Sex offender risk assessment in the Netherlands: Towards a risk need responsivity oriented approach
Smid, W.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
CHAPTER 4
A COMPARISON OF THE PREDICTIVE PROPERTIES OF NINE SEX OFFENDER RISK ASSESSMENT INSTRUMENTS
ABSTRACT

Sex offender treatment is most effective when tailored to the Risk Need Responsivity (RNR) principles, dictating that treatment levels should match risk levels as assessed by structured risk assessment instruments. The predictive properties, missing values, and inter rater agreement of the scores of nine structured risk assessment instruments were compared in a national sample of 397 Dutch convicted sex offenders. The instruments included RRASOR, Static-99, Static-99R, a slightly modified version of Static-2002 and Static-2002R, SACJ-Min, Risk Matrix 2000, SVR-20 and a modified version of SORAG; sexual- and violent (including sexual) recidivism was assessed over five and ten years fixed-, and variable follow-up periods. Generally, the instrument scores showed moderate to large predictive accuracy for the occurrence of reoffending and the number of reoffenses in this sample. Predictive accuracy regarding latency showed more variability across instrument scores. Static-2002(R) and Static-99R scores showed a slight but consistent advantage in predictive properties over the other instrument scores across outcome measures and follow-up periods in this sample. The results of SVR-20 and RRASOR scores were the least positive. A positive association between predictive accuracy and inter rater agreement at the item level was found for both sexual (r = .28, p = .01) and violent (including sexual) recidivism (r = .45, p < .001); no significant association was found between predictive accuracy and missing values at the item level. Results underscore the feasibility and utility of these instruments for informing treatment selection according to the RNR principles.
Accurate risk assessment is essential to optimal treatment selection for offenders. As dictated by the Risk Need Responsivity (RNR) principles, high-risk offenders should receive more intensive treatment, and lower risk offenders less intensive treatment to prevent recidivism (Andrews & Bonta, 2010; Andrews, Bonta & Hodge, 1990). Research (Hanson, Bourgon, Helmus, & Hodgson, 2009; Olver, Wong & Nicholaichuk, 2009) has provided support for this RNR notion regarding sex offenders. Furthermore, ample research has shown that structured risk assessment tools perform better than unsystematic clinical appraisals of sex offender recidivism risk (Andrews, Bonta & Wormith, 2006; Hanson & Morton-Bourgon, 2009; Janus & Prentky, 2003; Quinsey, Harris, Rice & Cormier, 2006).

Standard clinical practice in The Netherlands deviates from these guidelines. Specifically, use of structured risk assessment instruments is recommended, but typically serves as merely one input in a larger, individualized clinical decision-making process. Recent research by Smid et al. has demonstrated that these procedures can lead to treatment referrals (Smid, Kamphuis, Wever, & Van Beek, 2013) and treatment group composition (Smid, Kamphuis, Wever, & Verbruggen, in press) that are inconsistent with the RNR principles, and presumably, are associated with suboptimal outcome.

Research on structured sex offender risk assessment in Dutch samples is limited to three studies including either the Static-99 (Hanson & Thornton, 2000), the SVR-20 (Boer, Hart, Kropp & Webster, 1997) or both. Each of these studies (Koster, Van Lankveld & Spreen, 2006; Schönberger, Hildebrand, Spreen & Bloem, 2008; De Vogel, De Ruiter, Van Beek & Mead, 2004) was based on a (partly overlapping) selective sample of Dutch sex offenders referred to inpatient forensic psychiatric treatment; a referral limited to approximately 5% of convicted Dutch sex offenders (Brouwers & Smit, 2005). Results in this selective subgroup of Dutch sex offenders suggested moderate predictive accuracy for the Static-99 score and moderate to large predictive accuracy for the SVR-20 score considering violent (including sexual) recidivism. Results were inconclusive for the predictive accuracy of the SVR-20 score with regard to sexual recidivism. International research on the predictive accuracy of the formal risk assessment instrument scores is extensive (for a recent comprehensive overview of published and unpublished studies, see the meta-analysis of Hanson & Morton-Bourgon, 2009), and has yielded a more complex picture. Hanson and Morton-Bourgon conclude, based on their meta-analysis of 118 distinct samples including 45,398 offenders, that the overall predictive accuracy of the formal risk assessment instrument scores is moderate-to-large, and that there is no clear evidence for the superiority or inferiority of any individual instrument. The exact predictive values of the scores on the individual instruments, however, vary considerably per study.
Moderating variables commonly suggested to account for the inconsistent performance of instruments across studies, are inter rater agreement, missing values, sample selection bias, and widely differing outcome measures (e.g., in terms of the duration of follow-up periods, official versus unofficial sources utilized, inconsistent definitions for recidivism; see for example Craig, Beech and Browne, 2006; Falshaw, Bastes, Patel, Corbett & Friendship, 2003; Hanson, 2002; Marshall & Barbaree, 1988). While sample selection and outcome measures can be put on the same metric by studying the performance of alternative instruments in the same sample (Hanson & Morton-Bourgon, 2009; Langton, Barbaree, Seto, Peacock, Harkins & Hansen, 2007, and the present study), this does not hold for the issues of inter rater agreement and missing values.

**Inter rater agreement.** To make a prediction as accurate as possible, the inter rater agreement is of crucial importance (Weir & Roberts, 1994). Barbaree, Seto, Langton and Peacock (2001) analyzed the inter rater agreement on a subset of 30 cases for SORAG (Quinsey et al., 2006), RRASOR (Hanson, 1997), Static-99 and MnSOST-R scores (Epperson, Kaul & Hesselton, 1998) upon finding that, within the same sample, the MnSOST-R score was not significant in the prediction of violent and sexual recidivism, while the SORAG and the Static-99 scores were. Inter rater agreement for the SORAG, RRASOR, and Static-99 scores was significantly higher (Pearson $r > .90$) than for the MnSOST-R scores (Pearson $r = .80$). This provided a possible explanation for the observed weak predictive accuracy of the MnSOST-R scores: the MnSOST-R was more difficult to score than the other actuarial measures (Barbaree et al., 2001). In the recent review by Tully, Chou and Browne (2013) it was noted that less than half of the included studies reported good inter rater agreement for the scores on the instruments used. Hanson and Morton-Bourgon (2009) included the topic of inter rater agreement in their meta-analysis and found that studies reporting higher rater agreement showed larger effect sizes than studies reporting lower rater agreement, as reflected by a positive correlation between ICC and Cohen’s $d$ ($r = .21$, $p < .001$).

**Missing values.** The importance of complete data for the prediction of recidivism was stressed by Harris, Rice, Quinsey, Lalumière, Boer and Lang (2003). When cases with missing data were excluded from their analysis, the predictive accuracy of the SORAG and RRASOR scores increased. Other researchers also attributed lower predictive accuracy, specifically for the SORAG scores to omitted or altered items (e.g. Grann, Belfrage, & Tengström, 2000; Nunes, Firestone, Bradford, Greenberg & Broom, 2002; Sjöstedt & Långström, 2002). As noted by Harris et al. (2003), this omitting and altering of items fluctuated with the variability in the quantity and quality of information that was available to researchers, which lead to the recommendation that future studies of actuarial instruments should specifically describe omitted or approximated
items (Harris et al., 2003). Hanson and Morton-Bourgon (2009) included the topic of missing values in their meta-analysis and surprisingly found that studies reporting larger percentages of missing data reported larger effect sizes, as reflected in a significant correlation between percentage of total items not rated per subject and Cohen’s $d$ ($r = .11, p < .01$). However, only 149 of the 536 included studies reported the percentage of missing values. Tully et al. (2013) highlighted in their recent review that the way in which authors dealt with missing information was among the methodological areas that were particularly unclear in the reviewed studies.

