CRIMINOLOGY
RESEARCH GRANT

Testing the reliability and validity of the VERA-2R on individuals who have radicalised in Australia

Adrian Cherney
Emma Belton

Report to the Criminology Research Advisory Council
Grant: CRG 40/21–22

June 2024

This project was supported by a Criminology Research Grant.
The views expressed are the responsibility of the author and are not necessarily those of the Council.

Celebrating 50 years

© Australian Institute of Criminology 2024

Apart from any fair dealing for the purpose of private study, research, criticism or review, as permitted under the *Copyright Act 1968* (Cth), no part of this publication may in any form or by any means (electronic, mechanical, microcopying, photocopying, recording or otherwise) be reproduced, stored in a retrieval system or transmitted without prior written permission. Inquiries should be addressed to the publisher.

Published by the Australian Institute of Criminology
GPO Box 1936 Canberra ACT 2601
Tel: (02) 6268 7166
Email: front.desk@aic.gov.au
Website: www.aic.gov.au/crg

ISBN 978 1 922877 34 5 (Online)
https://doi.org/10.52922/crg77345

This research was supported by a Criminology Research Grant. The views expressed are those of the author and do not necessarily reflect the position of the Criminology Research Advisory Council or the Australian Government.

This report was reviewed through a double-blind peer review process.

Edited and typeset by the Australian Institute of Criminology.

Contents

v Acknowledgements

vi Acronyms and abbreviations

vii Abstract

viii Executive summary

viii Project aim

viii Method

viii Analysis

viii Results

ix Implications

1 Introduction

1 Project background and research aims

3 Violent Extremism Risk Assessment—Version 2 Revised

6 Research on violence risk assessment tools and the VERA-2R

12 Methodology

13 Methods

19 Results

19 Interrater reliability testing of the VERA-2R

23 Patterns across VERA-2R risk domains

27 Predictive validity of the VERA-2R

30 Discussion and conclusion

32 Observations and lessons when applying specific indicators

33 The issue of measuring predictive validity

33 Future research

34 References

41 Appendix: VERA-2R domains

Figures

23 Figure 1: Distribution of indicators across low, moderate and high risk ratings for beliefs, attitudes and ideology

24 Figure 2: Distribution of indicators across low, moderate and high risk ratings for social context and intention

24 Figure 3: Distribution of indicators across low, moderate and high risk ratings for history, action and capacity

25 Figure 4: Distribution of indicators across low, moderate and high risk ratings for commitment and motivation

25 Figure 5: Distribution of additional indicators across the sample

29 Figure 6: Area under the curve analysis for VERA-2R risk ratings (low vs high)

Tables

3 Table 1: VERA-2R domains and indicators

20 Table 2: Kappa and ICC values for beliefs, attitudes and ideology VERA-2R indicators

21 Table 3: Kappa and ICC values for social context and intention VERA-2R indicators

22 Table 4: Kappa and ICC values for history, action and capacity VERA-2R indicators

22 Table 5: Kappa and ICC values for commitment and motivation VERA-2R indicators

26 Table 6: Major risk VERA-2R indicators association with violent and non-violent extremists

28 Table 7: Risk judgement distribution across violent extremists, non-violent extremists, and total sample

41 Table A1: VERA-2R risk across the sample

43 Table A2: VERA-2R additional indicators across the sample

Acknowledgements

The project was funded through a Criminology Research Grant: CRG 40/21–22.

Acronyms and abbreviations

AUC	area under the curve
BA	beliefs, attitudes and ideology
CM	commitment and motivation
EDT	European Database of Convicted Terrorist Offenders
HAC	history, action and capacity
ICC	intraclass correlation coefficient
NPV	negative predictive value
PIRA	Profiles of Individual Radicalisation in Australia
PPV	positive predictive value
SCI	social context and intention
SPJ	structured professional judgement
TRAP-18	Terrorist Radicalization Assessment Protocol
VERA-2R	Violent Extremism Risk Assessment—Version 2 Revised

Abstract

Violent extremism risk assessment has become an important way of dealing with terrorism and violent extremists. One violent extremism risk assessment tool adopted for use in Australia is the Violent Extremism Risk Assessment—Version 2 Revised (VERA-2R). The VERA-2R captures risk and contextual indicators across five domains and background characteristics. Two trained assessors (the authors) completed VERA-2R risk assessments on a sample of 50 individuals identified as having radicalised to violent extremism in Australia. Patterns in risk factors across the sample were analysed, including testing the interrater reliability and predictive validity of the VERA-2R. Results showed differences in risk profiles between individuals who were violent and non-violent. It was found the VERA-2R had good interrater reliability but low predictive validity. Implications for research and the practice of risk assessment are considered. Limitations of the project design and sample are acknowledged.

