International Journal of
Offender Therapy and
Comparative Criminology
55(5) 716–731

DOI: 10.1177/0306624X09360662
http://ijo.sagepub.com

Risk Assessment Instruments in Repeat Offending: The Usefulness of FOTRES

Astrid Rossegger[1], Arja Laubacher[1], Konstantin Moskvitin[1], Thomas Villmar[2], George B. Palermo[3-5], and Jérôme Endrass[1]

Abstract

Research in the area of predicting recidivism has produced several well-validated standardized risk assessment instruments. The question arises, which instruments best serve which purposes? The objective of this study was to evaluate and compare several actuarial and dynamic risk assessment instruments as to their predictive accuracy and their usefulness in forensic practice. The sample consisted of 109 violent and sex offenders who had been released from prison in Switzerland between 1994 and 1999, and for whom the Psychopathy Checklist–Revised (PCL-R); Historical, Clinical, Risk Management–20; Level of Service Inventory–Revised; Violence Risk Appraisal Guide (VRAG); and the Swiss assessment instrument FOTRES were scored. Using bivariate logistic regression analyses, all instruments were able to discriminate between recidivists and nonrecidivists. The receiver operating characteristic analyses yielded area under the curve values between 0.70 (VRAG) and 0.84 (PCL-R). Furthermore, it was shown that solely examining AUC values does not suffice to determine usefulness. A comprehensive evaluation of an instrument's usefulness for forensic practice should also look at qualitative criteria such as area of application, specificity of risk assessed, and inclusion of dynamic items among others.

Keywords

risk assessment instrument, predictive validity, risk of recidivism, violent and sex offenders

[1]Criminal Justice System, Canton of Zurich, Switzerland
[2]Risk Assessment Centre, Correctional Facility Hannover, Germany
[3]University of Nevada Medical School, Las Vegas
[4]Medical College of Wisconsin, Milwaukee
[5]Marquette University, Milwaukee, Wisconsin

Corresponding Author:
Astrid Rossegger, Psychiatric/Psychological Service, Criminal Justice System, Canton of Zurich, Crime Prevention Research Unit, Feldstrasse 42, 8090 Zurich, Switzerland
Email: astrid.rossegger@ji.zh.ch

Introduction

Standardized risk assessment instruments are considered to be the method of choice in North America (Quinsey, Harris, Rice, & Cormier, 2006) as several studies have clearly demonstrated that using them leads to better predictive validity than clinical or intuitive approaches. Over the years, several standardized risk assessment instruments have been developed and published, such as the Psychopathy Checklist–Revised (PCL-R; Hare, 1991), the Violence Risk Appraisal Guide (VRAG; Quinsey et al., 2006), the Sex Offender Risk Appraisal Guide (Quinsey et al., 2006), the Static-99 (Hanson & Thornton, 1999), the Historical, Clinical, Risk Management–20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997), and the Level of Service Inventory–Revised (LSI-R; Andrews & Bonta, 2001).

All these instruments have been validated in numerous studies. In a typical validation study, reoffending is defined as the outcome variable and the scores of the risk assessment scales are treated as the predictor variables. Harris, Rice, and Quinsey (1993) stated that using receiver operating characteristic (ROC) analyses is the best methodological approach to assess the validity of a risk assessment instrument (and thus to compare the validity of competing instruments). The advantage of ROC analyses is that they produce an effect size (the area under the curve [AUC]), which is a measure for the discriminatory power of an instrument: In an ROC analysis, the scores of all "positives" (e.g., recidivists) are pairwise compared with the scores of all "negatives" (e.g., nonrecidivists). The relative frequency with which the recidivists achieved higher scores in the applied instrument (i.e., higher scores equal higher likelihood of reoffending as stated by the instrument) than the nonrecidivists in these pairwise comparisons illustrates the discriminatory power of the tested instrument. If all recidivists achieved a higher score than the nonrecidivists, the AUC value would be 1, which would mean that the instrument discriminates perfectly between recidivists and nonrecidivists. The AUC value thus allows determining how reliably persons with higher risks of reoffending can be differentiated from those with lower risks.

In this context, it is important to emphasize that although the AUC value is a measure for determining the discriminatory power of an instrument, it is not a measure with which to determine the accuracy of an instrument's calibration (Endrass, Urbaniok, Held, Vetter, & Rossegger, 2009). Calibration statistics analyze for each risk category (of an instrument) the match between predicted rates of recidivism (as predicted by the instrument) and the empirically observed rates of recidivism.

The following is an example of an instrument with poor calibration and perfect discrimination: If all recidivists in a sample had a risk of 10% (as calculated by the instrument to be validated) and all nonrecidivists a risk of 9%, the AUC value of the instrument in question would be 1 (= perfect discrimination), as in all the pairwise comparisons the recidivists would have a higher risk of reoffending than the nonrecidivists. However, the assessment of the risk of reoffending would be poor, because a recidivism rate of 100% is not to be expected for a group for which the calculated risk was 10%. Furthermore, the difference between a risk of 9% and 10% would be too small to be of importance in daily practice and would most likely be disregarded.

