

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

Klaus-Peter Dahle *

Charité Universitätsmedizin Berlin, Institute of Forensic Psychiatry, Limonenstr. 27, 12203 Berlin Germany

Received 5 February 2006; received in revised form 13 March 2006; accepted 15 March 2006

Abstract

Unlike many other countries, for many years, Germany disregarded structured instruments for assessing the risk of criminal reoffence. However, this negative attitude now seems to be gradually changing. An increasing number of contributions regarding structured instruments have been published in the German literature in the last years, and some instruments have already found their way into practice. However, studies that systematically examine the applicability of the mostly Anglo-American instruments to German criminals are still lacking. Therefore, the major objective of the current study was to test some internationally established procedures in a larger unselected sample from the German penal system. The following were included in the study: the Level of Service Inventory – Revised (LSI-R), the HCR-20 Scheme, and the Psychopathy Checklist – Revised (PCL-R). On the whole, the instruments proved to be applicable to German criminals with only a few adaptations to the German situation, and they achieved a predictive accuracy comparable to the values reported internationally. However, there were only minor differences in the predictive performance between the measures. Moreover, some basic limitations became apparent. Firstly, we found quite high percentages of criminals with medium scores and a correspondingly ambiguous prognosis. Furthermore, the predictive accuracy seemed to be dependent on demographic, criminological and psychopathological characteristics of the offenders. Finally, the instruments appeared to only partially utilize the empirical store of knowledge available regarding factors influencing the recidivism of criminals, since even a simple predictive model that only added a few further aspects besides the tested instruments (e. g. treatment yes or no) achieved systematically better predictions than the instruments alone. Altogether, the tested measures turned out to be useful instruments for risk assessments and may be conducive for a more systemized practice. However, due to the limitations inherent, they should be seen as a complement to a careful and clinically informed appraisal and not a substitute.

© 2006 Elsevier Inc. All rights reserved.

German penal law stipulates that prognostic assessments of offenders must be considered in many judicial decisions. In court, they influence the sentence severity for a criminal offence. Later, during a prison sentence, they help to determine whether an open or secure closed detention is given and affect mitigations of punishment. A favorable prognosis of future legal behavior is the prerequisite for terminating life prison sentences or granting remission for limited sentences. With the so-called “Maßregeln der Besserung und Sicherung” (correctional measure for improvement and security), German penal law also provides special measures for mentally ill or particularly dangerous offenders that are based exclusively on

* Tel.: +49 30 48851411; fax: +49 30 84451440.

E-mail address: klaus-peter.dahle@charite.de.

expectations regarding the future behavior of the persons involved. These measures can mean lifelong imprisonment, and the prerequisite for its termination is a fundamentally revised forecast of future legal behavior.

For decisions with severe consequences, the German penal code and jurisdiction require that psychological or psychiatric experts be consulted to support the judge in the prognostic appraisal of the case. Although the judge makes the ultimate decision, it remains the responsibility of these experts to give the judge's decision a scientific grounding in its prognostic aspects. However, no methodological specifications are made for this expertise; the experts are free in their choice of assessment methods. This has given rise to criticism, since the methodological diversity increases inconsistent legal practice. Indeed, unguided clinical appraisals of recidivism probability seem to be very frequent in this field, and the exact procedure used remains unclear and has barely been researched. Moreover, the quality of these assessments has not yet been systematically studied.

In the international literature on methods of criminal prediction, many authors regard the use of actuarial measures as the method of choice, and quite a few advocate the complete replacement of clinical assessments with such procedures (e.g., Quinsey, Harris, Rice, & Cormier, 1998). Particularly in English-speaking countries, numerous instruments have been developed, which have been tested in manifold studies substantiating their quality. Meta-analytic evaluations of predictive validity are available for some of these instruments, most of which yielded a predictive performance ranging between $r = .30$ and $.40$ (e.g., Gendreau, Little, & Goggin, 1996). Some of the instruments are purely statistical procedures based on a small number of static and easy to handle variables (e.g. the *Offender Group Recidivism Scale* of Copas & Marshall, 1998), while others include both static and dynamic recidivism predictors, which are much more difficult to handle. Several are based on explicit theoretical conceptions regarding the causes of criminality and recidivism (cf. Andrews & Bonta, 2003 or Palmer, 2001). A common factor of the latter instruments and those that are more complex is that they are assessment systems developed for clinically and diagnostically trained experts who are skilled in their use, and require the assessment of a fairly extensive amount of diagnostic information on the case. Some of the instruments were primarily designed as tools for guiding a systematic clinical assessment (e.g. the HCR-20 of Webster, Douglas, Eaves, & Hart, 1997), although there is also a great deal of empirical literature available that proves the predictive accuracy of their numerical scoring (for the HCR-20 see Douglas & Weir, 2003).

For some years now, the value of predictive scoring measures has also been increasingly discussed in German literature, and growing numbers of authors are demanding that psychological or psychiatric experts systematically consider such measures in risk assessments of criminal offenders (e.g. Endres, 2002). In fact, they are already used in connection with expert opinions, but examinations are lacking on the extent of their application. Above all, however, systematic examinations are lacking on the applicability of the mostly Anglo-American procedures to German conditions and their predictive power in German criminal populations.

Therefore, the major objective of this study was to gain fundamental information about the applicability of a selection of internationally established instruments in a sample of German prison inmates and to explore their predictive accuracy with this sample. If the instruments also proved to be useful, the second aim was to examine whether they are equally useful for all subjects examined or whether the subjects differ in their predictability and, if so, which ones show poor predictability. A final objective was to investigate whether the instruments exhaust the store of empirically confirmed knowledge regarding factors influencing the recidivism of criminal offenders or whether the prediction could be improved by considering factors not directly included in the measures.

