Psychological Assessment

A Comparison of the Predictive Properties of Nine Sex Offender Risk Assessment Instruments

Wineke J. Smid, Jan H. Kamphuis, Edwin C. Wever, and Daniël J. Van Beek
Online First Publication, April 28, 2014. http://dx.doi.org/10.1037/a0036616

Psychological Assessment

1040-3590/14/$12.00 DOI: 10.1037/a0036616

A Comparison of the Predictive Properties of Nine Sex Offender Risk Assessment Instruments

Wineke J. Smid
Forensic Care Specialists, Utrecht, the Netherlands

Jan H. Kamphuis
University of Amsterdam

Edwin C. Wever
Forensic Care Specialists, Utrecht, the Netherlands

Daniël J. Van Beek
Utrecht, the Netherlands

Sex offender treatment is most effective when tailored to risk-need-responsivity principles, which dictate that treatment levels should match risk levels as assessed by structured risk assessment instruments. The predictive properties, missing values, and interrater agreement of the scores of 9 structured risk assessment instruments were compared in a national sample of 397 Dutch convicted sex offenders. The instruments included the Rapid Risk Assessment for Sexual Offense Recidivism, Static-99, Static-99R, a slightly modified version of Static-2002 and Static-2002R, Structured Anchored Clinical Judgments Minimum, Risk Matrix 2000, Sexual Violence Risk 20, and a modified version of the Sex Offender Risk Appraisal Guide; sexual and violent (including sexual) recidivism was assessed over 5- and 10-year fixed and variable follow-up periods. In general, 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-2002R 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 Sexual Violence Risk 20 and Rapid Risk Assessment for Sexual Offense Recidivism scores were the least positive. A positive association between predictive accuracy and interrater agreement at the item level was found for both sexual recidivism ($r = .28$, $p = .01$) 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 risk-need-responsivity principles.

Keywords: sex offender risk assessment, predictive validity, comparison

Accurate risk assessment is essential to optimal treatment selection for offenders. As dictated by risk-need-responsivity (RNR) principles, high-risk offenders should receive more intensive treatment and lower risk offenders should receive 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. In particular, 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 the Static-99 (Hanson & Thornton, 2000), the Sexual Violence Risk 20 (SVR-20; Boer, Hart, Kropp, & Webster, 1997), or both. Each of these studies (de Vogel, de Ruiter, Van Beek, & Mead, 2004; Koster, Van Lankveld, & Spreen, 2006; Schönberger, Hildebrand, Spreen, & Bloem, 2008) was based on a (partly overlapping) selective sample

Wineke J. Smid, Research Department, Forensic Care Specialists, Utrecht, the Netherlands; Jan H. Kamphuis, Faculty of Social and Behavioral Sciences, Clinical Psychology, University of Amsterdam; Edwin C. Wever, Research Department, Forensic Care Specialists, Utrecht, the Netherlands; Daniël J. Van Beek, independent practice, Utrecht, the Netherlands.

The views expressed are those of the authors and not necessarily those of Forensic Care Specialists. This article was included in the doctoral dissertation of Wineke J. Smid, titled "Sex Offender Risk Assessment in the Netherlands: Toward a Risk Need Responsivity Oriented Approach" (University of Amsterdam, 2014). We thank Inge E. Hoebe for assisting with references and tables.

Correspondence concerning this article should be addressed to Wineke J. Smid, Forensic Care Specialists, Research Department, P.O. Box 174, 3500 AD Utrecht, the Netherlands. E-mail: wsmid@hoevenkliniek.nl

of Dutch sex offenders referred to inpatient forensic psychiatric treatment; such a referral was 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 concluded, 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 interrater agreement, missing values, sample selection bias, and widely differing outcome measures (e.g., in terms of the duration of follow-up periods, official vs. unofficial sources utilized, inconsistent definitions for recidivism; see, e.g., Craig, Beech, & Browne, 2006; Falshaw, Bastes, Patel, Corbett, & Friendship, 2003; Hanson, 2002; Marshall & Barbaree, 1988). Although 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 et al., 2007; and the present study), this does not hold for the issues of interrater agreement and missing values.

Interrater Agreement

The interrater agreement is of crucial importance to making a prediction as accurate as possible (Weir & Roberts, 1994). Barbaree, Seto, Langton, and Peacock (2001) analyzed the interrater agreement on a subset of 30 cases for the Sex Offfender Risk Appraisal Guide (SORAG; Quinsey et al., 2006), Rapid Risk Assessment for Sexual Offense Recidivism (RRASOR; Hanson, 1997), Static-99, and Minnesota Sex Offender Screening Tool–Revised (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 but the SORAG and the Static-99 scores were. Interrater 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 their recent review, Tully, Chou, and Browne (2013) noted that less than half of the included studies reported good interrater agreement for the scores on the instruments used. Hanson and Morton-Bourgon (2009) included the topic of interrater 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 intraclass correlation coefficient (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 et al. (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 led 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 found, surprisingly, 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 participant 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, interrater 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. The current study compares in particular the ability of nine of the internationally most commonly used sex offender risk assessment instruments 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 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 et al., 2003), a modified version of the SORAG (Quinsey et al., 2006), and the 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 interrater 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. The two samples added up to a total of 437 sex offenders, of whom 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 that they were supposed to start upon discharge from prison.