Executive summary

Project aim

The aim of this project was to examine the applicability and validity of the Violent Extremism Risk Assessment—Version 2 Revised (VERA-2R) against data on individuals who have radicalised in Australia.

Method

Drawing on a sample of 50 individuals selected from the Profiles of Individual Radicalisation in Australia (PIRA) dataset, VERA-2R risk assessments were completed on the sample. A range of approaches were adopted to ensure the quality of data sources used to inform assessments. Both authors received the official VERA-2R training and consulted with practitioners. They each assessed the 50 cases separately and blind.

Analysis

Standard tests relating to interrater reliability and predictive validity were completed based on several performance indicators. Trends in risk factors across the sample were analysed.

Results

The results showed that the VERA-2R had good interrater reliability, with interrater reliability being high across all domains and most specific VERA-2R indicators. Patterns across the sample showed differences in the risk profiles for violent and non-violent individuals, with a clustering of risk factors around the VERA-2R domains of *Beliefs, attitudes and ideology* and *Social context and intention.* When testing the validity of the VERA-2R, the analysis indicated the tool has some capacity to delineate between levels of risk but was not necessarily predictive.

Implications

Project findings show that the application of the VERA-2R can lead to consistent risk judgements. However, there are limitations to the tool. These included the scoring of specific indicators and overlap across specific indicators. Project findings indicate there can be differences between violent and non-violent individuals which can help to inform risk assessments and risk scenario planning. Challenges pertaining to testing the predictive validity of violent extremism risk assessment tools are discussed and as well as recommendations for future research.

Introduction

Project background and research aims

A key aim of violence risk assessment is to evaluate the propensity that a person will act in a dangerous and harmful way and then to develop strategies to mitigate or manage those behaviours and their impact (Borum 2015; Hart & Logan 2011). Several different risk assessment frameworks and approaches have been developed to help meet this objective (Borum 2015; Hart & Logan 2011; Monahan 2012). These approaches have included nondiscretionary and informal or subjective unstructured assessments, probabilistic actuarial measures, and structured evidence-based frameworks (Douglas & Kropp 2002; Hart & Logan 2011; Olver et al. 2009). Various specific risk assessment tools have been developed aligned with these approaches and frameworks (Hart & Logan 2011).

One approach to violence risk assessment is the use of the structured professional judgement (SPJ) framework. SPJ tools have been widely used for the assessment of criminogenic risks. This includes, for example, the Risk for Sexual Violence Protocol, the Historical Clinical Risk Management—20, and the juvenile-specific Structured Assessment of Violence Risk in Youth. SPJ is defined as an analytical method used to understand and mitigate the risk for interpersonal violence posed by individual people that is discretionary in essence but relies on evidence-based guidelines (Hart, Douglas & Guy 2016; Hart & Logan 2011). SPJ as a practice involves practitioners collecting and analysing offender case information according to risk factors as outlined in the applicable manual, evaluating the relevance of these factors to the specific case, formulating scenarios as to the likely outcomes of these risk factors and then developing management plans based on these scenarios (Hart & Logan 2011).

Final risk judgements are arrived at based on these steps. SPJ allows for a link to be drawn between evidence on behavioural risk factors and information on individually nuanced content (Shepherd & Lewis-Fernandez 2016). SPJ has been widely adopted for the assessment of radicalised offenders and extremist violence risk (Logan & Lloyd 2019; Monahan 2012). The suite of extremist violence assessment tools that have adopted the SPJ approach include the Violent Extremism Risk Assessment—Version 2 Revised (VERA-2R; Pressman et al. 2018), the Extremism Risk Guidelines 22+ (ERG 22+; Lloyd & Dean 2015), Radar (Corner & Taylor 2023b), and the Terrorist Radicalization Assessment Protocol (TRAP-18; Meloy 2017).

The Australian Government has endorsed the use of the VERA-2R (see Pressman et al. 2018) for adoption within Australia (Renwick 2018; Ripperger 2021). The VERA-2R is used in Australia across correctional and community-based settings to assist decision-making in relation to sentencing, custodial classification and placement, programs, case management and release. It is used to inform assessments of individuals subject to Commonwealth terrorist offender preventative legislation, such as continuing detention orders (ie High Risk Terrorist Offender scheme, or Division 105A of the Criminal Code). It is in this context that the use of the VERA-2R has come under intense scrutiny and debate (Cubitt & Wolbers 2023; Independent National Security Legislation Monitor 2023). More recently concerns have been raised about the validity and reliability of the tool and the evidence base underpinning the VERA-2R (Corner & Taylor 2023b; Independent National Security Legislation Monitor 2023). The consensus is that more research is needed on the VERA-2R and its application to violent extremists in Australia (Corner & Taylor 2023b; Cubitt & Wolbers 2023).