Thus, validation studies of risk assessment instruments should always also include an examination of the instruments' calibration aside from an analysis of its ability to discriminate. Unfortunately, most empirical studies continue to neglect this and only report AUC values.

What is even more problematic is that focusing exclusively on AUCs will inevitably influence the development of new (and the improvement of established) risk assessment instruments insofar as aspects of an instrument that are detrimental to achieving high AUC values may be dropped even if they would be important to forensic practitioners. Or aspects may get modified in such a way as to increase AUC values even though the modification might mean a decrease of practical usefulness. Such aspects important to forensic practitioners in daily practice include an instrument's area of application ("Which population can the instrument be applied to?"), its specificity ("What type of risk can the instrument assess?"), its practicability ("How are scores to be translated into practical decisions?"), its exhaustiveness ("Can all aspects important in assessing a particular offender be represented with the items of the instrument?"), and its ability to record change processes. In the following sections, these aspects are presented and discussed in more detail.

Area of Application

The "value" of an instrument also lies in its applicability to various populations: Is the instrument valid for a specific offender population or for a specific judicial or correctional setting? Although some instruments, such as the PCL-R (Hare, 1991), have been developed for offenders in general, other scales have been developed for specific offender groups, for instance the VRAG (Quinsey et al., 2006) for violent offenders and the Static-99 (Hanson & Thornton, 1999) for sex offenders. The more specific the target population, the narrower the area to which the respective instrument can be applied. Thus, more than one instrument is needed in order to deal with different types of offenders.

Specificity of Risk Assessed (Outcome Definition)

The definition of the outcome criteria determines whether the use of a specific risk assessment instrument is helpful in answering the question concerned or not. Risk assessment instruments differ considerably regarding the definition of the outcome criterion. Some risk assessment instruments do not specify the outcome criterion (e.g., the PCL-R; Hare, 1991); some predict violent offenses (e.g., VRAG; Quinsey et al., 2006; and HCR-20; Webster et al., 1997) or sexual offenses (e.g., Static-99; Hanson & Thornton, 1999). In the case of the VRAG, the term *violent recidivism* includes sexual offenses. As it is known that the base rates of the different types of violent and sexual offenses vary considerably, this seems unsatisfactory from a methodological point of view. Furthermore, for forensic practitioners, this may often not be specific enough, as for example the courts in Germany and Switzerland demand expert reports to state offense-specific estimates.

Type of Result (Interpretation of Sum Score)

Irrespective of the results obtained in ROC analyses, the score of a risk assessment instrument has to be "translated" in order for the criminal justice system to make any use of it. The following example shows how clinical usefulness and high AUC values can be diametrically opposed: Assume that there is a risk assessment instrument that performs very well in a validation study, resulting in an AUC of 0.95. This instrument consists of 200 dichotomous items, which by comparison is a very large risk assessment instrument and covers a wide range of scores (0-200). The impressive AUC score of this instrument is quite likely the result of the reliability and range of the scale. Simply put, it is the result of the statistical power. Statistical power refers to the probability that the test will reject the null hypothesis when the alternative hypothesis is true. The statistical power of a test is—among other factors—affected by the reliability of a scale, which in turn is a function of the length of a scale. A scale containing 20 items (e.g., the PCL-R) is by comparison an extensive scale and thus likely to be reliable and to produce statistically significant results.

The good performance of extensive scales comes along with a serious disadvantage, namely, the difficulty of interpretation. For example, in IQ testing, intelligence is measured using a standardized score, the IQ, with a mean of 100 and a standard deviation of 15. This score is reliable but practitioners will tend to break down the information into subcategories, such as above average ($IQ > 100$) and below average ($IQ < 100$) or cognitive impairment ($IQ \leq 70$). A similar difficulty can be found when applying the PCL-R (Hare, 1991) and, for instance, trying to make a clinically useful interpretation of a score of 20 points. A forensic practitioner might be interested to know whether an offender with a score of 20 points can be treated or released without the necessity of psychotherapy. The PCL-R in its current form, however, serves "only" as an unspecific indicator of the risk of committing any offense, without specifying the probability of the offense occurring within a certain time period (time at risk). The advantage of instruments like the VRAG (Harris et al., 1993) and the Static-99 (Hanson & Thornton, 1999) is that they do just that: They provide for each sum-score an estimate of the recidivism risk—the empirically observed rates of specific types of reoffending (e.g., sexual or violent) associated with the different risk categories within a certain time period.