A total of 40 records were discarded for various reasons: Eleven convicted sex offenders were excluded because they were female; two offenders were under 18 years of age at the time of discharge; and 6 offenders died 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 2 years from their discharge and another 5 had died within 5 years from their discharge. Follow-up data for 5 remaining offenders were 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 years ($SD = 12.2$), 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, and 6.8% ($n = 27$) had previous charges/convictions for a juvenile sexual offense. See Table 1 for further characteristics of the sample.

Table 1
Demographic and Judicial Characteristics of the Study Participants ($N = 397$)

Characteristic	Statistic
Age at discharge	$M = 38.7$ ($SD = 12.2$)
Caucasian	72.8% ($n = 289$)
Single	55.7% ($n = 221$)
Education	
None beyond primary school	22.5% ($n = 71$)
Vocational training	55.2% ($n = 174$)
Higher education	22.2% ($n = 70$)
Unemployed	37.0% ($n = 147$)
Victims	
Under 16 years old	45.8% ($n = 182$)
16 years or older	41.8% ($n = 166$)
Both	12.3% ($n = 49$)
Prison sentence in months	$M = 18.4$ ($SD = 18.3$)
Conditional part in months	$M = 3.3$ ($SD = 3.4$)
Probation in months	$M = 14.9$ ($SD = 13.5$)

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 offender's criminal offense record is generally available, digital storage of comprehensive file information and digital linkage to related case files have been introduced only recently (around 2005) and were not retroactively executed. With regard to 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 pretrial 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 were 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 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 10 trained coders scored the original 437 files, with 56 files doubly scored by independent raters to assess interrater 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 nonmissing items within each record. Predictive accuracy was calculated for cases with complete records. For those with up to 20% missing items, the raw scores and the prorated scores were used, respectively.

No items were missing for the RRASOR, Static-99R, and the RM2000 (Thornton et al., 2003). A total of two cases had one item missing from the Static-2002R. Prorating for these items did not change the composite score for either participant, as the weighted score for the specific domain remained the same. Hence, all Static-2002R composite scores were included in the analyses as complete. With regard to the SACJ-Min, 18 participants 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 $\leq .50$, the missing item was scored 0; when this average score was $> .50$, the missing item was scored 1. The prorated scores resulted in unchanged risk levels for 15 participants, and three participants were upped one level. With regard to the SORAG, 67 participants had three or more items missing and were excluded. In total, 74 participants 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 participants, 12 participants were adjusted one level up, and three participants were adjusted one level down. With regard to the SVR-20, 64 participants had five or more items missing and were excluded. In total, 248 participants 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 various risk assessment instruments had been coded, 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 and 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 child pornography 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 participants, from the date of discharge to the date of decease. Mean follow-up was 145 months (12 years and 1 month, $SD = 30$ months), with a minimum of 51 months (4 years and 3 months) and a maximum of 201 months (16 years and 9 months). The predictive accuracy of the scores of the nine risk assessment instruments was calculated over fixed follow-up periods of 5 and 10 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 these data.

Instruments

Rapid Risk Assessment for Sexual Offense Recidivism. 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, and 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 (area under the curve [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 participants: A median AUC of .68 (mean AUC = .66) was found for sexual recidivism and a median AUC of .58 (mean AUC = .59) was found 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. The SACJ was developed by Thornton (Grubin, 1998), based on the RRASOR, and was aimed to assess the risk of sexual and violent recidivism over time. The SACJ consists of three stages: (a) assessment of historical risk factors, (b) assessment of aggravating factors, (c) 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, with each item being dichotomously scored. 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) noncontact 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 an AUC of .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 that included almost 1,400 participants: A median AUC of .63 (mean AUC = .62) was calculated for sexual recidivism, and a median AUC of .60 (mean AUC = .60) was calculated for any violent (including sexual) recidivism.

