

The interrater reliability of violence risk assessment using HCR 20 Version 3 and SAPROF in a sample of Polish forensic psychiatric patients and prison inmates

Marlena Banasik¹, Agnieszka Welento-Nowacka²,
Agnieszka Wojtowicz³, Józef Krzysztof Gierowski⁴

¹ Faculty of Psychology in Krakow, SWPS University

² University of Gdansk

³ University of Physical Culture in Krakow

⁴ Faculty of Psychology in Katowice, SWPS University

Summary

The aim of the study was to validate the reliability of the Historical Clinical Risk Management-20 Version 3 (HCR-20 V3) and the Structured Assessment of Protective Factors (SAPROF) assessment tools for use in Polish forensic psychiatry and penitentiary settings. Despite their global popularity and Polish-language adaptations, these methods remain underutilized in Poland.

Data were collected with HCR-20 V3 and SAPROF from 401 male subjects, including forensic psychiatric patients and inmates, between 2020 and 2023. Assessments were conducted twice, six months apart, and they were independently analyzed by three trained investigators to calculate interrater reliability using Krippendorff's alpha.

The results demonstrated very good interrater reliability for historical risk assessments across groups. Reliability was slightly lower for clinical and dynamic risk factors, reflecting challenges in evaluating factors without direct contact with subjects. For inmates, limited psychological support and systemic post-release aid influenced the reliability of risk management evaluations. In contrast, external protection assessments were more reliable in inmates than forensic psychiatric patients, whose outcomes depended on uncertain post-release supervision and therapy availability.

In the study, the importance was highlighted of integrating protective factors into risk assessments and it was confirmed that HCR-20 V3 and SAPROF produce reliable results when used by trained professionals. These findings address concerns among Polish practitioners regarding potential biases in SPJ tools. High interrater reliability of final risk and protection assessments reinforces the potential for broader adoption of these tools in Poland.

Key words: protective factors, violence risk factors, interrater reliability

Introduction

The HCR-20 V3 is one of the most widely used tools in clinical and forensic practice for assessing and managing the risk of violence. In line with the Structured Professional Judgment (SPJ) stream, it was created, among others, to ensure that the risk assessment process is structured, to provide it with an empirical basis as well as to streamline and standardize the communication of its results on case prioritization. It is designed to enable the monitoring of unit risk concerning violence over time by taking risk factors of a dynamic nature into account. The HCR-20 V3 is suitable for short-term forecasting, over a period of six to 12 months, or whenever there is a significant change in context, e.g., acquisition of unsupervised passes, change in security level, etc.

Sing, Desmarais, Hurducas, Arbach-Lucioni, Condemarin et al. [1] empirically confirmed the prevalence of violence risk assessment in the daily practice of professionals, as well as the frequency of HCR-20 V3 use in this process, based on data collected from 2,135 respondents from 44 countries (located on six continents). The surveyed professionals reported using the HCR-20 V3 in more than half of the risk assessments developed in the past 12 months. The scale was used by psychologists more often than by psychiatrists or nurses.

Increasingly, as recommended by the scale's authors, the HCR-20 V3 is being used in conjunction with SAPROF, allowing the researcher's perspective to be broadened by including assessment of protective factors. This makes it possible to develop an integrated risk assessment approach, as well as an asset-focused plan for therapeutic/repair interventions. The psychometric properties of the HCR-20 V3 and SAPROF have been studied for different populations and contexts [2]. In the remainder of this article, evidence will be presented of at least good concordance of the judges' assessments gathered from studies involving correctional, forensic psychiatric, and mixed populations. The literature was also analyzed with regard to cases of HCR-20 V3 use in judicial assessments since publication in 2013. De Vogel et al. [2] analyzed 546 cases from Australia, Canada, Ireland, the Netherlands, New Zealand, the United Kingdom, and the U.S., proving that the results of assessments formulated based on the HCR-20 V3 were extremely rarely challenged (4.03%) by the court.

Meanwhile, Challinor, Ogundalu, McIntyre, Bramwell, and Nathan [3], looked for evidence of its clinical utility in publications on the HCR-20 V3. They noticed that published empirical analyses ($n = 21$) were most often (86%; $n = 18$) focused on determining the "presence of risk factors" (step 2 of the assessment procedure). They did not find any analyses dedicated to two assessment steps: "scenario planning" and "management". Frequently – in 64% of the publications, the predictive accuracy of the tool was evaluated. From this, the authors conclude that there is limited empirical evidence to support its effectiveness as a structured professional approach to violence risk assessment and management [3].

Despite the above-described worldwide popularity of the methods and the availability of their Polish language adaptations, they are still not widely used in the conditions of the Polish forensic psychiatry system or in penitentiary settings. The purpose of the undertaken research was to provide Polish professionals with data confirming

the reliability of measurements made using HCR-20 V3 and SAPROF. These data are unique in Polish conditions. They provide answers to doubts about interrater reliability in terms of summary risk assessments (SRRs), which are not a simple sum of factor assessments, but require analysis of the presence, relevance and interrelationships between perceived risk factors.