Exhaustiveness

The usefulness of an instrument also lies in its ability to provide detailed case information. Such detail will help in policy making and population management. Brevity—among other attributes—characterizes actuarial instruments. Although the predictive validity of most of these instruments is satisfactory, the scarcity of relevant information makes it almost impossible to develop an understanding of the offender's personality and the offense pattern. This can be illustrated using the following example: Assume that one must assess two young offenders who are both less than 25 years old, who always lived with their biological parents, and who never displayed any problematic behavior at school. These young offenders do not have a (family) history of alcohol abuse, no

prior convictions, never lived in a marital relationship, do not meet the criteria of any *DSM-IV* Axis I or II disorder, and for whom it was not possible to establish a reliable PCL-R score (Hare, 1991) because of lack of information. Assume that the first offender after a bad day at work had gone to a bar, drank too much, and after a verbal dispute, had hit his drinking buddy, knocking him out for a couple of seconds. Worried for his friend, the offender had called the paramedics who, after arrival, determined that no hospitalization was needed. The second offender had been living with deviant sexual fantasies about raping and killing a girl for more than 10 years. These fantasies had become so intense that the offender had abducted a girl, then raped and killed her. The first offender would score 0 on the VRAG (see www.zurichforensic.org for a computerized version of the VRAG), which would mean that on average 48% of offenders with these characteristics will reoffend within 10 years. The second offender, however, would score only minus 9 points, thus 9 points less than the first one, and the expected rate of reoffending for an offender with this profile is "only" 24% within 10 years. Applying clinical and common sense, it is evident that the first offender, presents less of a danger to society than the second one even though their profiles according to the VRAG would read very similarly. Furthermore, according to the VRAG total score, the second offender would be assessed as less problematic with a reoffending risk of only 24% compared to the 48% of the first offender. There are of course several good explanations why an actuarial instrument would assess a higher likelihood of reoffending for the first offender: The most salient would be that offenders who commit a capital offense (the second offender in the above example) are very likely to be incarcerated for life and would therefore no longer pose a risk to society, which would lead in any statistical modeling to lower probabilities. Although this example may show the limitations of actuarial risk models, it does not serve as an argument to question the validity of a statistical instrument. Exponents of actuarial instruments rightly emphasize the difficulty in applying statistical scores to an individual and therefore suggest statements like the following:

> Interpretation of Results (for the first offender): Using the VRAG, a total score of 0 was calculated. . . . This score is assigned to Risk Category 5. Among offenders in the development sample for the VRAG, 50% obtained higher scores, and approximately 35% in Mr. Lastname's category reoffended violently within an average of 7 years (48% within an average of 10 years) after release. (www.zurichforensic.org, 2009)

It is obviously difficult to develop a sense for the case and communicate the risk with actuarial instruments as they usually contain only few items, which rely on few coarse personal characteristics (e.g., growing up with both parents—yes/no) and offense characteristics (e.g., previous convictions, gender of victim). Unlike the reductionism of actuarial scales, clinical instruments attempt to assess the complexity of a case, which allows professionals in the field to better understand the offender and thus to better plan suitable interventions.

Inclusion of Dynamic Items (Ability to Document Change Processes)

In many constitutional states, the risk of reoffending needs to be assessed at different times during the judicial process, such as during remand detention prior to sentencing, during incarceration prior to granting furloughs or release from prison, and during the probation period. Also, most such states demand an expert assessment of treatment progress before releasing sex and violent offenders into society. Ideally, instruments used for monitoring offender treatment would provide information on the indicated goals of a treatment, the required level of security, recommendations for early release (Kroner & Mills, 2001), and information about the therapeutic progress in offense-oriented treatment plans (Ogloff, Wong, & Greenwood, 1990; Seto & Barbaree, 1999). Furthermore, it would seem practical if risk assessment instruments would also include items that assess therapeutically relevant factors, such as openness, awareness of one's own risk factors, and (emotional) authenticity.

Actuarial risk assessment instruments such as the VRAG (Quinsey et al., 2006) or the Static-99 (Hanson & Thornton, 1999) mainly consist of static variables and are therefore of very limited use in assessing any changes to the risk of reoffending following different correctional interventions. This is why certain instruments include dynamic items that allow the documentation of needs and changes. One of these instruments is the Level of Service Inventory–Revised (LSI-R; Andrews & Bonta, 2001). The LSI-R is useful in bringing together information on the offender's basic risk disposition, the offender's need for treatment and supervision, and the level of freedom that can be granted in a correctional setting. A limitation of the LSI-R is that it was not specifically developed for the standardized documentation of sex and violent offenders. In fact, the majority of risk assessment instruments were designed neither to assess the change potential of a sex or violent offender nor to track changes in offense-relevant aspects, such as sadism, narcissistic traits, or deviant sexual fantasies.

All in all, the advantage of the currently available risk assessment instruments lies in their predictive validity. Practical validity or other criteria of practical usefulness have barely been investigated. Clinicians who need to estimate specific risks, the potential for change, or treatment progress might be forced to fall back to the method of intuitive risk assessment as a result of inadequate standardized alternatives. Table 1 gives an overview over the different features of risk assessment instruments.