Given the influence of violent extremism risk assessment on decisions relating to pre- and post-detention, the management of extremist offenders and the forms of assistance provided to them, it is important that any tool provide valid and reliable assessments of risk. Briefly, validity as it pertains to violence risk assessment relates to whether an instrument (ie a specific tool) measures what it purports to measure. This includes, for example, whether a risk assessment tool can predict the risk of offending. Reliability on the other hand refers to the consistency of the tool in reaching risk judgements. This includes, for example, if there is consistency between assessors when it comes to their assessments of risk—termed interrater reliability (for more detail on specific validity and reliably criteria, see Cubitt & Wolbers 2023; Groth-Marnat & Wright 2016: 117; Singh 2013).

Like many areas of social and behavioural sciences, there will be debates and disagreements about the validity of instruments that aim to assess human behaviour and motivation. There will be strengths and weaknesses with various studies and the methodologies employed. Practitioners will have different perspectives about the accuracy of a risk assessment tool, as will researchers, policymakers, and other stakeholders, such as police and judicial officers. Given the high-stakes environment of continuing detention, the use of the VERA-2R within that context is highly contested, particularly when it is used to inform decisions to detain someone beyond their release date. This research did not set out to conclude whether the VERA-2R is more or less valid, or reliable, than other available tools. The aims of this project are:

- to examine whether the VERA-2R can help understand the types of factors that characterise violent extremists in Australia;
- to assess if the VERA-2R can discern between different risk levels in relation to violent and non-violent individuals; and
- to undertake a series of tests examining the interrater reliability and predictive validity of the VERA-2R.

Our findings do have implications for judgements and debates around the reliability and validity of the VERA-2R and the practice of violent extremism risk assessment. However, there are limitations to our approach and the generalisability of the results given the study's research sample. We return to these issues in the *Discussion and conclusion* section. Our primary objective is to build upon the emerging body of research on the VERA-2R.

Violent Extremism Risk Assessment—Version 2 Revised

The VERA represents one tool that was developed to assess dangerousness and the risks posed by ideologically motivated individuals. The tool was first published in 2009, and later revised to the VERA-2R (Pressman & Flockton 2012; Pressman et al. 2018). Iterations include the VERA, VERA-2 and now the VERA-2R. The VERA was developed based on existing literature and consultations with law enforcement, intelligence, national security and correctional professionals working in the field of violent extremism (Pressman et al. 2018). The VERA-2R expanded on earlier versions to include additional motivational and contextual indicators capturing various background characteristics. The VERA-2R is designed to capture contextual factors and risk indicators relevant to violent extremism across five domains, as shown in Table 1 below (Pressman et al. 2018). These domains are *Beliefs, attitudes and ideology* (BA); *Social context and intention* (SCI); *History, action and capacity* (HAC), *Commitment and motivation* (CM), and *Protective and risk mitigating* (P) factors (Pressman et al. 2018). Additional background characteristics are also included.

Table 1: VERA-2R domains and indicators
Beliefs, attitudes and ideology (BA)
BA 1 Commitment to ideology that justifies violence
BA 2 Perceived grievance and/or injustice
BA 3 Dehumanisation of designated targets associated with injustice
BA 4 Rejection of democratic society and values
BA 5 Expressed emotions in response to perceived injustice
BA 6 Hostility to national identity
BA 7 Lack of empathy and understanding for those outside one's own group
Social context and intention (SCI)
SCI 1 Seeker, user or developer of violent extremist materials
SCI 2 Target for attack identified (person, group, location)
SCI 3 Personal contact with violent extremists (informal or social context)
SCI 4 Expressed intention to commit acts of violent extremism
SCI 5 Expressed willingness and/or preparation to die for a cause or belief
SCI 6 Planning, preparation of acts of violent extremism
SCI 7 Susceptibility to influence, control or indoctrination

