Dienst Justitiële Inrichtingen
Ministerie van Justitie en Veiligheid

Discriminative and Divergent Validity of the Violent Extremism Risk Assessment tool

Annemaryn de Bruin | Nils Duits | Maaike Kempes | Merel Prinsen

Research report 2022-1

Annemaryn A. de Bruin, MSc. Junior researcher at the Netherlands Institute of Forensic Psychiatry and Psychology (NIFP) of the Dutch Ministry of Justice and Security.

Nils Duits, MD, PhD. Forensic child- and adolescent psychiatrist & senior researcher at the Netherlands Institute of Forensic Psychiatry and Psychology (NIFP) of the Dutch Ministry of Justice and Security. Co-author and supervisor-trainer of the Violent Extremism Risk Assessment tool (VERA-2R). Former research coordinator of the VERAPRO2EU Project.

Prof. dr. Maaike Kempes. Head of Department of Science and Education, Netherlands Institute of Forensic Psychiatry and Psychology (NIFP) of the Dutch Ministry of Justice and Security.

Merel Prinsen, MA, LLM, PhD. Research coordinator of the VERAPRO2EU Project.

Acknowledgements:

This work was supported by the European Commission under Grant (Call JUST-JCOO-AG-2020, Project 101007383). We want to thank them for their financial support. Furthermore, we would like to thank the Dutch Public Prosecution Service for providing the judicial files.

Disclosure Statement:

No potential financial interest or benefit was reported by the authors.

Correspondence:

Correspondence should be addressed to: Annemaryn A. de Bruin, Department of Science and Education, Netherlands Institute of Forensic Psychiatry and Psychology (NIFP), Dutch Ministry of Justice and Security, Utrecht 3511 EW, The Netherlands (email: annemaryndebruin@gmail.com).

Data Availability Statement:

The data are not publicly available due to its sensitive nature.

Co-funded by the European Union's Justice Programme (2014–2020)

This project was supported by the European Commission under Grant (Call JUST-JCOO-AG-2020, Project 101007383).

Table of content

1. Abstract

The Violent Extremism Risk Assessment - Version 2 Revised (VERA-2R) is an evidence-based structured professional judgement (SPJ) tool for ideologically motivated violence. Use of the tool can help professionals in risk assessment and risk management of terrorists and violent extremists. It is important that the tool leads to reliable and valid risk assessments. Therefore, we aimed to establish the validity of the VERA-2R, focusing on the discriminative validity and divergent validity. In order to do so, trained researchers assessed a sample of 33 convicted terrorist offenders and 33 convicted violent offenders with the VERA-2R for the discriminative validity study, and a sample of 66 convicted terrorist offenders with both the VERA-2R and HCR20^{V3} for the divergent validity study. With respect to the discriminative validity, the results show us that the VERA-2R in general incorporates indicators which are specifically relevant for terrorist offenders compared to violent offenders. Furthermore, by comparing the VERA-2R to the HCR20^{V3}, it became clear that in general the concepts represented by the VERA-2R indicators are not incorporated in the HCR20^{V3}. In conclusion it can be said that the results provide evidence for the fact that the VERA-2R is a specialized tool for ideologically motivated violence, which endorses the need for risk assessment instruments for special populations. Although the results of the validity study can, in combination with the results of the reliability study, be used to make well-founded recommendations on how to improve the VERA-2R, it still remains important to establish the predictive validity of the VERA-2R, since this is considered to be the gold standard for forensic risk assessment tools.

2. Introduction

The VERA-2R is used worldwide by trained professionals in judicial practice for violent extremism and terrorism risk assessment and risk management (Van der Heide, Van der Zwan, & Van Leyenhorst, 2019). In recent years, the number of VERA-2R trained professionals in Europe has increased considerably to more than 1000 professionals and the number of judicial purposes for which the VERA-2R is applied has expanded (Ministry of Justice and Security, 2021a, 2021b). Risk assessments are relevant for the security threats of individuals who are imprisoned for a violent extremist or terrorist offence, as well as for prisoners who are radicalized in prison, and should be carried out both during and after their release from prison (European Union, 2022). In the Netherlands VERA-2R risk assessments are used for these purposes and being shared more and more between judicial professionals (National Coordinator for Counterterrorism, 2022). For that reason, it is of paramount importance that the VERA-2R is reliable, meaning that the instrument is stable, consistent, predictable, accurate and free from random error (Groth-Marnet, 2009) and valid, meaning that the instrument measures what it purports to measure (Field, 2005).

We examined the interrater and intrarater reliability of the VERA-2R in the first part of the validation project (Bruin, Duits, Kempes, & Prinsen, 2023). In this study we examine the validity of the VERA-2R, focusing on the discriminative and divergent validity.

Discriminative validity refers to the degree to which the evaluated test is able to discriminate between groups that are known and expected to be different from each other (Gouttebarge, 2008). Following the statement that ideologically motivated violence and regular violence are known to be two distinct constructs (Clarke & Newman, 2006; Lafree & Dugan, 2004) with different underlying risk factors (Van der Heide et al., 2019), terrorist offenders and violent offenders can be said to be two different groups. Therefore, in order to evaluate the discriminative validity, we compared the test scores of terrorist offenders and violent offenders on the VERA-2R. Since the VERA-2R should contain specifically relevant indicators for terrorist offenders (Pressman, Duits, Rinne, & Flockton, 2018), we expect significant higher test scores for terrorist offenders in comparison to violent offenders.

Divergent validity is demonstrated by evidence that measures of constructs that theoretically should not be highly related to each other are, in fact, not found to be highly correlated to each other (Gouttebarge, 2008). In this study we chose to compare the VERA-2R, as an instrument for ideologically motivated violence, with an instrument for regular violence. We selected the Historical-Clinical-Risk Management-20 – Version 3 ($HCR20^{V3}$), because it is the most widely used risk assessment instrument for regular violence (Silva, 2020). Given the fact that ideologically motivated violence and regular violence are known to be two distinct constructs and that most of the indicators incorporated in $HCR20^{V3}$ are found not to be relevant to the idiosyncrasies of terrorists and violent extremists (Smith & Nolan, 2016), we hypothesize to find weak correlations between the VERA-2R and $HCR20^{V3}$. Hereby, high divergent validity will indicate that the VERA-2R assesses a different type of risk than the $HCR20^{V3}$, and therefore provides support for the VERA-2R its relative uniqueness (Holton, Bates, Bookter, & Yamkovenko, 2007).

Support for the discriminative validity and divergent validity of the VERA-2R confirms that the VERA-2R is a specialized tool for terrorism and violent extremism (Pressman et al., 2018), which accurately measures what its purports to (Holton et al., 2007). This endorses the need for risk assessment instruments for special populations.

3. Methods Discriminative Validity

Assessors and Cases

The assessors included in this study were two Dutch researchers with a master's degree in criminology. At the time of the study, the researchers were employed by the Netherlands Institute of Forensic Psychiatry and Psychology (NIFP) of the Dutch Ministry of Justice and Security. The assessors took part in a two-day training course to obtain an in-depth understanding of the instrument and to acquire experience in applying the VERA-2R indicators and forming (final structured) risk judgements.

To assess the discriminative validity, initially 34[1] cases of convicted terrorist offenders and 34 cases of convicted violent offenders were selected to be coded with the VERA-2R on the basis of extensive judicial files. These files were provided by the Dutch Public Prosecution Service, and can include a forensic mental health (FMH) assessment, a probation report, a transcript of the verdict, a police report, a criminal record and/or information from intelligence services.

In order to compare the group of terrorist offenders with the group of violent offenders on the VERA-2R, differences in background characteristics between both groups were reduced as much as possible. Therefore, we individually matched both groups on several relevant characteristics, among which type of index crime. If the person concerned was convicted for multiple criminal acts, we matched on the basis of the most severe criminal act in terms of the maximum possible sentence. This is referred to as the index crime. Although some of the terrorist crimes could be matched perfectly to a comparable violent crime, for some terrorist crimes, e.g., participation in a terrorist organization, training, etc., no related criminal code for regular violence exists. For these types of terrorist crimes, we had to choose the best alternative. In our opinion we did so by matching these crimes to "Public violence committed with other people" (see Table 1). The underlying motivation for this matching procedure is that both the concerned terrorist crimes and the crime 'Public violence committed with other people' are related to the involvement of violence in a group. Furthermore, although the concerned terrorist crimes do not include active violent acts, they at least incite or enable the use of violence, and therefore these offences eventually do lead to active violence.

In addition to this, both groups were also matched on age, sex, and the presence of a FMH assessment. With respect to age, the case with the closest age is selected in case an exact match on age was not possible. Hereby we tried to keep the mean age in both groups the same. Matching on FMH assessment is important, since most information about the risk indicators will be derived from these reports. Therefore, an unequal distribution of available reports among the groups impacts the chance to find risk indicators. A disadvantage of matching on FMH assessment lies in the fact that the judicial files of terrorist offenders more often include a FMH assessment than the judicial files of violent offenders. Therefore, by matching the groups on the presence of a FMH assessment, the sample of violent offenders

[1] A-priori power analysis estimated that 34 cases are required in both groups.

may not be representative for the population of violent offenders. Nevertheless, matching on FMH assessment will, compared to not matching, result in more reliable results.

Due to an erroneous match the final sample consisted of 33 cases of convicted terrorist offenders and 33 cases of convicted violent offenders. Both groups (see Table 1) comprised 30 men and 33 women who were aged between 15 and 59 years at the time of their offence (terrorist offenders: M_{age} = 26.5, SD = 9.7; violent offenders: M_{age} = 26.2, SD = 9.8). In both groups, a FHM assessment was present for about half of subjects.