1. Method

1.1. Subjects

The sample consisted of participants in a longitudinal study on the biographical development of criminal offenders. The study originally included 397 offenders imprisoned from February to May 1976 in former West Berlin; every fourth adult male who began serving a sentence during this period was selected. Since no selection was made regarding the type of crime (with the exception of pure traffic offense), the length of sentence, or the level of security, the sample represents a cross-section of adult male entrants of the prison system of West Berlin in the Spring of 1976. Participation in the study was voluntary, but participants were paid a fee of 10 DM and the basic examinations were scheduled early at the beginning of the sentence, when the prisoners had not yet been assigned to regular work or other routines. The refusal rate was therefore only 2%, although another 9% could not be included for technical reasons (e. g. release before the end of the examinations or an insufficient command of the German language). At the time of admission, the

participants were thoroughly examined, with the procedure comprising biographical, clinicopsychological and criminological interviews, psychological tests (including a general personality inventory, some clinical checklists, measures of intellectual performance and cognitive functioning, various attitude scales and a biographical inventory), medical and neurological examinations (including EEG examinations) and analysis of their files of the criminal inquiry. The age of the subjects at the time of admission ranged from 21 to 45 years ($M=29.83$; $SD=5.35$). The predominant offences were property offences (52%), followed by acts of violence (19%) and support violations (10%); the other offences were distributed in small percentages across different offence types. More than 75% of the subjects already had a criminal history, which included violent offences in 40% (further descriptions in Dahle, 2001).

The following analysis included 307 of the original 397 subjects, the inclusion criterion being survival for at least a ten-year observation period after release. The other 90 subjects did not survive this term. However, by 2004, a total of 126 participants had already died – causes of early death were most frequently alcohol and drug-related diseases and cancer, but also accidents and suicides were gravely above average compared to coeval German males in total.

1.2. Procedure

The prognoses were retrospectively established for the time when the prisoners were released from the prison sentence that they began in 1976. The assessments were carried out by criminologically trained psychologists with experience in the appraisal of offenders as well as in rating risk measures (including the assessed measures) who were blind to the recidivism of the subjects. The foundation of the analysis comprised the data recorded at the basic examination, supplemented by information from the prisoners' personal files on the course of imprisonment and the social situation at the time of release. Recidivism data were gathered from the criminal record of the time after release.

1.3. Measures

The risk assessments included the Level of Service Inventory – Revised (LSI-R, Andrews & Bonta, 1995), the HCR-20 Risk Assessment Scheme (HCR-20 Version 2, Webster et al., 1997) and the Psychopathy Checklist – Revised (PCL-R, Hare, 1991). The LSI-R is an instrument for risk/needs assessment with 54 items related to ten different risk areas. It was selected as a measure of general risk of reoffence as it appeared in multiple international comparative studies as one of the best predictors (e.g. Gendreau et al., 1996). Coding was largely carried out according to the English manual. Only the items on the school education level (items 15 and 16) had to be adapted to the German school system, and the question regarding the neighborhood crime level (item 29) was coded according to the criminal statistics of Berlin residential districts from the year 1976. The HCR-20 was selected as an instrument for estimating the risk of violence. It was originally developed for forensic psychiatric patients, but a number of studies support its suitability for prison populations as well (cf. Douglas & Weir, 2003). It includes three scales: The Historical (*H*) Scale comprises ten largely static variables from the previous history, while the Clinical (*C*) Scale has five items on the current mental, emotional and psychiatric status, and the Risk Management (*R*) Scale also has five items with appraisals of potentially destabilizing future living conditions. It should be pointed out that the *H* items were largely rated according to the basic examinations at the beginning of the sentence, while the ratings of the *C* and *R* items relied primarily on behavioral descriptions that were taken from the prison personal files. An adapted German version of the HCR-20 is available (Müller-Isberner, Jöckel, & Cabeza, 1998) and was used for this study. The PCL-R is a 20-item instrument for measuring the construct of a psychopathic personality, as originally described by Cleckley (1941). It is therefore not an instrument for criminal predictions per se, but psychopathy has often been shown to be predictive for persistent delinquency and future violence (e. g., Hare, 1999).

It should be noted that the selected instruments are rather designed for risk/needs and structured clinical assessments, and therefore do not constitute actuarial measures for risk of reoffence in a literal sense. However, they were treated here in an actuarial way, since there is a great deal of international empirical evidence regarding the predictive validity of their scoring and the primary objective of this study was to find out whether the instruments manage to achieve a comparable predictive performance with German offenders.

1.4. Overview of analyses

Reliability analyses were based on homogeneity estimations (Cronbach's alpha) and the determination of interrater agreement between two independent raters (intraclass correlations) for a random subsample ($n=30$), while

validity testing was based on correlations with different recidivism events. Validity comparisons were made on the basis of AUC values from ROC analyses, using the ROCKit program version 0.9.1 (Metz, 1998) for dependent samples (comparisons of the instruments) and some χ^2 techniques according to McClish (1992) for independent samples (comparison of different subgroups). If the predictive accuracy of the instruments did not differ significantly, the subgroup comparisons were made with weighted common AUC values of the three instruments (according to formulae 1 and 2¹ in McClish, 1992) to avoid redundant tests. Cut-off values for classifying the subjects into appropriate risk groups were determined using the CHAID algorithm². Subjects with inaccurate prediction by the instruments were identified using a procedure introduced by Ghiselli (1960) as the “prediction of predictability”. In a partial random sample ($n=200$), the squared residual values of logistic regressions (with the scores from LSI-R, HCR-20 and PCL-R as predictors and the recidivism behavior as the criterion) were first used to form a scale, measuring the inaccuracy of the prediction. A further step-by-step regression analysis then served to identify independent characteristics that give a proper estimation of this unreliability measure. For this purpose, selected criminological and demographic variables, some behavior variables from the term of imprisonment, and the subscales of a general personality inventory (FPI³, Fahrenberg, Selg, & Hempel, 1970), which was applied in the basic assessment of the subjects, were included. A new scale was created from the significant variables and cross-validated in the remaining subjects ($n=107$).