In sum, predictive accuracy is not a sample independent property of risk assessment instrument scores. It is contingent upon a number of factors, including the composition of the sample, inter rater agreement, and the available information (missing values). Accordingly, it seems advisable (Rettenberger & Eher, 2006), if not obligate (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) to conduct local / national validity studies prior to routine implementation of tools developed in other countries. The current study aims, rather than looking for the best instrument per se, to assess the best match between a structured sex offenders risk assessment instrument and the unambiguous information that is retrievable from Dutch sex offenders’ judicial files. Specifically, the current study compares the use of nine of the internationally most commonly used sex offender risk assessment instruments in their ability to predict reoffending in a large national sample of convicted Dutch sex offenders. Following Olver, Nicholaichuk, Gu and Wong (2013), we included other salient outcome measures that provide more context to the dichotomous occurrence of recidivism, i.e. the number of reoffenses and latency for the first reoffense.

The nine instruments include the Rapid Risk Assessment for Sex Offender Recidivism, RRASOR (Hanson, 1997), STATIC-99 (Hanson & Thornton, 2000), STATIC-99R (Helmus, Thornton, Hanson & Babchishin, 2012), a slightly modified version of STATIC-2002 (Hanson & Thornton, 2003) and STATIC-2002R (Helmus, et al., 2012), Structured Anchored Clinical Judgments Minimum, SACJ-Min (Grubin, 1998), Risk Matrix 2000, RM2000 (Thornton, Mann, Webster, Blud, Travers, Friendship & Erikson, 2003), a modified version of the Sex Offender Risk Appraisal Guide, SORAG (Quinsey et al., 2006) and Sexual Violence Risk 20, SVR-20 (Boer et al., 1997). It is hypothesized that the scores on all the included instruments will predict occurrence, number of reoffenses and latency of reoffending in this sample significantly better than chance. Additionally, it is hypothesized that inter rater agreement will be positively related to predictive accuracy and the percentage of missing values will be negatively related to predictive accuracy.
Method

Participants

The study sample consisted of two subsamples of men convicted for a contact sexual offense (sexual assault, rape or sexual abuse) in The Netherlands. The first subsample included all offenders who were discharged from inpatient treatment between 1996 and 2002 ($n = 96$). The second subsample consisted of a random selection of 25% of male sex offenders who were discharged from prison between 1996 and 2002 ($n = 341$). This subsample was obtained by selecting half of all Dutch jurisdictions, stratified for urban and rural areas, and subsequently randomly selecting half of the obtainable files from each selected jurisdiction. Both samples added up to a total of 437 sex offenders of which 203 (46.5%) were convicted to a prison sentence without any form of treatment, 96 (21.7%) were imposed with inpatient treatment, and 138 (31.6%) were imposed with community-based treatment which they were suppose to start upon discharge from prison.

A total of 40 records were discarded for various reasons: 11 convicted sex offenders were excluded because they were female; two offenders were under 18 years old at the time of discharge and 6 offenders deceased before discharge. For 21 offenders we were unable to find a new extract of criminal records to assess new offenses. Of those 21 offenders, 16 had died before 2006 and had their record deleted from the judicial system, as was customary up to that date. Of these 16 deceased offenders, 6 had died within two years from their discharge and another 5 had died within five years from their discharge. Follow-up data for 5 remaining offenders was missing for unknown reasons (it is possible that the offender was not a Dutch citizen and/or left the country or that the offender changed his name). This resulted in a final sample of 397 offenders who were included in the subsequent analyses. The mean age at discharge for this sample was $38.7 \ (SD = 12.2)$, with a minimum age of 18 and a maximum age of 78. Of the offenders 69.0 $\% \ (n = 274)$ had had previous contact with the justice system, 30.0 $\% \ (n = 119)$ had previous charges/convictions for a sexual offense, 6.8 $\% \ (n = 27)$ had previous charges/convictions for a juvenile sexual offense. See Table 1 for further characteristics of the sample.

Procedure

Judicial files. All participants’ criminal files as compiled for the index conviction (index files) were studied to retrospectively code the items of all risk assessment instruments. Dutch criminal files are built around specific offenses and are not aggregated per individual. Although an extract of an offenders’ criminal offense record is generally available, digital storage of comprehensive file information and digital linkage to related case files has been introduced only recently (around 2005) and was not retroactively
executed. Regarding the current study, this meant that extensive information regarding offenses prior to the index offense was likely to be stored in separate files, sometimes at different locations in different jurisdictions, or had already been destroyed. For the current study only the comprehensive files regarding the index offense were retrieved, and not all comprehensive files regarding all prior offenses.

The index files generally contained offenders’ criminal record at the time of conviction, minutes concerning the index offense (with statements of both offenders and victims), and the exact indictment, verdict and sentence concerning the index offense. The majority (81.9%) of the files also contained a report concerning some sort of pre-trial assessment. The size and content of these reports varied widely, as they were between three and 68 pages long and alternately written by psychologists, psychiatrists or social workers. However, all assessments were unstructured, in the sense that none of the offenders was assessed for risk level with a structured risk assessment instrument at the time of conviction. In the rare occasion when a file did not contain an extract of the offender’s criminal record at the time of conviction, a current extract of his criminal record was requested from the Ministry of Justice. Because the digital criminal record database only provides complete extracts of lifetime criminal records, records were printed and subsequently stripped of any information beyond the date of the index offense by an independent third party, before being viewed by the raters.
Scoring the items. The retrospective scoring of the various risk assessment instruments was achieved by means of scoring a combination-list that contained all items of each of the individual instruments ordered by content area rather than by instrument. The combination-list included a schedule of the offenders’ criminal history at the time of conviction. A total of ten trained coders scored the original 437 files, with 56 files doubly scored by independent raters to assess inter rater agreement. The items were coded based on the file information available at conviction (index files). Items referring to the offenders’ age at the time of the risk assessment were scored according to the offenders' age at the time of discharge. Information on the exact time of discharge was obtained from the Ministry of Justice and consisted of two separate files: one file containing dates of discharge from mandatory (inpatient) treatment for the first subsample and the other file containing dates of discharge from prison for the second subsample. The composite scores on the various instruments were calculated, according to their respective manuals, from the (SPSS) data file containing all individual items through a syntax file.

Missing values. The percentage of missing items per instrument was calculated by dividing the total number of missing items by the total number of items to be scored, across records. For the assessment of the predictive accuracy of each of the instrument scores, records with more than 20% of the items missing were excluded from subsequent analyses. For records with up to 20% missing items, scores were prorated by imputing the average of the non-missing items within each record. Predictive accuracy was calculated for cases with complete records, and for those with up to 20% missing items, using the raw scores and the prorated scores respectively.

No items were missing for the RRASOR, Static-99(R) and the RM2000 (Thornton et al., 2003). A total of two cases had one item missing from the Static-2002(R). Prorating for these items did not change the composite score for either subject as the weighted score for the specific domain remained the same. Hence all Static-2002(R) composite scores were included in the analyses as complete. Regarding the SACJ-Min, 18 subjects had one missing item; in all cases this concerned one of the ‘aggravating’ items. Prorating was achieved by assessing the average score on the eight completed items. When this average score was ≤ .50, the missing item was scored ‘0’; when this average score > .50, the missing item was scored ‘1’. The prorated scores resulted in unchanged risk levels for 15 subjects and three subjects were upped one level. Regarding the SORAG, 67 subjects had three or more items missing and were excluded. In total 74 subjects had one or two SORAG items missing. Prorating was achieved by assessing the percentage of the maximum score attained over the completed items and calculating the same percentage of the possible score range on the missing items. The prorated scores resulted in unchanged risk levels for 59 subjects, 12 subjects were
adjusted one level up and three subjects were adjusted one level down. Regarding the SVR-20, 64 subjects had five or more items missing and were excluded. In total 248 subjects had one to four SVR-20 items missing. Prorating was achieved by assessing the average score over the completed items and adding this average score for each of the missing items.

