Am Stat [Internet]. 2001 Feb;55(1):19–24. Available from: https://doi.org/10.1198/000313001300339897

84.   Senn SJ. Letter to the editor: power is indeed irrelevant in interpreting completed studies. BMJ [Internet]. 2002 Nov;325(7375):1304. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12458264

85.   Gelman A. Letter to the editor: don't calculate post-hoc power using observed estimate of effect size. Ann Surg [Internet]. 2019 Jan;269(1):e9–10. Available from: https://doi.org/10.1097/SLA.0000000000002908

86.   Goodman SN. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med [Internet]. 1994 Aug;121(3):200–6. Available from: https://doi.org/10.7326/0003-4819-121-3-199408010-00008

87.   American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-III-R). 3rd, revised ed. Washington, DC: American Psychiatric Association; 1987. 567 p.

88.   American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-IV). 4th ed. Washington, DC: American Psychiatric Association; 1994. 886 p.

89.   The Swedish National Council for Crime Prevention. Brott som begås under rättspsykiatrisk vård [Internet]. Stockholm (SE): Ministry of Justice; 2019. Available from: https://bra.se/publikationer/arkiv/publikationer/2019-04-02-brott-som-begas-under-rattspsykiatrisk-vard.html

90.   Machin D, Cheung YB, Parmar MK. Survival analysis: a practical approach [Internet]. 2nd ed. Chichester, UK: John Wiley & Sons, Ltd; 2006. 266 p. Available from: https://doi.org/10.1002/0470034572

91.   Rossegger A, Gerth J, Seewald K, Urbaniok F, Singh JP, Endrass J. Current obstacles in replicating risk assessment findings: a systematic review of commonly used actuarial instruments. Behav Sci Law [Internet]. 2013 Jan;31(1):154–64. Available from: https://doi.org/10.1002/bsl.2044

92.   Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med [Internet]. 2016 Mar;15(2):155–63. Available from: https://doi.org/10.1016/j.jcm.2016.02.012

93.   Gross KD. Measurement revisited: reliability and validity statistics. In: Portney LG, editor. Foundations of clinical research: applications to evidence-based practice. 4th ed. Philadelphia, PA: F.A. Davis Company; 2020. p. 486–508.

94.   McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods [Internet]. 1996;1(1):30–46. Available from: https://doi.org/10.1037/1082-989X.1.1.30

95.   Portney LG. Diagnostic Accuracy. In: Portney LG, editor. Foundations of clinical research: applications to evidence-based practice. 4th ed. Philadelphia, PA: F.A. Davis Company; 2020. p. 509–28.

96.   Gonçalves LC, Gerth J, Rossegger A, Noll T, Endrass J. Predictive validity of the Static-99 and Static-99R in Switzerland. Sex Abus A J Res Treat [Internet]. 2020 Mar;32(2):203–19. Available from: https://doi.org/10.1177/1079063218821117

97.   Hanson RK. Assessing the calibration of actuarial risk scales: a primer on the E/O Index. Crim Justice Behav [Internet]. 2017;44(1):26–39. Available from: https://doi.org/10.1177/0093854816683956

98.   Rufibach K. Use of Brier score to assess binary predictions. J Clin Epidemiol [Internet]. 2010 Aug;63(8):938–9. Available from: https://doi.org/10.1016/j.jclinepi.2009.11.009

99.   Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol [Internet]. 2007 Mar;46(1):1–18. Available from: https://doi.org/10.1348/014466506X96931

100.   Masters GN. A Rasch model for partial credit scoring. Psychometrika [Internet]. 1982 Jun;47(2):149–74. Available from: https://doi.org/10.1007/BF02296272

101.   Bond TG, Fox CM. The partial credit Rasch model. In: Bond TG, Fox CM, editors. Applying the Rasch model: fundamental measurement in the human sciences. 1st ed. Psychology Press; 2001. p. 88–93.

102.   Rasch G. Probabilistic models for some intelligence and attainment tests. 1st ed. Chicago: University of Chicago Press; 1960. 199 p.