STATIC-99R. The STATIC-99 is the most widely used and most extensively researched 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 10 items, including all four items from the RRASOR and seven items 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 dichotomously scored. The composite 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 of .71 for sexual recidivism and an AUC of .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 published and unpublished studies that included over 20,000 participants: A median AUC of .70 (mean AUC = .68) for sexual recidivism and a median AUC of .64 (mean AUC = .66) for any violent (including sexual) recidivism were observed, 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 describes 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-2002R. The STATIC-2002 was developed by Hanson and Thornton (2003) in an effort to increase coherence and conceptual clarity. Their aim was to assess items referring to theoretically more meaningful characteristics presumed to be the cause of recidivism 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 General 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 similar levels of predictive accuracy to the Static-99 score for the prediction of sexual recidivism (AUC = .76 for both) but the Static-2002 score predicted violent (including sexual) recidivism better than the Static-99 score (AUC = .76 vs. .73). The most recent meta-analysis (Hanson & Morton-Bourgon, 2009) summarizes 10 findings from published and unpublished studies that included over 3,000 participants: A median AUC of .69 (mean AUC = .69) for sexual recidivism and a median AUC of .67 (mean AUC = .68) for any violent (including sexual) recidivism were found, 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-2002R was slightly adjusted for this study. Due to Dutch judicial archiving procedures, it proved to be impossible for nearly all participants 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 cutoff 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 interrater 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. The RM2000 was developed by Thornton et al. (2003) following a reexamination of the SACJ-Min. Thornton et al. created a two-dimensional risk assessment system for sex offenders assessing risk of sexual recidivism and risk of nonsexual 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 1 is translated into a risk category (low, medium, high, or very high). Step 2 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 U.K. samples of treated ($n = 647$) and untreated ($n = 429$) sex offenders. The study reported AUCs of .77 for the treated and .75 for the untreated sample regarding sexual recidivism and AUCs 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 10 RM2000/s findings from published and unpublished studies that included over 2,700 participants: A median AUC of .68 (mean AUC = .68) for sexual recidivism was found. Moreover, five studies including over 1,400 participants for the RM2000/c resulted in a median AUC of .69 and a mean AUC of .70 for any violent (including sexual) recidivism.

Sex Offender Risk Appraisal Guide. 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, martial 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 participants for the SORAG: A median AUC = .66 (mean AUC = .67) for sexual recidivism and a median AUC = .69 (mean AUC = .71) for any violent (including sexual) recidivism were calculated, 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 of which refer to actual modifications. First, as phallometric testing for sex offenders is virtually nonexistent in the Netherlands, Item 13 of the SORAG was, as dictated by the manual, replaced by the offender's 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 available for any of the offenders. Guy and Douglas (2006) found high correspondence between the PCL:SV and PCL–R scores, with intermeasure 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 *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed.; *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 pretrial assessment reports. This also included personality disorders not otherwise specified. 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. 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 Historical, Clinical and Risk Management Scales (HCR-20) and was aimed specifically at the risk assessment of sex offenders with psychiatric disorders in clinical forensic settings. Its use has been mandatory in inpatient forensic treatment settings in the 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 3-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 aims not only to assess risk but also to 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 10 findings from published and unpublished studies that included over 1,700 participants for the SVR-20 using the mechanical approach (adding up the items): It resulted in a median AUC of .66 (mean AUC = .68) for sexual recidivism and a median AUC of .58 (mean AUC = .61) for any violent (including sexual) recidivism.

International research regarding structured professional judgment (SPJ) based on the SVR-20 (and other SPJ instruments) is falling behind. Only three studies with 245 participants were included in Hanson and Morton-Bourgon's (2009) meta-analysis, 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 nonsexual 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

We assessed interrater agreement by means of the intraclass correlation coefficient, 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 cutoff score of the instrument. The resulting area under the curve can be interpreted as the probability that a randomly selected recidivist would score higher on the instrument than a randomly selected nonrecidivist. 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, we used the software program ROCTools (V1.0.2, September 2007; Allaire & Cismaru, 2007). 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 prorated composite scores were used for the analysis, depending on which of these scores resulted in the highest AUC values in the prior analysis.

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

The relationship between interrater 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 5- and 10-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 5- and 10-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 child pornography or exhibitionism), 3 cases concerned only exhibitionism, and 1 case concerned only child pornography possession. Of the 102 nonsexual 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 nonsignificant 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 5-year follow-up.

The results regarding the additional outcome measures (i.e., number of offenses and latency to first offense) are presented in Tables 4 and 5. In general, moderate to large effect sizes were found for the number of offenses (see Table 4) when nonrecidivists 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 nonrecidivists were excluded, the effects were smaller and reached significance only 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 nonrecidivists were included (latency = follow-up time) and generally small effect sizes when nonrecidivists were excluded. Results regarding the latency for the first sexual reoffense were opposite: In general, small effect sizes were found when nonrecidivists were included (latency = follow-up time), whereas effect sizes ranged between none and large when nonrecidivists 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 5-year follow-up period, $\chi^2(8, N = 309) = 30.93$, $p < .001$; a fixed 10-year

Table 2
Predictive Accuracy Regarding Sexual Reoffending (Charge or Conviction for a New Sexual Offense) Expressed in AUC Values and Corresponding Confidence Intervals