A New Instrument for Risk/Needs Assessment Integrating Scales to Estimate Treatment Progress

The aforementioned limitations of current risk assessment instruments were the principal reason why the forensic service of the largest Criminal Justice System in Switzerland decided to routinely use the Forensic Operationalized Therapy/Risk Evaluation System (FOTRES; Urbaniok, 2007) for assessing risk as well as managing and monitoring treatment progress and quality, as part of a risk assessment battery.

Table 1. Different Aspects of Risk Assessment Instruments Relevant to Forensic Practice

Aspects of instruments	PCL-R (Hare, 1991)	VRAG (Quinsey et al., 2006)	HCR-20 (Webster et al., 1997)	FOTRES (Urbaniok, 2007)	LSI-R (Andrews & Bonta, 2001)
Area of application	Any offender	Violent and sex offender	Mentally disordered violent offender	Any offender	Any offender
Specificity of risk assessed	Unspecific	Violent and sex	Violent	Specific, specified by user	Unspecific
Type of result					
Risk categories	No	Yes	Yes	Yes	Yes
Calibrated reoffending probabilities	No	Yes, within 7 and 10 years	No	No	Yes, within 1 year
Exhaustiveness	Semiexhaustive	Scarce	Semiexhaustive	Exhaustive	Semiexhaustive
Inclusion of dynamic items	Semidynamic	Static	Semidynamic	Dynamic	Dynamic

Note: PCL-R = Psychopathy Checklist–Revised; VRAG = Violence Risk Appraisal Guide; HCR-20 = Historical, Clinical, Risk Management–20; FOTRES = Forensic Operationalized Therapy/Risk Evaluation System; LSI-R = Level of Service Inventory–Revised.

FOTRES (Urbaniok, 2007) can be applied to all types of offenses; however, the offense to be evaluated must be specified before the application. Accordingly, the risk of reoffending will only be estimated for the specified offense ("target offense"). If an offender committed different offenses, a different recidivism risk would be estimated for each specified offense. FOTRES has two main levels: the Risk-Needs Assessment level and the Risk Management level. In the former, the risk of reoffending is estimated and a prognosis regarding the treatment outcome is established. In the latter, the treatment progress, as well as changes to the recidivism risk caused by situational circumstances, is documented. Whereas the Risk-Needs Assessment is only assessed once (at the time of the index offense), the Risk Management level is scored repeatedly, that is, each time the current risk of reoffending needs to be assessed.

The Risk-Needs Assessment consists of two sublevels: the Structural Risk of Recidivism and Mutability. The items of the Structural Risk of Recidivism sublevel explore the offender's personality disposition to delinquency (e.g., versatile offending, juvenile delinquency). These items also cover specific areas of concern relevant to the offense (e.g., propensity to violence, pedophilia, need for dominance, or personality disorders) as well as the pattern of the offense itself (e.g., degree of planning, consequences of the offense for the victim, weapon use). The second sublevel, Mutability, assesses the mutability of the offender's risk disposition through therapy and/or coping strategies. Here items such as clarity of change focus, openness, resistance to change, or capacity to control impulses are being evaluated (Urbaniok, 2007).

The Risk Management level in FOTRES (Urbaniok, 2007) measures the actual risk reduction achieved through therapy progress, through the implementation of coping strategies, and through the identification and management of offense-related personality patterns. Therefore, FOTRES serves not only as an instrument of prognosis but also as a tool for planning, documenting, and assessing therapy progress.

From a clinical perspective, unlike the majority of the currently available risk assessment instruments, FOTRES (Urbaniok, 2007) offers a combination of risk-needs assessment and risk management scales that could serve as a valuable addition to a battery of risk assessment instruments. Despite its practical advantages, there is no published study available to date that has investigated the outcome validity of this instrument.

The objective of this study was to evaluate the predictive accuracy of FOTRES (Urbaniok, 2007) regarding repeat offending in comparison to several risk assessment instruments, including instruments such as the VRAG (Harris et al., 1993) and the PCL-R (Hare, 1991).

Method

Sample

The study sample consisted of all violent and sex offenders released from the maximum-security unit of the state penitentiary, Pöschwies, between January 1994 and November 1999. Foreigners were only included in the sample if they had a permanent residence status (and could therefore not be deported to the country of origin after having served their prison sentence). One hundred twenty-eight offenders met the aforementioned inclusion criteria. Eleven offenders died during the follow-up period, and eight offenders had to be excluded from the study as their files contained insufficient information to score the selected risk assessment instruments. The final sample consisted of 109 participants.

Characteristics of the Penitentiary

The Pöschwies is the largest and most modern penitentiary in Switzerland, housing up to 436 male inmates. The penitentiary contains three units: The main unit is a maximum-security unit, housing offenders serving a sentence of at least 2 years. The second unit is a medium-security unit, housing offenders who are serving short prison sentences. The last unit is a minimum-security unit, with mostly first-time offenders and offenders released from the maximum-security unit before being discharged to a halfway house. As a general rule, offenders are released on probation after having served two thirds of their sentence. Therapies started during incarceration are continued by the same therapist in a forensic outpatient clinic.