Table 1: VERA-2R domains and indicators (cont.)
History, action and capacity (HAC)
HAC 1 Early exposure to violence-promoting, militant ideology
HAC 2 Network of family and friends involved in violent extremism
HAC 3 Violent criminal history
HAC 4 Strategic, paramilitary and/or explosives training
HAC 5 Training in extremist ideology in own country or abroad
HAC 6 Organisational skills and access to funding and sources of help
Commitment and motivation (CM)
CM 1 Motivated by perceived religious obligation and/or glorification
CM 2 Motivated by criminal opportunism
CM 3 Motivated by camaraderie, group belonging
CM 4 Motivated by moral obligation, moral superiority
CM 5 Motivated by excitement and adventure
CM 6 Forced participation in violent extremism
CM 7 Motivated by acquisition of status
CM 8 Motivated by a search for meaning and significance in life
Protective and risk mitigating (P)
P 1 Reinterpretation of the ideology
P 2 Rejection of violence as a means to achieve goals
P 3 Change in concept of the enemy
P 4 Participant in programmes against violent extremism
P 5 Support from the community for non-violence
P 6 Support from family members, other important persons for non-violence
Additional indicators
CH Criminal history
CH 1 Client of the juvenile system/convicted non-violent offence(s)
CH 2 Non-compliance with conditions or supervision
PH Personal history
PH 1 Violence in family
PH 2 Problematic upbringing and/or placed in juvenile care
PH 3 Problems with school and work
MD Mental disorder
MD 1 Personality disorder
MD 2 Depressive disorder and/or suicide attempts
MD 3 Psychotic and schizophrenic disorder
MD 4 Autism spectrum disorder
MD 5 Post-traumatic stress disorder
MD 6 Substance use disorder

When applying the VERA-2R, users are required to score each indicator listed in the four main risk domains (BA, SCI, HAC and CM) according to the ratings of 'high', 'moderate' and 'low'. The VERA-2R manual has accompanying text and descriptions for each rating pertaining to the specific indicator. Authors of the tool explain that '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' (Pressman et al. 2018: 28; see also de Bruin et al. 2022: 9).

For the indicators listed in the *Protective and risk mitigating* domain (P), users are required to score them in reverse, in that low scores indicate the absence of a protective factor. The authors of the tool state that '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, that is, when there is some positive change in the direction away from violent extremism; and as "high" if clear risk-mitigating indicator characteristics or information are present, that is, when there is a clear, positive change in the direction away from violent extremism' (Pressman et al. 2018: 28–29; see also de Bruin et al. 2022).

When it comes to the *Additional indicators* they are rated as either 'present' or 'absent'. The authors of the VERA-2R state that a range of information sources should be used when applying the VERA-2R, ranging from clinical interviews to police and intelligence reports, legal or judicial files and psychological and psychiatric evaluations. If no information relating to a particular indicator is present, then it should not be rated. Drawing on this information and through the use of structured professional judgement, VERA-2R users are advised to interpret the weighting of the indicators to arrive at an overall risk assessment and case formulation. The authors of the VERA-2R state that this does not comprise a numerical score relating to a risk level. In the VERA-2R manual this overall risk score is stated as either 'low', 'moderate' or 'high' or as suggested by the authors it can also include 'low-moderate', and 'moderate-high' risk ratings. Suggestions or guidance for how a user of the tool might arrive at a low, low-moderate, moderate, moderate-high or high risk rating is not outlined in the manual.

Research on violence risk assessment tools and the VERA-2R

The field of violence risk assessment is highly contested. The very concept of risk as an effective way of understanding and managing violent offending behaviour has been debated (eg Whitehead et al. 2007). There are debates around methods and approaches to measuring the validity and reliability of various risk assessment tools (Singh 2013). Many different methods and approaches have been advocated and adopted, each with its own strengths and limitations (Hassan et al. 2022; Singh et al. 2011; Singh 2013). The very issue of assessing predictive validity and whether this is possible has been debated within the field (Singh 2013). There is not the space here to review these issues. Importantly, a number of gaps in research on violent extremism risk assessment have been raised. For example, this includes the lack of validation studies, difficulties in accessing data on known terrorists, challenges in assessing the varied nature of radicalisation risk factors and terrorist behaviour, the lack of focus on non-violent individuals displaying indicators of radicalisation, and lack of data on end-user experiences (Cherney, Grossman & Khalil 2022; Corner & Pyszora 2022; Hassan et al. 2022; Herzog-Evans 2018; Gill et al. 2020; Sarma 2017). Also, testing the predictive validity of violent extremism risk assessment tools and what this means for the prediction of reoffending risk is difficult given the low base rate of terrorist offending and reoffending—compared with other forms of criminality—in a country such as Australia.