**Recidivism data.** After the coding of the various risk assessment instruments, the offenders’ current criminal records were requested in order to acquire data on their reoffending. The current criminal records were coded by two independent raters who were blind to the initial risk assessment results. Outcome measures were ‘sexual recidivism’ and ‘violent including sexual recidivism’. Following the meta-analysis of Hanson & Morton-Bourgon (2009), sexual recidivism was defined as any charge or conviction for offenses with a sexual motivation, such as child sexual abuse, rape, exhibitionism, and child pornography offenses. (These offenses are defined as sexual according to the Dutch penal code). Offenses against public morals without an identifiable victim (e.g., prostitution related offenses) were not included. Additionally, precise classification of the sexual recidivism types was recorded. Violent recidivism was defined as any charge or conviction for a new offense that involved any actual, attempted, or threatened physical harm to the victim and included all sexual offenses. Violent offenses included explicit display of violence, such as assault, murder or robbery and also included more covert forms of violence, such as threats, extortion and stalking. Precise classification of the violent recidivism types was recorded. Violent (including sexual) recidivism would arguably form the outcome of principal concern to the Dutch system. This outcome includes all offenses that directly target a victim, sexually or violently, supplemented with CP offenses, which are penalized with comparable sentences.

Follow-up time was counted in months from the date of discharge from prison or inpatient treatment, to the day the follow-up criminal record was retrieved; or, in the case of deceased subjects, from the date of discharge to the date of decease. Mean follow-up was 145 months (12 yrs and one month, \(SD = 30\) months), with a minimum of 51 months (four yrs and three mos) and a maximum of 201 months (16 yrs and nine mos). The predictive accuracy of the scores of the nine risk assessment instruments was calculated over fixed follow-up periods of five and ten years, as well as over the variable follow-up periods. Latency was counted in months from the date of discharge from prison or inpatient treatment until the day the reoffense was committed. The number of reoffenses was assessed by counting the total number of legal facts included in new charges and convictions during the follow-up period. No prior research has been published on this data.
**Instruments**

**Rapid Risk Assessment for Sexual Offense Recidivism (RRASOR).** Hanson (1997) developed this brief actuarial scale through a stepwise regression procedure over the seven best predictors found by Hanson and Bussière (1996) in a sample of 2,592 sex offenders. The RRASOR consists of the four variables that significantly increased predictive accuracy. These items refer to: a) prior sexual offenses, b) age at release, c) victim gender and d) the relationship between victim and offender. The scoring is weighted such that the ‘prior sexual offenses’ item yields a score between ‘0’ and ‘3’, while the other items are dichotomous. The composite score is calculated by adding up the item scores and ranges between zero and six.

In the original study (Hanson, 1997), the RRASOR score showed moderate to large predictive accuracy for sexual recidivism (AUC = .71). Several validation studies have also found the RRASOR score to be a moderate predictor of sexual recidivism and a significant but weak predictor of violent (including sexual) recidivism. The most recent meta-analysis (Hanson & Morton-Bourgon, 2009) reported on accumulated findings from 34 published and unpublished studies that included over 11,000 subjects: A median AUC = .68 (mean AUC = .66) was found for sexual recidivism, and a median AUC = .58 (mean AUC = .59) for any violent (including sexual) recidivism, respectively. The $d$-values reported by Hanson and Morton-Bourgon were converted to their equivalent AUC-values, following Ruscio (2008), for comparison purposes.

**Structured Anchored Clinical Judgments Minimum (SACJ-Min).** The SACJ was developed by Thornton (Grubin, 1998) based on the RRASOR and aimed to assess the risk of sexual and violent recidivism over time. The SACJ consists of three stages: 1) assessment of historical risk factors, 2) assessment of aggravating factors, 3) assessment of current behavior and response to treatment programs. The first two stages are referred to as SACJ-Min. The historical risk factors consist of five items referring to prior and current sentencing occasions for sexual, violent and general criminal offenses, each item is scored dichotomous. The historical factors lead up to an initial risk category (low, medium or high) based on the sum of the scores. The aggravating factors consist of four items referring to a) victim gender, b) relationship between victim and offender, c) non-contact sexual offenses and d) relationship history. For each two aggravating factors present, the offender is upped one risk category.

The original chapter (Grubin, 1998), did not report AUC values, but a validation study by Hanson and Thornton (2000) found an AUC of .67 for the prediction of sexual recidivism and .64 for the prediction of violent (including sexual) recidivism, respectively. The Hanson and Morton-Bourgon (2009) meta-analysis reported on accumulated findings from 6 published and unpublished studies which included almost 1,400 subjects: A median AUC = .63 (mean AUC = .62) was calculated for sexual recidivism,
and a median AUC = .60 (mean AUC = .60) for any violent (including sexual) recidi-

**STATIC-99(R).** The STATIC-99 is the most widely used and most extensively re-
searched actuarial risk assessment instrument for sex offenders (Hanson & Morton-
Bourgon, 2009). It was developed by Hanson and Thornton (2000) as a combination
of the RRASOR (Hanson, 1997) and the SACJ-Min (Grubin, 1998). The aim was to
develop a widely applicable risk tool for the prediction of sexual recidivism that could
be reliably scored based on objective and relatively easily obtainable information. The
STATIC-99 consists of ten items including all four items from the RRASOR and seven
from the SACJ-Min (one overlapping item). The item referring to prior sexual offenses
yields a score between ‘0’ and ‘3’; all other items are scored dichotomous. The com-
posite score is calculated by adding up the items scores (ranging from zero to 12),
yielding four risk levels: low, low-moderate, moderate-high or high.

In the original study (Hanson & Thornton, 2000), the STATIC-99 score showed
moderate predictive accuracy with an AUC = .71 for sexual recidivism and an AUC = .69
for violent including sexual recidivism (Hanson & Thornton, 2000), respectively. The
predictive accuracy of the STATIC-99 score has been widely replicated. The most recent
meta-analysis (Hanson & Morton-Bourgon, 2009) summarized 63 findings from pub-
lished and unpublished studies that included over 20,000 subjects: A median AUC =
.70 (mean AUC = .68) for sexual recidivism was observed, and a median AUC = .64
(mean AUC = .66) for any violent (including sexual) recidivism, respectively. Recent
research found age to add incrementally to the predictive accuracy of the Static-99
score, resulting in a revised scoring system for the Static-99 that more accurately de-
scribes older offenders’ risk of recidivism (Helmus et al., 2012). The Static-99R age
item distinguishes four age categories yielding a score between ‘-3’ and ‘+1’.

**STATIC-2002(R).** The STATIC-2002 was developed by Hanson and Thornton (2003)
in an effort to increase coherence and conceptual clarity, aiming to assess items refer-
ing to theoretically more meaningful characteristics presumed to be the cause of re-
cidivism risk. The 14 items of the STATIC-2002 show some overlap with the STATIC-99
(five items). To better understand what is being measured, the Static-2002 items are
grouped into five domains: Age (one item), Persistence of Sex Offending (three items),
Deviant Sexual Interests (three items), Relationship to Victims (two items), and Gen-
eral Criminality (five items). For each cluster a weighted score is calculated, based on
the item scores. The composite STATIC-2002 score is the sum of the cluster scores and
corresponds to five risk levels: low, low-moderate, moderate, moderate-high or high.