Table 1. Descriptive statistics

Terrorist offenders		**Violent offenders**	
Age	26.7	Age	26.2
Sex		Sex	
Male	30	Male	30
Female	3	Female	3
FMH assessment present		FMH assessment present	
Yes	19	Yes	19
No	14	No	14
Type of crime		Type of crime	
Participation in terrorist organization	8	Public violence committed with other	8
Attempted participation in a terrorist organization	2	Public violence committed with other	2
Training	1	Public violence committed with other	1
Incitement	3	Public violence committed with other	3
Preparing terrorist act	3	Public violence committed with other	3
Preparing homicide	5	Attempted homicide	5
Threatening	9	Threatening	9
Arson	2	Arson	2

Materials

The Violent Extremism Risk Assessment – Version 2 Revised (VERA-2R) contains 34 risk and protective indicators specifically related to the risk of violent extremism and terrorism (Pressman et al., 2018). The VERA-2R indicators are divided into five domains: Beliefs, Attitudes and Ideology (BA), Social Context and Intention (SCI), History, Action and Capacity (HAC), Commitment and Motivation (CM), and Protective and risk-mitigating indicators (P). The scientific basis for each indicator is explained, along with the underlying criteria for the three rating levels: low, moderate or high. A risk indicator is rated as 'low' if the risk-promoting indicator characteristics are objectively not present, as 'moderate' if the risk-promoting indicator characteristics are present to a specified level, and as 'high' if the risk-promoting indicator

characteristics are clearly present or present to a high level. The protective indicators are rated in reverse, which is to say that lower ratings indicate a higher level of risk (Pressman et al., 2018). A protective indicator is rated as 'low' if no risk-mitigating indicator characteristics are present, as 'moderate' if some risk-mitigating indicator characteristics are present, and as 'high' if clear risk-mitigating indicator characteristics are present (Pressman et al., 2018). It is important to stress here that the VERA-2R does not provide a numerical score for the ratings (Pressman et al., 2018). However, for the purposes of this study we assigned the numerical scores '0', '1' and '2' to the ratings 'low', 'moderate' and 'high', respectively. Furthermore, we decided to assign the numerical score '-99' (missing value) if the judicial file did not contain information about an indicator.

The VERA-2R also includes 11 additional indicators, which may contribute to a person's vulnerability to engage in future acts of violent extremism and terrorism, when combined with the presence of ideological, contextual, and motivational indicators identified in the VERA-2R (Pressman et al., 2018). These additional indicators are divided into three domains: Criminal History (CH), Personal History (PH) and Mental Disorder (MD). The scientific basis for each indicator is explained, along with the criteria for the two rating levels: not present or present. The rating 'not present' (0) corresponds to the absence of the additional indicator characteristics, while the rating 'present' (1) corresponds to the presence of the additional indicator characteristics (Pressman et al., 2018). We decided to assign the numerical score '-99' (missing value) if the judicial file did not contain information about an indicator.

After carefully considering the relevant indicators, the assessor then assigns risk judgements to the VERA-2R domains and weighs the domains. Subsequently, a final structured risk judgement is made in terms of the likelihood of an individual engaging in ideologically motivated violence. The (final structured) risk judgements are formulated in a risk narrative, as well as rated on a scale of low (0), moderate (1), and high (2) (Pressman et al., 2018). Furthermore, different risk scenarios are identified with a risk management strategy for each of these scenarios (Douglas et al., 2014; Logan, 2017).

Research Design

Our research design includes trained researchers and extensive judicial files, and therefore closely resembles VERA-2R assessments in practice.

Since the VERA-2R is developed specifically for terrorist offenders, the encoding rules of the tool do not focus on how the indicators should be assessed for violent offenders. Therefore, we expand the existing encoding rules with criteria for violent offenders (See Appendix A). Hereby it is important to emphasize that the encoding rules provided for the indicators within the 'Commitment and Motivation' (CM) domain were not completely in line with the working procedure of the VERA-2R, since for the violent offenders we shifted the focus from motives for ideologically motivated violence to motives for regular violence in order to be able to make a more meaningful comparison.

In order to determine whether the use of multiple assessors would impact the research findings, we established the interrater reliability between assessor-1 (gold standard) and assessor-2 on the indicators. This enabled us to: a) evaluate the extent to which the assessors score the same ratings for the feature that is being observed or measured, and b) evaluate whether assessor-2 is able to rate the VERA-2R indicators in an adequate way. The reliability analysis was based on testcases (3 testcases of terrorist offenders and 3 testcases of violent offenders).

The distributions of the observed ratings frequently fell under one category of ratings. As a result, kappa estimates appeared to be unrepresentatively low (Eugenio & Glass, 2004). Therefore, an alternative kappa was calculated based on the percentage of agreement between the evaluators, and corrected for agreement based merely on chance, which depends on the number of answer options available. To establish the strength of the agreement, Landis & Koch's cut-off points were used: $\kappa \leq .20$ = slight, $.20 < \kappa \leq .40$ = fair, $.40 < \kappa \leq .60$ = moderate, $.60 < \kappa \leq .80$ = good and $.80 < \kappa \leq .1.00$ = excellent (Landis & Koch, 1977).

The results show that the average amount of agreement between the assessors can be classified as excellent ($\kappa = .90$). This indicates that the assessment is independent of the raters or professional assessors (Jonsson & Svingby, 2007) and that both assessors are able to evaluate the VERA-2R indicators in an adequate way. Based on these promising results, we agreed to the use of multiple raters. As a result, assessor-1 evaluated all the cases of the terrorist offenders and assessor-2 evaluated all the cases of the violent offenders.

Although we generally found high interrater reliability, for five indicators we found an interrater reliability equal to or below the critical value of .60. Since we agreed to the use of multiple raters, we had to make sure that this would not have an impact on the research findings. Therefore, all the indicators that were found to have low interrater reliability were elaborately discussed beforehand. In addition to this, given that assessor-1 was found to be the gold standard, assessor-1 verified all the ratings that assessor-2 had assigned to the indicators with low interrater reliability. In case of disagreement between the two assessors, the perspective of assessor-1 was leading.

With respect to the (final structured) risk judgements, we dealt with the use of multiple raters by letting assessor-1 verify all the (final structured) risk judgements that assessor-2 has assigned.

Security and Privacy

To ensure that there were no risks to the privacy of the subjects, we anonymized the data. Moreover, with regard to data protection, we stored the anonymized dataset in a secure digital environment, in order to protect the information against misuse, unauthorized access, disclosure and theft.

Statistical Analysis

Before analysing the data using IBM SPSS Statistics for Mac Version 25.0, some data processing steps had to be taken. First of all, reverse coding was applied to the 'Protective and risk-mitigating indicators' domain (P), so that higher ratings

correspond to a higher risk status. Secondly, total scores had to be calculated for research purposes. Hereby it is important to stress that total scores have no relevance in professional SPJ-practice. To calculate the total scores of the domains, we added the numerical scores assigned to the indicators representing the concerned domain, divided this by the number of indicators for which a rating was present (i.e., excluding the indicators for which no information was available), and multiplied it with the number of indicators representing the concerned domain. To calculate the final total scores, we added the total scores of the domains. Hereby two different total scores were calculated: 1) a final total score solely based on the ratings assigned to the VERA-2R indicators 2) a final total score based on the ratings assigned to both the VERA-2R indicators and the additional indicators.

Afterwards, in order to determine whether the VERA-2R is able to discriminate between a group of terrorist offenders and a group of violent offenders, a Mann-Whitney U test was carried out with the VERA-2R indicators, (final structured) risk judgements and total scores as dependent variables and the distinction between terrorist offenders and violent offenders as independent variable. A p-value of $< .05$ is considered statistically significant. The correlation coefficient effect size (r) was calculated for scores that differed significantly. The effect sizes were interpreted in accordance with the guidelines outlined by Cohen (1988): $.10 < r < .30$ = weak, $.30 < r < .50$ = moderate, $r \geq .50$ = strong.

In addition to this, a Fisher's exact test was carried out for the additional indicators, because extant literature has showed that the Fisher's exact test is the most adequate statistic method when using a small sample size (Bower 2003). In order to determine the strength of the relationship effect size Cramer's V was calculated. The guidelines of Cohen (1988) were used to interpret the effect sizes (see Table 2).

Table 2. Interpretations for Cramer's V

df	Negligible	Small	Medium	Large
1	$0 < V < .10$	$.10 < V < .30$	$.30 < V < .50$	$V \geq .50$
2	$0 < V < .07$	$.07 < V < .21$	$.21 < V < .35$	$V \geq .35$
4	$0 < V < .06$	$.06 < V < .17$	$.17 < V < .29$	$V \geq .29$
4	$0 < V < .05$	$.05 < V < .15$	$.15 < V < .25$	$V \geq .25$
5	$0 < V < .05$	$.05 < V < .13$	$.13 < V < .22$	$V \geq .22$

4. Results Discriminative Validity

Discriminative Validity of the VERA-2R Indicators

In order to determine whether the VERA-2R is able to discriminate between terrorist offenders and violent offenders, multiple analyses were carried out. The results of the Mann-Whitney U tests on the VERA-2R indicators are shown in Table 3.

With respect to the 'Beliefs, Attitudes and Ideology' (BA) domain we see that significant higher average scores are found for the terrorist offender group in comparison to the violent offender group on almost all the indicators (6 of 7). An exception in this domain is the indicator 'BA.5', for which no significant differences are found between the terrorist offenders and the violent offenders. This indicator is about expressed emotions of anger, moral outrage or hatred towards a particular person, a group, or an institution such as the government.

When we look at the 'Social Context and Intention' (SCI) domain we see significant differences between the terrorist offender group and violent offender group on all the indicators. More specifically, most of the indicators (6 of 7) reveal significant higher scores for the terrorist offenders in comparison to the violent offenders. However, for one of the indicators, indicator 'SCI.2', significant higher scores are found for the violent offender group. For all of the indicators in this domain strong effect sizes are found ($r \geq .50$).

Within the 'History, Action and Capacity' (HAC) domain we find significant higher average scores for the terrorist offenders in comparison to the violent offenders on almost all the indicators (5 of 6). However, with respect to violent criminal history (indicator HAC.3), no significant differences are found between the terrorist offender group and violent offender group.

For half of the motivations (4 of 8) of the 'Commitment and Motivation' (CM) domain no significant differences are found between the terrorist offenders and the violent offenders. Terrorist offenders and violent offenders seem to the same extent motivated by criminal opportunism (CM.2), excitement and adventure (CM.5), and the acquisition of status (CM.7) when we compare the mean scores of the groups. There is also no significant difference found on the indicator that looks at whether the offenders were in some way forced to participate in the crime (CM.6). However, regarding the other motivations the results reveal significant higher scores for the terrorist offender group in comparison to the violent offender group.

Furthermore, for the last VERA-2R domain, which includes the 'Protective and risk-mitigating indicators' (P), we see that significant higher average scores are found for the terrorist offender group in comparison to the violent offender group on the indicators 'P.1', 'P.2' and 'P.3'. With respect to the indicators 'P.5' and 'P.6' no significant differences are found. This means there is no significant difference when we compare the mean scores of the received support for non-violence from the community, family members or other important persons between the terrorist offender group and the violent offender group.