Instrument	5-year follow-up [95% CI]	10-year follow-up [95% CI]	Variable follow-up [95% CI]
RRASOR	.68*** [.59, .77]	.69*** [.60, .77]	.67*** [.59, .75]
	n = 396	*n* = 323	*n* = 397
SACJ-Min, complete	.69*** [.60, .78]	.71*** [.63, .79]	.69*** [.61, .77]
	n = 378	*n* = 306	*n* = 379
SACJ-Min, ≤20% missing, raw	.68*** [.58, .77]	.70*** [.62, .78]	.69*** [.61, .77]
	n = 396	*n* = 323	*n* = 397
SACJ-Min, ≤20% missing, prorated	.67*** [.58, .76]	.69*** [.61, .78]	.69*** [.61, .77]
	n = 396	*n* = 323	*n* = 397
Static-99	.72*** [.64, .81]	.73*** [.65, .81]	.72*** [.65, .80]
	n = 396	*n* = 323	*n* = 397
Static-99R	.74*** [.65, .82]	.74*** [.66, .82]	.74*** [.67, .81]
	n = 396	*n* = 323	*n* = 397
Static-2002	.76*** [.67, .85]	.75*** [.67, .83]	.74*** [.66, .81]
	n = 396	*n* = 323	*n* = 397
Static-2002R	.77*** [.68, .85]	.75*** [.67, .83]	.74*** [.67, .82]
	n = 396	*n* = 323	*n* = 397
RM2000/s	.72*** [.64, .81]	.71*** [.63, .79]	.71*** [.64, .79]
	n = 396	*n* = 323	*n* = 397
SORAG, Complete	.63** [.52, .75]	.64** [.54, .75]	.64** [.54, .74]
	n = 256	*n* = 211	*n* = 256
SORAG, ≤20% missing, raw	.66** [.56, .76]	.68*** [.59, .77]	.67*** [.59, .76]
	n = 329	*n* = 272	*n* = 330
SORAG, ≤20% missing, prorated	.66** [.56, .76]	.68*** [.59, .77]	.67*** [.59, .76]
	n = 329	*n* = 272	*n* = 330
SVR-20, complete	.53 [.34, .72]	.48 [.30, .66]	.50 [.33, .66]
	n = 85	*n* = 68	*n* = 85
SVR-20, ≤20% missing, raw	.58 [.48, .67]	.61* [.52, .69]	.61* [.53, .69]
	n = 332	*n* = 268	*n* = 333
SVR-20, ≤20% missing, prorated	.58 [.48, .67]	.61* [.52, .70]	.61* [53, .69]
	n = 332	*n* = 268	*n* = 333

Note. AUC = area under the curve; CI = confidence interval; RRASOR = Rapid Risk Assessment for Sex Offender Recidivism; SACJ-Min = Structured Anchored Clinical Judgments Minimum; RM2000/s = Risk Matrix 2000/Sexual; SORAG = Sex Offender Risk Appraisal Guide; SVR-20 = Sexual Violence Risk 20.
* $p < .05$. ** $p < .01$. *** $p < .001$.

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 5-year follow-up period, $\chi^2(8, N = 309) = 24.46$, $p < .001$; the fixed 10-year follow-up period, $\chi^2(8, N = 252) = 43.27$, $p < .001$; and the variable follow-up period, $\chi^2(8, N = 309) = 50.86$, $p < .001$. Pairwise 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 5-year follow-up period, 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 10-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 5-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 10-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 significantly better than the instrument with the lowest performance on four of the six 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%). ICCs 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, interrater agreement, and percentage of miss-

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

Instrument	5-year follow-up [95% CI]	10-year follow-up [95% CI]	Variable follow-up [95% CI]
RRASOR	.59** [.53, .66] n = 396	.63*** [.56, .69] n = 323	.60** [.54, .66] n = 397
SACJ-Min, complete	.67*** [.60, .73] n = 378	.69*** [.62, .75] n = 306	.67*** [.61, .73] n = 379
SACJ-Min, ≤20% missing, raw	.65*** [.59, .72] n = 396	.67*** [.61, .74] n = 323	.66*** [.61, .72] n = 397
SACJ-Min, ≤20% missing, prorated	.65*** [.59, .72] n = 396	.68*** [.61, .74] n = 323	.67*** [.61, .72v n = 397
Static-99	.68*** [.62, .74] n = 396	.71*** [.65, .77] n = 323	.70*** [.64, .75] n = 397
Static-99R	.71*** [.65, .76] n = 396	.74*** [.68, .79] n = 323	.73*** [.68, .79] n = 397
Static-2002	.71*** [.66, .77] n = 396	.76*** [.70, .81] n = 323	.73*** [.68, .78] n = 397
Static-2002R	.72*** [.66, .77] n = 396	.75*** [.70, .81] n = 323	.74*** [.69, .79] n = 397
RM2000/c	.70*** [.64, .76] n = 396	.73*** [.68, .79] n = 323	.73*** [.68, .78] n = 397
SORAG, Complete	.71*** [.63, .78] n = 256	.75*** [.67, .82] n = 211	.73*** [.66, .80] n = 256
SORAG, ≤20% missing, raw	.70*** [.64, .77] n = 329	.75*** [.68, .81] n = 272	.73*** [.67, .78] n = 330
SORAG, ≤20% missing, prorated	.70*** [.64, .77] n = 329	.75*** [.68, .81] n = 272	.73*** [.67, .78] n = 330
SVR-20, complete	.58 [.44, .72] n = 85	.64 [.50, .78] n = 68	.64 [.51, .76] n = 85
SVR-20, ≤20% missing, raw	.64*** [.57, .71] n = 332	.65*** [.58, .72] n = 268	.65*** [.59, .71] n = 333
SVR-20, ≤20% missing, prorated	.65*** [.59, .71] n = 332	.67*** [.60, .73] n = 268	.67*** [.61, .73] n = 333