User qualifications

The multi-step nature of the risk assessment procedure and the qualitative nature of the final risk assessment may suggest to possible users of HCR-20 V3 and SAPROF the need for extensive clinical experience to realize their full potential. The tools' authors dispel doubts in this regard on the pages of the HCR-20 V3 user manual. In their view, evaluators should meet three basic criteria: (a) familiarity with the professional and scientific literature on the nature, causes, and management of violence; (b) experience in individual assessment, such as conducting interviews, reviewing documentation and, although not necessary but helpful, understanding how to administer and interpret standardized tests; and (c) have experience in assessing and diagnosing not only mental illnesses, but also personality and substance-related disorders [4]. Users who do not meet the last criterion should use the HCR-20 V3 after consultation or with the supervision of a person experienced in the diagnosis of mental disorders. Thus, they can, for example, be part of a team making a common diagnosis, or they can base their risk assessment on previously established diagnoses. The developers of the HCR-20 V3 recommend that potential users undergo training under the guidance of a qualified trainer. Use of the scale in a legal context also requires knowledge of the laws and procedures related to risk assessment. The criteria cited above do not imply the need for an advanced medical or scientific degree or a specific professional position. The HCR-20 V3 can be and is used not only by psychologists or psychiatrists, but also by nurses, law enforcement officers, social workers, and other such professionals [5]. Nevertheless, according to de Vogel et al. [2], the level of proficiency among HCR-20 V3 users is crucial, especially in the context of formulating a forensic psychiatric opinion. They recognize that an evaluation using the HCR-20 V3 can only constitute expert evidence if it represents a study prepared by trained and experienced evaluators in an appropriate legal or clinical context. Evaluators should be aware of the risk of bias in evaluations and document in opinions as to what strategies have been implemented to minimize this bias [2]. De Vogel et al. [2] also emphasized that evaluators should always transparently describe the strengths and limitations of their findings.

Possible sources and causes of bias in raters

De Vogel et al. [2] noted that the majority of studies published to date on bias in risk assessment have been based on actuarial tools (ARAI), while few studies have been devoted to the issue of assessments using the HCR-20. Essentially, SPJ tools were developed to avoid certain shortcomings of actuarial tools; in particular, their lack of flexibility, ignoring context, and/or the unique characteristics of each individual.

The objective of tool developers in the SPJ stream was to create conditions for more individualized risk assessment. The consequence of this approach, based on a kind of clinical freedom, is that it becomes more susceptible to potential assessor bias. Evaluator bias refers to the unique way in which the evaluator may perceive the subject, which can lead to disagreement among evaluators [6]. Assessments formulated with the HCR-20 V3 should be seen as dynamic and strongly dependent on (future) context. As a consequence, the scale is more applicable for short-term violence risk assessment and parole decisions than for long-term prognosis [2].

Murray et al. [7] pointed out the possibility of attribution bias in the assessment process using the HCR-20 in their study. The students they studied ($n = 40$) formed different HCR-20 ratings, depending on whether they attributed internal (i.e., individual characteristics) or external (e.g., due to circumstances) attribution to the behavior of those being assessed. Historical and clinical risk ratings were higher when internal attribution was triggered compared to external attribution. In contrast, Kamorowski and de Ruiter et al. [2] showed that raters with a more negative attitude towards offenders rated clinical risk and severity of risk management problems higher, and formulated higher ratings of overall HCR-20 V3 risk than those with a more positive attitude towards offenders.

Venner et al. [6] also indicated two other sources of variance in risk-coding professionals' assessments using the HCR-20 V3. These are adversarial allegiance and professional supremacy. The term "adversarial allegiance" describes the phenomenon of changing risk assessments as a consequence of the fact that the assessor and the assessed represent opposing parties in legal proceedings. The phenomenon of "professional supremacy", on the other hand, directs attention towards the bias that arises as a result of cross-cultural differences between the evaluator and the evaluated. There is no denying that there is a lack of data on the consistency of assessments made in different cultures or on the cultural bias revealed in assessments regarding the risk of violence. It is also no secret that information about an offender's cultural background can be relevant, if only when deciding on his/her conditional release from an institution. Certain cultural groups are stereotypically considered more likely to experience greater socioeconomic disadvantage, which can translate into higher scores on historical items and dynamic factors such as education/employment and family/marriage issues. Whiteness and "westernness" invariably remain the standard to which we relate the behavior of non-white and non-western perpetrators. Both types of described biases appear to affect the reliability of risk assessments and require further research so that the tools of the SPJ stream can be safely applied across cultural groups.

Results of empirical analyses in the area of conformity of raters

Interrater reliability is considered the most important indicator of HCR-20 V3 reliability and that of other scales developed in the Structured Professional Judgement (SPJ) trend. Despite the long – lasting almost decades – history of using tools from the SPJ strand around the world, empirical analyses aimed at assessing the compatibility of judges can safely be considered few.