Table 3. VERA-2R indicators

	Terrorist offenders		Violent offenders		Mann-Whitney U-test			
Indicators	*N*	Mean Rank	*N*	Mean Rank	*U*	*Z*	*p*	*r*
BA. Beliefs, Attitudes and Ideology								
BA.1 Commitment to ideology that justifies violence	31	41.50	31	21.50	170.50	-5.29	≤ .001	-.67
BA.2 Perceived grievances and/or perceived injustice*	22	29.82	25	18.68	147.00	-2.91	≤ .01	-.42
BA.3 Dehumanization of designated targets associated with injustice*	10	28.10	33	20.15	104.00	-3.16	≤ .01	-.55
BA.4 Rejection of democratic society and values	24	32.50	30	23.50	240.00	-3.38	≤ .001	-.46
BA.5 Expressed emotions in response to perceived injustice*	17	25.12	25	19.04	151.00	-1.73	.08	-.27
BA.6 Hostility to national identity	28	32.50	25	24.50	280.00	-3.02	≤ .01	-.41
BA.7 Lack of empathy and understanding for those outside one's own group	24	36.40	31	21.50	170.50	-4.61	≤ .001	-.62
SCI. Social Context and Intention								
SCI.1 Seeker, user or developer of violent extremist materials	29	44.59	33	20.00	99.00	-6.26	≤ .001	-.80
SCI.2 Target for attack identified (person, group, location) *	27	21.50	33	37.86	202.50	-4.60	≤ .001	-.59
SCI.3 Personal contact with violent extremists*	28	38.13	27	17.50	94.50	-5.61	≤ .001	-.76
SCI.4 Expressed intention to commit acts of violent extremism	30	45.20	33	20.00	99.00	-6.42	≤ .001	-.81
SCI.5 Expressed willingness and/or preparation to die for a cause or belief*	19	32.58	33	23.00	198.00	-3.70	≤ .001	-.51
SCI.6 Planning, preparation of acts of violent extremism	29	43.69	32	19.50	96.00	-6.27	≤ .001	-.80
SCI.7 Susceptibility to influence, control or indoctrination*	23	36.93	31	20.50	139.50	-4.95	≤ .001	-.66
HAC. History, Action and Capacity								
HAC.1 Early exposure to violence-promoting, militant ideology	26	30.79	29	25.50	304.50	-2.45	≤ .01	-.33
HAC.2 Network of family and friends involved in violent extremism	30	32.15	27	25.50	310.50	-2.66	≤ .01	-.35
HAC.3 Violent criminal history	33	29.56	32	36.55	414.50	-1.62	.11	-.20
HAC.4 Strategic, paramilitary and/or explosives training	16	21.50	18	13.94	80.00	-2.89	≤ .01	-.50
HAC.5 Training in extremist ideology in own country or abroad*	16	29.81	31	21.00	155.00	-3.60	≤ .001	-.52
HAC.6 Organizational skills and access to funding and sources of help	29	39.83	30	20.50	150.00	-4.68	≤ .001	-.61
CM. Commitment and Motivation								
CM.1 Motivated by perceived religious obligation and/or glorification	27	39.48	32	22.00	176.00	-5.00	≤ .001	-.65
CM.2 Motivated by criminal opportunism*	24	25.56	29	28.19	313.50	-.75	.45	-.10
CM.3 Motivated by camaraderie, group belonging	18	30.06	29	20.24	152.00	-2.68	≤ .01	-.39
CM.4 Motivated by moral obligation, moral superiority*	26	31.98	31	26.50	325.50	-2.53	≤ .01	-.34
CM.5 Motivated by excitement and adventure*	10	18.75	25	17.70	117.50	-.68	.50	-.11
CM.6 Forced participation	31	33.03	32	31.00	464.00	-1.45	.15	-.18
CM.7 Motivated by acquisition of status*	14	26.64	28	19.93	152.00	-1.93	.05	-.30
CM.8 Motivated by a search for meaning and significance in life*	16	31.16	29	18.50	101.50	-4.44	≤ .001	-.66
P. Protective and risk-mitigating indicators								
P.1 Reinterpretation of the ideology*	25	34.78	33	25.50	280.50	-3.46	≤ .001	-.45
P.2 Rejection of violence as a means to achieve goals	24	33.81	33	22.50	280.50	-3.28	≤ .001	-.43
P.3 Change in concept of the enemy	19	33.45	33	22.50	181.50	-4.00	≤ .001	-.55
P.4 Participant in programmes against violent extremism	8	4.50	0	0	-	-	-	-
P.5 Support from the community for non-violence	27	28.46	26	25.48	311.50	-1.39	.17	-.19
P.6 Support from family members, other important persons for non-violence	29	26.26	27	30.91	326.50	-1.16	.25	-.19

Note. *N* represents the number of cases included in the analysis. If the judicial file did not contain information about an indicator (-99), the case was excluded from the analysis.

*For these indicators, the elimination of the missing values resulted in differences between the groups on one or several matchings variables. In response to this, we performed an additional analysis, including only the cases for which a rating of low, moderate or high was assigned to both the terrorist case and the corresponding violent offender case. For one indicator the original analysis and additional analysis revealed different results: the additional analysis established that indicator 'BA.5' was significant more often, to a certain extent, present in the terrorist offender group in comparison to the violent offender group (n = 19, mean rank terrorist offenders =19.17, mean rank violent offenders = 11.83, U = 57.50, p = .02).

Since programs against violent extremism were not offered to individuals within the violent offender group, this item was not applicable to them. Therefore indicator 'P.4' was excluded from the analyses.

Discriminative Validity of the Additional Indicators

In table 4 the results of the Fisher's exact tests on the additional indicators are displayed. There is only one additional indicator with a significant result. The violent offenders are significantly more often convicted for non-violent offences before the index crime in comparison to the terrorist offenders (67% vs 33%; $p = 0.01$). The strength of this association, as represented by the Cramer's V effect size, is moderately strong ($V = 0.34$). On all the other additional indicators, no significant differences are found between the terrorist offender and violent offender group. Although there is no significant difference, a striking result was found on the indicator 'Problems with school and work' (indicator PH.3). This indicator stands out in terms of high rates in both groups; almost 70% of the terrorist offenders and 67% of the violent offenders have had problems with school and/or work before the index crime. When we look at the mental disorders, we see that in both groups' personality disorders are present the most.

Table 4. Additional indicators

Additional indicators	Terrorist offenders (N = 33) N	%	Violent offenders (N = 33) N	%	Fisher's exact *p*	Cramer's V *v*	*p*
CH. Criminal History							
CH.1 Convicted for non-violent offence(s)	21	63.63%	10	30.30%			
No	11	33.33%	22	66.67%	.01**	.34	≤ .01
Yes							
CH.2 Non-compliance with conditions or supervision*							
No	3	9.10%	12	36.36%			
Yes	4	12.12%	6	18.18%	.38	.22	.29
PH. Personal History							
PH.1 Violence in family							
No	19	57.58%	13	39.39%			
Yes	8	24.24%	11	33.33%	.26	.17	.23
PH.2 Problematic upbringing and/or placed in juvenile care*							
No	12	36.36%	8	24.24%			
Yes	16	48.48%	17	51.52%	.57	.11	.42
PH.3 Problems with school and work							
No	5	15.15%	6	18.18%	1.00	.05	.74
Yes	23	69.70%	22	66.67%			
MD. Mental Disorder							
MD.1 Personality disorder*							
No	11	33.33%	6	18.18%	.21	.24	.12
Yes	10	30.30%	15	45.45%			
MD.2 Depressive disorder and/or suicide attempts*							
No	11	33.33%	14	42.42%	.75	.09	.57
Yes	8	24.24%	7	21.21%			
MD.3 Psychotic and schizophrenic disorder							
No	16	48.48%	15	45.45%	.33	.16	.28
Yes	4	12.12%	8	24.24%			
MD.4 Autism spectrum disorder							
No	13	39.39%	19	57.58%	.43	.18	.27
Yes	5	15.15%	3	9.10%			
MD.5 Post-traumatic stress disorder							
No	15	45.45%	17	51.52%	.67	.10	.54
Yes	2	6.10%	4	12.12%			
MD.6 Substance use disorder							
No	18	54.55%	16	48.48%	.39	.13	.34
Yes	7	21.21%	11	33.33%			

Note. *N* represents the number of cases included in the analysis. If the judicial file did not contain information about an indicator (-99), the case was excluded from the analysis.

Discriminative Validity of the (Final Structured) Risk Judgements

The results of the (final structured) risk judgements are shown in table 5. On a domain-level, higher risk judgements are assigned to the terrorist offenders compared to the violent offenders. An exception is the 'Protective and risk-mitigating indicators' (P) domain. In this domain we found no significant differences between the risk judgements of the terrorist offenders and violent offenders. With respect to the final structured risk judgement, the results reveal higher scores for the terrorist offenders in comparison to the violent offenders.

Table 5. (Final structured) risk judgements

	Terrorist offenders		Violent offenders		Mann-Whitney U test			
	N	Mean Rank	*N*	Mean Rank	*U*	*Z*	*p*	*r*
Risk judgement domain 'BA'	33	43.73	33	23.27	207.00	-4.91	≤ .001	-.60
Risk judgement domain 'SCI'	32	48.98	33	17.50	16.50	-7.45	≤ .001	-.92
Risk judgement domain 'HAC'	33	45.30	33	21.70	155.00	-5.68	≤ .001	-.70
Risk judgement domain 'CM'	32	45.28	33	21.09	135.00	-5.70	≤ .001	-.70
Risk judgement domain 'P'	33	30.88	33	35.06	460.00	-.99	.32	-.12
Final structured risk judgement	33	47.00	33	20.00	99.00	-6.50	≤ .001	-.80

Note. $\bar{N}$ represents the number of cases included in the analysis. If no information was present (-99) for all the indicators within a certain domain, a risk judgement could not be assigned to that domain. As a result, the case was excluded from the analysis.

Discriminative Validity of the Total Scores

In table 6 the analyses regarding the total scores are displayed. First of all, significantly higher total scores are found for the terrorist offender group in comparison to the violent offender group on each of the VERA-2R domains. However, on the domain-level of the additional indicators, no significant differences are found between the terrorist offender group and violent offender group. With respect to the final total scores, both including and excluding the additional indicators, the results reveal higher total scores for the terrorist offenders in comparison to the violent offenders.