The following procedure was performed to test whether the instruments make optimal use of empirical knowledge regarding recidivism and factors influencing it. Firstly, the base rate of future recidivism was estimated for each subject using an algorithm developed by Beck and Shipley (1997). This algorithm is the result of a logit analysis of a few simple variables to estimate the risk of being rearrested after release from a prison term within a period of three years, which the authors performed in a large recidivism study of more than 16,000 released prisoners in the USA. It includes the following predictor variables: (1.) age, (2.) the number of adulthood arrests, (3.) the type of initial offence, (4.) earlier escapes or probation failures, and (5.) precustody experience. The inverse natural logarithm of the weighted (with the logit coefficients) sum of these factors enables a direct estimation of recidivism probability. The algorithm appeared to be fairly effective for our German subjects as well, since their estimated probabilities did not differ significantly from their de facto recidivism rates (see Fig. 1). Therefore, for the purpose of this study, it was assumed that the result of this algorithm represents the best estimation of the recidivism probability, given the condition that the subject is a perfect “statistically average case”. In further steps, each subject was then examined to test the hypothesis whether the case involved was really a statistically average case or not. The first question was the extent of personal risk factors as measured by risk instruments. Here, only the LSI-R was included, as the authors of the LSI-R offer direct estimations of effect sizes of low and high scores on recidivism rates in their manual (Andrews & Bonta, 1995). According to these norms, for a below-average value of the LSI total score with a cut-off at 13 points, 35% of the basic expectation was subtracted, while 15% was added for an above average parameter value (cut-off at 34). All other cases were treated as “average”; the average hypothesis was therefore not rejected and the subjects retained their primal estimation score. The next question was the possible membership of a well-known high-risk group. A PCL-R score of 20 points or above was evaluated or, for sexual offenders, the diagnosis of an ingrained sexual deviance, as these factors are well described high-risk factors. With reference to effect sizes found in the meta-analyses of Gendreau et al. (1996) and Hanson and Morton-Bourgon (2004), a value of +25% was assumed as the effect size on recidivism probability in the presence of membership of one of these high-risk groups⁴. If no high-risk factor was present, the next step was to examine the question of a low-risk constellation. This evaluation was made for a single aggression offence in a typical partnership conflict constellation or for support violations, if the perpetrators had no previous convictions or additional offences. Since these constellations are criminologically described, but no suitable study has been found to estimate adequate effect values, the effect was assumed as being –25% according to the effect of high-risk membership. The final question covered possible treatment effects during imprisonment. Since we were dealing here with treatments from the 1970s,

¹ $AUC_{\text{common}} = \sum W_i AUC_i / \sum W_i$ and $\text{Var}(AUC_{\text{common}}) = 1 / \sum W_i$ with $W_i = 1 / \text{Var}(AUC_i)$.

² CHAID (“chi-squared automatic interaction detector”) is an algorithm for generating classification trees to predict a variable on the basis of one or more predictor variables and to identify appropriate cut-off scores (for details see e.g. Hartigan, 1975).

³ *Freiburg Personality Inventory*: A well established general personality inventory in Germany, which is based on a factor analytic approach and frequently used in assessments of prisoners (e.g. Steller, 1983).

⁴ For an interpretation of correlative effect measures as a binomial effect measure, cf. Rosenthal and Rubin (1982) as well as Falk and Well (1997).

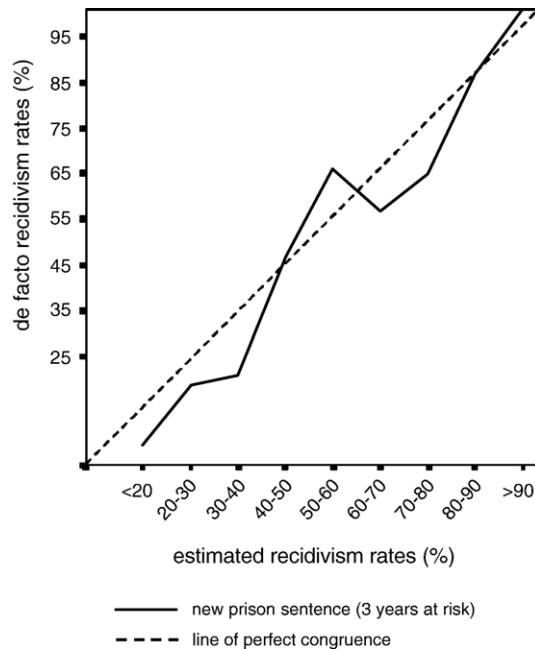


Fig. 1. Estimated recidivism probability using an algorithm by Beck and Shipley (1997) and de facto recidivism rates.

they were regarded as “unspecific treatment” according to Andrews et al. (1990). Thus, an effect size of -10% reduction of recidivism probability was assumed, as this was found as a mean effect size in various meta-analyses (e.g., Lösel, 1995). The prerequisite was regular completion of therapy and non-membership of high- or low-risk groups. In the case of premature termination for disciplinary or motivational reasons, a risk-increasing effect of $+10\%$ was assumed, as found in various studies (cf. Lösel, 1995). Using the described procedures, the integrative estimation of recidivism probability determined in this way was compared with the estimation of the instruments in order to examine whether there was a substantial gain in predictive accuracy. Fig. 2 provides a comprehensive overview of the procedure.