Prior to the release of the latest, third version of the HCR-20, the psychometric properties of the revised tool were evaluated. The evaluation process included eight samples (five forensic-psychiatric, one general-psychiatric, two prison samples) from six countries (Canada, the Netherlands, Norway, Sweden, the UK – England and Wales, and the United States), giving a total population of 782 subjects. Data on the concordance of assessments were provided by Douglas and Belfrage [8], Strub, Douglas, Nicholls [9], de Vries Robbé, de Vogel, and Douglas [10], and Doyle et al. [11]. These authors generally showed excellent/nearly excellent interrater reliability in terms of the sum of assessments concerning the presence of individual risk factors (ICC values were between 0.92 and 0.94). Interrater reliability in terms of historical risk severity was shown to range from ‘good/significant’ to ‘excellent/excellent’ (ICC between 0.72 and 0.94). Interrater reliability in clinical risk assessment also ranged from ‘good/significant’ to ‘excellent/excellent’ (ICC between 0.69 and 0.90). A similar consistency indicator was shown in assessing the severity of risk management problems (ICC between 0.69 and 0.93). Since SPJ instruments do not provide a numerical score, compliance indicators for summary risk ratings (SRR) are often provided in screening reports. The SRR reflects the final risk classification determined by the rater, taking the presence and clinical relevance of perceived risk factors into account. Summative risk ratings form the basis of intervention decisions, thus compliance indicators in this regard are assigned particular significance. In the HCR-20 V3 validation process, it was determined that interrater reliability for SRR could be considered ‘moderate’ to ‘high’ (Cohen’s κ between 0.72 and 0.81) [8, 4, 11].

Douglas and Shaffer [5] conducted a literature review looking for data on the interrater reliability of evaluations performed by using the HCR-20 V3. They were able to find 17 published reports. The underlying data were collected in different populations – psychiatric inpatients, forensic psychiatric inpatients, and correctional groups. The published results of analyses were obtained based on groups of various sizes, ranging from 10 subjects [12] to as many as 409 [11]. The assessment process involved raters with different levels of experience (including professionally-trained students) and in different numbers (from two to five). A review of available reports allows conclusions to be drawn about at least acceptable (‘moderate’) interrater reliability. Most of the published ICC rates were between 0.70 and 0.80, indicating ‘high’/‘very high’ concordance [2]. Selected examples of the results of empirical analyses on the interrater reliability obtained in different populations are presented in Table 1.

Table 1. Selected results of empirical analyses on interrater reliability conducted in different populations

Authors	Subjects	Number/ characteristic of raters	Interrater reliability indicators for historical risk assessment, clinical risk assessment, and risk management	Interrater reliability indicators for summary risk ratings (SRR)
Forensic population				
Smith et al. [13]	N = 15 (USA)	N = 2 (after professional training)	HCR-H ICC ₁ = 0.92 HCR-C ICC ₁ = 0.67 HCR-R In (intramural context) ICC ₁ = 0.68 R Scale Out (extramural context) ICC ₁ = 0.88	$\kappa = 0.66^{**}$
Sea, Hart, Douglas [14]	N = 80 (Korea)	N = 3	HCR-H ICC ₁ = 0.73; ICC ₂ = 0.91 HCR-C ICC ₁ = 0.89; ICC ₂ = 0.97 HCR-R ICC ₁ = 0.86; ICC ₂ = 0.96	ICC ₁ = 0.83 ICC ₂ = 0.97
Forensic-psychiatric population				
Kötter et al. [15]	N = 30 (Germany)	N = 5 (after professional training, without experience)	HCR-H ICC ₁ = 0.65 HCR-C ICC ₁ = 0.66 HCR-R-Out (extramural context) ICC ₁ = 0.73	ICC ₁ = 0.86
Doyle et al. [11]	N = 409 (UK)	N = 4	HCR-H ICC**** = 0.91, C Scale ICC = 0.90, HCR-R-Out (extramural context) ICC = 0.93	N/A
Haag, Hogan & Cheng [16]	N = 32 (Canada)	N = 4	N/A	SRR Case Prioritization ICC ₁ = 0.88
Cabeldue et al. [17]	N = 74 (USA)	N = 2 (after professional training)	HCR-H ICC ₁ = 0.87; HCR – C ICC ₁ = 0.88; HCR – R ICC ₁ = 0.87.	N/A
Cook et al. [18]	N = 11 (Canada)	N = 5 (after professional training)	N/A	SRR Case Prioritization ICC ₁ = 0.80

table continued on the next page

Mixed population (forensic-psychiatric + forensic)				
Persson et al. [12]	N = 10 (Sweden)	N = 2	N/A	ICC ₁ = 0.80
Chen et al. [19]	N = 33 (China)	N = 3	HCR-H ICC ₁ = 0.90; HCR-C ICC ₁ = 0.63; HCR-R-Out ICC ₁ = 0.69. HCR-H ICC ₂ = 0.96; HCR-C ICC ₂ = 0.83; HCR-R ICC ₂ = 0.87.	ICC ₁ = 0.90 ICC ₂ = 0.96

* The ICC is a value between 0 and 1, where values below 0.5 indicate 'poor reliability', between 0.5 and 0.75 'moderate reliability', between 0.75 and 0.9 'good reliability', and any values above 0.9 indicate 'excellent reliability'

** Cohen's Kappa

*** Six combined SRRs: Future Violence-Institution, Serious Physical Harm-Institution, Imminent Violence-Institution, Future Violence-Community, Serious Physical Harm-Community, and Imminent Violence-Community.