103.   Bessing B, Honan CA, van der Mei I, Taylor B V., Claflin SB. Development and psychometric properties of the Multiple Sclerosis Knowledge Assessment Scale: Rasch analysis of a novel tool for evaluating MS knowledge. Mult Scler J [Internet]. 2021;7(5):767–77. Available from: https://doi.org/10.1177/1352458520929626

104.   Steyer R. Classical (psychometric) test theory. In: Smelser NJ, Baltes PB, editors. International encyclopedia of the social & behavioral sciences [Internet]. 1st ed. Elsevier; 2001. p. 1955–62. Available from:

https://doi.org/10.1016/B0-08-043076-7/00721-X

105. Andrich D. A rating formulation for ordered response categories. Psychometrika [Internet]. 1978 Dec;43(4):561–73. Available from: https://doi.org/10.1007/BF02293814

106. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum [Internet]. 2007 Dec;57(8):1358–62. Available from: https://doi.org/10.1002/art.23108

107. Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. 3rd ed. Bond TG, Fox CM, editors. Routledge; 2015. 406 p.

108. Linacre JM. Sample size and item calibration stability [Internet]. Institute for Objective Measurement, Inc. 1994 [cited 2021 Mar 9]. Available from: https://www.rasch.org/rmt/rmt74m.htm

109. Kleppang AL, Steigen AM, Finbråten HS. Using Rasch measurement theory to assess the psychometric properties of a depressive symptoms scale in Norwegian adolescents. Health Qual Life Outcomes [Internet]. 2020 May;18(Article 127):1–8. Available from: https://doi.org/10.1186/s12955-020-01373-5

110. Robinson M, Johnson AM, Walton DM, MacDermid JC. A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/eRm/TAM/lordif). BMC Med Res Methodol [Internet]. 2019 Feb;19(Article 36):1–12. Available from: https://doi.org/10.1186/s12874-019-0680-5

111. Andrich D, Humphry SM, Marais I. Quantifying local, response dependence between two polytomous items using the Rasch model. Appl Psychol Meas [Internet]. 2012 Jun;36(4):309–24. Available from: https://doi.org/10.1177/0146621612441858

112. Hagell P. Testing rating scale unidimensionality using the principal component analysis (PCA)/t-test protocol with the Rasch model: the primacy of theory over statistics. Open J Stat [Internet]. 2014;4(6):456–65. Available from: https://doi.org/10.4236/ojs.2014.46044

113. Ho Yu C. A simple guide to the Item Response Theory (IRT) and Rasch modeling [Internet]. Creative Wisdom. 2017 [cited 2020 May 4]. Available from: https://www.creative-wisdom.com/computer/sas/IRT.pdf

114. Hagquist C, Bruce M, Gustavsson JP. Using the Rasch model in nursing research: an introduction and illustrative example. Int J Nurs Stud [Internet].

2009 Mar;46(3):380–93. Available from:
https://doi.org/10.1016/j.ijnurstu.2008.10.007

115. Berthelsen H, Westerlund H, Pejtersen JH, Hadzibajramovic E. Construct validity of a global scale for workplace social capital based on COPSOQ III. PLoS One [Internet]. 2019 Aug;14(8):e0221893. Available from: https://doi.org/10.1371/journal.pone.0221893

116. Marshall E, Miller HA, Cortoni F, Helmus LM. The Static-99R is not valid for women: predictive validity in 739 females who have sexually offended. Sex Abus A J Res Treat [Internet]. 2020 Jul;1–23. Available from: https://doi.org/10.1177/1079063220940303

117. Långström N, Babchishin KM, Fazel S, Lichtenstein P, Frisell T. Sexual offending runs in families: a 37-year nationwide study. Int J Epidemiol [Internet]. 2015;44(2):713–20. Available from: https://doi.org/10.1093/ije/dyv029

118. Rice ME, Harris GT. Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. Law Hum Behav [Internet]. 2005;29(5):615–20. Available from: https://doi.org/10.1007/s10979-005-6832-7

119. Hanson RK, Thornton D, Helmus L-M, Babchishin KM. What sexual recidivism rates are associated with Static-99R and Static-2002R scores? Sex Abus A J Res Treat [Internet]. 2016 Apr;28(3):218–52. Available from: https://doi.org/10.1177/1079063215574710