In the first study (Helmus & Hanson, 2007), the Static-2002 score showed simi-
lar levels of predictive accuracy to the Static-99 score for the prediction of sexual re-
cidivism (AUC = .76 for both) but the Static-2002 score predicted violent (including
sexual) recidivism better than the Static-99 score (AUC = .76 versus .73). The most recent meta-analysis (Hanson & Morton-Bourgon, 2009) summarizes ten findings from published and unpublished studies that included over 3,000 subjects: A median AUC = .69 (mean AUC = .69) for sexual recidivism was found, and a median AUC = .67 (mean AUC = .68) for any violent (including sexual) recidivism, respectively. Recent research found age to add incrementally to the predictive accuracy of the Static-2002 score, resulting in a revised scoring system for the Static-2002 that more accurately describes older offenders’ risk of recidivism (Helmus et al., 2012). The Static-2002R age item distinguishes four age categories yielding a score between ‘-2’ and ‘+2’.

The scoring of item 13 of the Static-2002(R) was slightly adjusted for this study. Due to Dutch judicial archiving procedures, it proved to be impossible for nearly all subjects to assess the date of their release from prison prior to committing the index offense. The reference point for the item was therefore adjusted to the most recent prior conviction date for any offense. The cut-off for the number of months between the prior conviction date and the date of (onset of) perpetration of the index offense was set at 48 months rather than 36 months, accounting for time spent in prison after the prior conviction. This adjustment proved to be adequate. Only two out of the 437 values were missing, associated inter rater agreement was excellent (ICC = .87), and the adjusted item score showed substantial positive correlation with both sexual ($r = .25, p < .001$) and violent (including sexual) recidivism ($r = .18, p < .001$).

Risk Matrix 2000 (RM2000). The RM2000 was developed by Thornton et al. (2003) following a re-examination of the SACJ-Min. Thornton et al. (2003) created a two-dimensional risk assessment system for sex offenders assessing risk of sexual recidivism and risk of non-sexual violent recidivism with separate scales; the Risk Matrix 2000/Sexual (RM2000/s) and the Risk Matrix 2000/Violence (RM2000/v). These scales can be combined in the Risk Matrix/Combined (RM2000/c) scale for the prediction of violent (including sexual) recidivism. The RM2000/s comprises two steps; the first step consists of three risk items (age, sexual appearances, and criminal appearances). The sum score of step one is translated into a risk category (low, medium, high or very high). Step two contains four aggravating factors referring to victim characteristics and marital status. For every two aggravating factors present, the risk category is upped one step. The RM2000/v consists of three risk items (age, violent appearances and burglary) and the sum score on these items is translated into a risk category (low, medium, high or very high). Both risk categories on the RM2000/s and RM2000/v can be added up and translated into one of four RM2000/c risk categories (low, medium, high or very high).

The RM2000 was first validated by Thornton et al. (2003) on two UK samples of treated ($n = 647$) and untreated ($n = 429$) sex offenders. The study reported an AUC of
A COMPARISON OF THE PREDICTIVE PROPERTIES OF NINE SEX OFFENDER RISK ASSESSMENT INSTRUMENTS

.77 for the treated and .75 for the untreated sample regarding sexual recidivism, and an AUC of .81 for the treated and .74 for the untreated sample regarding sexual (including violent) recidivism predicted by the RM2000/c score. The Hanson and Morton-Bourgon (2009) meta-analysis summarized ten RM2000/s findings from published and unpublished studies that included over 2,700 subjects: A median AUC = .68 (mean AUC = .68) for sexual recidivism was found. Moreover, five studies including over 1,400 subjects for the RM2000/c, resulted in a median AUC = .69 and mean AUC = .70 for any violent (including sexual) recidivism.

Sex Offender Risk Appraisal Guide (SORAG). The SORAG (Quinsey, Rice & Harris, 1995) was designed for the prediction of violent (including sexual) recidivism for sex offenders and constitutes the sex offender version of the Violence Risk Appraisal Guide (VRAG; Harris, Rice & Quinsey, 1993). The SORAG contains 10 of the 12 VRAG items referring to childhood maladjustment, alcohol problems, age, marital status, offense history, disorders and psychopathy. These items were complemented with four specific sex offender items referring to violent and sexual offense history, victim characteristics, and phallometrically diagnosed sexual deviance. Each answer option to each item is weighted differently based on the optimal weight in the creation sample; item scores vary between -5 and +12. The composite score ranges from –27 to +51 and refers to one of nine risk categories.

The original study (Quinsey et al., 1995) did not report AUC values, but found a large linear relationship between SORAG scores and sexual recidivism (r = .45) and violent (including sexual) recidivism (r = .46). The Hanson and Morton-Bourgon (2009) meta-analysis reported on 12 findings from published and unpublished studies that included over 3,000 subjects for the SORAG: A median AUC = .66 (mean AUC = .67) was calculated for sexual recidivism, and a median AUC = .69 (mean AUC = .71) for any violent (including sexual) recidivism, respectively.

Four changes to the standard coding of the SORAG in the present study deserve mention, two of which refer to permissible substitutions following the manual and two other refer to actual modifications. First, as phallometric testing for sex offenders is virtually non-existent in the Netherlands, item 13 of the SORAG was, as dictated by the manual, replaced by the offenders’ score on the Screening Scale for Pedophilic Interests (SSPI). The SSPI was developed by Seto and Lalumière (2001) to identify possible pedophilic interests among child molesters. Individuals with the highest SSPI score were more than five times as likely to show pedophilic preference on phallometric assessment than individuals receiving the lowest score.

Second, for item 14, Psychopathy Checklist scores (PCL-R; Hare, 2003) were substituted with scores on the Psychopathy Checklist: Screening Version (PCL:SV; Hart, Cox & Hare, 1995) assessed from file information, because PCL-R scores were not
Guy and Douglas (2006) found high correspondence between the PCL:SV and PCL–R scores, with inter-measure correlations above .90 for both forensic and correctional samples and similar correlations with violent recidivism ($r = .42$ for PCL–R and $r = .37$ for PCL:SV). PCL:SV scores in this study were, following the manual, used to calculate corresponding PCL-R scores by multiplying them by 1.67.

Third, the coding rules for item 11 of the SORAG were adjusted. The manual requires official DSM-III (American Psychiatric Association, 1980) coded personality disorders. Due to the lack of structured assessment of personality disorders this was changed to a statement regarding (any) personality disorder in the pre-trial assessment reports. This also included personality disorders Not Otherwise Specified (NOS). Schizophrenia was occasionally reported in reference to psychodiagnoses received (years) prior to the index offense. Reference to offenders’ mental state at the time of the offense was predominantly put in terms of psychotic symptoms.

Finally, for item 12, the required official DSM-III (American Psychiatric Association, 1980) coding for schizophrenic disorders was therefore replaced by statements regarding clear and significant psychotic problems (e.g. ‘chronic psychosis’, ‘severe psychosis’, ‘schizophrenic psychosis’).