Instruments

The PCL-R (Hare, 1991), HCR-20 (Webster et al., 1997), LSI-R (Andrews & Bonta, 2001), VRAG (Harris, et al., 1993), and FOTRES (Urbaniok, 2007) were scored solely on the basis of the files of the penitentiary; there was no direct contact with any of the offenders. These files contained psychiatric expert assessments, verdicts, information on intramural infractions, and resulting disciplinary actions. They also provided extensive information on the offenders' history, including criminal history and type as well as the exact circumstances of the offense.

The sum scores of the PCL-R and the LSI-R were determined according to the guidelines of the English manual (Andrews & Bonta, 2001; Hare, 2003). The sum scores of the HCR-20 and the VRAG were assessed on the basis of the German version of the instruments (Müller-Isberner, Jöckel, & Gonzalez Cabeza, 1998; Rossegger, Urbaniok, Danielsson, & Endrass, 2009). The sum score of the FOTRES was determined according to the original German version (Urbaniok, 2007) and using an algorithm developed for validation studies of the instrument. The algorithm is posted on the Web page of FOTRES (www.fotres.ch).

Interrater Agreement

In a pilot study, interrater agreement for all instruments was assessed. The interrater reliability (intraclass correlation) was assessed among three raters for $n = 20$ cases and can be considered excellent. In detail, interrater reliability for the HCR-20 was ICC = .98; for the PCL-R, ICC = .93; for the VRAG, ICC = .95, for the LSI, ICC = .97; and for FOTRES, kappa > .65.

Assessing Recidivism

Data on recidivism was gathered from the criminal records in November 2006. Two types of reoffending were analyzed: (a) *General recidivism* was defined as an entry in the criminal record due to any offense having been committed after the index offense. (b) *Repeat offending* was defined as an entry in the criminal record for an offense in the same offense category as the index offense. Recidivism was determined by a psychologist who was blind to the offender's score in the risk assessment instruments.

Statistical Analysis

Logistic regression and ROC analyses were used to estimate the predictive accuracy of the instruments. The AUC value is a measure that assesses an instrument's discriminatory power, that is, its ability to discriminate between recidivists and nonrecidivists. The choice of ROC analysis as a measure of effect size was made according to current recommendation (Swets, Dawes, & Monahan, 2000) as well as based on the thought

that using Cohen's *d* or *r* is not suitable because of the characteristics of the studied sample (Rice & Harris, 2005). All models were estimated using STATA SE 10.0

Results

Sample Description

The mean age of the offenders at the time of their release from prison was 34.7 years. One fourth (26%, *n* = 28) were married at the time of the offense, 70% (*n* = 76) had completed a school education, and 50% (*n* = 55) had completed vocational training. According to ICD-10, the diagnostic criteria for the diagnosis of a psychiatric disorder was met by 68% (*n* = 74) of the participants. In detail, 36% (*n* = 39) of the offenders were diagnosed with a personality disorder, 35% (*n* = 38) with substance use disorder, 6% (*n* = 6) with paraphilia, and 4% (*n* = 4) with major depression. Furthermore, one offender met the diagnostic criteria for schizophrenia.

The index offenses were violent offenses in 80% (*n* = 87) and sex offenses in 20% (*n* = 22) of the cases: In total, 48% (*n* = 52) of the offenders were convicted for robbery, 17% (*n* = 19) for endangering human life, 14% (*n* = 15) for homicide, and 8% (*n* = 9) for abduction and illegal restraint. With 14% (*n* = 15) each, the most frequent sex offenses were child molestation and rape. Sixty percent of the offenders (*n* = 65) had a criminal record entry for a prior offense (prior to the index offense) and 28% (*n* = 30) had entries for a violent or sex offense.

Risk Assessment Instruments

Sum scores and risk categories. PCL-R: The mean PCL-R sum score was 15 points, with a range from 1 to 27.8 points (*SD* = 6.8); no offender had a score above the cutoff of 30 points for the diagnosis of psychopathy. HCR-20: In the HCR-20, the mean sum score was 8 points (*SD* = 17.1, range: 2 to 33). The HCR-20 does not allow categorization of the sum score into risk bins. VRAG: The mean sum score in the VRAG was 5 points (*SD* = 9.4, range: –14 to 28). Transforming the sum score into one of the nine risk categories, 9% (*n* = 10) of offenders were assigned to Risk Category 3, 21% (*n* = 23) to Category 4, 26% (*n* = 28) to Category 5, 24% (*n* = 26) to Category 6, 15% (*n* = 16) to Category 7, 5% (*n* = 5) to Category 8, and 1% (*n* = 9) to Risk Category 9. None of the offenders were assigned to Risk Category 1 or 2. LSI-R: The mean LSI-R score was 21 points (*SD* = 8.9, range: 4 to 41). Twenty-nine percent (*n* = 32) of the offenders were assigned to the lowest risk bin, 30% (*n* = 33) to the second, and 33.9% (*n* = 37) to the third risk bin. Six percent (*n* = 6) were assigned to risk bin 4 and one person to the highest risk bin. FOTRES: To obtain one FOTRES score, the authors decided to use an algorithm to combine the scales of Structural Risk of Recidivism and Mutability, and which was developed by the author of the instrument and published online (Urbaniok, 2009). Adding up the scores of the three subscales, the mean sum score