****Unspecified type

Since the publication of Douglas and Shaffer's review [5], at least two new studies have appeared in the literature on the topic of assessment concordance regarding SPJ stream tools. Wolf, Mayer, Steiner, Franke, Klein, Streb, and Dudeck [20] conducted a compliance assessment of judges in a group of 452 female patients with substance use disorders from German forensic psychiatry departments. Risk assessments were performed by five investigators in parallel. 'Moderate' concordance was found for risk assessments using the HCR-20 V3 (ICC = 0.606, 95%CI = 0.345; 0.845) and 'good' concordance for assessments formulated based on the FAM (ICC = 0.818, 95%CI = 0.638; 0.938). On the other hand, Sideli, Carpentieri, Capri, Bracalenti, Cavallo, Schimmenti, and Caretti [21] undertook validation of the Italian adaptation of Historical-Clinical-Risk Management-20 Version 3 in a group of adolescents – wards of correctional institutions (the youth justice system). Interrater reliability was assessed in a selected subsample of 10 participants (out of $n = 61$). 'High' to 'near-perfect' concordance (Cohen's κ ranged from 0.69 to 1.00) of item presence scores was observed.

Interrater reliability analyses on the presence of protective factors and the severity of protection are published as infrequently as the reports on the HCR-20 V3 described above. Burghart, de Ruiter, Hynes, Krishnan, Levtova, and Uyar [22] conducted a meta-analysis of validation studies regarding the SAPROF scale, considering all versions published to date: "classic", used for assessment in adolescents (SAPROF-YV) and in sex offenders (SAPROF-SO). Eighteen of the pool of 39 publications analyzed by the authors provided data on the interrater reliability of the basic version of the SAPROF. In the majority of them, the interclass correlation (ICC) index was used. The average ICC index for the SAPROF was 0.80 (IQR = 0.74–0.86), which can be considered

evidence of ‘good’ agreement among the referees. In two articles, the Spearman’s rho coefficient ($\rho = 0.83$) [23] and Pearson’s r ($r = 0.86$) [24] were presented for the total SAPROF score (sum of factor scores). Similarly, they demonstrate ‘good’ to ‘excellent’ concordance. Concordance was also assessed for the SAPROF FPJ score (ICC = 0.72; IQR = 0.66–0.77), as well as for all SAPROF subscales. The average ICC coefficients for these ranged from ‘moderate’ to ‘good’ (SAPROF Internal: ICC = 0.7; IQR = 0.74–0.86; SAPROF Motivational: ICC = 0.79; IQR = 0.71–0.84; SAPROF External: ICC = 0.69; IQR = 0.58–0.76) and were, for the most part, slightly lower than the ICC averages for total scores.

Materials and methods

The data used in the analysis were obtained between 2020 and 2023 from 401 men in institutional settings. These participants comprised 201 patients hospitalized in forensic psychiatric wards and 200 individuals serving time in prison. The mean age of the subjects was 37.16 years ($SD = 10.87$). A similar mean age was observed in the subgroup of inmates (PI; $M = 37.05$; $SD = 10.90$) and in the subgroup of patients from forensic psychiatric wards (FP; $M = 37.29$; $SD = 10.87$). The criterion for selection of the study group was male gender, the fact of serving a prison sentence of at least six months (in the group of inmates) or being in a forensic psychiatry ward (in the group of patients), and consent to participate in the study. The data essential for statistical analysis was acquired through the utilization of the following methods: HCR-20 V3 and SAPROF.

HCR-20v3

The Historical Clinical Risk Management-20 (HCR-20) represents a widely utilized, structured professional judgment (SPJ) tool within the field of violence risk assessment for adults. The present iteration, HCR-20 V3, is currently in use [4]. Characterized as a semi-structured and comprehensive risk assessment instrument, the HCR-20 V3 entails an in-depth interview with the subject, integrating information from diverse sources such as medical records, informant interviews and forensic psychiatric reports. Comprising three scales – historical, clinical, and management of future violence risk – the HCR-20 V3 encompasses a total of 20 risk factors.

The assessment process involves a 7-step procedure encompassing information collection, determining the presence and relevance of risk factors, risk formulation, development of risk scenarios, devising a risk management plan, and concludes with a summary rating of risk. Reliability testing of the HCR-20 V3 was conducted through expert-rater methods involving the evaluation of 32 forensic patients. The results demonstrated high conformity of assessments across subscales (0.82–0.93) and satisfactory internal reliability (0.60–0.80), as assessed by Cronbach’s alpha [8]. The widespread deployment of the HCR-20 V3 spans across various settings, including determining the level of security required for the studied subject and setting therapeutic goals in case of psychiatric ward patients.

SAPROF

The Structured Assessment of Protective Factors (SAPROF) is a scale utilized for evaluating protective factors, and its second edition was translated to Polish in 2016 [25]. Often employed in conjunction with the HCR-20 V3, SAPROF is specifically designed to incorporate protective factors into the assessment of violence risk. Administration of the SAPROF involves a thorough interview, encompassing the analysis of medical documentation, forensic psychiatry reports and background surveys. Developed by experts at Van der Hoeven Kliniek in Utrecht, Pompe Kliniek in Nijmegen, and Maastricht University, it is applicable for assessing protective factors in individuals with a history of sexual and physical violence. Comprising three scales – internal, motivational, and external factors – with a total of 17 items, assessors are tasked with coding these items to identify key protective factors present in the patient and pinpointing areas requiring further therapeutic attention. While demonstrating good ‘internal’ reliability and validity, it is important to note that these characteristics specifically pertain to surveys conducted within the male population.