**Sexual Violence Risk 20 (SVR-20).** The SVR-20 (Boer et al., 1997) is a structured professional judgment instrument. The SVR-20 was developed as the sex offender counterpart of the HCR-20 and was aimed specifically at the risk assessment of sex offenders with psychiatric disorders in clinical forensic settings. Its use is mandatory in inpatient forensic treatment settings in Netherlands since 2005. Contrary to the HCR-20, the SVR-20 includes, next to empirically derived risk factors, a number of factors added on clinical indication. In total, the SVR-20 contains 20 items subdivided in three clusters: psychosocial adjustment (11 items), sexual offenses (seven items) and future plans (two items). The items are scored on a three point scale: absent (‘0’), maybe or somewhat present (‘1’), or definitely present (‘2’). Scores can be added up (‘mechanical approach’), or a structured professional judgment can be made based on the item scores. In that case the risk level is summarized as ‘low’, ‘moderate’ or ‘high’ according to the clinical weighing and combining of the items and possibly inclusion of case specific risk- or protective factors by the rater(s). Although the SVR-20 aimed not only to assess risk, but also provide treatment indications and assess change in risk over time, only four out of the 20 items are more or less dynamic in nature.

The most recent meta-analysis (Hanson & Morton-Bourgon, 2009) summarized ten findings from published and unpublished studies that included over 1,700 subjects for the SVR-20 using the mechanical approach (adding up the items): It resulted in a median AUC = .66 (mean AUC = .68) for sexual recidivism, and a median AUC = .58 (mean AUC = .61) for any violent (including sexual) recidivism.
International research regarding structured professional judgment based on the SVR-20 (and other SPJ instruments) is falling behind. Only three studies with 245 subjects were included in Hanson and Morton-Bourgon’s meta-analysis (2009) which showed widely varying results. Predictive accuracy of the SPJ score (low, moderate, high) for sexual recidivism was excellent in studies by Dempster (1998, unpublished thesis) (AUC = .85) and De Vogel et al. (2004) (AUC = .83), but failed to show significance in a study by Sjöstedt and Långström (2002) (AUC = .49). Both De Vogel et al. (2004) and Sjöstedt and Långström (2002) found significant AUC values for the prediction of non-sexual violent recidivism (AUC = .64 for both). The current study employed the mechanical approach to the SVR-20, i.e., adding up the item scores.

**Statistical analyses**

Inter rater agreement was assessed by means of the intra class correlation coefficient (ICC), using the two-way random effect variance model and consistency type (McGraw & Wong, 1996).

The predictive accuracy of the instrument scores was established with receiver operating characteristics (ROC) analyses (Mossman, 1994; Rice & Harris, 1995). The advantage of this statistical method is its relative insensitivity to base rates. The ROC analysis plots the true positive rate (sensitivity) against the false positive rate (1 minus specificity) for every possible cut-off score of the instrument. The resulting area under the curve (AUC) can be interpreted as the probability that a randomly selected recidivist would score higher on the instrument than a randomly selected non-recidivist. An AUC of .50 represents prediction at chance level and an AUC of 1.0 represents perfect prediction. AUC values > 0.50 indicate a prediction level better than chance. AUC values correspond to Cohen’s $d$ effect sizes, such that an AUC value of .56 corresponds to Cohen’s $d$ of .20 (small effect size), an AUC value of .64 corresponds to Cohen’s $d$ of .50 (moderate effect size), and an AUC value of .71 corresponds to Cohen’s $d$ of .80 (large effect size) (Rice & Harris, 2005). As all instruments were scored on the same offender sample, the various risk scores were not independent and the analysis to assess for significant differences in predictive accuracy between the various instrument scores would have to take this interdependence into account. Paired sample statistical techniques have been developed for the comparison of two instrument scores administered on the same sample (DeLong, DeLong & Clarke-Pearson, 1988). This nonparametric approach to the analysis of areas under correlated ROC-curves uses the theory on generalized U-statistics to generate an estimated covariance matrix. To compare the obtained AUC values, the software program ROCTools (V1.0.2 - September 2007) (Allaire & Cismaru, 2007) was used. This analysis required the exclusion of missing values, therefore all cases with >20% missing values on any
one of the included instruments were excluded. For the cases with up to 20% missing values on any one of the included instruments, the raw composite scores or the pro-rated composite scores were used for the analysis, depending on which of these scores resulted in the highest AUC values in the prior analysis.

To assess the relationship between the instrument scores and the dependent variables (including number of offenses, months of latency), Spearman rank correlations were computed. We opted for non-parametric associations as both the independent and dependent variables failed to satisfy the assumption of normality. Spearman’s \( \rho \) values correspond to Cohen’s \( d \) effect sizes, such that a Spearman’s \( \rho \) value of .10 corresponds to Cohen’s \( d \) of .20 (small effect size), a Spearman’s \( \rho \) value of .24 corresponds to Cohen’s \( d \) of .50 (moderate effect size), and a Spearman’s \( \rho \) value of .37 corresponds to Cohen’s \( d \) of .80 (large effect size) (Rosenthal, 1994).

The relationship between inter rater agreement and predictive accuracy was assessed through the estimation of the bivariate Pearson correlation coefficient between the AUC and ICC values of the 64 original items that formed the building-blocks for all nine included instruments. The relationship between missing values and predictive accuracy was assessed through the estimation of the bivariate Pearson correlation coefficient between the AUC values and the percentages of missing values of the 64 original items.

**Results**

The observed sexual recidivism rate was 14.1% (\( n = 56 \)) overall, and rates of 10.1% (\( n = 40 \)) and 14.6% (\( n = 47 \)) were observed for the five and ten year fixed follow-up periods, respectively. The violent (including sexual) recidivism rate was 33.5% (\( n = 133 \)) overall, and rates of 24.7% (\( n = 98 \)) and 32.2% (\( n = 104 \)) were observed for the five and ten year fixed follow-up periods, respectively. Of the 56 sexual recidivism cases, 32 cases concerned rape or sexual assault (occasionally including exhibitionism), 20 cases concerned child sexual abuse (occasionally including CP or exhibitionism), 3 cases concerned only exhibitionism, and one case concerned only child pornography possession. Of the 102 non-sexual violent recidivism cases, 9 cases concerned murder or manslaughter, 66 cases concerned physical assault, and 27 cases concerned threats of violence. The average number of reoffenses was 0.39 (\( SD = 1.61 \), ranging between zero and 25) for sexual reoffenses and 1.17 (\( SD = 2.88 \), ranging between zero and 32) for violent (including sexual) reoffenses. In case of recidivism, the average latency was 41.93 months (\( SD = 39.87 \), ranging between zero and 160) for the first sexual reoffense and 41.98 months (\( SD = 40.85 \), ranging between zero and 173) for the first violent (including sexual) reoffense.
Predictive accuracy of the scores on the included instruments regarding the occurrence of sexual reoffending (See Table 2) and violent (including sexual) reoffending in the current sample (See Table 3) was moderate to large, with the exception of the SVR-20 and RRASOR scores. The SVR-20 score showed non-significant to small predictive accuracy for sexual recidivism and the RRASOR score showed small predictive accuracy for violent (including sexual) recidivism. All AUC confidence intervals were overlapping, except the intervals of the SVR-20 score, AUC = .58; 95% CI [.48, .67] and the Static-2002R score, AUC = .77; 95% CI [.68, .85] regarding the prediction of sexual recidivism within a fixed five-year follow-up.

The results regarding the additional outcome measures, i.e., number of offenses and latency to first offense are presented in Table 4 and 5. Generally, moderate to large effect sizes were found for the number of offenses (see Table 4) when non-recidivists were included in the analyses (number of offenses = 0). Scores on the Static variants showed the largest effect sizes whereas RRASOR and SVR-20 scores showed the smallest effect sizes. When non-recidivists were excluded, the effects were smaller, and only reached significance for the Static variants and SACJ-Min. The relationship to latency (see Table 5) for the first violent (including sexual) reoffense showed a similar pattern with generally moderate effect sizes when non-recidivists were included (latency = follow-up time), and generally small effect sizes when non-recidivists were excluded. Results regarding the latency for the first sexual reoffense were opposite: generally small effect sizes were found when non-recidivists were included (latency = follow-up time), whereas effect sizes ranged between none and large when non-recidivists were excluded.