2. Results

2.1. Descriptive analyses

The mean values of the total score of the actuarial instruments was $M=24.65$ ($SD=7.35$) for the LSI-R, $M=16.52$ ($SD=6.31$) for the HCR-20 and $M=12.03$ ($SD=4.70$) for the PCL-R. The distribution did not deviate significantly from a normal distribution for the HCR-20 (Kolmogorov–Smirnov test with $z=0.99$, $p>.05$), while a slight positive skew was found for the LSI-R ($z=1.40$, $p<.05$) and a slight negative skew for the PCL-R ($z=1.57$, $p<.05$).

2.2. Reliability

The internal consistencies (Cronbach’s alpha) of the total score were $\alpha=.84$ for the LSI-R as well as for the HCR-20 scheme and $\alpha=.71$ for the PCL-R. Agreement among raters was $ICC=.93$ ($CI^{95\%}: .86-.97$) for the LSI-R (the subscales ranged between .51 and 1 with a mean of .75), $ICC=.91$ ($CI^{95\%}: .83-.96$) for the HCR-20 (.92 for *H* items, .82 for *C* items and .78 for *R* items), and $ICC=.94$ ($CI^{95\%}: .88-.97$) for the PCL-R.

2.3. Concurrent validity

The intercorrelations (Pearson) showed substantial congruence between the instruments. The coefficients were $r=.80$ for LSI-R and HCR-20, $r=.61$ for LSI-R and PCL-R, and $r=.76$ for HCR-20 and PCL-R.

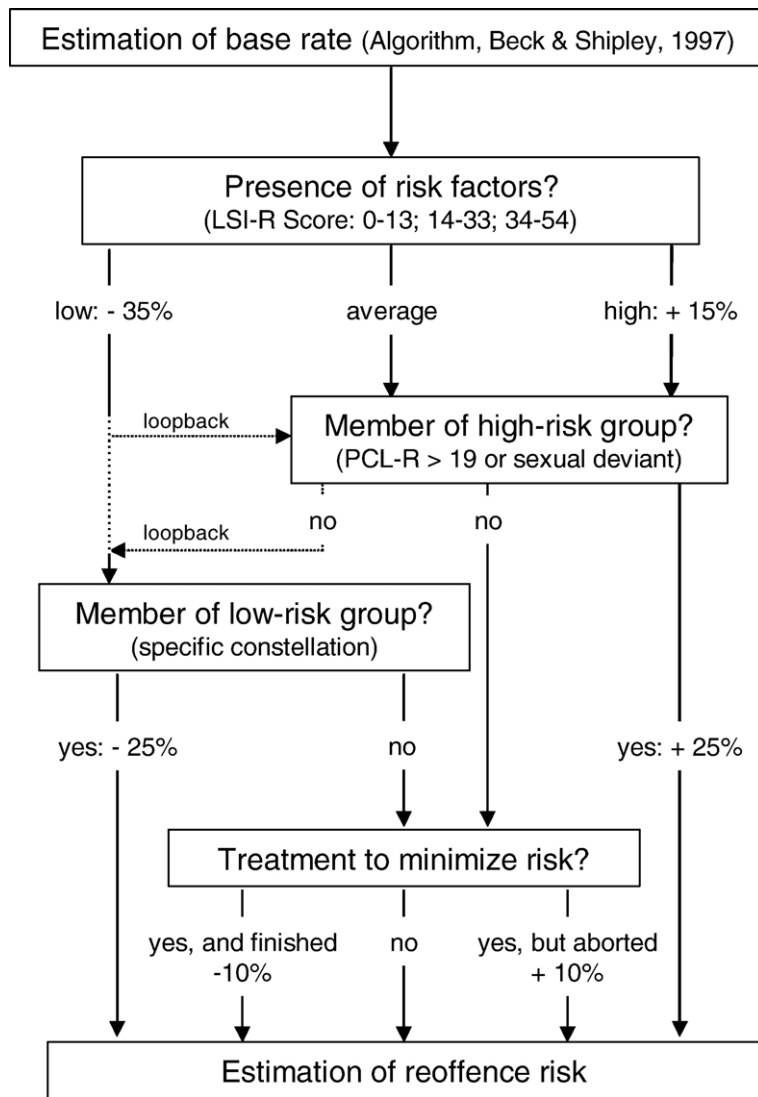


Fig. 2. Flow diagram on the integrative estimation of the recidivism risk.

2.4. Predictive validity

The survey in Table 1 correlates the total score of the instruments with various recidivism events for short (2 years at risk), medium (5 years at risk) and long (10 years at risk) observation periods. It is evident that – except with the PCL-R – the prediction of general recidivism events gradually decreases in accuracy as the observation period increases, whereas the prediction of violent recidivisms apparently improves as the length of time increases. HCR-20 and PCL-R appear to have advantages over the LSI-R for a long-term violence prediction, whereas short-term general recidivism prediction was best achieved with the LSI-R. However, most differences were not significant. Only for the 10-year violence prediction did a comparison of the AUC scores yield a significant superiority of the HCR-20 scheme over the LSI-R ($AUC_{LSI-R} = .65$ [SD = .03]; $AUC_{HCR} = .71$ [SD = .03]; $z = 2.17, p < .05$), but this was lost after adjustment of the significance level for three independent tests.