Note. AUC = area under the curve; CI = confidence interval; RRASOR = Rapid Risk Assessment for Sex Offender Recidivism; SACJ-Min = Structured Anchored Clinical Judgments Minimum; RM2000/c = Risk Matrix 2000/Combined; SORAG = Sex Offender Risk Appraisal Guide; SVR-20 = Sexual Violence Risk 20.
** $p < .01$. *** $p < .001$.

ing 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

Our aim in 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, Helmus, & Thornton, 2010; Correctional Service of Canada sample: 8.7% sexual and 23.4% violent including sexual recidivism) but were 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 all offenses are not 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

Table 4
Relationship Between the Number of Sexual and Violent (Including Sexual) Reoffenses Expressed in Spearman Correlation Coefficients

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

Note. Violent refers to violent including sexual reoffenses. RRASOR = Rapid Risk Assessment for Sex Offender Recidivism; SACJ-Min = Structured Anchored Clinical Judgments Minimum; RM2000s/c = Risk Matrix 2000, Sexual/Combined; SORAG = Sex Offender Risk Appraisal Guide; SVR-20 = Sexual Violence Risk 20.
[a] n = 330. [b] n = 49. [c] n = 107. [d] n = 333. [e] n = 50. [f] n = 110.
* $p < .05$. ** $p < .01$.

originally designed as a structured professional judgment (i.e., guiding clinical judgment) instrument for use in clinical forensic psychiatric treatment settings, not as a comprehensive actuarial risk marker based on file information. This possibility 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. Although this finding 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 men-

Table 5
Relationship Between Latency for the First Sexual and Violent (Including Sexual) Reoffense Expressed in Spearman Correlation Coefficients

Instrument	Sexual nonrecidivists included (n = 397)	Sexual nonrecidivists excluded (n = 56)	Violent nonrecidivists included (n = 397)	Violent nonrecidivists excluded (n = 133)
RRASOR	−.18**	−.14	−.20**	−.16
SACJ-Min, ≤20% missing, prorated	−.21**	−.14	−.27**	−.15
Static-99	−.23**	−.24	−.32**	−.21*
Static-99R	−.19**	−.29*	−.33**	−.23**
Static-2002	−.21**	−.35**	−.35**	−.21*
Static-2002R	−.20**	−.37**	−.34**	−.21*
RM2000s/c	−.19**	−.38**	−.29**	−.22*
SORAG, ≤20% missing, prorated	−.14[a]*	−.06[b]	−.31[a]**	−.11[c]
SVR-20, ≤20% missing, prorated	−.04[d]	−.03[e]	.15[d]**	.00[f]

Note. Violent refers to violent including sexual reoffenses. Latency for nonrecidivists consisted of their follow-up time. RRASOR = Rapid Risk Assessment for Sex Offender Recidivism; SACJ-Min = Structured Anchored Clinical Judgments Minimum; RM2000s/c = Risk Matrix 2000, Sexual/Combined; SORAG = Sex Offender Risk Appraisal Guide; SVR-20 = Sexual Violence Risk 20.
[a] n = 330. [b] n = 49. [c] n = 107. [d] n = 333. [e] n = 50. [f] n = 110.
* $p < .05$. ** $p < .01$.

tioning that some of the requisite information is not readily available in Dutch judicial record keeping. In particular, 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 reoffending 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 nonviolent sexual reoffending, but we note that the occurrence of nonviolent 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 interrater agreement and predictive accuracy. The findings of the current study, similar to Hanson and Morton-Bourgon's (2009) meta-analytic findings that studies reporting higher interrater agreement also reported higher predictive accuracy, 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 interrater agreement of their scores regarding both sexual ($r = .28$, $p = .03$) and violent (including sexual) recidivism ($r = .45$, $p < .001$). Like those of 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 the use of variables that can consistently be scored (or roughly estimated) but whose scores lack reliability.