Comparison of the AUC values of all nine risk assessment tools regarding the (dichotomous) occurrence of recidivism, following the Delong et al. (1988) method accounting for their interdependence, showed significant differences between the AUC values with respect to sexual recidivism within a fixed five year follow-up period ($\chi^2 (8, N = 309) = 30.93, p < .001$), a fixed ten year follow-up period ($\chi^2 (8, N = 252) = 17.43, p = .026$), and variable follow-up ($\chi^2 (8, N = 309) = 23.25, p = .003$). Differences were also found with respect to violent (including sexual) recidivism within the fixed five year follow-up period ($\chi^2 (8, N = 309) = 24.46, p < .001$), the fixed ten year follow-up period ($\chi^2 (8, N = 252) = 43.27, p < .001$), and the variable follow-up period ($\chi^2 (8, N = 309) = 50.86, p < .001$). Pair wise test results with Bonferroni corrections regarding the prediction of sexual recidivism showed that, in the current sample, the AUC values of the Static-99, Static-99R, Static-2002 and the Static-2002R scores were significantly larger than the AUC value of the SVR-20 (prorated) scores within the fixed five year follow-up period,
### Table 2: Predictive Accuracy Regarding Sexual Reoffending (Charge or Conviction for a New Sexual Offense) Expressed in AUC Values and Corresponding Confidence Intervals

<table>
<thead>
<tr>
<th>Instrument</th>
<th>5 year follow-up (95% CI)</th>
<th>10 year follow-up (95% CI)</th>
<th>Variable follow-up (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRASOR</td>
<td>.68*** (.59 - .77)</td>
<td>.69*** (.60 - .77)</td>
<td>.67*** (.59 - .75)</td>
</tr>
<tr>
<td>SACJ-Min, complete</td>
<td>.69*** (.60 - .78)</td>
<td>.71*** (.63 - .79)</td>
<td>.69*** (.61 - .77)</td>
</tr>
<tr>
<td>SACJ-Min, ≤ 20% missing, raw</td>
<td>.68*** (.58 - .77)</td>
<td>.70*** (.62 - .78)</td>
<td>.69*** (.61 - .77)</td>
</tr>
<tr>
<td>SACJ-Min, ≤ 20% missing, prorated</td>
<td>.67*** (.58 - .76)</td>
<td>.69*** (.61 - .78)</td>
<td>.69*** (.61 - .77)</td>
</tr>
<tr>
<td>Static-99</td>
<td>.72*** (.64 - .81)</td>
<td>.73*** (.65 - .81)</td>
<td>.72*** (.65 - .80)</td>
</tr>
<tr>
<td>Static-99R</td>
<td>.74*** (.65 - .82)</td>
<td>.74*** (.66 - .82)</td>
<td>.74*** (.67 - .81)</td>
</tr>
<tr>
<td>Static-2002</td>
<td>.76*** (.67 - .85)</td>
<td>.75*** (.67 - .83)</td>
<td>.74*** (.66 - .81)</td>
</tr>
<tr>
<td>Static-2002R</td>
<td>.77*** (.68 - .85)</td>
<td>.75*** (.67 - .83)</td>
<td>.74*** (.67 - .82)</td>
</tr>
<tr>
<td>RM2000/s</td>
<td>.72*** (.64 - .81)</td>
<td>.71*** (.63 - .79)</td>
<td>.71*** (.64 - .79)</td>
</tr>
<tr>
<td>SORAG, Complete</td>
<td>.63** (.52 - .75)</td>
<td>.64** (.54 - .75)</td>
<td>.64** (.54 - .74)</td>
</tr>
<tr>
<td>SORAG, ≤ 20% missing, raw</td>
<td>.66** (.56 - .76)</td>
<td>.68*** (.59 - .77)</td>
<td>.67*** (.59 - .76)</td>
</tr>
<tr>
<td>SORAG, ≤ 20% missing, prorated</td>
<td>.66** (.56 - .76)</td>
<td>.68*** (.59 - .77)</td>
<td>.67*** (.59 - .76)</td>
</tr>
<tr>
<td>SVR-20, complete</td>
<td>.53 (.34 - .72)</td>
<td>.48 (.30 - .66)</td>
<td>.50 (.33 - .66)</td>
</tr>
<tr>
<td>SVR-20, ≤ 20% missing, raw</td>
<td>.58 (.48 - .67)</td>
<td>.61* (.52 - .69)</td>
<td>.61* (.53 - .69)</td>
</tr>
<tr>
<td>SVR-20, ≤ 20% missing, prorated</td>
<td>.58 (.48 - .67)</td>
<td>.61* (.52 - .70)</td>
<td>.61* (.53 - .69)</td>
</tr>
</tbody>
</table>

*Note: * p < .05, ** p < .01, *** p < .001*
### Table 3: Predictive Accuracy Regarding Violent (Including Sexual) Reoffending (Charge or Conviction for a New Violent or Sexual Offense) Expressed in AUC Values and Corresponding Confidence Intervals