Table 6. Total scores

	Terrorist offenders		Violent offenders		Mann-Whitney U-test			
	N	Mean Rank	*N*	Mean Rank	*U*	*Z*	*p*	*r*
Domain 'BA'	33	44.41	33	22.59	184.50	-4.79	≤ .001	-.59
Domain 'SCI'	33	49.98	33	17.02	.50	-7.20	≤ .001	-.89
Domain 'HAC'	33	42.67	33	24.33	242.00	-3.91	≤ .001	-.48
Domain 'CM'	33	45.79	33	21.21	139.00	-5.29	≤ .001	-.65
Domain 'P'	32	43.05	33	23.26	206.50	-4.24	≤ .001	-.52
Final VERA-2R indicators*	33	49.30	33	17.70	23.00	-6.69	≤ .001	-.82
Domain 'CH'	32	27.34	32	37.66	347.00	-2.42	0.02	-.28
Domain 'PH'	30	28.93	29	31.10	403.00	-.51	.61	-.10
Domain 'MD'	27	27.59	27	27.41	362.00	-.04	.97	-.01
Additional indicators	33	30.65	33	36.35	450.50	-1.21	.23	-.15
Final VERA-2R & Additional indicators*	33	47.12	33	19.88	95.00	-5.77	≤ .001	-.71

Note. *N* represents the number of cases included in the analysis. If no information was present (-99) for all the indicators within a certain domain, a total score for that domain could not be calculated. As a result, the case was excluded from the analysis.

*If no information was present (-99) for all indicators within a certain domain, as said before, a total score for that domain could not be calculated. However, in order to be able to calculate a final total score, we used an imputation technique to deal with this missing value: we added the available total scores on a domain-level, divided this by the number of domains for which a total score could be calculated and multiplied it with the number of domains within the instrument. In order to determine whether the imputation had an effect on the results, we performed an additional analysis in which we included only the cases for which a total score could be calculated for all the domains. However, no significant differences were found in comparison to the original analysis, meaning that the imputation did not had an effect on the results.

5. Discussion Discriminative Validity

This part of the study aimed to get insight into the degree to which the VERA-2R is able to discriminate between groups that are known and expected to be different from each other. To this end, we examined the relevance of the risk and risk-mitigating indicators from the VERA-2R in a terrorist offender group and a violent offender group. In line with our hypothesis, since the VERA-2R states to incorporate indicators which are specifically related to the risk of terrorism and violent extremism, the results reveal significant higher scores for the terrorist offender group in comparison to the violent offender group on almost all the VERA-2R indicators. This indicates that the VERA-2R indicators are more relevant for terrorist offenders than for violent offenders. However, for the additional indicators the results revealed almost no significant differences between the terrorist offenders and the violent offenders. Yet, by following the statement that the additional indicators are not specifically related to the risk of terrorism and violent extremism, but can only contribute to a person's vulnerability to engage in future acts of ideologically motivated violence when combined with the presence of VERA-2R indicators, this finding was also in line with our expectations. However, although the additional indicators seem to be relevant for different types of violent offenders, one should bear in mind that these indicators still might play a different role in the terrorist offender group compared to the violent offender group in terms of the interaction with other risk- and protective indicators.

Combining the above results, it can be said that the VERA-2R incorporates indicators specifically related to the risk of terrorism and violent extremism, providing evidence for the fact that the VERA-2R is a specialized tool for ideologically motivated violence and endorsing the need for risk assessment instruments for special populations. However, in addition to this, the tool also includes indicators which seem to be pertinent to the engagement of violence in general.

Interpretations can be made to the different indicators of the domains. In the domain 'Beliefs, Attitudes and Ideology' (BA) 6 of the 7 indicators seem to be specifically related to the risk of terrorism and violent extremism. Only the indicator 'Expressed emotions in response to perceived injustice' (BA.5) appears to be equally relevant for terrorist offenders and violent offenders. This is an interesting finding, because indicator 'BA.5' often relates to the indicator 'Perceived grievances and/or injustice' (BA.2) in terrorist offenders (Duits, Alberda, & Kempes, 2022). In order to get more insight into this relationship, we examined for both groups the type of events that led to the perceived injustice. We found that for the terrorist offender group discrimination is mostly the cause of the perceived injustice, while for the violent offenders more general causes are found, such as a breakup or being treated unfairly by their parents. This illustrates that although terrorist offenders and violent offenders seem to experience emotions in response to perceived injustice to the same extent, the underlying cause for the perceived injustice differs.

In the domain 'Social Context and Intention' (SCI) 6 of the 7 indicators appear to be specifically relevant for terrorist offenders. However, we found significant higher scores for the indicator 'Target for attack identified' (SCI.2) for the violent offender group in comparison to the terrorist offender group. This could be explained by the fact that the

indicator focuses on whether the person concerned identifies a target in general, instead of specifically for an extremist attack. Therefore, this was not a surprising finding, since all the individuals within the violent offender group attacked someone or something, or at least threatened to do so, contrary to the individuals within the terrorist offender group. A shift in focus, by choosing specifically a target for an extremist attack as focal point, will directly be accompanied by a shift in the relationship for indicator 'SCI.2'.

In the domain 'History, Action and Capacity' (HAC) most of the indicators (5 of 6) have a clear relevance solely for terrorist offenders. The exception is the indicator 'Previous criminal violence' (HAC.3), which seems to be equally relevant for both terrorist offenders and violent offenders. This finding is in line with our expectations, since previous studies revealed that a violent criminal history is one of the strongest predictors for future violence in general. Therefore, previous criminal violence should be included in the risk assessments of both terrorist offenders and violent offenders. A quick inspection of the commonly used risk assessment tools for regular violence revealed that previous violent acts are incorporated in most, if not all, instruments (e.g., Douglas et al., 2014; Lodewijks, Doreleijers, & de Ruiter, 2008).

Moreover, half of the motivations within the 'Commitment and Motivation' (CM) domain are specifically related to the engagement in acts of violent extremism and terrorism. The other 4 motivations seem to be drivers for both ideologically motivated violence and regular violence: Motivated by criminal opportunism (CM.2), Motivated by excitement and adventure (CM.5), Forced participation (CM.6), and Motivation by acquisition of status (CM.7). As expected, since most of the terrorist offenders were supporters of the Salafi-jihadi ideology, the motivation 'Motivated by a perceived religious obligation and/or glorification' (CM.1) is found to be more relevant for terrorist offenders compared to violent offenders. However, quite possible due to the fact that terrorist offenders withhold information about their motives, it should be mentioned that little is known about the motives of the terrorist offenders, especially regarding the following indicators: 'Motivated by excitement and adventure' (CM.5), 'Motivated by acquisition of status' (CM.7), and 'Motivated by a search for meaning and significance in life' (CM.8). While little is known about the motives of the terrorist offenders, we found that a lot of the motives incorporated in the VERA-2R tool do not play a role in the involvement in acts of regular violence. A more detailed inspection of this finding revealed that the offences committed by violent offenders are often not preceded by clear motives or drivers, but are mostly an impulsive act in response to the situation. This finding is in line with the study of Bakker, Drost and Roeleveld (2010) which state that perpetraters of regular violence often do not have clear motives for their acts. According to them, physical violence often is the result of conflicted interactions or is impulsive in nature (Bakker, Drost, & Roeleveld, 2010). While interpreting the above findings, one should bear in mind that the encoding rules provided for the violent offenders regarding the indicators within the 'Commitment and Motivation' (CM) domain were not completely in line with the working procedure of the VERA-2R, since we shifted the focus from motives for violent extremism and terrorism to motives for regular violence for the violent offenders. We expect that this had a significant effect on the research findings, given that motives for terrorist and violent extremist acts are presumably not often applicable to the violent offender group.

Furthermore, we found that most of the indicators (4 of 6) within the 'Protective and risk-mitigating indicators' (P) domain appear to be specifically relevant for terrorist offenders. However, for both terrorist offenders and violent offenders support to relinquish violence is available to the same extent. Therefore, support of important others should be included in the risk assessments of both terrorist offenders and violent offenders. A quick inspection of the available risk assessment tools for regular violence revealed that social support is incorporated in most, if not all, instruments (e.g., Douglas et al., 2014; Lodewijks et al., 2008).

Lastly, with respect to the additional indicators, as said before, for most of the indicators related to criminal history, personal history and mental disorders (11 of 12) no significant differences are found between the terrorist offenders and the violent offenders. However, against our expectations, since previous studies found evidence for an important link between petty crime and ideologically motivated violence, the indicator 'Convicted for non-violent offences' (CH.1) seems to be more relevant for violent offenders. This indicates that violent offenders are more often convicted for non-violent offences in the past in comparison to terrorist offenders, which, in turn, provides evidence for the fact that attention should be paid to previous non-violent crimes while assessing the risk status of violent offenders. Luckily, a quick inspection of the commonly used risk assessment tools for regular violence revealed that previous non-violent acts are incorporated in most, if not all, instruments (e.g., Douglas et al., 2014; Lodewijks et al, 2008). Furthermore, it can be said that a lot of terrorist offenders and violent offenders have problems with school and/or work (PH.3) and that personality disorders are the most common mental disorders in both groups (MD.1).

Although our findings undoubtedly provide valuable insights into the discriminative validity of the VERA-2R, there are some limitations. The first limitation is that, although the research design simulates the practice of structured professional risk assessment as closely as possible, it needs to be defined as a research setting. In this study, the VERA-2R risk assessments were carried out by trained researchers on the basis of judicial files, while, in practice, these risk assessments are carried out by professionals who have experience in undertaking individual risk assessments, preferably on the basis of judicial files and direct interviews with the person concerned. Since assessors must use their professional judgement to arrive at the (final structured) risk judgements (Pressman et al., 2018), the use of research assessors can be criticized on the grounds of their ability to form adequate (final structured) risk judgements. However, we (partially) overcame this limitation by providing a two-day training course for the assessors, in which they acquired experience in forming (final structured) risk judgements. With regard to the exclusion of interviews, it's important to state that the inclusion of personal interviews is not a requirement for the use of the VERA-2R. If the person concerned is absent or refuses to co-operate, VERA-2R risk assessments can and will be carried out without the information that would be obtained from a direct interview (Pressman et al., 2018).

A second limitation pertains to the fact that the results are based on a small sample size. Besides this, due to some methodological issues, in both groups one extra case had to be excluded from the sample. As a result, we were not able to achieve the required sample size of 34. Therefore, follow-up research should be carried out to determine if the results can be replicated in a larger sample.

In addition to above limitations, we also have to acknowledge that, although we were able to match both groups in the most feasible way, the terrorist offenders and violent offenders could not always be perfectly matched. Furthermore, specifically the terrorist offender group seems to withhold information, which could have impacted the research findings. Above this, while (almost) all VERA-2R indicators are elaborately discussed in the judicial files of terrorist offenders, for the violent offenders, these indicators are almost never considered to be relevant. Therefore, we have to be aware that less concrete information was available for the violent offender group, which resulted in making more assumptions.