Table 2. Repeat Offenders Versus Non-Repeat Offenders: Sum Scores

	All (N = 109)		Repeat offenders (n = 10)		Non-repeat offenders (n = 99)	
Instrument	*M*	*SD*	*M*	*SD*	*M*	*SD*
PCL-R	14.7	6.8	22.1	4.14	13.9	6.59
VRAG	5.0	9.4	11.5	9.62	4.4	9.14
HCR-20	7.7	17.1	24.3	5.23	16.4	7.58
LSI-R	20.5	8.9	27.4	7.14	19.9	8.78
FOTRES	7.1	2.2	9.4	1.54	6.9	2.33

Note: PCL-R = Psychopathy Checklist–Revised; VRAG = Violence Risk Appraisal Guide; HCR-20 = Historical, Clinical, Risk Management–20; LSI-R = Level of Service Inventory–Revised; FOTRES = Forensic Operationalized Therapy/Risk Evaluation System.

was 7 points (SD = 2.2, range: 2 to 12). The mean score in the Mutability (MU) was 1.5 (SD = 0.7, range: 0 to 3.5).

In Table 2 the mean scores of the different instruments are summarized for recidivists and nonrecidivists.

Recidivism Rates

The base rate for general recidivism (any reconviction) was 56% (n = 61), whereas the base rate for repeat offending (same offense category) was 9% (n = 10).

Follow-Up Time and Time at Risk

The average follow-up time was 9.0 years (SD = 1.9 years), ranging from 4.5 to 12.5 years. It was not possible to estimate time at risk because of limitations of the Swiss criminal register. The Swiss criminal register only records date of conviction, penalty or sentence, and type of offense. It does not, however, record whether an offender has spent time in an institution; thus, there is no information available on the actual time of incarceration. This limitation made analyses using time-to-event data impossible.

Predicting Recidivism

Sum score. Bivariate logistic regression analyses using the sum score as the predictor variable and recidivism as the outcome variable showed that the odds of repeat offending were increased per point on the scale of the different instruments between 9% (for the VRAG) and 74% (for the FOTRES; see Table 3). Furthermore, the regression analyses revealed that all instruments were at least moderately able to discriminate between offenders who recidivated and those who did not, with p values between .03

Table 3. Repeat Offending: Predictive Validity of Sum Scores

Instrument	exp(*B*)	OR	*p*	95% CI (OR)		AUC	95% CI (AUC)	
PCL-R	0.148	1.25	.00	1.09	1.44	0.84	0.76	0.90
VRAG	0.007	1.09	.03	1.01	1.18	0.70	0.60	0.78
HCR-20	0.111	1.16	.00	1.05	1.29	0.80	0.71	0.87
LSI-R	0.008	1.11	.02	1.02	1.21	0.75	0.65	0.82
FOTRES	0.375	1.74	.00	1.19	2.5	0.81	0.72	0.88

Note: OR = odds ratio; CI = confidence interval; AUC = area under the curve; PCL-R = Psychopathy Checklist–Revised; VRAG = Violence Risk Appraisal Guide; HCR-20 = Historical, Clinical, Risk Management–20; LSI-R = Level of Service Inventory–Revised; FOTRES = Forensic Operationalized Therapy/Risk Evaluation System.

Table 4. Repeat Offending: Predictive Validity of the Risk Categories

Instrument	exp(*B*)	OR	*p*	95% CI (OR)		AUC	95% CI (AUC)	
PCL-R	NA	NA	NA	NA	NA	NA	NA	NA
VRAG	0.051	1.91	.02	1.11	3.30	0.72	0.62	0.80
HCR-20	NA	NA	NA	NA	NA	NA	NA	NA
LSI-R	0.067	2.35	.04	1.07	5.14	0.70	0.60	0.78
FOTRES	0.102	3.73	.01	1.35	10.28	0.76	0.67	0.84

Note: OR = odds ratio; CI = confidence interval; AUC = area under the curve; PCL-R = Psychopathy Checklist–Revised; VRAG = Violence Risk Appraisal Guide; HCR-20 = Historical, Clinical, Risk Management–20; LSI-R = Level of Service Inventory–Revised; FOTRES = Forensic Operationalized Therapy/Risk Evaluation System.

(VRAG) and .00 (PCL-R). The ROC analyses yielded AUC values for all the instruments between 0.70 (VRAG) and 0.84 (PCL-R).

Risk categories. Bivariate logistic regression analyses using the risk categories as predictor variables and recidivism as the outcome criterion could only be carried out for those instruments that allow the sum scores to be transferred to risk categories (VRAG, FOTRES, LSI-R). The analyses showed that the odds of repeat offending were almost doubled per risk category (for the VRAG) and 3.7 times higher per risk category (for the FOTRES). Analyses further revealed AUC values of 0.70 for the LSI-R, 0.72 for the VRAG, and 0.76 for FOTRES (see Table 4).