120. Hagquist C, Andrich D. Recent advances in analysis of differential item functioning in health research using the Rasch model. Health Qual Life Outcomes [Internet]. 2017 Dec;15(1):Article 181. Available from: https://doi.org/10.1186/s12955-017-0755-0

121. Lund C, Hofvander B, Forsman A, Anckarsäter H, Nilsson T. Violent criminal recidivism in mentally disordered offenders: a follow-up study of 13-20 years through different sanctions. Int J Law Psychiatry [Internet]. 2013;36(3–4):250–7. Available from: https://doi.org/10.1016/j.ijlp.2013.04.015

122. Fazel S, Sjöstedt G, Långström N, Grann M. Severe mental illness and risk of sexual offending in men. J Clin Psychiatry [Internet]. 2007;68(4):588–96. Available from: https://pubmed.ncbi.nlm.nih.gov/17474815/

123. Harsch S, Bergk JE, Steinert T, Keller F, Jockusch U. Prevalence of mental disorders among sexual offenders in forensic psychiatry and prison. Int J Law Psychiatry [Internet]. 2006;29(5):443–9. Available from: https://doi.org/10.1016/j.ijlp.2005.11.001

124. Cuddeback GS, Grady M, Blank Wilson A, Van Deinse T, Morrissey JP. Persons with severe mental illnesses and sex offenses: recidivism after prison release. Int J Offender Ther Comp Criminol [Internet]. 2019;63(12):2157–70.

Available from: https://doi.org/10.1177/0306624X19842994

125.   Novak B, McDermott BE, Scott CL, Guillory S. Sex offenders and insanity: an examination of 42 individuals found not guilty by reason of insanity. J Am Acad Psychiatry Law [Internet]. 2007;35(4):444–50. Available from: https://pubmed.ncbi.nlm.nih.gov/18086736/

126.   Bonta J, Blais J, Wilson HA. A theoretically informed meta-analysis of the risk for general and violent recidivism for mentally disordered offenders. Aggress Violent Behav [Internet]. 2014 May;19(3):278–87. Available from: https://doi.org/10.1016/j.avb.2014.04.014

127.   Craig LA, Giotakos O. Sexual offending in psychotic patients. In: Boer DP, Eher R, Craig LA, Miner MH, Pfäfflin F, editors. International perspectives on the assessment and treatment of sexual offenders [Internet]. 1st ed. Chichester, UK: John Wiley & Sons, Ltd; 2011. p. 463–78. Available from: https://doi.org/10.1002/9781119990420.ch24

128.   Hanson RK, Harris AJR, Helmus L, Thornton D. High-risk sex offenders may not be high risk forever. J Interpers Violence [Internet]. 2014 Oct;29(15):2792–813. Available from: https://doi.org/10.1177/0886260514526062

129.   Wormith JS. The risks of communicating sexual offender risk. J Threat Assess Manag [Internet]. 2014;1(3):162–78. Available from: https://doi.org/10.1037/tam0000019

130.   Hadžibajramović E, Schaufeli W, De Witte H. A Rasch analysis of the Burnout Assessment Tool (BAT). Hoefer S, editor. PLoS One [Internet]. 2020 Nov;15(11):e0242241. Available from: https://doi.org/10.1371/journal.pone.0242241

131.   Kleppang AL, Hagquist C. The psychometric properties of the Hopkins Symptom Checklist-10: a Rasch analysis based on adolescent data from Norway. Fam Pract [Internet]. 2016 Dec;33(6):740–5. Available from: https://doi.org/10.1093/fampra/cmw091

132.   Helmus L-M, Thornton D. Stability and predictive and incremental accuracy of the individual items of Static-99R and Static-2002R in predicting sexual recidivism. Crim Justice Behav [Internet]. 2015 Sep;42(9):917–37. Available from: https://doi.org/10.1177/0093854814568891

133.   Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas [Internet]. 2002;3(1):85–106. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11997586

134.   Barabas C, Dinakar K, Ito J, Virza M, Zittrain J. Interventions over

predictions: reframing the ethical debate for actuarial risk assessment. Proc Mach Learn Res [Internet]. 2018 Dec;81:1–15. Available from: http://arxiv.org/abs/1712.08238