CHAID analyses were then performed to determine suitable cut-off scores for classifying subjects in appropriate risk levels. The criterion variable was the recidivism severity index from Table 1. It was evident for all three instruments that, with an observation period of 5 years or more, CHAID analysis only differentiated three risk groups as selective. For the 5-year period, the cut-off scores were 19 and 32 points for the LSI-R ($\chi^2(2, 307) = 42.64, p < .001$), 9 and 20 points for the

Table 1
 Predictive validity (correlations) of LSI-R, HCR-20 and PCL-R for various recidivism events and observation periods

	Reimprisonment	Violent crime	Recidivism severity ^a
2 years time at risk			
LSI-R	$r = .41$	$r = .15$	$\rho = .43$
HCR-20	$r = .37$	$r = .21$	$\rho = .42$
PCL-R	$r = .31$	$r = .14$	$\rho = .34$
5 years time at risk			
LSI-R	$r = .34$	$r = .21$	$\rho = .43$
HCR-20	$r = .34$	$r = .28$	$\rho = .42$
PCL-R	$r = .32$	$r = .25$	$\rho = .34$
10 years time at risk			
LSI-R	$r = .29$	$r = .23$	$\rho = .33$
HCR-20	$r = .31$	$r = .31$	$\rho = .40$
PCL-R	$r = .34$	$r = .32$	$\rho = .39$

^a Index with the grades 0 (no recidivism), 1 (only fine[s]), 2 (imprisonment[s] for up to 2 years), 3 (imprisonment[s] for more than 2 years) and 4 (severe violent crime).

HCR-20 scheme ($\chi^2(2, 307) = 62.16, p < .001$), and 10 and 16 points for the PCL-R ($\chi^2(2, 307) = 48.47, p < .001$). Table 2 provides an overview of the respective recidivism rates in the risk groups formed in this way.

It is evident that recidivism rates differ markedly between the risk levels. Particularly the low-risk group of the HCR-20 had only a very small percentage of subjects who became criminally active again, and only 10% of them had to serve another sentence within 5 years after release from prison. However, it is also evident that a large percentage of the subjects were classified in the particularly unspecific middle group and that their recidivism behavior hardly differed from the base rates in the total group. In this sense, for a great number of subjects, the predictive information gained was minimal. It is also apparent that the rates of false positive assessments were quite high for recidivism events with a low base rate (here: violent reoffences). In the high-risk groups, the corresponding recidivisms were markedly above the base rate, but the percentage of false positive assessments ranged between 70 and 80% for this recidivism criterion (cf. Table 2).

2.5. Predictable and unpredictable subjects

To examine the question of how the instruments handle different groups of subjects, the total group was first divided into different subgroups according to criminological, demographic and psychopathological variables, and comparisons of the predictive accuracy were made in each case (AUC values⁵; the criterion was another prison sentence within a 5-year observation period). Differentiation according to variables in the criminal record (age at first prison sentence, record of violent acts, variability and severity of predelinquency) did not reveal statistically significant differences in any case between the subgroups formed in this way. However, the level of the AUC values dropped as a whole. For example, they were $AUC = .64$ ($SD = .02$) in subjects with a record of violence versus $AUC = .67$ ($SD = .02$) in those without acts of violence ($\chi^2(1, 307) = 1.04, p > .05$); in relation to age at the first prison sentence, the values were $AUC = .62$ ($SD = .03$; serious before age 18), $AUC = .66$ ($SD = .03$; serious at age 18 to 21) and $AUC = .65$ ($SD = .02$; serious after age 21; $\chi^2(2, 307) = 1.01, p > .05$). For the undifferentiated total group, however, the corresponding values were $AUC = .70$ ($SD = .03$), and therefore higher than in any of the subgroups. It appears that part of the predictive power of the instruments is based on their ability to separate subjects with different severity degrees of criminal history.

Differentiation according to the age of subjects at the time of the prognosis, on the other hand, yielded a drop in predictive accuracy only for the 27–33-year-olds ($AUC = .63, SD = .02$) compared to the older ($AUC = .70, SD = .02$) and younger subjects ($AUC = .78, SD = .02$); the differences were highly significant ($\chi^2(2, 307) = 22.88, p < .001$). Differences were also found when subjects were grouped according to psychopathological variables, e.g. after addictive drug or alcohol abuse (without addiction: $AUC = .68$ [$SD = .03$]; with addiction: $AUC = .73$ [$SD = .02$], $\chi^2(1, 307) = 4.73, p < .05$) or various biographical variables (e.g., antisocial behavior in childhood with $AUC = .62$ [$SD = .04$] compared to unobtrusive behavior with $AUC = .72$ [$SD = .03$]; $\chi^2(1, 307) = 3.91, p < .05$).

⁵ Since the AUC values did not differ significantly between the instruments, the comparisons were carried out with weighted average areas according to the formulae from Footnote 1.

Table 2

Frequency of various recidivism events in different risk groups of LSI-R, HCR-20 and PCL-R (5 years time at risk)

	No recidivism	Reimprisonment	Violent crime
LSI-R score			
0–19 (<i>n</i> =71)	59%	23%	3%
20–32 (<i>n</i> =182)	35%	52%	15%
over 32 (<i>n</i> =54)	17%	74%	21%
HCR-20 score			
0–9 (<i>n</i> =39)	85%	10%	0%
10–20 (<i>n</i> =192)	36%	50%	9%
over 20 (<i>n</i> =76)	16%	68%	30%
PCL-R score			
0–10 (<i>n</i> =121)	55%	29%	5%
11–16 (<i>n</i> =140)	32%	54%	16%
over 16 (<i>n</i> =46)	4%	74%	28%
Base rate (<i>N</i> =307)	37%	49%	14%