6. Methods Divergent Validity

Assessors and Cases

The assessor included in this study were three Dutch researchers with a master's degree in criminology or psychology. At the time of the study, the researchers were employed by the Netherlands Institute of Forensic Psychiatry and Psychology (NIFP) of the Dutch Ministry of Justice and Security. To obtain an in-dept understanding of the instruments and acquire experience in applying the indicators and forming (final structured) risk judgements, the assessors took part in a two-day training course for both the VERA-2R and HCR20^{V3}.

To assess the divergent validity, 67[2] cases of convicted terrorist offenders were selected to be coded with both the VERA-2R and HCR20^{V3}. These files were provided by the Dutch Public Prosecution Service, and could include a FMH assessment, a probation report, a transcript of the verdict, a police report, a criminal record and/or information from intelligence services.

In one case the assessor was aware that the person concerned re-engaged in a terrorist act in the future. Therefore, this case was excluded. This resulted in a final sample of 59 men and 7 women who were convicted of terrorist offences in the Netherlands between 2012 and 2019. The subjects' ages ranged from 15 – 59 at the time of their terrorist act (M_{age} = 28.26, *SD* = 9.89). Consequently, the sample appears to be quite representative of the target population with respect to age and gender, since similar descriptive statistics were found in previous studies on Dutch jihadists (Bakker & De Bont, 2016; Weenink, 2019). For most of the subjects the index crime was threatening (25.80%), followed by participation in a terrorist organization (15.20%) and preparation of murder with terrorist intent (13.60%).

Materials

VERA-2R.

A detailed description of the VERA-2R is included in the material section of the discriminative validity.

HCR20^{V3}.

The HCR20^{V3} can be used to assess the risk status of individuals who may pose a risk for future interpersonal violence (Douglas et al., 2014). The HCR20^{V3} contains 20 risk factors. Since they represent complex or nebulous constructs, some items include subitems. The HCR20^{V3} focuses on the past, present, and future through the inclusion of three scales: Historical Scale (*n* = 10), Clinical scale (*n* = 5), Risk management scale (*n* = 5). To evaluate the risk factors, a 3-point scale (no, possibly or partially, and yes) is used to indicate the extent of their presence in the individual case. For the purposes of this study we assigned the numerical scores '0', '1' and '2' to the ratings 'no', 'possibly or partly' and 'yes', respectively. Furthermore, we decided to assign the numerical score '-99' (missing value) if the judicial file did not contain information

[2] A-priori power analysis estimated that 67 cases are required.

about an indicator. Subsequently, evaluators should try to determine which of the risk factors that were rated as present or partially present may play a causal role in violence, at the individual level. Relevance is rated on a three-point scale: low, moderate, or high (Douglas et al., 2014).

After carefully considering the indicators, final structured risk judgements are made for 1) the risk of future violence generally, or case prioritization, 2) the risk for serious physical violence, and 3) the risk for imminent violence (Douglas et al., 2014). The final structured risk judgements are rated on a scale of low (0), moderate (1) and high (2). "Low" means the person is not considered in need of any special intervention or supervision strategies designed to manage violence risk, and that there is no need to monitor the person closely for changes in risk. "Moderate" means that the person requires some special management strategies, including at the very least, an increased frequency of monitoring. "High" suggests that there is a pressing need to develop a risk management plan for the person. Based on the final structured risk judgements, different risk scenarios are identified with a risk management strategy for each of these scenarios (Douglas et al., 2014).

Extensive research supports the validity and reliability of the HCR20^{V3} and its previous versions, including inter-rater reliability (Belfrage & Douglas, 2012; De Vogel, Van den Broek, & De Vries Robbé, 2014; Douglas & Belfrage, 2014; Doyle, Shaw & Coid, 2013), predictive validity (Douglas & Strub, 2013; Doyle, Shaw & Coid, 2013) and concurrent validity Belfrage & Douglas, 2012; Eidhammer, Selmer, & Bjørkly, 2013).

Research Design

Also, within this study, our research design included trained researchers and extensive judicial files, and therefore closely resembles risk assessments in practice.

With respect to the HCR20^{V3}, based on an advice given by one of the HCR20^{V3} authors, we decided to focus solely on the presence of the risk factors (and not on the relevance of the risk factors) and on the final structured risk judgement regarding the risk of future violence generally (and not on the final structured risk judgements regarding the risk for serious physical violence and the risk for imminent violence). Furthermore, the HCR20^{V3} author advised to only include the HCR20^{V3} main indicators (and not the subitems). Hereby, in line with the procedure of the HCR20^{V3}, the rating of the main indicators is determined by the highest rating given to one of the related sub-indicators.

Furthermore, it is important to state that the HCR20^{V3} is not specifically designed and not extensively investigated for violent extremism and terrorism (De Vogel, De Beuf, Shepherd, & Schneider, 2022). As a result, the instrument does not pay attention to how the HCR20^{V3} should be used within the terrorist offender population. More specifically, it remains unknown which terrorist acts should be seen as interpersonal violent acts. In order to deal with this lack of information as best as possible, after consultation with one of the HCR20^{V3} authors, we agreed to the use of two definitions of interpersonal violence. The first definition, referred to as 'All terrorist offences', is based on a broad interpretation of the definition of interpersonal violence used in the HCR20^{V3} and classifies all types of terrorist acts as

interpersonal violent acts. The second definition, referred to as 'Specific terrorist offences', is based on a strict interpretation of the definition of interpersonal violence used in the HCR20^{V3} and classifies specific types of terrorist acts as interpersonal violent acts: (attempt of) threatening, (attempt of/preparation of) murder, preparation of manslaughter, preparation of a terrorist offence, (attempt of/preparation of) arson with danger for other individuals as consequence, (attempt of/preparation of) traveling to conflict zones. Using different definitions first of all had an effect on the total score of the 'Historical' (H) domain and on the final total score of the HCR20^{V3}, since the item 'History of problems with violence' had to be assessed in a different way while using the different definitions of interpersonal violence. Secondly, it had an effect on the final structured risk judgement, since the estimation of the likelihood that an individual will engage in future acts of interpersonal violence depends on which type of acts are classified as interpersonal violent acts.

In order to determine whether the use of multiple assessors would impact the research findings, we established the interrater reliability between assessor-1 (gold standard) and assessor-2, and between assessor-1 (gold standard) and assessor-3. This enabled us to: a) evaluate the extent to which the assessors score the same ratings for the feature that is being observed or measured, and b) evaluate whether assessor-2 and assessor-3 are able to rate the VERA-2R and HCR20^{V3} indicators in an adequate way. The reliability analyses were based on 5 testcases of terrorist offenders, which were assessed with both the VERA-2R and HCR20^{V3}.

The distributions of the observed ratings frequently fell under one category of ratings. As a result, kappa estimates appeared to be unrepresentatively low (Eugenio & Glass, 2004). Therefore, an alternative kappa was calculated based on the percentage of agreement between the evaluators, and corrected for agreement based merely on chance, which depends on the number of answer options available. To establish the strength of the agreement, Landis & Koch's cut-off points were used: $\kappa \leq .20$ = slight, $.20 < \kappa \leq .40$ = fair, $.40 < \kappa \leq .60$ = moderate, $.60 < \kappa \leq .80$ = good and $.80 < \kappa \leq .1.00$ = excellent (Landis & Koch, 1977).

Focusing on the VERA-2R, the interrater reliability between assessor-1 and assessor-2 ($\kappa = .73$) and the interrater reliability between assessor-1 and assessor-3 ($\kappa = .70$) can be classified as good**.** With respect to the HCR20^{V3}, two different kappa-values were calculated, since two different definitions of interpersonal violence were used. For both assessor-2 ($\kappa_{\text{All terrorist offences}} = .71$ $\kappa_{\text{Specific terrorist offences}} = .70$) and assessor-3 ($\kappa_{\text{All terrorist offences}} = .73$, $\kappa_{\text{Specific terrorist offences}} = .79$) the calculated kappa-values can be classified as good.

The results above indicate that the assessments are reasonably independent of the raters or professional assessors (Jonsson & Svingby, 2007). Furthermore, it indicates that all the assessors are able to evaluate the VERA-2R and HCR20^{V3} indicators in an adequate way. Based on these promising results, we agreed to the use of multiple raters. As a result, assessor-1 evaluated all the cases of the terrorist offenders with the VERA-2R, and assessor-2 and assessor-3 evaluated teamwise all the cases of the terrorist offenders with the HCR20^{V3}.

Although we generally found high interrater reliability, we also found indicators with an interrater reliability equal to or below the critical value of .60. Since we agreed to the use of multiple raters, we had to make sure that this

would not have an impact on the research findings. Therefore, all the indicators that were found to have low interrater reliability were elaborately discussed beforehand. Furthermore, with respect to the HCR20^{V3}, a consensus rating was assigned to the indicators with a low interrater reliability, based on the perspectives of both assessor-2 and assessor-3. Moreover, halfway through the data collection phase of the HCR20^{V3}, assessor-2 and assessor-3 both assessed the same six cases in order to evaluate the state of affairs. The outcoming results were subsequently used for training purposes (e.g., additional explanation) in order to increase the level of agreement between the assessors.

With respect to the (final structured) risk judgements, we dealt with the use of multiple raters by letting assessor-1 verify all the (final structured) risk judgements that assessor-2 and assessor-3 have assigned.

Security and Privacy

To ensure that there were no risks to the privacy of the subjects, we anonymized the data. Moreover, with regard to data protection, we stored the anonymized dataset in a secure digital environment, in order to protect the information against misuse, unauthorized access, disclosure and theft.

Statistical Analysis

Before analysing the data using IBM SPSS Statistics for Mac Version 25.0, some data processing steps had to be taken. First of all, reverse coding was applied to the 'Protective and risk-mitigating indicators' (P) domain of the VERA-2R, so that higher ratings correspond to a higher risk status. Secondly, total scores had to be calculated for research purposes. Hereby it is important to stress that total scores have no relevance in professional SPJ-practice. To calculate the total scores of the domains, we added the numerical scores assigned to the indicators representing the concerned domain, divided this by the number of indicators for which a rating was present (i.e., excluding the indicators for which no information was available), and multiplied it with the number of indicators representing the concerned domain. To calculate the final total scores, we first of all added the total scores of the domains. Secondly, in order to be able to compare the final total score of the VERA-2R with the final total score of the HCR20^{V3}, this score was divided by the number of indicators within the concerned instrument. Hereby it is important to mention that for the VERA-2R two different final total scores were calculated: 1) a final total score solely based on the ratings assigned to the VERA-2R indicators 2) a final total score based on the ratings assigned to both the VERA-2R indicators and the additional indicators.