Discussion

In a first step ROC analyses were conducted for the outcome repeat offending on four widely used and validated risk assessment instruments (PCL-R: Hare, 1991; VRAG: Harris et al., 1993; LSI-R: Andrews & Bonta, 2001; HCR-20: Webster et al., 1997)

and on the FOTRES (Urbaniok, 2007), a newly developed instrument. The examination revealed AUC values of 0.70 to 0.84 for the five instruments indicating that they all were at least satisfactorily able to discriminate between recidivists and nonrecidivists. In a second step, we investigated other aspects significant to forensic practice such as area of application, specificity of risk assessed, etc.

The instrument that produced the highest AUC score was the PCL-R (Hare, 1991). The performance of the PCL-R is remarkable, especially as it was not specifically developed for the estimation of repeat offending but rather for assessing the personality trait of psychopathy. Apart from that, the PCL-R is a rather unspecific instrument: Although it is flexible in its application and can be applied to any offender, its outcome or risk assessed is unspecific, meaning that even though there is a linear relationship between high PCL-R scores and any kind of reoffending (which might be "only" misdemeanors), it is not useful in determining the probability of specific types of reoffending within certain time periods.

Even though the PCL-R (Hare, 1991) showed good discrimination in our sample, it needs to be mentioned that the sample was relatively nonpsychopathic, at least as measured by North American standards. In Europe, studies have frequently found lower average PCL-R scores in their samples (Cooke & Michie, 1999; Cooke, Michie, Hart, & Clark, 2005; Dahle, 2006; Dietiker, Dittmann, & Graf, 2007; Urbaniok, 2007) than studies conducted in North America (Hare, 1991). However, this intercultural variability does not seem to curtail the instrument's ability to discriminate.

In the current study, the VRAG (Quinsey et al., 2006) and the HCR-20 (Webster et al., 1997) did not perform as well as the PCL-R (Hare, 1991). This is further accentuated by the fact that the results of both the VRAG and the HCR-20 are "improved" through the inclusion of the PCL-R—using the PCL-R's sum score as one of their items—this being especially true for the VRAG. One might ask if this indicates the use of the VRAG to be redundant. Clearly it is not, seeing as the VRAG is the only one of the investigated instruments that provides calibrated probabilities for violent or sexual recidivism for a 10-year time at risk period (e.g., scores 7-13 [Risk Category 6] have a 58% risk of violent or sexual reoffending within 10 years).

Apart from the VRAG's (Quinsey et al., 2006) limited predictive validity in this study, its usefulness is also reduced by the fact that it is the most static of all the instruments investigated. In forensic practice, this would mean that once offenders have been assigned to a risk category, they would very likely remain in the same one. Thus, change processes designed to lower recidivism risk, for instance old age, advancing physical illness, learned coping strategies, or personality changes achieved through intensive offense-oriented treatment programs, will not result in a lower score in the VRAG.

The PCL-R (Hare, 1991), on the other hand, is a less static instrument than the VRAG and its sum score can change over time as the instrument assesses personality traits, which may, at least theoretically, be altered given intensive treatment. However, to regard the PCL-R as a dynamic instrument would not seem appropriate in view of the robustness of the measured personality trait. More appropriately, the instrument may be classified as a "semidynamic" instrument.

The HCR-20 (Webster et al., 1997), the LSI-R (Andrews & Bonta, 2001), and the FOTRES (Urbaniok, 2007) are in a better position to meet the needs of forensic practitioners who need to measure treatment progress in order to report on the risk management process. Of these instruments (and of all instruments compared), the highest odds ratio was found for FOTRES. The instrument offers some additional advantages, for example, the possibility of focusing the assessment on a specific offense instead of on the whole criminal history, thus making any prognosis of future criminal behavior more specific. Of these instruments, the LSI-R is the only one to offer the possibility of assigning scores to probabilities, but only for unspecific and short-term reoffending. It is a shortcoming of the HCR-20 and the FOTRES that they do not allow scores to be assigned to probabilities of reoffending. The FOTRES at least offers the possibility of categorizing the sum scores and provides an explanation regarding the level of risk for these categories. These explanations facilitate risk management decisions but are not estimates of risk, such as those provided by the VRAG or the LSI-R, and thus do not permit the running of calibration statistics that compare the actual prevalence rates across the risk categories with the hypothesized categories. Apart from this, both the LSI-R and the FOTRES offer new possibilities for mental health professionals responsible for managing offender risk. The principal difference between FOTRES and the LSI-R is that although the former was developed to monitor the risk management processes of sex and violent offenders by forensic mental health experts, the latter was validated for the short-term assessment of the risks and needs of general offender populations, which better serves the everyday needs of probation officers.