A number of additional subgroup comparisons were beyond the scope of the current study, but are certainly 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 should 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 pretrial 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. Although all instruments performed at better than chance levels, the Static-2002R and Static-99R 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 pretrial assessment but should form the basis of the assessment and guide decisions regarding treatment referral.

References

Allaire, J. F., & Cismaru, D. (2007). ROCTools: A statistical software for the analysis of ROC curves [Computer software]. Montreal, Ontario, Canada: Statistics Consulting Group of Centre de recherche de l'Institut Philippe-Pinel de Montréal.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). Newark, NJ: LexisNexis.

Andrews, D. A., Bonta, J., & Hodge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17,* 19–52. doi:10.1177/0093854890017001004

Andrews, D. A., Bonta, J., & Wormith, S. J. (2006). The recent past and near future of risk/need assessment. *Crime & Delinquency, 52,* 7–27. doi:10.1177/0011128705281756

Bagley, C., & Pritchard, C. (2000). Criminality and violence in intra- and extra-familial child abusers in a 2-year cohort of convicted perpetrators. *Child Abuse Review, 9,* 264–274. doi:10.1002/1099-0852(200007/08)9:4<264::AID-CAR635>3.0.CO;2-6

Barbaree, H. E., Seto, M. C., Langton, C. M., & Peacock, E. J. (2001). Evaluating the predictive validity of six risk assessment instruments for adult sex offenders. *Criminal Justice and Behavior, 28,* 490–521. doi:10.1177/009385480102800406

Bartosh, D. L., Garby, T., Lewis, D., & Gray, S. (2003). Differences in the predictive validity of actuarial risk assessments in relation to sex offender type. *International Journal of Offender Therapy and Comparative Criminology, 47,* 422–438. doi:10.1177/0306624X03253850

Bengtson, S. (2008). Is newer better? A cross-validation of the Static-2002 and the Risk Matrix 2000 in a Danish sample of sexual offenders. *Psychology, Crime & Law, 14,* 85–106. doi:10.1080/10683160701483104

Boer, D. P., Hart, S. D., Kropp, P. R., & Webster, C. D. (1997). *Manual for the Sexual Violence Risk-20: Professional guidelines for assessing risk of sexual violence*. Vancouver, British Columbia, Canada: British Columbia Institute Against Family Violence.

Brouwers, M., & Smit, P. R. (2005). Prevalentie van seksuele delinquentie door de jaren heen [Prevalence of sexual offending throughout the years]. *Justitiële Verkenningen, 31,* 37–47.

Corbett, C., Patel, V., Erikson, M., & Friendship, C. (2003). The violent reconvictions of sexual offenders. *Journal of Sexual Aggression, 9,* 31–39. doi:10.1080/1355260031000137959

Craig, L. A., Beech, A. R., & Browne, K. D. (2006). Evaluating the predictive validity of sex offender risk assessment measures on UK samples: A cross-validation of the Risk Matrix 2000 scales. *Sexual Offender Treatment, 1*. Retrieved from http://www.sexual-offender-treatment.org/19.98.html

Craig, L. A., Browne, K. D., & Stringer, I. (2004). Limitations in actuarial risk assessment of sexual offenders: A methodological note. *British Journal of Forensic Practice, 6,* 16–32. doi:10.1108/14636646200400003

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44,* 837–845. doi:10.2307/2531595

Dempster, R. J. (1998). *Prediction of sexually violent recidivism: A comparison of risk assessment instruments* (Unpublished master's thesis). Simon Fraser University, Burnaby, British Columbia, Canada.

de Vogel, V., de Ruiter, C., Van Beek, D. J., & Mead, G. (2004). Predictive validity of the SVR-20 and Static-99 in a Dutch sample of treated sex offenders. *Law and Human Behavior, 28,* 235–251. doi:10.1023/B:LAHU.0000029137.41974.eb

Epperson, D. L., Kaul, J. D., & Hesselton, D. (1998, October). *Final report of the development of the Minnesota Sex Offender Screening Tool–Revised (MnSOST-R)*. Presentation at the conference of the Association for the Treatment of Sexual Abusers, Vancouver, British Columbia, Canada.

Falshaw, L., Bastes, A., Patel, V., Corbett, C., & Friendship, C. (2003). Assessing reconviction, reoffending and recidivism in a sample of UK sexual offenders. *Legal and Criminological Psychology, 8,* 207–215. doi:10.1348/135532503322362979

Grann, M., Belfrage, H., & Tengström, A. (2000). Actuarial assessment of risk for violence: Predictive validity of the VRAG and the historical part of the HCR-20. *Criminal Justice and Behavior, 27,* 97–114. doi:10.1177/0093854800027001006

Grubin, D. (1998). *Sex offending against children: Understanding the risk* (Police Research Series Paper 99). London, United Kingdom: Home Office.