One year post-assessment, the validity coefficient of the scale stood at 0.85 for physical violence and 0.83 for sexual violence, decreasing slightly to 0.74 and 0.77, respectively, after three years. When combined with the HCR-20 V3, the validity coefficient was 0.87 after one year, decreasing to 0.76 after three years. The internal reliability of the SAPROF for physical violence was 0.88, and for sexual violence, it totaled 0.85 [25].

The procedure involved collecting information on subjects twice, and measuring risk (according to HCR-20 V3, HCR-SRR Case Prioritization) and protection (according to SAPROF guidelines, SAPROF FPJ, SAPROF final protection judgement) twice, six months apart. The collected data were analyzed independently by three investigators, which allowed each investigator to independently formulate a risk assessment (SRR Case Prioritization) and protection assessment for each subject. This provided material for analyzing the concordance of the judges’ assessments. The evaluators had no direct contact with the subjects, but only analyzed the material provided by the information collectors (step 1 of the SPJ procedure). The researchers made assessments taking the non-institutional context into account, considering the level of risk and protection if the subject left the institution where s/he currently resided. The researchers received professional training to prepare them to use the instruments.

Data were processed using JASP 0.18.2.0 and IBM SPSS Statistics with macro A.F. Hayes (2018 version) to calculate the Krippendorff’s alpha judges’ concordance coefficient (procedure described in: [26]). The level of historical and clinical risk, as well as risk management problems and the final risk score (overall violence risk severity) were reflected as a continuous variable. Similarly, the level of internal, motivation-related, and external protection, as well as the overall level of protection were continuous variables. For each variable, Krippendorff’s alpha (α) coefficient was calculated. ‘Low’ interrater reliability was inferred based on $\alpha < 0.67$, ‘good’/‘average’ at $0.67 \leq \alpha \leq 0.79$; and ‘very good’ when $\alpha \geq 0.8$ [27]. Krippendorff’s α coefficient is considered the most universal interrater reliability coefficient.

Results

First, interrater reliability coefficients were calculated for the assessments made in the group of inmates ($n = 200$; three judges). The results are presented in Table 2. Coefficients indicating ‘good’ interrater reliability were recorded for assessments of: general risk (HCR-SRR Case Prioritization) in both measurements, while ‘very good’ interrater reliability was observed for assessments of: historical risk (HCR-H) in both measurements. ‘Low’ interrater reliability characterized assessments of clinical risk (HCR-C) and severity of risk management problems (HCR-R) in both measurements.

The analysis proved ‘good’ interrater reliability among judges in the area of evaluation of external protection (SAPROF external) in both measurements. ‘Very good’ interrater reliability was also found among judges in the area of assessing: internal (SAPROF internal), motivation-related (SAPROF motivation), and general protection (SAPROF general) in both measurements.

Table 2. Interrater reliability coefficients in prison inmates group (Krippendorff’s alfa coefficient)

Prison inmates group ($n = 200$)		
Variable	1 st measurement	2 nd measurement (after 6 months)
HCR-H	$\alpha = 0.84$	$\alpha = 0.84$
HCR-C	$\alpha = 0.52$	$\alpha = 0.66$
HCR-R	$\alpha = 0.55$	$\alpha = 0.53$
HCR-SRR Case Prioritization	$\alpha = 0.77$	$\alpha = 0.79$
SAPROF internal	$\alpha = 0.83$	$\alpha = 0.83$
SAPROF motivation	$\alpha = 0.85$	$\alpha = 0.83$
SAPROF external	$\alpha = 0.77$	$\alpha = 0.76$
SAPROF FPJ	$\alpha = 0.89$	$\alpha = 0.88$

Interrater reliability coefficients were further calculated for assessments made in a group of forensic psychiatric patients ($n = 201$; three judges). The results are presented in Table 3. Coefficients indicating ‘good’ interrater reliability were recorded for assessments of: severity of risk management problems (HCR-R) in both measurements and clinical risk (HCR-C) in the second measurement. ‘Very good’ interrater reliability was observed for assessments of: historical risk (HCR-H), overall risk (HCR SRR Case Prioritisation) in both measurements. ‘Low’ interrater reliability characterized clinical risk assessments (HCR-C) in the first measurement. This means that between the first and second measurements, the interrater reliability of clinical risk assessments increased from ‘low’ to ‘good’.

The analysis proved ‘low’ interrater reliability in the area of evaluating external protection (SAPROF external) in both measurements. ‘Good’ interrater reliability was also seen in the area of evaluating: internal protection (SAPROF internal) and protection related to motivation (SAPROF motivation) in both measurements. ‘High’

interrater reliability was observed for assessments of general protection (SAPROF general) in both measurements.