A key concern and criticism made of the VERA-2R is that there is a lack of research on the validity and reliability of the tool. Compared with other tools such as the TRAP-18—which is also used in the Australian context (see Corner & Pyszora 2022)—the number of studies on the VERA, VERA-2 and VERA-2R is small. For example, research on the TRAP-18 has many studies supporting its validity across several categories of violent extremists, including Islamist terrorists (eg Böckler et al. 2020), members of the Sovereign Citizens movements (eg Challacombe & Lucas 2019), patients with severe mental illness, lone-actor terrorists, and violent and non-violent individuals (eg Allely & Wicks 2022; Meloy et al. 2019). Studies have also shown the TRAP-18 to score from good to excellent for tests measuring interrater reliability (eg Challacombe & Lucas 2019; Meloy et al. 2015). Corner and Pyszora (2022) evaluated the face and content validity of the TRAP-18 using a cohort of Australian practitioners and found that, while implementation problems were experienced when using the tool, the TRAP-18 was rated by their sample of practitioners as having face and content validity. A synthesis of research on the TRAP-18 concluded it was empirically based and a useful approach to early detection and case management (Allely & Wicks 2022).

However, currently a similar number of studies does not exist relating to various versions of the VERA. Pressman (2016) suggests that the tool contains face validity since it is applicable and useful across various contexts for differing practitioner needs. An early study by Beardsley and Beech (2013) rated factors captured in the first version of the VERA across a sample of five cases of known terrorist offenders spanning different ideological spectrums. Interrater reliability was found to be high across their two raters, and they concluded that 'the majority of factors in the VERA seem to be relevant and important to risk assessment and could be easily applied across the variety of terrorists in the sample' (Beardsley & Beech 2013: 12). A study by Pressman and Flockton (2014) on the VERA-2 concluded the tool had construct validity by comparing the risk assessment scores between a sample of convicted violent extremist (*n*=11) and non-ideologically-motivated offenders (*n*=11). They found that the terrorist cohort scored on average lower levels of risk on assessment tools targeting violence more generally and scored higher levels of risk on the VERA-2 (see also Pressman 2016). Hart et al. (2017) examined the degree of overlap between the VERA-2 and other risk assessment tools (ie the Multi-Level Guidelines: see Cook, Hart & Kropp 2014; and the Structured Risk Guidance/ Extremism Risk Guidelines 22+: see Lloyd & Dean 2015). They found that, while overlap existed, the VERA-2 was useful in that it captured 'different facets or aspects of extremist desires, beliefs, and attitudes' (Hart et al. 2017: 38).

Specifically in relation to the VERA-2R, de Bruin et al. (2022) set out to examine the interrater reliability of the tool. Two assessors, both Dutch researchers (one male, one female) with a bachelor's and master's degree in psychology and/or criminology and who were trained in the VERA-2R, independently rated 30 terrorist cases derived from judicial files. These files included mental health assessments, a probation report, a transcript of the verdict, a police report, a criminal record and/or information from intelligence services. The assessors rated each indicator within the various VERA-2R risk and protective domains according to high, moderate and low, and the 11 additional indicators as present or absent. Final risk judgements were also assigned by the assessors.

De Bruin et al. (2022) found good to excellent interrater reliability estimates for a majority of indicators within the risk domains of *Beliefs, attitudes and ideology*; *Social context and intention*; *History, action and capacity*; and *Commitment and motivation*. That is, they were equal to, or above, scores of acceptable levels of reliability as indicated in the literature, or more simply there were high levels of agreement across the two assessors. However, for a minority of risk indicators, fair to poor interrater reliability was found. This included SCI 6 (planning or preparation of acts of violent extremism) within the *Social context and intention* domain, HAC 6 (organisational skills and access to funding and sources of help) within the *History, action and capacity* domain, and CM 5 (motivated by excitement and adventure) within the *Commitment and motivation* domain. The authors conclude one possible reason for these low levels of reliability may be due to a lack of clarity and clear delineation within the VERA-2R manual on how these indicators are to be assessed. Good to excellent interrater reliability scores were found for most indicators within the *Protective and risk-mitigating* (P) domain, except for P 3 (change in concept of the enemy) and P 6 (support from family members, other important persons for non-violence). When it came to the final risk judgement, high estimates for interrater reliability were found. Also, the overall level of agreement across the additional indicators was high across the two assessors.