To establish whether the risk measured by the VERA-2R differs significantly from the risk measured by the HCR20^{V3}, multiple correlation analyses were performed. Due through violation of the assumptions, Kendall's tau-b was chosen as correlation coefficient. A p-value of $< .05$ is considered statistically significant and the effect sizes were interpreted in accordance with the guidelines outlined by Botsch (2011): $\tau b < .10$ = very weak, $.10 \leq \tau b < .20$ = weak, $.20 \leq \tau b < .30$ = moderate, $\tau b \geq .30$ = strong.

In addition to this, depending on whether or not assumptions for the test were violated, we performed a Paired sample t-test, a Wilcoxon signed-rank test or a Sign test with the final total scores and final structured risk judgements as dependent variables and type of instrument (VERA-2R or HCR20^{V3}) as independent variable. A *p*-value of < .05 is considered statistically significant. For the Paired sample t-test Cohen's *d* was calculated to determine the strength of the relationship. The guidelines of Cohen (1988) were used to interpret the effect sizes: the effect size is classified as small if *d* = .02, the effect size is classified as medium if *d* = .05, and the effect size is classified as large if *d* = .08. Furthermore, for the Wilcoxon signed-rank test and the Sign test we calculated the correlation coefficient effect size (*r*). These effect sizes were interpreted in accordance with the following guidelines, also outlined by Cohen (1988): .10 < *r* < .30 = weak, .30 < *r* < .50 = moderate, *r* ≥ .50 = strong.

7. Results Divergent Validity

Correlations Total Scores

In order to determine whether the VERA-2R assesses a different type of risk than the HCR20^{V3}, multiple correlation analyses were carried out, starting with the correlations between the total scores of the VERA-2R and the total scores of the HCR20^{V3} (see Table 7). This with the realization that the HCR20^{V3} is not specifically designed and not extensively investigated for violent extremism and terrorism (De Vogel., et al., 2022).

The results first of all illustrate that in general the *domains* of the VERA-2R do not correlate significantly with the *domains* of the HCR20^{V3}, with the exception of the significant weak negative correlation between the 'Commitment and Motivation' (CM) domain of the VERA-2R and the 'Clinical' (C) domain of the HCR20^{V3}, and the significant weak positive correlation between the 'Protective and risk-mitigating indicators' (P) domain of the VERA-2R and the 'Risk' (R) domain of the HCR20^{V3}.

On the other side, a significant positive correlation was found between almost all *additional domains* of the VERA-2R and the *domains* of the HCR20^{V3}, meaning that higher ratings on the additional indicators are associated with higher ratings on the HCR20^{V3} indicators. Only the correlation between the 'Criminal History' (CH) domain of the VERA-2R and the 'Clinical' (C) domain of the HCR20^{V3} was found to be not significant.

With respect to the *final total scores*, it can be said that if the final total score of the VERA-2R is only based on the VERA-2R indicators, the results reveal no significant correlations between the final total score of the VERA-2R and the final total scores of the HCR20^{V3}. However, when the additional indicators are included in the final total score of the VERA-2R significant weak positive correlations are found between the final total score of the VERA-2R and the final total scores of the HCR20^{V3}, meaning that these variables tend to increase together. Furthermore, a more detailed inspection of these relationships revealed that the final total scores of the VERA-2R are significantly lower than the final total scores of the HCR20^{V3}. This result is found both when the additional indicators are excluded from the final total score of the VERA-2R (VERA-2R & HCR20^{V3} -All terrorist offences: $t(65) = -6.30$, $p \leq .001$, $d = -.78$; VERA-2R & HCR20^{V3}–Specific terrorist offences: $t(65) = -6.30$, $p \leq .001$, $d = -.77$), and when the additional indicators are included in the final total score of the VERA-2R (VERA-2R & HCR20^{V3}-All terrorist offences: $Z = -6.00$, $p \leq .001$, $r = -.74$; VERA-2R & HCR20^{V3}–Specific terrorist offences: $Z = -6.00$, $p \leq .001$, $r = -.74$).

Table 7. Correlations total scores

	HCR20^{V3}					
	Domain 'H' – All terrorist offences	Domain 'H' – Specific terrorist offences	Domain 'C'	Domain 'R'	Final HCR20^{V3} indicators - All terrorist offences	Final HCR20^{V3} indicators - Specific terrorist offences
VERA-2R						
Domain 'BA'	.11	.10	.13	.15	.15	.14
Domain 'SCI'	-.04	-.05	.01	-.02	-.01	-.01
Domain 'HAC'	.09	.11	.03	.05	.08	.09
Domain 'CM'	-.00	-.04	-.22*	-.10	-.10	-.12
Domain 'P'	.11	.11	.04	.23*	.15	.15
Domain 'CH'	.44***	.44***	.19	.32***	.38***	.37***
Domain 'PH'	.38***	.38***	.25*	.36***	.39***	.40***
Domain 'MD'	.31**	.32**	.27**	.35***	.36***	.37***
Final VERA-2R indicators	.09	.08	.01	.09	.10	.09
Final VERA-2R & Additional indicators	.22*	.21*	.08	.20*	.21*	.20*

Note. *$p \leq .05$. **$p \leq .01$. ***$p \leq .001$ (two-tailed).

If no information was present (-99) for all indicators within a certain domain, a total score for that domain could not be calculated. As a result, the case was excluded from the analysis. However, in order to be able to calculate a final total score, we used an imputation technique to deal with this missing value: we added the available total scores on a domain-level, divided this by the number of domains for which a total score could be calculated and multiplied it with the number of domains within the instrument.

Reverse coding is applied to the 'Protective and risk-mitigating indicators' (P) domain. As a result, higher ratings correspond with a higher risk status.

Correlations (Final Structured) Risk Judgements

We also examined the correlations between the (final structured) risk judgements of the VERA-2R and HCR20^{V3} (see Table 8). We found that all the *risk judgements* of the risk-increasing domains of the VERA-2R do not correlate significantly with the *final structured risk judgement* of the HCR20^{V3}, with the exception of the risk judgement belonging to domain 'Beliefs and Attitudes' (BA) for which a significant weak positive correlation was found with the final structured risk judgement of the HCR20^{V3}. While in general insignificant correlations are found for the risk-increasing domains, we found only significant correlations for the relationship between the *risk judgement* of the risk-reducing domain of the VERA-2R (Protective and risk-mitigating indicators (P)) and the *final structured risk judgements* of the HCR20^{V3}, indicating that these variables tend to increase together. Hereby it is good to state that a positive correlation is found, because, for research purposes, reverse coding is applied to the domain 'Protective and risk-mitigating indicators' (P). However, originally this relationship will be negative, since the presence of protective indicators reduces the risk status of an individual.

Furthermore, the results illustrate that the *final structured risk judgement* of the VERA-2R correlates significantly and positively with the *final structured risk judgement* of the HCR20^{V3}, but only when the final structured risk judgement of the HCR20^{V3} includes all types of terrorist offences while determining whether an individual will engage in future acts of ideologically motivated violence. A more detailed inspection of these relationships revealed that there are no significant differences between the final structured risk judgements assigned on the basis of the VERA-2R and the final structured risk judgements assigned on the basis of the HCR20^{V3}. This result is found both when the HCR20^{V3} focuses on all types of terrorist offences ($Z = -1.95$, $p = .05$, $r = -.24$), and when the instrument focuses solely on specific terrorist offences ($Z = -.22$, $p = .82$, $r = -.03$).

Table 8. Correlations (final structured) risk judgements

	Final structured risk judgement HCR20^{V3}– All terrorist offences	Final structured risk judgement HCR20^{V3} – Specific terrorist offences
Risk judgement domain 'BA'	.25*	.07
Risk judgement domain 'SCI'	-.08	.22
Risk judgement domain 'HAC'	.07	.06
Risk judgement domain 'CM'	.11	-.09
Risk judgement domain 'P'	.33**	.26*
Final structured risk judgement VERA-2R	.23*	.11

Note. *$p \leq .05$. **$p \leq .01$. ***$p \leq .001$ (two-tailed).
If no information was present (-99) for all the indicators within a certain domain, a risk judgement could not be assigned to that domain. As a result, the case was excluded from the analysis.
Reverse coding is applied to the domain 'Protective and risk-mitigating indicators' (P). As a result, higher ratings correspond with a higher risk status.

8. Discussion Divergent Validity

This part of the study aimed to get insight in whether the VERA-2R assesses a different type of risk than the HCR20^{V3}, while bearing in mind that the HCR20^{V3} is not specifically designed and not extensively investigated for violent extremism and terrorism (De Vogel et al., 2022). In sum, we could state that the results regarding the final total scores and final structured risk judgements give the impression that both instruments can be used to determine the risk status of terrorist offenders. However, by inspecting the results on a domain-level, it becomes clear that the concepts represented by the VERA-2R indicators are not incorporated in the HCR20^{V3}. Given the fact that the concepts represented by the VERA-2R indicators should be incorporated in risk management strategies in order to be able to accurately manage the risk, we must conclude that the risk status of terrorist offenders can only correctly be determined by the VERA-2R (or when investigated by another specialized tool for ideologically motivated violence).

Above conclusions can first of all be supported by the results concerning the final total scores. In line with our expectations, since the VERA-2R and HCR20^{V3} indicators aim to measure different concepts, the results illustrate that no significant correlation is found between the final total score of the VERA-2R and the final total score of the HCR20^{V3} if the final total score of the VERA-2R only includes the original VERA-2R indicators. This indicates that high ratings on the VERA-2R indicators are not associated with high ratings on the HCR20^{V3} indicators. However, when the final total score of the VERA-2R includes the additional indicators, higher ratings on the indicators within the VERA-2R tool are associated with higher ratings on the HCR20^{V3} indicators. This relationship is largely based on the association between the additional indicators and the HCR20^{V3} indicators.