<table>
<thead>
<tr>
<th>Instrument</th>
<th>5 year follow-up (95% CI)</th>
<th>10 year follow-up (95% CI)</th>
<th>Variable follow-up (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRASOR</td>
<td>0.59** (.53 - .66)</td>
<td>0.63*** (.56 - .69)</td>
<td>0.60** (.54 - .66)</td>
</tr>
<tr>
<td>n = 396</td>
<td>n = 323</td>
<td>n = 397</td>
<td></td>
</tr>
<tr>
<td>SACJ-Min, complete</td>
<td>0.67*** (.60 - .73)</td>
<td>0.69*** (.62 - .75)</td>
<td>0.67*** (.61 - .73)</td>
</tr>
<tr>
<td>n = 378</td>
<td>n = 306</td>
<td>n = 379</td>
<td></td>
</tr>
<tr>
<td>SACJ-Min, ≤ 20% missing, raw</td>
<td>0.65*** (.59 - .72)</td>
<td>0.67*** (.61 - .74)</td>
<td>0.66*** (.61 - .72)</td>
</tr>
<tr>
<td>n = 396</td>
<td>n = 323</td>
<td>n = 397</td>
<td></td>
</tr>
<tr>
<td>SACJ-Min, ≤ 20% missing, prorated</td>
<td>0.65*** (.59 - .72)</td>
<td>0.68*** (.61 - .74)</td>
<td>0.67*** (.61 - .72)</td>
</tr>
<tr>
<td>n = 396</td>
<td>n = 323</td>
<td>n = 397</td>
<td></td>
</tr>
<tr>
<td>Static-99</td>
<td>0.68*** (.62 - .74)</td>
<td>0.71*** (.65 - .77)</td>
<td>0.70*** (.64 - .75)</td>
</tr>
<tr>
<td>n = 396</td>
<td>n = 323</td>
<td>n = 397</td>
<td></td>
</tr>
<tr>
<td>Static-99R</td>
<td>0.71*** (.65 - .76)</td>
<td>0.74*** (.68 - .79)</td>
<td>0.73*** (.68 - .79)</td>
</tr>
<tr>
<td>n = 396</td>
<td>n = 323</td>
<td>n = 397</td>
<td></td>
</tr>
<tr>
<td>Static-2002</td>
<td>0.71*** (.66 - .77)</td>
<td>0.76*** (.70 - .81)</td>
<td>0.73*** (.68 - .78)</td>
</tr>
<tr>
<td>n = 396</td>
<td>n = 323</td>
<td>n = 397</td>
<td></td>
</tr>
<tr>
<td>Static-2002R</td>
<td>0.72*** (.66 - .77)</td>
<td>0.75*** (.70 - .81)</td>
<td>0.74*** (.69 - .79)</td>
</tr>
<tr>
<td>n = 396</td>
<td>n = 323</td>
<td>n = 397</td>
<td></td>
</tr>
<tr>
<td>RM2000/c</td>
<td>0.70*** (.64 - .76)</td>
<td>0.73*** (.68 - .79)</td>
<td>0.73*** (.68 - .78)</td>
</tr>
<tr>
<td>n = 396</td>
<td>n = 323</td>
<td>n = 397</td>
<td></td>
</tr>
<tr>
<td>SORAG, Complete</td>
<td>0.71*** (.63 - .78)</td>
<td>0.75*** (.67 - .82)</td>
<td>0.73*** (.66 - .80)</td>
</tr>
<tr>
<td>n = 256</td>
<td>n = 211</td>
<td>n = 256</td>
<td></td>
</tr>
<tr>
<td>SORAG, ≤ 20% missing, raw</td>
<td>0.70*** (.64 - .77)</td>
<td>0.75*** (.68 - .81)</td>
<td>0.73*** (.67 - .78)</td>
</tr>
<tr>
<td>n = 329</td>
<td>n = 272</td>
<td>n = 330</td>
<td></td>
</tr>
<tr>
<td>SORAG, ≤ 20% missing, prorated</td>
<td>0.70*** (.64 - .77)</td>
<td>0.75*** (.68 - .81)</td>
<td>0.73*** (.67 - .78)</td>
</tr>
<tr>
<td>n = 329</td>
<td>n = 272</td>
<td>n = 330</td>
<td></td>
</tr>
<tr>
<td>SVR-20, complete</td>
<td>0.58 (.44 - .72)</td>
<td>0.64 (.50 - .78)</td>
<td>0.64 (.51 - .76)</td>
</tr>
<tr>
<td>n = 85</td>
<td>n = 68</td>
<td>n = 85</td>
<td></td>
</tr>
<tr>
<td>SVR-20, ≤ 20% missing, raw</td>
<td>0.64*** (.57 - .71)</td>
<td>0.65*** (.58 - .72)</td>
<td>0.65*** (.59 - .71)</td>
</tr>
<tr>
<td>n = 332</td>
<td>n = 268</td>
<td>n = 333</td>
<td></td>
</tr>
<tr>
<td>SVR-20, ≤ 20% missing, prorated</td>
<td>0.65*** (.59 - .71)</td>
<td>0.67*** (.60 - .73)</td>
<td>0.67*** (.61 - .73)</td>
</tr>
<tr>
<td>n = 332</td>
<td>n = 268</td>
<td>n = 333</td>
<td></td>
</tr>
</tbody>
</table>

Note: **p < .01. ***p < .001
and the AUC values of the Static-99R, Static-2002 and the Static-2002R scores were significantly larger than the AUC value of the SVR-20 (prorated) scores for the variable follow-up period; no significant post hoc differences were found concerning the fixed ten year follow-up period. Post hoc cross comparisons with Bonferroni corrections regarding the prediction of violent (including sexual) recidivism showed that, in the current sample, the AUC values of the Static-99, Static-99R, Static-2002 and the Static-2002R scores were significantly larger than the AUC value of the RRASOR score within the fixed five year follow-up period, and the AUC values of the Static-99, Static-99R, Static-2002, Static-2002R and the SORAG (raw) scores were significantly larger than the AUC value of the RRASOR score within the fixed ten year follow-up period. Finally, the AUC values of the Static-99, Static-99R, Static-2002, Static-2002R, SORAG (raw) scores and the RM2000/c scores were significantly larger than the AUC value of the RRASOR score for the variable follow-up period. Stated differently, the Static-2002R, Static-2002 and the Static-99R performed significantly better than the instrument with the lowest performance on five of the six outcome measures. The Static-99 performed

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Sexual non-recidivists included (n = 397)</th>
<th>Sexual non-recidivists excluded (n = 56)</th>
<th>Violent non-recidivists included (n = 397)</th>
<th>Violent non-recidivists excluded (n = 133)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRASOR</td>
<td>.22**</td>
<td>.22</td>
<td>.18**</td>
<td>.10</td>
</tr>
<tr>
<td>SACJ-Min, ≤ 20% missing, prorated</td>
<td>.24**</td>
<td>.11</td>
<td>.31**</td>
<td>.22*</td>
</tr>
<tr>
<td>Static-99</td>
<td>.28**</td>
<td>.23</td>
<td>.35**</td>
<td>.23**</td>
</tr>
<tr>
<td>Static-99R</td>
<td>.30**</td>
<td>.23</td>
<td>.41**</td>
<td>.26**</td>
</tr>
<tr>
<td>Static-2002</td>
<td>.30**</td>
<td>.31*</td>
<td>.40**</td>
<td>.24**</td>
</tr>
<tr>
<td>Static-2002R</td>
<td>.30**</td>
<td>.29*</td>
<td>.41**</td>
<td>.26**</td>
</tr>
<tr>
<td>RM2000s/c</td>
<td>.28**</td>
<td>.24</td>
<td>.38**</td>
<td>.14</td>
</tr>
<tr>
<td>SORAG, ≤ 20% missing, prorated</td>
<td>.22** a</td>
<td>-.02 b</td>
<td>.38** a</td>
<td>.18 c</td>
</tr>
<tr>
<td>SVR-20, ≤ 20% missing, prorated</td>
<td>.14** d</td>
<td>.02 e</td>
<td>.28** d</td>
<td>.15 f</td>
</tr>
</tbody>
</table>

Note: * p < .05. ** p < .01
‘Violent’ refers to violent including sexual reoffenses
*a n = 330.  b n = 49.  c n = 107.  d n = 333.  e n = 50.  f n = 110.
outcome measures. The SORAG performed significantly better than the instrument with the lowest performance, regarding two of the six outcome measures. Finally, the RM2000/c performed significantly better than the instrument with the lowest performance, regarding one of the six outcome measures.

The percentage of missing values was generally low (under 1%), with the exception of the SORAG (6.0%) and the SVR-20 (10.1 %). ICC’s for classification were consistently satisfactory, and ranged from .85 (SVR-20 composite score) to .95 (Static-2002R composite score) with a median ICC of .90. Correlational analysis of the AUC values, inter rater agreement and percentage of missing values of the 64 original comprising items, showed a significant positive correlation between the ICC values and AUC values for sexual reoffending \( r(64) = .28, p = .03 \) and between the ICC values and AUC values for violent (including sexual) reoffending \( r(64) = .45, p < .001 \). Correlations between the percentages of missing values and the AUC values were not significant, (sexual reoffending, \( r(64) = -.12, p = .34 \); violent (including sexual) reoffending, \( r(64) = -.11, p = .38 \).
Discussion
The aim of this study was to compare the accuracy of the scores of nine internationally widely used sex offender risk assessment instruments in predicting the occurrence, number of reoffenses and latency for sexual and violent (including sexual) reoffending in a Dutch national sex offender sample. The recidivism rates observed in the sample (14.1% for sexual recidivism, 33.5% for violent (including sexual) recidivism) were slightly higher than unselected/routine samples with comparable follow-up periods (Hanson et al., 2010, Correctional Service of Canada sample: 8.7% sexual and 23.4% violent including sexual recidivism), but considerably lower than completely psychiatric samples with comparable follow-up periods (Bengtson, 2008: 28% sexual- and 46% violent including sexual recidivism; De Vogel et al., 2004: 39% sexual- and 60% violent including sexual recidivism). The observed recidivism rate may reflect the mixed composition of the current sample, with a slight overrepresentation of forensic psychiatric patients. Observed recidivism rates should be considered low estimates of the true recidivism rates (Falshaw et al., 2003), because not all offenses are detected (Hanson & Morton-Bourgon, 2009). Observed recidivism rates regarding sexual reoffenses may have been further reduced through plea-bargaining (Bagley & Pritchard, 2000; Rice & Harris, 1999) or mislabeling, as Corbett, Patel, Erikson and Friendship (2003) found that 12% of violent convictions were actually sexually motivated.