Corner and Taylor (2023b) assessed the VERA-2R in relation to its empirical and theoretical foundations, as well as its validity and reliability. This study examined both the VERA-2R and a tool labelled Radar, given their widespread use in Australia, and drew on quantitative and qualitative methods including an experimental design. In reference to the approach adopted by Corner and Taylor (2023b) to examining the VERA-2R, they assessed the content and empirical basis of the VERA-2R manual by reviewing all known publications investigating the tool, completed a thematic analysis on the VERA-2R manual and associated training materials and a systematic review of the existing evidence base on radicalisation and extremist behaviours, and compared the degree to which the VERA-2R was underpinned by empirically validated variables and existing evidence.

The same study also tested the validity and reliability of the VERA-2R through the formulation of 60 cases (Corner & Taylor 2023b). These cases comprised individuals that differed across intensities and outcomes of radicalisation, varying forms of extremism, and perpetrator groups (which included a small group of non-radicalised individuals). The cases were derived from a mix of open and closed sources. This sample comprised a mix of individual cases spanning different levels of radicalisation and terrorist behaviour from Australia, Europe and Canada. The study recruited 30 participants (assessors) across three groups referred to as novices (university students, with some background knowledge in risk assessment), experts (academics, Commonwealth and state government agencies who supplied case vignettes) and trained individuals (personnel trained in the VERA-2R). The 60 cases were randomly allocated to research participants for assessment (Corner & Taylor 2023b).

This study (Corner and Taylor 2023b) also assessed the face validity of VERA-2R by looking at the number of factors participants found most pertinent during their assessments, tested for sensitivity and specificity by examining the proportion of individuals correctly identified as high or low risk, and measured the tool's predictive validity by conducting what is termed an area under the curve (AUC) test. They examined the tool's reliability by testing for interrater reliability across a selection of cases, and also tested for equity in the application of assessments by investigating if the assessment of risk was influenced by ideology or an individual's background as a violent extremist or non-extremist.

Corner and Taylor (2023b) concluded that VERA-2R was best described as 'SPJ lite', in that its application as described in the VERA-2R manual did not accord to key principles underpinning the SPJ approach (see above for the SPJ principles and also Hart & Logan 2011). They raised concern that few of the VERA-2R indicators are supported by theoretically and empirically valid evidence. They concluded the tool lacked reliability and validity, finding for example poor interrater reliability and also poor predictive validity (see Corner & Taylor 2023b). Corner and Taylor (2023b) draw attention to limitations of their study and state that tools such as VERA-2R require further investigation.

Cubitt and Wolbers (2023) of the Australian Institute of Criminology undertook on behalf of the Commonwealth Department of Home Affairs a review of the utilisation and suitability of risk assessment tools for convicted terrorist offenders in Australia. This included the VERA-2R. One aim of the study was to assess the suitability of various tools in relation to applications for control orders and continuing detention or extended supervision orders, targeting convicted terrorists. The study included a literature review and semi-structured interviews with experts and practitioners involved in violence risk assessment. The study did not set out to measure the reliability and validity of various tools, but it noted the lack of research in this area. Most interview participants reported experience with the VERA-2R. In relation to the VERA-2R, participants reported it was applied to various forms of extremism and perpetrator groups. The lack of research supporting the VERA-2R was acknowledged by interview participants and the need for independent research to be undertaken. Across the interview sample there was the majority view—but not all agreed—that the VERA-2R was an appropriate tool to inform risk assessments and a useful aid in decision-making, particularly relating to control orders and post-sentence orders.

Several scholars have drawn on the VERA-2R and its associated indicators as a way of understanding extremist risk and to identify patterns in risk and protective factors across radicalised samples of individuals. It should be emphasised that these studies did not set out to draw conclusions about the validity and reliability of the VERA-2R. For example, Duits, Alberda and Kempes (2022), drawing on the European Database of Convicted Terrorist Offenders (EDT), examined the prevalence of psychopathological factors (eg mental disorders and psychological problems) among a sample of adults and young terrorist offenders (aged 15–21) and their relationship with grievances and anger about perceived injustices. The EDT is a compiled database from open and closed sources of data on convicted terrorists across Europe and, among other variables, captures indicators within the risk and protective domains as outlined in the VERA-2R. Trained assessors across the various European partners involved in the project score all available information with the use of a codebook (see Alberda et al. 2021).