When comparing the final total scores of the VERA-2R with the final total scores of the HCR20^{V3}, we found that the final total score of the HCR20^{V3} is generally significantly higher than the final total score of the VERA-2R, both when including and excluding the additional indicators within the final total score of the VERA-2R. This may illustrate that the HCR20^{V3} indicators are more often present than the VERA-2R indicators within the terrorist offender population. This could be due to the fact that the HCR20^{V3} includes more general indicators which are relevant for different types of violent offenders, among which terrorist offenders. Examples of these risk indicators are 'History of problems with violence', 'Violent attitudes' and '(symptoms of) A mental disorder'. Contrary to this, the VERA-2R risk- and protective indicators are specifically related to terrorism and violent extremism. This enables not only proper risk assessment and risk management, but also research into risks between different types of terrorist offenders, e.g., between terrorist offenders who commit homicide and terrorist offenders who commit another offence (Alberda, Duits, Van den Bos, Autsema, & Kempes, 2022). Above this, the HCR20^{V3} focuses on the past situation of an individual (Historical domain), the present situation (Clinical domain) and the future situation (Risk Management domain), whereby a couple of relevant indicators are incorporated in all of these three domains. If these indicators are expected to be present during different time periods, they will be included in the final total score multiple times, which quickly can lead to a high final total score.

Despite the fact that interesting results are found in regard to the final total scores, we must bear in mind that both the VERA-2R and HCR20^{V3} adhere to the SPJ methodology. Therefore, special attention has to be paid to the relationship between the final structured risk judgements of the instruments. The results first of all show us that when only specific terrorist acts are classified as interpersonal violent acts while using the HCR20^{V3}, we found no significant correlation and no significant differences between the final structured risk judgement based on the VERA-2R and the final structured risk judgement based on the HCR20^{V3}. In order to get more insight in this relationship, we examined it more in depth. We found that in some of the cases a high final structured risk judgement on the basis of the VERA-2R is associated with a low final structured risk judgement on the basis of the HCR20^{V3}, and in other cases the exact opposite relation is found. This could explain why no significant correlation is found, and why on average the instruments will arrive at the same level of final structured risk judgement.

However, when we assume for the sake of the argument that the VERA-2R and HCR20^{V3} both relate to all types of terrorist acts, because the HCR20^{V3} classifies them all as interpersonal violent acts, the results reveal a positive significant correlation between the final structured risk judgement of the VERA-2R and the final structured risk judgement of the HCR20^{V3}. This indicates that when the risk status of an individual is assessed as high on the basis of the VERA-2R in general this will also be the case on the basis of the HCR20^{V3}. However, again, no significant differences were found between the final structured risk judgements, meaning that when the VERA-2R and HCR20^{V3} both relate to all types of terrorist acts, the instruments also will arrive at the same risk status for an individual.

Although the finding that both instruments will arrive at the same level of final structured risk judgement can give the impression that both instruments can be used to determine the risk status of terrorist offenders, the results on a domain-level illustrate why this idea is inaccurate. In line with our expectations, since the results related to the discriminative validity pointed out that the additional indicators seem to be risk indicators for violence in general, we found that the additional domains of the VERA-2R seem to correlate significantly with the domains of the HCR20^{V3}. This gives the impression that the concepts measured by the additional domains are also represented in the HCR20^{V3}. An inspection of the indicators related to the concerned domains verifies this impression. Despite the fact that the concepts of the additional indicators seem to be incorporated in the HCR20^{V3}, one should bear in mind that these indicators still might play a different role within the VERA-2R tool compared to the HCR20^{V3} tool in terms of the interaction with the other risk- and protective indicators.

However, contrary to the results related to the additional domains, we found that the VERA-2R domains in general do not correlate significantly with the HCR20^{V3} domains. This indicates that the concepts represented by the VERA-2R domains are not incorporated in the HCR20^{V3} and therefore are distinctive for the VERA-2R. This, in turn, illustrates that the VERA-2R assesses a different type of risk than the HCR20^{V3}, and provides evidence for its relative uniqueness (Holton et al., 2007). Since the VERA-2R indicators are supported empirically by research to be relevant for the risk of terrorism and violent extremism (Monahan, 2012; Pressman et al., 2018; Sarma, 2017), it is important that these indicators are incorporated in risk assessments for terrorist offenders. Therefore, although both instruments seem to

arrive at the same level of final structured risk judgement, the HCR20^{V3} cannot be used to determine the risk status of violent extremists and terrorists.

The above mentioned is especially the case within in the context of risk scenarios and risk management. Within risk scenarios all the relevant information related to the indicators is elaborately discussed, resulting in a substantiation of the risk status. This in turn provides essential information for risk management. Since the indicators within the VERA-2R seem to measure different concepts than the indicators within the HCR20^{V3}, logically, both instruments will arrive at different risk scenarios and risk management strategies. However, given that the VERA-2R indicators are supported empirically by research to be relevant for the risk of terrorism and violent extremism (Monahan, 2012; Pressman et al., 2018; Sarma, 2017), in order to be able to accurately manage the risks of terrorist offenders, risk management strategies should be based on the VERA-2R indicators.

Although our findings undoubtedly provide valuable insights into the divergent validity of the VERA-2R, there are some limitations. An important limitation, which is also mentioned in relation to the discriminative study, is that although the research design simulates the practice of structured professional risk assessment as closely as possible, it still needs to be defined as a research setting since it differs in terms of type of assessor and type of information source used. Therefore our findings cannot be generalized to risk assessments in practice.

A second limitation pertains to the fact that the results are based on a relatively small sample size. Besides this, due to some methodological issues, one extra case had to be excluded from the sample. As a result, we were not able to achieve the required sample size of 67. Therefore, follow-up research should be carried out to determine if the results can be replicated in a larger sample.

9. Overall Discussion

Given that risk assessments have a significant role to play in the fight against violent extremism and terrorism (European Union, 2022; National Coordinator for Counterterrorism, 2022), it is important that risk assessment tools for ideologically motivated violence provide reliable and valid risk assessments. The present study investigated the validity of the VERA-2R, focusing on the discriminative- and divergent validity. In line with our hypothesis, since the VERA-2R states to incorporate indicators which are specifically related to the risk of terrorism and violent extremism, the results reveal that in general the VERA-2R indicators are more relevant for terrorist offenders than for violent offenders. However, the additional indicators seem to be equally relevant for terrorist offenders and violent offenders. Yet, by following the statement that the additional indicators are not specifically related to the risk of terrorism and violent extremism, but can only contribute to a person's vulnerability to engage in future acts of ideologically motivated violence when combined with the presence of VERA-2R indicators, this finding was also in line with our expectations. With respect to the divergent validity, in sum we could state that the results regarding the final total scores and final structured risk judgements give the impression that both instruments can be used to determine the risk status of terrorist offenders. However, by inspecting the results on a domain-level, it becomes clear that in general the concepts represented by the VERA-2R indicators are not incorporated in the HCR20^{V3}. Given the fact that the concepts represented by the VERA-2R indicators should be incorporated in risk management strategies in order to be able to accurately manage the risk, we must conclude that the risk status of terrorist offenders can only correctly be determined by the VERA-2R (or when investigated by another specialized tool for ideologically motivated violence).

In addition to clarifying the validity of the VERA-2R, the present study also showed how the validity of a structured professional risk assessment tool can be investigated with trained assessors based on extensive judicial files. Furthermore, the results of the present study can, in combination with the results of the reliability study, be used to make well-founded recommendations on how to improve the VERA-2R. As with most research, the obtained knowledge can be deepened and strengthened by carrying out further research. In addition to this, it is also necessary to strengthen the empirical foundations of the VERA-2R. Due to both the limited access to (primary) data (Sageman, 2014) and the ethical barriers in conducting research on sensitive topics (Horgan, 2012), the evidence-base underpinning the risk-promoting and risk-mitigating indicators for ideologically motivated violence is scant at best (Sarma, 2017). In order to obtain a more evidence-based professional approach to conducting violent extremism and terrorism risk assessments, the European Database of Convicted Terrorists (EDT) was developed (Alberda et al., 2021). The EDT is based on judicial documents and contains personal and contextual information about convicted (and deceased) terrorists and violent extremists. By analysing this data, relevant insights could be obtained into the underlying indicators that drive individuals' engagement, continuation or disengagement in violent extremism and terrorism. This would enable the validation of the VERA-2R indicators on the basis of a larger sample size. Furthermore, this will lead to the identification of other relevant indicators vis-à-vis the risk of ideologically motivated violence. Lastly, it is of utmost importance to

establish the predictive validity of the VERA-2R, since this is considered to be the gold standard for forensic risk assessment instruments (Lloyd & Dean, 2015). Upon completion of this research more insight could be obtained on how to further improve the VERA-2R, which, in turn, will lead to more accurate violent extremism and terrorism risk assessments and risk management strategies, and, most importantly, a safer society.

10. References

Alberda, D., Duits, N., Van den Bos, K., Ayanian, A. H., Zick, A., & Kempes, M. (2021). The European Database of Terrorist Offenders (EDT). *Perspectives on Terrorism, 15*(2), 77 – 99.

Alberda, D.L., Duits, N., van den Bos, K., Autsema, A. & Kempes, M., (2022). Identifying risk factors for Jihadist terrorist offenders committing homicide: An explorative analysis using the European Database of Terrorist offenders. *Frontiers in Psychology*, 13. DOI 10.3389/fpsyg.2022.1000186.

Bakker, E., & De Bont, R. (2016). Belgian and Dutch jihadist foreign fighters (2012–2015): Characteristics, motivations, and roles in the war in Syria and Iraq. *Small Wars & Insurgencies, 27*(5), 837 – 857. https://doi.org/10.1080/09592318.2016.1209806

Bakker, I., Drost, L., & Roeleveld, W. (2010). *Wat hebben geweldplegers gemeen? Een typologie van plegers van geweld tegen de publieke taak en van publiek geweld.* Utrecht: Verwey-Jonker Instituut.

Belfrage, H. & Douglas, K. S. (2012). *Interrater reliability and concurrent validity of HCR-20 Version 3.* Unpublished data analyses. Mid-Sweden University, Sundsvall, Sweden.

Botsch, R. E. (2011). *Significance and measures of association.* Retrieved from http://www.usca.edu/polisci/apls301/Text/Chapter%2012.%20Significance%20and%20Measures%20of%20Association.htm

Bower, K. M. (2003). When to use Fisher's exact test. *American Society for Quality, Six Sigma Forum Magazine, 2*(4), 35 – 37.

Bruin, A. A., Duits, N., Kempes, M., & Prinsen, M. (2023). *Interrater and Intrarater Reliability of the Violent Extremism Risk Assessment tool.* Article in preparation.

Clarke, R. V. G., & Newman, G. R. (2006). *Outsmarting the terrorists.* Portsmouth, N.H: Greenwood Publishing Group.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, N.J: L. Erlbaum Associates.

De Vogel, V., De Beuf, T., Shepherd, S., & Schneider, R. D. (2022). Violence risk assessment with the HCR-20V3 in legal contexts: *A critical reflection. Journal of personality assessment, 104*(2), 252 – 264.