Guy, L. S., & Douglas, K. S. (2006). Examining the utility of the PCL:SV as a screening measure using competing factor models of psychopathy. *Psychological Assessment, 18,* 225–230. doi:10.1037/1040-3590.18.2.225

Hanson, R. K. (1997). *The development of a brief actuarial risk scale for sexual offense recidivism* (User Report 97–04). Ottawa, Ontario, Canada: Department of the Solicitor General of Canada.

Hanson, R. K. (2002). Recidivism and age: Follow-up data from 4,673 sexual offenders. *Journal of Interpersonal Violence, 17,* 1046–1062. doi:10.1177/088626002236659

Hanson, R. K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). The principles of effective correctional treatment also apply to sexual offenders: A meta-analysis. *Criminal Justice and Behavior, 36,* 865–891. doi:10.1177/0093854809338545

Hanson, R. K., & Bussière, M. T. (1996). *Predictors of sexual offender recidivism: A meta-analysis* (User Report No. 1996–04). Ottawa, Ontario, Canada: Department of the Solicitor General of Canada.

Hanson, R. K., Helmus, L. M. D., & Thornton, D. (2010). Predicting recidivism amongst sexual offenders: A multi-site study of Static-2002. *Law and Human Behavior, 34,* 198–211. doi:10.1007/s10979-009-9180-1

Hanson, R. K., & Morton-Bourgon, K. E. (2005). The characteristics of persistent sexual offenders: A meta-analysis of recidivism studies. *Journal of Consulting and Clinical Psychology, 73,* 1154–1163. doi:10.1037/0022-006X.73.6.1154

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21,* 1–21. doi:10.1037/a0014421

Hanson, R. K., & Thornton, D. (2000). Improving risk assessment for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior, 24,* 119–136. doi:10.1023/A:1005482921333

Hanson, R. K., & Thornton, D. (2003). *Notes on the development of Static-2002* (Corrections Research User Report No. 2003–01). Ottawa, Ontario, Canada: Department of the Solicitor General of Canada.

Hare, R. D. (2003). *The Hare Psychopathy Checklist—Revised* (2nd ed.). Toronto, Ontario, Canada: Multi- Health Systems.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior, 20,* 315–335. doi:10.1177/0093854893020004001

Harris, G. T., Rice, M. E., Quinsey, V. L., Lalumière, M. L., Boer, D. P., & Lang, C. (2003). A multisite comparison of actuarial risk instruments for sex offenders. *Psychological Assessment, 15,* 413–425. doi:10.1037/1040-3590.15.3.413

Hart, S. D., Cox, D. N., & Hare, R. D. (1995). *The Hare Psychopathy Checklist: Screening Version*. Toronto, Ontario, Canada: Multi-Health Systems.

Helmus, L. M. D., & Hanson, R. K. (2007). Predictive validity of the Static-99 and Static-2002 for sex offenders on community supervision. *Sexual Offender Treatment, 2,* 1–14.

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Improving the predictive validity of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse: A Journal of Research and Treatment, 24,* 64–101. doi:10.1177/1079063211409951

Hood, R., Shute, S., Feilzer, M., & Wilcox, A. (2002). Sex offenders emerging from long-term imprisonment: A study of their long-term reconviction rates and of parole board members' judgments of their risk. *British Journal of Criminology, 42,* 371–394. doi:10.1093/bjc/42.2.371

Janus, E. S., & Prentky, R. (2003). Forensic use of actuarial risk assessment with sex offenders: Accuracy, admissibility, and accountability. *American Criminal Law Review, 40,* 1443–1499.

Koster, K., van Lankveld, J., & Spreen, M. (2006). Voorspelling van recidive bij zedendelinquenten met behulp van retrospectief gebruik van de PCL-R en SVR-20 [Retrospective prediction of recidivism among sex offenders, using PCL-R and SVR-20]. *Tijdschrift voor Seksuologie, 30,* 204–214.

Langton, C. M., Barbaree, H. E., Seto, M. C., Peacock, E. J., Harkins, L., & Hansen, K. T. (2007). Actuarial assessment of risk for reoffense among adult sex offenders: Evaluating the predictive accuracy of the Static-2002 and five other instruments. *Criminal Justice and Behavior, 34,* 37–59. doi:10.1177/0093854806291157

Marshall, W. L., & Barbaree, H. E. (1988). The long-term evaluation of a behavioural treatment program for child molesters. *Behaviour Research and Therapy, 26,* 499–511. doi:10.1016/0005-7967(88)90146-5

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46. doi:10.1037/1082-989X.1.1.30

Michel, S. F., Riaz, M., Webster, C., Hart, S. D., Levander, S., Müller-Isberner, R., . . . Hodgins, S. (2013). Using the HCR-20 to predict aggressive behavior among men with schizophrenia living in the community: Accuracy of prediction, general and forensic settings, and dynamic risk factors. *International Journal of Forensic Mental Health, 12,* 1–13. doi:10.1080/14999013.2012.760182

Morton, K. E. (2003). *Psychometric properties of four risk assessment measures with male adolescent sexual offenders* (Unpublished master's thesis). Carleton University, Ottawa, Ontario, Canada.

Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology, 62,* 783–792. doi:10.1037/0022-006X.62.4.783

Nunes, K. L., Firestone, P., Bradford, J. M., Greenberg, D. M., & Broom, I. (2002). A comparison of modified versions of the Static-99 and Sex Offender Risk Appraisal Guide. *Sexual Abuse: A Journal of Research and Treatment, 14,* 253–269. doi:10.1023/A:1015322509304

Olver, M. E., Nicholaichuk, T. P., Gu, D., & Wong, S. C. (2013). Sex offender treatment outcome, actuarial risk, and the aging sex offender in Canadian corrections: A long-term follow-up. *Sexual Abuse: A Journal of Research and Treatment, 25,* 396–422. doi:10.1177/1079063212464399

Olver, M. E., Wong, S. C. P., & Nicholaichuk, T. P. (2009). Outcome evaluation of a high-intensity inpatient sex offender treatment program. *Journal of Interpersonal Violence, 24,* 522–536. doi:10.1177/0886260508317196

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). *Violent offenders: Appraising and managing risk* (2nd ed.). doi:10.1037/11367-000

Quinsey, V. L., Rice, M. E., & Harris, G. T. (1995). Actuarial prediction of sexual recidivism. *Journal of Interpersonal Violence, 10,* 85–105. doi:10.1177/088626095010001006

Rettenberger, M., & Eher, R. (2006). Actuarial assessment of sex offender recidivism risk: A validation of the German version of the Static-99. *Sexual Offender Treatment, 1,* 1–11.

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63,* 737–748. doi:10.1037/0022-006X.63.5.737

Rice, M. E., & Harris, G. T. (1999). Sexual aggressors. In D. L. Faigman, D. H. Kaye, M. J. Saks, & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony* (pp. 89–121). St. Paul, MN: West.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d,* and *r*. *Law and Human Behavior, 29,* 615–620. doi:10.1007/s10979-005-6832-7

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York, NY: Sage.

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13,* 19–30. doi:10.1037/1082-989X.13.1.19

Schönberger, H. J. M., Hildebrand, M., Spreen, M., & Bloem, O. (2008). *De Waarde van Gestructureerde Risicotaxatie en van de Diagnose Psychopathie bij Seksuele Delinquenten* [The value of structured risk assessment and psychopathy diagnosis for sex offenders]. Utrecht, Netherlands: Expertisecentrum Forensische Psychiatrie.

Seto, M. C., & Lalumière, M. L. (2001). A brief screening scale to identify pedophilic interests among child molesters. *Sexual Abuse: A Journal of Research and Treatment, 13,* 15–25. doi:10.1023/A:1009510328588

Sjöstedt, G., & Långström, N. (2002). Assessment of risk for criminal recidivism among rapists: A comparison of four different measures. *Psychology, Crime & Law, 8,* 25–40. doi:10.1080/10683160208401807

Smid, W. J., Kamphuis, J. H., Wever, E. C., & Van Beek, D. J. (2013). Treatment selection for sex offenders based on clinical judgment vs. actuarial risk assessment: Match and analysis of mismatch. *Journal of Interpersonal Violence, 28,* 2273–2289. doi:10.1177/0886260512475313

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

Thornton, D., Mann, R., Webster, S., Blud, L., Travers, R., Friendship, C., & Erikson, M. (2003). Distinguishing and combining risks for sexual and violent recidivism. In R. A. Prentky, E. S. Janus, & M. C. Seto (Eds.), *Sexually coercive behavior: Understanding and management* (pp. 225–235). New York, NY: New York Academy of Sciences.

Tully, R. J., Chou, S., & Browne, K. D. (2013). A systematic review on the effectiveness of sex offender risk assessment tools in predicting sexual recidivism of adult male sex offenders. *Clinical Psychology Review, 33,* 287–316. doi:10.1016/j.cpr.2012.12.002

Weir, C., & Roberts, J. (1994). *Evaluation in ELT*. Oxford, United Kingdom: Blackwell.

Wormith, J. S., Hogg, S., & Guzzo, L. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. *Criminal Justice and Behavior, 39,* 1511–1538. doi:10.1177/0093854812455741

Received June 10, 2013
Revision received February 7, 2014
Accepted February 7, 2014 ■