The fact that the predictive accuracy was not entirely independent of target group characteristics raised the question as to the possibility of identifying subjects with better and worse predictability. For this purpose, a scale of the predictive accuracy was first formed in a partial random sample (*n*=200) according to the procedure suggested by Ghiselli (1960) from the squared residual values of a regression analysis with recidivism as the criterion and the score values of the instruments as predictors. Next, a search was made for characteristics associated with this scale. Testing was carried out to examine a number of characteristics including data from the *Freiburg Personality Inventory* (FPI), several variables from the criminal record, some biographic characteristics (early risk factors, age, professional development, marital status), and some data from the course of imprisonment (escapes, disciplinary incidents, treatment). Significant correlations with the quality scale were found, particularly for variables from the criminal record as well as for age, the aggression scales of the FPI, and several imprisonment variables. However, since these variables were correlated, a further regression analysis (backward stepwise method) was performed to identify variables with independent effects. This analysis yielded a residual set of five variables to be considered for the prediction of poor predictability (multiple correlation with $R=.37$): no disciplinary incidents in prison ($\beta=.21$), age between 27 and 33 ($\beta=.14$), no violent crimes in adolescence ($\beta=.14$), record of average offence variance (two to three offence variants; $\beta=.14$), and average prison experience (record of one to two prison sentences, $\beta=.13$). In short, prediction apparently proved to be difficult for middle-aged subjects with an unobtrusive (i.e. neither particularly serious nor wholly absent) criminal history and inconspicuous behavior in prison. Subdivision of the remaining cross-validation random sample (*n*=107) based on these characteristics (median split with 0 to 2 versus 3 to 5 characteristics) yielded an AUC=.79 (SD=.02) for the subjects with good predictability compared to AUC=.61 (SD=.02) for those with poor predictability ($\chi^2(1, 107)=55.24, p<.001$).

2.6. Saturation of empirical evidence

The integrative procedure described above (cf. Fig. 2) was performed to investigate the final question of whether the assessed instruments make optimal use of empirical knowledge regarding parameters influencing recidivism. The model led to a scale with a mean value of $M=45.07$ (SD=31.67) and right-skewed distribution characteristics (Kolmogorov–Smirnov's $z=1.76; p<.01$). In the general recidivism prediction, it achieved a markedly higher accuracy than the instruments alone (cf. Table 3). In contrast, the accuracy of violence prediction was only at about the level of HCR-20

Table 3

Predictive validity (correlations) of the integrative prediction model for various recidivism events and observation periods

	Reimprisonment	Violent crime	Recidivism severity ^a
2 years time at risk	$r=.48$	$r=.19$	rho=.53
5 years time at risk	$r=.49$	$r=.27$	rho=.53
10 years time at risk	$r=.44$	$r=.30$	rho=.51

^a Index with the grades 0 (no recidivism), 1 (only fine[s]), 2 (imprisonment[s] for up to 2 years), 3 (imprisonment[s] for more than 2 years) and 4 (severe violent crimes).

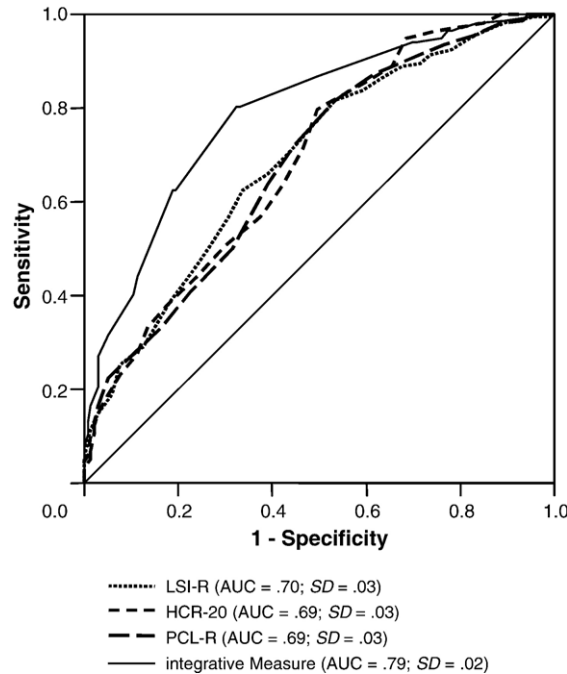


Fig. 3. ROC analysis of LSI-R, HCR-20, PCL-R and integrative recidivism risk estimation (Criterion: reimprisonment within 5 years after discharge).

and PCL-R, but the model’s underlying algorithm for base rate estimation according to Beck and Shipley (1997) was designed for general recidivism and not for estimating the rate of violent reoffence.

The AUC was .79 (SD=.03) for the recidivism criterion of another prison sentence within an observation period of five years and was thus markedly above the corresponding values of all three instruments (cf. Fig. 3). Tests of the differences with adjustment of the alpha level were highly significant (z between 4.02 and 4.08, $p < .001$). However, the integrative assessment of recidivism probability was also strongly dependent on the predictability of the subjects in terms of the previous section, even though it was on an altogether higher level. The AUC was an excellent .86 (SD=.03) for subjects with good predictability, but only .72 (SD=.04) for those with poor predictability and thus significantly worse ($\chi^2(1, 307)=7.84, p < .01$).