Table 3. **Interrater reliability coefficients in forensic patients group (Krippendorff's alpha coefficient)**

Forensic patients group (n = 201)		
Variable	1 st measurement	2 nd measurement (after 6 months)
HCR-H	$\alpha = 0.81$	$\alpha = 0.80$
HCR-C	$\alpha = 0.62$	$\alpha = 0.79$
HCR-R	$\alpha = 0.75$	$\alpha = 0.78$
HCR-SRR Case Prioritization	$\alpha = 0.86$	$\alpha = 0.83$
SAPROF internal	$\alpha = 0.78$	$\alpha = 0.77$
SAPROF motivation	$\alpha = 0.76$	$\alpha = 0.75$
SAPROF external	$\alpha = 0.55$	$\alpha = 0.56$
SAPROF FPJ	$\alpha = 0.82$	$\alpha = 0.82$

Discussion

Polish practitioners associated with the judiciary system often formulate objections to the presented methods in the SPJ approach, exposing concerns about the bias of risk assessments, especially final evaluations, which do not directly reflect the sum of the present risk factors, but require an assessment of their relevance. They are concerned about the dependence of final assessments on different experiences, views and, characteristics of the assessors' formulation of attributions. The results of the present study, for the most part, dispel these concerns.

The interrater reliability in the inmate group was 'good' for SRR Case Prioritization in both measurements. Higher interrater reliability was shown for SRR Case Prioritization scores in the forensic psychiatric group compared to the inmate group. The interrater reliability for historical risk assessment was 'very good' in both groups (in both measurements). Historical factors are stable in nature. Data to establish their presence are easily available in records, and are relatively efficient and readily disclosed by patients and inmates, especially those with a long history of incarceration. Similar observations were made by Smith et al. [13]. In their sample, as in the present study, judges agreed most strongly in their assessments of historical factors, and least strongly in their assessments of clinical risk. This discrepancy is not surprising. Assessing the level of clinical factors seems objectively more difficult in the absence of personal contact with the assessed person than in the case of historical risk. While the HCR-20 V3 scale user manual offers precise guidelines in this regard, the assessment requires strong consideration of observation and documentation data, not just that provided by the assessed person. Thus evaluation by a person, based on a description of the patient's condition, without contact with the patient is clearly difficult. Improvement

in interrater reliability was observed between the first and second measurements in the group of forensic psychiatric patients. It is highly likely that more information from mental health professionals appeared in the medical records between measurements in comparison to the personal files in the inmate group, making it easier for judges to make assessments. It is worth noting that from the very beginning, there was already more clinical data than in the inmates personal files. The surveyed inmates were incarcerated in closed-type units, not in therapeutic ones, therefore, the possibility for them to maintain regular contact with psychologists or to undertake therapeutic interventions was significantly hampered. In the conditions of Polish prisons, on average, a psychologist has about 100 inmates under his/her care, thus systematic consultations and psychological support are low in such a situation. As an effect, the clinical data were limited. Consequently, those collecting/updating the data in the second measurement did not have such rich knowledge from the documentation of the subject's functioning over the past six months as those analyzing the current situation of forensic psychiatric patients.

As mentioned in the context of the 'high' interrater reliability in historical risk assessment, it may be that dynamic factors on the clinical and risk management scales are simply more difficult to assess by those who do not have constant contact with the subject being evaluated as staff of the unit in which they reside. A similar difficulty was recognized by Vincent et al. [28] in a study using the SAVRY scale on a group of adolescents. The researchers noted lower ('poor' to 'moderate') ICC coefficients for dynamic factors than for static factors ('moderate' to 'good').

The severity of risk management problems was assessed in a non-institutional context. It is possible that the 'low' interrater reliability of the risk management scale scores among inmates could be associated with the difficulty of reliably and objectively determining their future plans. In Polish conditions, the system of post-penitentiary support is not extensive. The offer of systemic assistance is basically limited to a small amount of financial support, while the possibility of covering the costs of housing or organizing a place to live after leaving the unit is available to few inmates, only through the actions of NGOs. Data on the living situation in this group was based, due to the aforementioned limitations, only on an interview from the inmate himself. The possibility of systemic control regarding the situation in an inmate's place of living or cooperating with mental health specialists is present only in the case of inmates leaving prison in a situation of conditional early release, and no longer in the case of inmates serving their sentence to the end. In such a situation, there may have been a kind of uncertainty among the risk assessment judges about the actual presence of social support, the characteristics of the living situation or coping with stress, as well as the future plans declared by the inmates.

Interestingly, the interrater reliability of assessments concerning the severity of external protection in inmates, which were based on similar data as risk management issues, appeared to be 'good'. In contrast, 'low' interrater reliability of the external protection assessments in forensic psychiatric patients was observed at the same time. Assessments were focused on the availability and size of supervision after leaving the hospital, having a social network, an intimate relationship or being able to receive

specialist care in a non-institutional setting. In the group of forensic psychiatric patients, it is difficult to predict what supervision measures will ultimately be applied, since in Poland, this is decided by the court after consulting experts. It occurs that the solutions requested on the basis of HCR-20 and SAPROF do not find the approval of the court or cannot be applied to a particular patient as a consequence of their low availability, for example, in the case of sheltered housing. Also, the extent of specialized care availability is difficult to assess and depends on many factors. Despite initial declarations by patients about their willingness to continue therapy, factors such as the distance of the specialist's office from their place of residence or the cost of therapy prove to be decisive in practice for the realization of these promises. The low interrater reliability may reflect important differences in the extent of the judge's confidence in patients' declarations.