De Vogel, V., Van den Broek, E., & De Vries Robbé, M (2014). The use of the HCR-20V3 in Dutch forensic psychiatric practice. *International Journal of Forensic Mental Health, 13*(2), 109 – 121. https://doi:10.1080/14999013.2014.906518

Duits, N., Alberda, D. L., & Kempes, M. (2022). Psychopathology of Young Terrorist Offenders, and the Interaction With Ideology and Grievances. *Frontiers in Psychiatry, 13*, 1-14. https://doi.org/10.3389/fpsyt.2022.801751.

Douglas, K., & Belfrage, H (2014). Interrater reliability and concurrent validity of the HCR-20 Version 3. *International Journal of Forensic Mental Health, 13*(2), 130 – 139.

Douglas, K. S., Hart, S. D., Webster, C. D., Belfrage, H., Guy, L. S., & Wilson, C. M. (2014). Historical-clinical-risk management-20, version 3 (HCR-20V3): Development and overview. *International Journal of Forensic Mental Health, 13*(2), 93 – 108. https://doi.org/10.1080/14999013.2014.906519

Douglas, K. S., & Strub, D. S. (2013). *Predictive validity of HCR-20 (Version 3) amongst civil psychiatric patients and criminal offenders.* Unpublished data analyses. Simon Fraser University, Burnaby, Canada

Doyle, M., Shaw, J., & Coid, J. (2013). *Validation of risk assessments in medium secure services (VORAMSS): Preliminary HCR-20 (Version 3) analyses.* Unpublished data analyses. University of Manchester, UK, and Queen Mary University of London, UK.

Eidhammer, G., Selmer, L. E., & Bjørkly, S. (2013). *A brief report of a comparison of the HCR-20 (Version 2) and HCR-20 (Version 3) concerning internal consistency and clinical utility.* Centre for Research and Education in Forensic Psychiatry, Oslo University Hospital, Norway.

Eugenio, B. D., & Glass, M. (2004). The kappa statistic: A second look. *Computational linguistics, 30*(1), 95– 101.

European Commission. (2020). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on the EU Security Union Strategy.* COM(2020) 605 final. Retrieved from https://eur-lex.europa.eu/legalcontent/EN/TXT/PDF/?uri=CELEX:52020DC0605&

European Union (2022). *European Union Terrorism Situation and Trend Report (Te-Sat) 2022.* Retrieved from https://www.europol.europa.eu/publication-events/main-reports/european-union-terrorism-situation-and-trend-report-2022-te-sat

Field, A. P. (2005). *Discovering statistics using SPSS* (2nd ed.). London, England: Sage

Gouttebarge, V. (2008). *Quality of Functional Capacity Evaluation Tests: A Clinimetric Approach.* Amsterdam: University of Amsterdam

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: John Wiley & Sons.

Holton III, E. F., Bates, R. A., Bookter, A. I., & Yamkovenko, V. B. (2007). Convergent and divergent validity of the learning transfer system inventory. *Human Resource Development Quarterly, 18*(3), 385 – 419.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130 - 144. https://doi.org/10.1016/j.edurev.2007.05.002

LaFree, G., & Dugan, L. (2004). How does studying terrorism compare to studying crime. *Terrorism and counter-terrorism: Criminological perspectives, 5*, 53 – 74.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*(2), 363 – 374. https://doi.org/10.2307/2529786

Lloyd, M., & Dean, C. (2015). The development of structured guidelines for assessing risk in extremist offenders. *Journal of Threat Assessment and Management, 2*(1), 40 – 52.

Lodewijks, H. P. B., Doreleijers, T. A. H., & De Ruiter, C. (2008). SAVRY risk assessment in violent Dutch adolescents: Relation to sentencing and recidivism. *Criminal Justice and Behavior, 35*(6), 696 – 709.

Logan, C. (2017). Reporting structured professional judgement. In E. Bowen, S. Brown & D. Prescott (Eds.), *The forensic psychologist's report writing guide* (pp. 82 – 93). Chichester, UK: Wiley Blackwell.

Ministry of Justice and Security. (2021a, 22 september). *Research. Vera-2r*. Retrieved from https://www.vera-2r.nl/research

Ministry of Justice and Security. (2021b, mei 10). *VERA-PRO2EU. Research | Vera-2r*. Retrieved from https://www.vera-2r.nl/research/vera-pro2eu

Monahan, J. (2012). The individual risk assessment of terrorism. *Psychology, Public Policy, and Law, 18*(2), 167–205. https://doi.org/10.1037/a0025792

National Coordinator for Counterterrorism (2022). *National Counterterrorism Strategy for 2022-2026.* Retrieved from https://english.nctv.nl/documents/publications/2022/06/the-national-counterterrorism-strategy-for-2022-2026/the-national-counterterrorism-strategy-for-2022-2026

Pressman, D. E., Duits, N., Rinne, T., & Flockton, J. S. (2018). VERA– 2R Violent Extremism Risk Assessment – Version 2 Revised: A structured professional approach. Utrecht, The Netherlands: The Netherlands Institute of Forensic Psychiatry and Psychology (NIFP).

Sageman, M. (2014). The stagnation in terrorism research. *Terrorism and Political Violence, 26*(4), 565 – 580. https://doi.org/10.1080/09546553.2014.895649

Sarma, K. M. (2017). Risk assessment and the prevention of radicalization from nonviolence into terrorism. *American Psychologist, 72*(3), 278 – 288. https://doi.org/10.1037/amp0000121

Silva, E. (2020). The HCR-20 and violence risk assessment–will a peak of inflated expectations turn to a trough of disillusionment?. *BJPsych bulletin, 44*(6), 269 – 271.

Smith, C., & Nolan, M. (2016). Post-sentence continued detention of high-risk terrorist offenders in Australia. *Criminal Law Journal, 40*(3), 163 – 179.

Van der Heide, L., Van der Zwan, M., & Van Leyenhorst, M. (2019). *The Practitioner's guide to the galaxy - A comparison of risk assessment tools for violent extremism.* Geraadpleegd van https://icct.nl/app/uploads/2019/09/29Aug19_Formatted_ThePractitionersGuideto theGalaxy-2.pdf

Weenink, A. W. (2019). Adversity, criminality, and mental health problems in Jihadis in Dutch police files. *Perspectives on Terrorism, 13*(5), 130 – 142.

11. Appendix

Appendix A - Encoding rules for violent offenders

Domain and Indicator	Explanation
BA. Beliefs, Attitudes and Ideology	
BA.1 Commitment to ideology that justifies violence	Assessed in same way as for terrorist offenders
BA.2 Perceived grievances and/or perceived injustice*	Assessed in same way as for terrorist offenders
BA.3 Dehumanization of designated targets associated with injustice*	Assessed in same way as for terrorist offenders
BA.4 Rejection of democratic society and values	Assessed in same way as for terrorist offenders
BA.5 Expressed emotions in response to perceived injustice*	Assessed in same way as for terrorist offenders
BA.6 Hostility to national identity	Assessed in same way as for terrorist offenders
BA.7 Lack of empathy and understanding for those outside one's own group	Assessed in same way as for terrorist offenders
SCI. Social Context and Intention	
SCI.1 Seeker, user or developer of violent extremist materials	Assessed in same way as for terrorist offenders
SCI.2 Target for attack identified (person, group, location)*	Focus on whether the person concerned identifies a target for a violent attack instead of for a violent extremist attack
SCI.3 Personal contact with violent extremists*	Assessed in same way as for terrorist offenders
SCI.4 Expressed intention to commit acts of violent extremism	Assessed in same way as for terrorist offenders
SCI.5 Expressed willingness and/or preparation to die for a cause or belief*	Assessed in same way as for terrorist offenders. The item is rated as 'Low' if the person concerned has no cause or belief.
SCI.6 Planning, preparation of acts of violent extremism	Assessed in same way as for terrorist offenders
SCI.7 Susceptibility to influence, control or indoctrination*	Assessed in same way as for terrorist offenders. The item is rated as 'Low' if the person concerned is not exposed to a leader or person that advocates acts of violent extremism or terrorism
HAC. History, Action and Capacity	
HAC.1 Early exposure to violence-promoting, militant ideology	Assessed in same way as for terrorist offenders
HAC.2 Network of family and friends involved in violent extremism	Assessed in same way as for terrorist offenders
HAC.3 Violent criminal history	Assessed in same way as for terrorist offenders
HAC.4 Strategic, paramilitary and/or explosives training	Assessed in same way as for terrorist offenders
HAC.5 Training in extremist ideology in own country or abroad*	Assessed in same way as for terrorist offenders
HAC.6 Organizational skills and access to funding and sources of help	Assessed in same way as for terrorist offenders
CM. Commitment and Motivation	

CM.1 Motivated by perceived religious obligation and/or glorification	Focus on motives for regular violence instead of motives for ideologically motivated violence
CM.2 Motivated by criminal opportunism*	Focus on motives for regular violence instead of motives for ideologically motivated violence
CM.3 Motivated by camaraderie, group belonging	Focus on motives for regular violence instead of motives for ideologically motivated violence
CM.4 Motivated by moral obligation, moral superiority*	Focus on motives for regular violence instead of motives for ideologically motivated violence
CM.5 Motivated by excitement and adventure*	Focus on motives for regular violence instead of motives for ideologically motivated violence
CM.6 Forced participation	Focus on motives for regular violence instead of motives for ideologically motivated violence
CM.7 Motivated by acquisition of status*	Focus on motives for regular violence instead of motives for ideologically motivated violence
CM.8 Motivated by a search for meaning and significance in life*	Focus on motives for regular violence instead of motives for ideologically motivated violence
P. Protective and risk-mitigating indicators	
P.1 Reinterpretation of the ideology*	Assessed in same way as for terrorist offenders. The item is rated as 'High' if person concerned has no extremist ideology.
P.2 Rejection of violence as a means to achieve goals	Assessed in same way as for terrorist offenders. The item is rated as 'High' if person concerned never wanted to use violence to achieve ideological goals.
P.3 Change in concept of the enemy	Assessed in same way as for terrorist offenders. The item is rated as 'High' if person concerned has no enemy construct.
P.4 Participant in programs against violent extremism	Assessed in same way as for terrorist offenders, but item is rated as 'Not applicable' if no programs against violent extremism are offered to person concerned.
P.5 Support from the community for non-violence	Focus on support to relinquish use of violence, in terms of regular violence instead of violent extremism
P.6 Support from family members, other important persons for non-violence	Focus on support to relinquish use of violence, in terms of regular violence instead of violent extremism