Finally, cut-off scores for classifying subjects into appropriate risk groups were also determined for the integrative recidivism probability scale with the aid of CHAID analyses. Again, the analysis yielded a classification into only three groups ($\chi^2(2, 307)=77, 88; p < .001$); the cut-off points were 35 and 70. Table 4 provides an overview of the recidivism rates of the resultant risk groups. Thus, the risk groups classified according to the integrative model also showed a good distribution of the recidivism rates. The high-risk group in particular had a very high recidivism risk, with a 92% rate of new prison sentences within 5 years after release. In contrast, the rates of the low-risk group were only at the level of the LSI-R and PCL-R; here, the HCR-20 scheme achieved much more conclusive results. It must be noted, however, that the integrative procedure classified about 40% of all subjects into the low-risk group, and this showed a recidivism that was altogether markedly below the general base rate. The middle group, on the other hand, already showed markedly increased recidivism tendencies. In this sense, middle field problems were also recognizable in the integrative procedure, but they were less pronounced compared to those encountered with the single instruments.

Table 4
Frequency of various recidivism events in the risk groups of the integrative prediction model (5 years time at risk)

Estimated recidivism rate	No recidivism	Reimprisonment	Violent crime
Up to 35 ($n=123$)	67%	26%	2%
36–70 ($n=133$)	23%	70%	18%
over 70 ($n=51$)	2%	92%	27%
Base rate ($N=307$)	37%	49%	14%

3. Discussion

To begin with the good news: The examination of some internationally well-established instruments for assessing the risk of criminal recidivism in an unselected sample of German convicts showed that the procedures appear to be readily applicable with only a small number of adjustments to German conditions. When applied by trained raters, they showed excellent interrater reliability and achieved a prediction accuracy that was comparable to international findings. For example, the meta-analyses of Gendreau et al. (1996) yielded a mean predictive validity of $M(r) = .35$ ($SD = .08$) for the LSI-R and $M(r) = .28$ ($SD = .09$) for the PCL-R, and Douglas and Weir (2003) found AUC scores between .66 and .78 in various studies of the predictive validity of the HCR-20 in correctional settings. Our findings are on a fairly comparable level.

However, since the basic study was originally conducted in the West Berlin of 1976, and therefore in a quite specific political situation that was most certainly different from the former West Germany as a whole and especially different from East Germany, the question arises of whether the findings can be generalized to other German prison populations. Furthermore, due to the long observation period, the mortality rate was quite high, and it is very likely that the resulting loss of follow-up was selective. However, there is another study relating to the predictive power of the HCR-20 and the PCL-R in a mixed (partial forensic psychiatric) German sample of offenders that has recently been published by Stadtland and Nedopil (2005). Although their ratings are based on data gathered retrospectively from assessments of criminal responsibility preliminary to conviction and a following sentence term, the authors report a predictive accuracy ranging from $AUC = .64$ to $.71$ for the HCR-20 and from $AUC = .64$ to $.72$ for the PCL-R for general and violent reoffences after release (time at risk ranged from 1 to 138 mo with a mean of 59). The scores were slightly lower than our findings, but the differences are very small and not significant. Thus, two independent studies with different methodology and samples but comparable results provide good reason to assume that the assessed instruments, at least the HCR-20 and the PLC-R, are transferable to German offender populations.

The bad news is that this study also revealed a number of limitations of the examined instruments, particularly if treated as actuarial measures. These included, first of all, a considerable middle field problem with large proportions of subjects who achieved average results near to the mean value with the instruments and whose recidivism behavior did not differ appreciably from the total base rate. This middle field group accounted in the study for around 2/3 of the total sample with the HCR-20 and comprised 45 to 60% with the other instruments. The predictive information obtainable through the instruments was therefore minimal for many of the subjects examined. This middle field problem is not often discussed in the literature, and unfortunately, detailed information about distribution characteristics with related recidivism rates is rarely published. However, this problem also appears to be evident elsewhere. For example, in the risk norms for LSI-R that are provided in the manual (Andrews & Bonta, 1995), about 40% of subjects from a total of five differentiated risk groups are assigned to the middle group with a “moderate risk” and their recidivism rate seems quite close to the base rate.

The second limitation that should be noted is that when applying recidivism criteria with a low base rate (in this study, for example, violent recidivisms within a 5-year observation period), the examined instruments yielded rather high rates of false positive classifications. This phenomenon is to be expected for reasons relating to decision-making theory (cf. Wiggins, 1973). However, it particularly affects prediction scores with relatively symmetrical distribution characteristics, since these characteristics do not fit well with rare events.

A further weakness of the instruments was found when differentiating criminal offenders in homogeneous subgroups. There was an overall loss of predictive accuracy for differentiations with respect to the type and scope of the criminal record. This indicates that part of the instruments' validity is apparently based on their differentiation ability with regard to the criminal history. Since the history is really a strong predictor of recidivism behavior, its consideration appears to be well grounded. However, the subtle methods of the instruments are probably not necessary to identify criminal offenders with a serious history; a glance at the criminal record is sufficient. Differentiations according to psychopathological, demographic and criminological variables revealed a certain fluctuation of predictive accuracy in relation to the specific characteristics of the subjects examined. More detailed analyses to identify subjects with poor predictability showed that these were particularly middle-aged low-profile offenders with an average criminal history and unobtrusive behavior in prison.

Finally, the study demonstrated that the examined instruments apparently do not yet optimally consider the currently available empirical knowledge regarding factors influencing recidivism. Even a simple additive model for integrating different stores of knowledge in the prognostic assessment process yielded a marked improvement of the predictive accuracy compared to the instruments alone. Here, there is surely some room for future methodological enhancement of

actuarial prediction, especially since the integrative procedure described seemed to reduce some of the outlined weaknesses of the assessed instruments.