A similar, 'high' interrater reliability of the final protection assessment was established in both the inmate group and the forensic psychiatric patients group. Internal and motivation-related protection assessments were characterized by higher interrater reliability in the inmate group ('very good') than in the psychiatric patients group ('good'). This result may be related to the higher availability of objective evidence with regard to coping in non-institutional settings for inmates who may engage in gainful employment or take unaccompanied leave during their sentences. Similar data may have been missing in the records of forensic psychiatric patients. As a result, assessments in this group may have been accompanied by less certainty or higher bias. This kind of assessment in the forensic psychiatric group is also more complicated than in the inmates group due to the need to predict the stability of motivation to use pharmacotherapy. Non-adherence to pharmacotherapy, for example, can exacerbate psychotic symptoms, which significantly increases the risk of violence while minimizing the impact of other protective factors [25].

Particularly important in the context of increasing the level of trust of Polish professionals in assessments made with the SPJ tools, and thus, the possibility of increasing their prevalence in forensic psychiatric opinions and periodic assessments in penitentiary units, is the high interrater reliability of the final assessments of risk and level of protection. What is noteworthy is that the data on 'good'/'very' good interrater reliability obtained in the referred study allow to formulate similar conclusions to those achieved by the tool developers themselves, both in theoretical and empirical areas.

Conclusions

The conducted analyses provided conclusions on:

- the relevance of considering dynamic factors in assessing the risk of violence;
- the relevance of considering, in addition to risk factors, protective factors in the process of assessing the risk of violence;
- the accessibility of the SPJ methods, which makes them applicable to various professional groups associated with the penitentiary and psychiatric populations, regardless of experience;

- the need for researchers to be aware of the risk of biased assessments, despite the steps taken by the developers of HCR-20 and SAPROF to reduce it, e.g., by structuring the assessment process and recommending the use of the so-called consensus decision-making model.

Acknowledgments

*The study was conducted as a part of the project entitled **The Role of Risk and Protective Factors in Assessing and Monitoring the Probability of Committing a Criminal Act in Offenders Subjected to an Isolated System of Treatment and Restraint Measures and Serving a Prison Sentence** carried out at SWPS University and funded by the National Science Centre (NCN; grant: OPUS 15, No. 2018/29/B/HS5/00474).*

References

1. Singh JP, Desmarais SL, Hurducas C, Arbach-Lucioni K, Condemarin C, Dean K et al. *International perspectives on the practical application of violence risk assessment: A global survey of 44 countries*. Int J Forensic Ment Health. 2014; 13(3): 193–206. doi:10.1080/14999013.2014.922141.
2. de Vogel V, De Beuf T, Shepherd S, Schneider RD. *Violence risk assessment with the HCR-20V3 in legal contexts: A critical reflection*. J Pers Assess. 2022; 104(2): 252–264. doi:10.1080/00223891.2021.2021925.
3. Challinor A, Ogundalu A, McIntyre JC, Bramwell V, Nathan R. *The empirical evidence base for the use of the HCR-20: A narrative review of study designs and transferability of results to clinical practice*. Int J Law Psychiatry. 2021; 78: 101729. doi:10.1016/j.ijlp.2021.101729.
4. Douglas KS, Hart SD, Webster CD, Belfrage H. *HCR-20v3: Assessing risk for violence – User guide*. Burnaby: Mental Health, Law and Policy Institute, Simon Fraser University; 2013.
5. Douglas KS, Shaffer CS. *The science of and practice with the HCR-20 V3 (Historical-Clinical-Risk Management-20, Version 3)*. In: Douglas KS, Otto RK, editors. *Handbook of violence risk assessment*. New York: Routledge/Taylor & Francis Group; 2021. p. 253–293. doi:10.4324/9781315518374-16.
6. Venner S, Sivasubramaniam D, Luebbers S, Shepherd SM. *Cross-cultural reliability and rater bias in forensic risk assessment: A review of the literature*. Psychol Crime Law. 2020. doi:10.1080/1068316X.2020.1775829.
7. Murray J, Charles KE, Cooke DJ, Thomson ME. *Investigating the influence of causal attributions on both the worksheet and checklist versions of the HCR-20*. Int J Forensic Ment Health. 2014;13(1):8–17. doi:10.1080/14999013.2014.890978.
8. Douglas KS, Belfrage H. *Interrater reliability and concurrent validity of the HCR-20 Version 3*. Int J Forensic Ment Health. 2014;13(2):130–139. doi:10.1080/14999013.2014.908429.
9. Strub DS, Douglas KS, Nicholls TL. *The validity of Version 3 of the HCR-20 violence risk assessment scheme amongst offenders and civil psychiatric patients*. Int J Forensic Ment Health. 2014; 13(2): 148–159. doi:10.1080/14999013.2014.911785.
10. Robbé M de V, de Vogel V, Douglas KS. *Risk factors and protective factors: A two-sided dynamic approach to violence risk assessment*. J Forensic Psychiatry Psychol. 2013; 24(4): 440–457. doi:10.1080/14789949.2013.818162.