To measure violent ideology, grievances and anger, Duits, Alberda and Kempes (2022) based their measures on three VERA-2R indicators. This included BA 1 encompassing violent ideology, BA 2 capturing perceived grievances or injustice, and BA 5 encompassing emotional responses such as anger in reaction to a perceived injustice. Each indicator was captured as either low, moderate or high. They found that among their sample nearly all of their youth category supported a violent ideology, over half expressed grievances and injustice and a third expressed anger in response to a perceived injustice. The strength of associations between the VERA-2R measures and psychopathological factors in the adult and youth sample varied from weak to strong, and the authors did find that VERA-2R measures of grievance and anger were strongly associated with relationship problems and depressive symptoms. Alberda et al. (2022), also drawing on the EDT, looked at differences in risk factors using the VERA-2R indicators between 21 jihadist offenders who were convicted for homicide, and a comparison group of 30 jihadist offenders convicted for other terrorist offences. They did not code the sample according to the high, medium and low categories, but rather used a numerical coding scale to capture how explicitly and implicitly the indicator was present based on the levels of objectivity and accuracy in information sources (for more detail, see Alberda et al. 2022). Alberda et al. (2022) explored the presence of various VERA-2R derived indicators and their associations with the homicide and non-homicide groups. Among the homicide group they found that emotional responses relating to perceived grievances (BA 5), expressed intention to commit acts of terrorism (SCI 4), expressed willingness or preparation to die for a cause or belief (SCI 5), planning, preparation of acts of violent extremism (SCI 6) and early exposure to violence-promoting, militant ideology (HAC 1) were strongly associated with the homicide group compared with the non-homicide group. Protective factors were found to be more strongly associated with the non-homicide group compared with the homicide group.

Two studies by Thijssen et al. (2022, 2023) also draw on VERA-2R indicators to examine risk profiles and background characteristics for a sample of male jihadi detainees placed in terrorism wings in the Netherlands. The Thijssen et al. (2022) sample comprised 121 individuals who were scored for the presence of VERA-2R risk and protective indicators across the categories of high, moderate and low. This scoring was only completed by the first author using a variety of primary and secondary data sources (see Thijssen et al. 2022 for more detail). In summary the results indicated that, for those rated as high risk, the most prevalent risk factors included attachment to an ideology that justifies violence (BA 1), perception of injustice and grievances (BA 2), personal contact with violent extremists (SCI 3), network of family and friends involved in violent acts (HAC 2) and access to finance, resources or organisational skills (HAC 6). Thijssen et al. (2022) also examined differences across subgroups of individuals who were and were not convicted of terrorism within their sample. There were some differences between the two groups. For example, detainees convicted of terrorism scored significantly higher on a larger number of risk indicators, and lower on one protective factor (community support for non-violence) compared with those who were not convicted of terrorism.

Thijssen et al. (2023), drawing on the same detainee sample and approach to scoring the VERA-2R indicators as the previous Thijssen study, set out to examine if it was possible to identify distinct motivational groups (through latent class analysis) using eight motivational indicators (ie those captured in the VERA-2R *Commitment and motivation* domain). They were able to identify three groups (labelled classes) termed the low motivated class (class 1), morally motivated class (class 2) and the hardened ideologically driven class (class 3). They then examined differences across these classes. The most significant differences included the finding that individuals within the hardened ideologically driven class were more likely to be violent, have violent backgrounds and had previous contact with violent extremists. The other two lower risk classes typically had higher mean scores for the presence of protective factors.

The above research portrays a varied picture when it comes to the validity and reliability of the VERA-2R. Overwhelmingly it can be concluded there is a gap in research testing the VERA-2R compared with other violent extremism risk assessment tools. The rigour of previous studies on various iterations of the VERA can be questioned. The research seems to indicate the tool has some utility (eg Cubitt & Wolbers 2023; de Bruin et al. 2022), with it also having value as a way to understand and distinguish between levels of risk and the presence and absence of various risk and protective factors across different cohorts (eg Alberda et al. 2022; Duits, Alberda & Kempes 2022; Thijssen et al. 2022, 2023). However, there still remain questions over the reliability and validity of the VERA-2R and the evidence base underpinning it (eg Corner & Taylor 2023b).

Methodology

The research sample for this project was selected from the Profiles of Individual Radicalisation in Australia (PIRA) database. PIRA is modelled on the Profiles of Individual Radicalisation in the United States (PIRUS) dataset, an open-source database developed by the University of Maryland at the National Consortium for the Study of Terrorism and Responses to Terrorism (see https://www.start.umd.edu/data-tools/profiles-individual-radicalization-united-states-pirus). The PIRA captures data on individuals who have radicalised in Australia from 1985 to 2022. PIRA includes individuals who adhere to Islamist, far-right, far-left, and single-issue ideologies. Individuals are included in PIRA for either committing ideologically motivated illegal violent or non-violent acts, joining a designated terrorist organisation, or associating with an extremist group or organisation. To be eligible for inclusion into the database, each person must have resided in Australia when they radicalised and there is a clear link to ideological motives or behaviours. To date, a total of 254 individuals are included in the database.