These latter findings in particular need to be backed up with further empirical evidence and cross-validation. However, on the whole, they do seem to point to some limitations of scoring systems for criminal predictions and perhaps to some intrinsic limitations of the actuarial approach per se (cf. Ross & Pfäfflin, 2005). The value of predictive scoring systems is based on their pure rationality, their extensive empirical foundation and their ability to utilize empirical experience for predictive purposes. This seems sufficient reason to implement them systematically in prognostic assessment procedures, particularly as they performed fairly well in specific subgroups (the “predictables”). On the other hand, there are also limitations, and it is questionable whether an approach that is based only on scoring will be able to substitute a careful and clinically well informed risk assessment in the foreseeable future. However, since the assessed measures are innately not only scoring systems but also tools for risk/needs and guided clinical assessments, they can also be valuable in the context of comprehensive clinical surveys.

Acknowledgement

The presented study is based on a comprehensive survey that was initiated and conducted by the former director of the Berlin Institute of Forensic Psychiatry, Prof. Wilfried Rasch († 27.8.2000) in 1976. Prof. Rasch left the data to me for further analyses and follow-up studies. Special thanks also go to Dr. Katja Erdmann and Dipl.-Psych. Vera Schneider for their thorough ratings of the assessed instruments.

References

- Andrews, D. A., & Bonta, J. (1995). *LSI-R: The level of service inventory-revised*. Toronto, CA: Multi-Health Systems.
- Andrews, D. A., & Bonta, J. (2003). *The psychology of criminal conduct*, 3rd ed. Cincinnati, OH: Anderson.
- Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology*, 28, 369–404.
- Beck, A. J., & Shipley, B. E. (1997). *Recidivism of prisoners released in 1983*. Washington, DC: U.S. Department of Justice.
- Cleckley, H. (1941). *The mask of sanity*. St. Louis, MO: Mosby.
- Copas, J., & Marshall, P. (1998). The offender group reconviction scale: A statistical reconviction score for use by probation officers. *Applied Statistics*, 47, 159–171.
- Dahle, K.-P. (2001). Violent crime and offending trajectories in the course of life: An empirical life span developmental typology of criminal careers. In D. P. Farrington, C. R. Hollin, & M. McMurrin (Eds.), *Sex and violence: The psychology of crime and risk assessment* (pp. 197–209). London, UK: Routledge.
- Douglas, K. S., & Weir, J. (2003). *HCR-20 violence risk assessment scheme: Overview and annotated bibliography*. Retrieved from <http://www.violence-risk.com/hcr20annotated.pdf>
- Endres, J. (2002). Gutachten zur Gefährlichkeit von Strafgefangenen: Probleme und aktuelle Streitfragen der Kriminalprognose (Surveys of the dangerousness of convicts: Problems and current disputes of criminal prediction). *Praxis der Rechtspsychologie*, 12, 161–181.
- Fahrenberg, J., Selg, H., & Hempel, R. (1970). *Das Freiburger Persönlichkeitsinventar (The Freiburg Personality Inventory)*. Göttingen, GER: Hogrefe.
- Falk, R., & Well, A. D. (1997). Many faces of the correlation coefficient. *Journal of Statistics Education*, 5(3), 1–14.
- Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology*, 34, 575–607.
- Ghiselli, E. E. (1960). The prediction of predictability. *Educational and Psychological Measurement*, 20, 675–684.
- Hanson, R. K., & Morton-Bourgon, K. (2004). *Predictors of sexual recidivism: An updated meta-analysis*. Ottawa, CA: Public Works and Government Services.
- Hare, R. D. (1991). *The hare psychopathy checklist – Revised: Manual*. Toronto CA: Multi-Health Systems.
- Hare, R. D. (1999). Psychopathy as a risk factor for violence. *Psychiatric Quarterly*, 70, 181–197.
- Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.
- Lösel, F. (1995). The efficacy of correctional treatment: A review and synthesis of meta-evaluations. In J. McGuire (Ed.), *What works: Reducing reoffending* (pp. 79–111). Chichester, UK: Wiley.
- McClish, D. K. (1992). Combining and comparing area estimates across studies or strata. *Medical Decision Making*, 12, 274–279.
- Metz, C. E. (1998). *ROCKit 0.9.1 – IBM compatible ROCKIT user's guide*. Chicago: The University of Chicago/Department of Radiology.
- Müller-Isberner, R., Jöckel, D., & Cabeza, S. G. (1998). *Die Vorhersage von Gewalttaten mit dem HCR-20 (The prediction of violence with the HCR-20 scheme)*. Haina, GER: Institut für Forensische sychiatrie.
- Palmer, E. J. (2001). Risk assessment: Review of psychometric measures. In D. P. Farrington, C. R. Hollin, & M. McMurrin (Eds.), *Sex and violence: The psychology of crime and risk assessment* (pp. 7–22). London, UK: Routledge.
- Quinsey, V., Harris, G. T., Rice, M., & Cormier, C. (1998). *Violent offenders: Appraising and managing risk*. Washington, DC: American Psychological Association.

- Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74, 166–169.
- Ross, T., & Pfäfflin, F. (2005). Risk Assessment im Maßregelvollzug: Grenzen psychometrischer Gefährlichkeitsprognose im therapeutischen Umfeld (Risk assessment in forensic hospitals: Limits of actuarial prediction of violence in therapeutic settings). *Monatsschrift für Kriminologie und Strafrechtsreform*, 88, 1–11.
- Steller, M. (1983). Haftdauereinflüsse auf Selbstbeschreibungen von Delinquenten: Bezugsgruppeneffekte? (The influence of the duration of a prison term on self-descriptions of offenders: An effect of changing references?). *Zeitschrift für Experimentelle und Angewandte Psychologie*, 30, 474–499.
- Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR-20: Assessing risk of violence (version 2)*. Vancouver, CA: Mental Health Law and Policy, Institute Simon Fraser University.
- Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.