11. Doyle M, Power LA, Coid J, Kallis C, Ullrich S, Shaw J. *Predicting post-discharge community violence in England and Wales using the HCR-20V3*. Int J Forensic Ment Health. 2014; 13(2): 140–147. doi:10.1080/14999013.2014.906517.
12. Persson M, Belfrage H, Fredriksson B, Kristiansson M. *Violence during imprisonment, forensic psychiatric care, and probation: Correlations and predictive validity of the risk assessment instruments COVR, LSI-R, HCR-20V3, and SAPROF*. Int J Forensic Ment Health. 2017; 16(2): 117–129. doi:10.1080/14999013.2016.1266420.
13. Smith ST, Kelley SE, Rulseh A, Sörman K, Edens JF. *Adapting the HCR-20V3 for pre-trial settings*. Int J Forensic Ment Health. 2014; 13(2): 160–171. doi:10.1080/14999013.2014.906520.
14. Sea J, Bang S. *The interrater reliability and concurrent validity of the HCR-20 Version 3 in South Korea*. J Forensic Psychiatry Psychol. 2021; 32(6): 879–901. doi:10.1080/14789949.2021.1937279.
15. Kötter S, von Franqué F, Bolzmacher M, et al. *The HCR-20V3 in Germany*. Int J Forensic Ment Health. 2014; 13(2): 122–129. doi:10.1080/14999013.2014.911784.
16. Cheng J, Haag AM, Olver ME. *Predictors of Historical Clinical Risk Management-20 Version 3 (HCR-20:V3) summary risk ratings*. Psychiatry Psychol Law. 2019; 26(4): 682–692. doi:10.1080/13218719.2019.1618753.
17. Cabeldue M, Green D, Griswold H, Schneider M, Smith J, Belfi B, Kunz M. *Using the HCR-20V3 to differentiate insanity acquittees based on opinions of readiness for transfer*. J Am Acad Psychiatry Law. 2018; 46(3): 339–350. doi:10.29158/JAAPL.003769-18.
18. Cook AN, Moulden HM, Mamak M, Lalani S, Messina K, Chaimowitz G. *Validating the Hamilton Anatomy of Risk Management–Forensic Version and the Aggressive Incidents Scale*. Assessment. 2018; 25(4): 432–445. doi:10.1177/1073191116653828.
19. Chen Y, Douglas KS, Zhang Z, Xiao C, Wang H, Wang Y et al. *Evaluating the HCR-20V3 violence risk assessment measure with mentally disordered offenders and civil psychiatric patients in China*. Behav Sci Law. 2023; 1–21. doi:10.1002/bsl.2614.
20. Wolf V, Mayer J, Steiner I, Franke I, Klein V, Streb J et al. *Predicting violence in female forensic inpatients with substance use disorders – The utility of a gender-responsive assessment*. Front Psychiatry. 2024; 15: 1346815. doi:10.3389/fpsy.2024.1346815.
21. Sideli L, Carpentieri R, Capri C, Bracalenti R, Cavallo J, Schimmenti A et al. *Italian validation of the Historical-Clinical-Risk Management-20 Version 3 among teenagers admitted to the youth justice system*. Mediterr J Clin Psychol. 2024; 12(2). doi:10.13129/2282-1619/mjcp-4083.
22. Burghart M, de Ruiter C, Hynes SE, Krishnan N, Levtova Y, Uyar A. *The Structured Assessment of Protective Factors for Violence Risk (SAPROF): A meta-analysis of its predictive and incremental validity*. Psychol Assess. 2022. doi:10.1037/pas0001184.
23. Abidin Z, Davoren M, Naughton L, Gibbons O, Nulty A, Kennedy HG. *Susceptibility (risk and protective) factors for in-patient violence and self-harm: Prospective study of structured professional judgement instruments START and SAPROF, DUNDRUM-3 and DUNDRUM-4 in forensic mental health services*. BMC Psychiatry. 2013; 13: 197. doi:10.1186/1471-244X-13-197.
24. Oziel S, Marshall LA, Day DM. *Validating the SAPROF with forensic mental health patients*. J Forensic Psychiatry Psychol. 2020; 31(5): 667–698. doi:10.1080/14789949.2020.1785525.
25. de Vogel V, de Ruiter C, Bouman YHA, de Vries Robbé M. *SAPROF – Wytyczne do oceny czynników ochronnych dotyczących ryzyka przemocy [Guidelines for Assessing Protective Factors for the Risk of Violence]*. Transl.: M. Nowopolski, A. Walento-Nowacka, L. Ciszewski, A. Axer. Utrecht: Van der Hoeven Kliniek; 2016 (in Polish).

26. Hayes AF, Krippendorff K. *Answering the call for a standard reliability measure for coding data*. Commun Methods Meas. 2007; 1(1): 77–89. doi:10.1080/19312450709336664.
27. Krippendorff K. *Content Analysis: An Introduction to Its Methodology*. 4th ed. SAGE Publications; 2019. doi:10.4135/9781071878781.
28. Vincent GM, Guy LS, Grisso T. *Risk assessment in juvenile justice: A guidebook for implementation*. Chicago: Models for Change, Juvenile Justice Initiative, National Center for Mental Health and Juvenile Justice; 2012.

Corresponding author: Marlena Banasik
e-mail: marlaa@poczta.fm