The scores of all risk assessment instruments included in the current study generally showed moderate to large predictive accuracy for the occurrence and number of both types of reoffending, with two exceptions: The SVR-20 score predicted sexual recidivism poorly and the RRASOR score predicted violent (including sexual) recidivism poorly. The weaker results of the SVR-20 scores may have been due to the fact that the SVR-20 was originally designed as a Structured Professional Judgment instrument (SPJ, guiding clinical judgment) for use in clinical forensic psychiatric treatment settings, not as a comprehensive actuarial risk marker based on file information. This was also reflected by a substantial percentage of missing values (10.1%). Nevertheless, though the mechanical use of the SVR-20 generally showed better results than observed in this study (Hanson & Morton-Bourgon, 2009), poor predictive accuracy for the SVR-20 score, specifically regarding sexual recidivism has been found in several other European studies as well (Koster et al., 2006; Schönberger et al., 2008; Sjöstedt & Långström, 2002). It is unlikely that the use of the SPJ method would have improved the current results. Although some studies have found the ‘final risk judgment’ based on the SVR-20 scores to increase the predictive accuracy of its mechanical score (De Vogel et al., 2004), it has more often been found that clinical adjustment of any risk assessment instrument score conferred no benefit or even decreased the predictive accuracy of the scores (e.g. Dempster, 1998; Morton, 2003; Michel et al., 2013;
Wormith, Hogg & Guzzo, 2012). The weaker results for the RRASOR regarding violent (including sexual) recidivism concurs with the summarized results of prior research including the RRASOR (Hanson & Morton-Bourgon, 2009).

As expected, the SORAG scores were particularly valid for violent (including sexual) recidivism, for which SORAG was specifically designed. The SORAG scores attained that result despite a substantial number of missing values (6.0 %) permissible substitutions of two items and extensive adjustments to two of the original items. While this seems to hold promise for the validity of the interpretation of the SORAG scores when fully completed as originally intended by its developers, it bears mentioning that some of the requisite information is not readily available in Dutch judicial record keeping. Specifically, a number of items refer to formal psychodiagnoses (schizophrenia, personality disorder, psychopathy) that are preferably achieved through careful, structured and evidence based assessment.

Direct statistical comparison of the included measures showed that the Static-2002R, Static-2002 and the Static-99R scores performed significantly better than the lowest performing instrument score regarding the occurrence of recidivism on five out of six outcome measures. Scores on these instruments (including the RM2000s/c) also showed the largest correlation to the number of reoffenses and latency of reoffending in the current sample. This observation suggests that these instruments may be particularly useful for actuarial risk assessment in The Netherlands.

The same pattern of findings found for the dichotomous occurrence of re offending, generally held for the additional outcome measures, i.e. number of reoffenses and latency of reoffending. Results for the number of reoffenses were more consistent than for latency, probably due to the fact that the estimate of the latter was inherently based on smaller numbers (only those who reoffended instead of all offenders). Taken together, our findings suggest that the risk assessment instruments helped identify those offenders who were likely to continue their offending behavior, irrespective of the specific outcome index used. It should be acknowledged that not all of the included instruments were originally designed to predict non-violent sexual reoffending, but we note that the occurrence of non-violent sexual reoffending in the current sample was extremely low (n = 1) and is unlikely to have significantly influenced the results.

Results provided additional support for the hypothesized positive association between inter rater agreement and predictive accuracy. Similar to Hanson and Morton-Bourgon’s (2009) meta-analytic findings that studies reporting higher inter rater agreement also reported higher predictive accuracy, the current study showed the same result at the item level. The predictive accuracy of the original 64 comprising items of the various instruments correlated positively and significantly with the inter rater agreement of their scores regarding both sexual (r = .28, p = .03) and violent (including
sexual) recidivism ($r = .45$, $p < .001$). Like Hanson and Morton-Bourgon (2009), the results of the current study did not provide clear support for the hypothesized negative association between missing values and predictive accuracy. The correlation between the predictive accuracy of the 64 original items and the percentage of missing values was in the expected (negative) direction but did not reach statistical significance. However, it should be noted that the prevalence of missing values was quite low (1.5%). Findings regarding missing values also showed that the inclusion of incomplete cases (up to 20% missing variables) did not reduce the predictive accuracy of the instrument scores, nor did prorating the missing scores increase predictive accuracy. These results suggest that the use of variables that are not always retrievable, but when retrievable, can be scored unequivocally, may be preferable to variables that can consistently be scored (or roughly estimated) but which scores lack reliability.

A number of additional subgroup comparisons was beyond the scope of the current study, but certainly is worthy of subsequent research on this (and other) sample(s). For instance, different mechanisms may apply to different subgroups of sexual offenders (Hanson, 2002; Hood, Shute, Feilzer & Wilcox, 2002) who are known to often reoffend at different rates and ages (Bartosh, Garby, Lewis, & Gray, 2003; Craig, Browne & Stringer, 2004). Furthermore, the present study did not distinguish between participants who received treatment and those who did not. Hanson and Morton-Bourgon (2005) suggested that the factors that initiate sexual offending may not be the same as the factors associated with persistence, and, in turn, that persistence after treatment might be influenced by different factors than persistence after a prison sentence. Again, the scores of the nine risk assessment instruments may differ in their predictive power across these various subgroups, and future research on this sample may include such moderator analyses.

Some methodological limitations need to be recognized in interpreting our results. First, the administration of judicial files, compiled by case instead of by individual, made detailed information regarding prior offenses difficult to obtain. Most notably, a lack of uniformity in the reports of pre-trial assessments led to variability in available information across participants. Second, the limited time period during which judicial files of deceased offenders were accessible rendered recidivism data of some offenders off limits. Each of these research problems may be remedied by more consistent standards for reporting and better computerized administration by the Dutch judicial system. Finally, regarding the generalizability of the study results to the Dutch sex offender population as a whole, we must reckon with the fact that the subgroup of forensic psychiatric inpatients was overrepresented in our sample, which may, for instance, have somewhat inflated the observed recidivism rates. On the other hand, to
our knowledge, the present sample is a better representation of the Dutch population of convicted sex offenders than any other published study to date.

Notwithstanding these limitations, our findings show that the scores on nine actuarial risk assessment instruments demonstrated moderate to large predictive accuracy with regard to both sexual recidivism and violent (including sexual) recidivism estimates in a large national sample from the Netherlands. While all instruments performed at better than chance levels, the Static 2000(R) and Static99R were most consistent across outcome measures and follow-up periods in this sample. These results underscore the feasibility and utility of these instruments for informing treatment selection according to the RNR principles. Given the results of this study, it becomes increasingly apparent that Dutch policies are in need of revision. Standard risk assessment should not merely be included in pre-trial assessment, but should form the basis of the assessment and guide decisions regarding treatment referral.
References


Smid, W. J., Kamphuis, J. H., Wever, E. C., & Verbruggen, M. C. F. M. (in press). Risk levels, treatment duration, and drop out in a clinically com-
posed outpatient sex offender treatment group. *Journal of Interpersonal Violence.*