An obvious limitation of the present study was the small *N* for repeat offending. However, as the analytic strategy of the investigation was based on both a quantitative as well as a qualitative approach, it was possible to compensate this shortcoming somewhat and demonstrate that all investigated instruments had advantages and disadvantages. Furthermore, it has been shown that solely examining the AUC values of risk assessment instruments is not adequate to decide which instrument best serves which purpose. Aside from the fact that AUC values represent only a measure of discrimination and not of calibration (and are therefore not able to analyze the quality of the result produced by the VRAG and the LSI-R), the usefulness and validity of forensic risk assessment instruments should also be evaluated, considering aspects such as the area of application, specificity of risk assessed, type of result, exhaustiveness, and inclusion of dynamic items—allowing forensic practitioners to determine calibrated probabilities of specific risk, assemble an exhaustive profile, and document change processes relevant to recidivism risk.

Declaration of Conflicting Interests

The authors declared no conflicts of interests with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

References

Andrews, D. A., & Bonta, J. (2001). *LSI-R. The Level of Service Inventory–Revised. User's manual*. Toronto, Ontario, Canada: Multi-Health Systems.

Cooke, D. J., & Michie, C. (1999). Psychopathy across cultures: North America and Scotland compared. *Journal of Abnormal Psychology, 108*, 58-68.

Cooke, D. J., Michie, C., Hart, S. D., & Clark, D. (2005). Assessing psychopathy in the UK: Concerns about cross-cultural generalisability. *The British Journal of Psychiatry, 186*, 335-341.

Dahle, K.-P. (2006). Strengths and limitations of actuarial prediction of criminal reoffence in a German prison sample: A comparative study of LSI-R, HCR-20 and PCL-R. *International Journal of Law and Psychiatry, 29*, 431-442.

Dietiker, J., Dittmann, V., & Graf, M. (2007). Risk assessment of sex offenders in a German-speaking sample. Applicability of PCL-SV, HCR-20+3, and SVR-20. *Der Nervenarzt, 78*, 53-61.

Endrass, J., Urbaniok, F., Held, L., Vetter, S., & Rossegger, A. (2009). Accuracy of the Static-99 in predicting recidivism in Switzerland. *International Journal of Offender Therapy and Comparative Criminology, 53*, 482-490.

Hanson, R. K., & Thornton, D. (1999). *Static 99: Improving actuarial risk assessments for sex offenders* (No. 1999-02). Ottawa, Ontario, Canada: Solicitor General Canada.

Hare, R. D. (1991). *The Hare Psychopathy Checklist–Revised*. Toronto, Ontario, Canada: Multi-Health Systems.

Hare, R. D. (2003). *Manual for the Revised Psychopathy Checklist* (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.

Kroner, D. G., & Mills, J. F. (2001). The accuracy of five risk appraisal instruments in predicting institutional misconduct and new convictions. *Criminal Justice and Behavior, 28*, 471-489.

Müller-Isberner, R., Jöckel, D., & Gonzalez Cabeza, S. (1998). Die Vorhersage von Gewalttaten mit dem HCR-20. Institut für forensische Psychiatrie Haina.

Ogloff, J. D., Wong, S., & Greenwood, A. (1990). Treating criminal psychopaths in a therapeutic community program. *Behavioral Sciences & the Law, 8*, 181-190.

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). Violent offenders: Appraising and managing risk. In V. L. Quinsey, G. T. Harris, M. E. Rice, & C. A. Cormier (Eds.), *Violent offenders: Appraising and managing risk* (2nd ed., pp. 155-196). Washington, DC: American Psychological Association.

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.

Rossegger, A., Urbaniok, F., Danielsson, C., & Endrass, J. (2009). Der Violence Risk Appraisal Guide (VRAG)–Ein Instrument zur Kriminalprognose bei Gewaltstraftätern: Übersichtsarbeit und autorisierte deutsche Übersetzung [Violence Risk Appraisal Guide (VRAG)—a risk assessment instrument for violent offenders: Review and authorized German translation]. *Fortschritte der Neurologie-Psychiatrie, 77*, 577-584.

Seto, M. C., & Barbaree, H. E. (1999). Psychopathy, treatment behavior, and sex offender recidivism. *Journal of Interpersonal Violence, 14*, 1235-1248.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1-26.

Urbaniok, F. (2007). *FOTRES: Forensisches Operationalisiertes Therapie-Risiko-Evaluations-System* [FOTRES: Forensic Operationalized Therapy/Risk Evaluation-System]. Bern, Switzerland: Zytglogge.

Urbaniok, F. (2009). FOTRES 2.0. Available from http://www.fotres.ch/

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR-20: Assessing risk for violence (Version 2)* (Vol. 2). Burnaby, British Columbia, Canada: Simon Fraser University.

www.zurichforensic.org (2009). http://www.fotres.ch/index.cfm?&content=9010&CFID=23226710&CFToken=9b033566fe31bb9d-0012BC03-13D4-FEFB-5EA619B6386F6AD7