135. Hannah-Moffat K. The uncertainties of risk assessment. Fed Sentencing Report [Internet]. 2015 Apr;27(4):244–7. Available from: https://doi.org/10.1525/fsr.2015.27.4.244

136. Munthe C, Radovic S, Anckarsäter H. Ethical issues in forensic psychiatric research on mentally disordered offenders. Bioethics [Internet]. 2010;24(1):35–44. Available from: https://doi.org/10.1111/j.1467-8519.2009.01773.x

137. SBU. Treatment and evaluation methods used in forensic psychiatry: mapping of systematic reviews (Behandling och bedömning i rättspsykiatrisk vård: en kartläggning av systematiska översikter) (SBU 264) [Internet]. Stockholm (SE): Swedish Agency for Health Technology Assessment and Assessment of Social Services; 2017. Available from: https://www.sbu.se/264e

138. Harcourt BE. Risk as a proxy for race. Fed Sentencing Report [Internet]. 2015 Apr;27(4):237–43. Available from: https://doi.org/10.1525/fsr.2015.27.4.237

139. Oswald M, Grace J, Urwin S, Barnes GC. Algorithmic risk assessment policing models: lessons from the Durham HART model and "experimental" proportionality. Inf Commun Technol Law [Internet]. 2018;27(2):223–50. Available from: https://doi.org/10.1080/13600834.2018.1458455

140. Babuta A. Innocent until predicted guilty? Artificial intelligence and police decision-making. RUSI Newsbrief [Internet]. 2018 Mar;38(2):1–4. Available from: https://rusi.org/sites/default/files/20180329_rusi_newsbrief_vol.38_no.2_bab uta_web.pdf

141. Donaldson T, Abbott BR, Michie C. Problems with the Static-99R Prediction Estimates and Confidence Intervals. Open Access J Forensic Psychol [Internet]. 2012;4:1–23. Available from: https://www.researchgate.net/publication/320626733_Problems_with_the_St atic-99R_Prediction_Estimates_and_Confidence_Intervals

142. Doren DM. Recidivism base rates, predictions of sex offender recidivism, and the "sexual predator" commitment laws. Behav Sci Law [Internet]. 1998;16(1):97–114. Available from: https://doi.org/10.1002/(SICI)1099-0798(199824)16:1%3C97::AID-BSL295%3E3.0.CO;2-K

143. Wollert R. Low base rates limit expert certainty when current actuarials are used to identify sexually violent predators: an application of Bayes's theorem. Psychol Public Policy, Law [Internet]. 2006;12(1):56–85. Available from: https://doi.org/10.1037/1076-8971.12.1.56

144. Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol [Internet]. 2005 Jun;58(5):475–83. Available from: https://doi.org/10.1016/j.jclinepi.2004.06.017

145. Bengtson S. Is newer better? A cross-validation of the Static-2002 and the Risk Matrix 2000 in a Danish sample of sexual offenders. Psychol Crime Law [Internet]. 2008 Apr;14(2):85–106. Available from: https://doi.org/10.1080/10683160701483104

146. Sandbukt IJ, Skardhamar T, Kristoffersen R, Friestad C. Testing the Static-99R as a global screen for risk of sex crime recidivism in a Norwegian routine sample. Sex Abus A J Res Treat [Internet]. 2020 Aug;1–18. Available from: https://doi.org/10.1177/1079063220951194

147. Rice AK. Predictive validity of Static-99 and Static-99R scores among offenders scored on multiple occasions [dissertation]. Huntsville, TX: Sam Houston State University; 2016.

148. Boccaccini MT, Rice AK, Maaike Helmus L, Murrie DC, Harris PB. Field validity of Static-99/R scores in a statewide sample of 34,687 convicted sexual offenders. Psychol Assess [Internet]. 2017;29(6):611–23. Available from: https://doi.org/10.1037/pas0000377

149. Polaschek DLL. An appraisal of the risk-need-responsivity (RNR) model of offender rehabilitation and its application in correctional treatment. Leg Criminol Psychol [Internet]. 2012 Feb;17(1):1–17. Available from: https://doi.org/10.1111/j.2044-8333.2011.02038.x