To identify individuals for inclusion in the PIRA database, publicly available sources have been searched, including court documents, coronial inquest reports, online news articles, terrorist blogs, newspaper archives, open-source non-government reports, terrorist monitoring research institutes and terrorist attack databases. There are limitations in using open-source materials to compile profiles of individuals who have radicalised to extremism, relating to impartiality, accuracy and completeness. There can be missing information, for instance, compared with closed sources or psychological evaluations. However, open sources can be just as detailed as information derived from closed sources—for example, police files (Gill et al. 2019). Based on assessments of the reliability of collected source materials and according to a codebook (see Belton & Cherney 2023), individuals in PIRA are coded across 122 variables relating to background, demographic, group affiliation, and contextual information. The coding of the PIRA data has shown to have good interrater reliability (see Belton, Cherney & Zahnow 2023 for more detail). Data are first qualitatively captured and then transferred into quantitative indicators. It is this mix of data that was used to conduct VERA-2R assessments on the selected sample.

Methods

Sample and data

VERA-2R risk assessments were completed on a total of 50 cases captured in the PIRA sample. These 50 cases consisted of individuals who had participated in a range of terrorist-related acts. The sample of 50 cases was selected due to time constraints in completing the project and to also minimise the likelihood of assessor fatigue given the first and second author would be coding all 50 cases each across the VERA-2R indicators. We chose cases with the largest number of data sources and the most reliable information to enhance the accuracy of the assessments by minimising the possibility of missing data. To identify PIRA cases, they were first ranked according to the number of sources used, then by those with available court data (eg sentencing transcripts), which is considered to be the most reliable of open-source data (Gill et al. 2020). On average, cases were compiled using 29 separate data sources (*SD*=13.04), with 88 percent (*n*=44) having available court data.

The above selection process resulted in a sample comprising 37 violent and 13 non-violent individuals (*N*=50). Violent extremists in PIRA are defined as individuals that attempted or actually engaged in ideologically motivated behaviour intending to cause harm, injury or death. To be considered violent, there had to be clear evidence of operational plans that were aimed at engaging in violent acts (ie collection of weapons, dry runs of potential targets). Those considered to be non-violent were individuals who came to hold radical views but did not engage in violent action to aid or encourage terrorism, but did undertake non-violent activities that were motivated by an extremist ideology (eg this can include acts of financing a terrorist organisation, and viewing, compiling and disseminating extremist material). Hence the categorisation of 'violent' includes those conducting violent acts themselves and individuals who engaged in preparatory acts of violence. Existing studies have made similar definitional distinctions between violent and non-violent extremists when analysing comparable datasets (see Becker 2019; LaFree et al. 2018; Schuurman & Carthy 2023). In our sample non-violent cases often contained less source information (69% had available court data compared with 95% of violent cases), making coding against the VERA-2R risk indicators more challenging. However, the inclusion of non-violent extremists ensured the sample represented individuals from a range of backgrounds who participated in ideologically motivated behaviours that are not always characterised by violence. This sample was selected to ensure the data as close as possible reflected the variation and complexity of risk profiles, and hence levels of risk based on behavioural outcomes. Islamist and far-right aligned individuals were included to further ensure the data captured a range of ideological motivations. Most cases, however, were aligned with Islamist or jihadist ideologies (88%), with only six far-right cases included (12%).

The above approach led to the data being skewed towards more prolific offenders, those charged with a terrorist offence, or cases more frequently reported in open sources (Chermak et al. 2012). The authors were careful not to include large numbers of individuals from similar terrorist plots to increase the diversity of the sample. We also included a small number of individuals that did not have an associated court transcript but still had large numbers of sources. This was to incorporate those involved in violent extremism but who have not been charged with a terrorist-related offence (eg individuals who were prominent recruiters or had evaded law enforcement). Although this influenced the reliability of the information used, it also allowed for the inclusion of a wider group of offender profiles.

Coding VERA-2R indicators

Qualitative information in the PIRA database was used to measure the VERA-2R risk indicators as being either low, moderate or high. PIRA data were manually assessed against the VERA-2R risk domains. Identical to what is stated in the VERA-2R manual, specific indicators for the main five risk domains *Beliefs, attitudes and ideology* (BA), *Social context and intention* (SCI), *History, action and capacity* (HAC), and *Commitment and motivation* (CM) were scored—that is, coded—on a rating scale as 0=low, 1=moderate or 2=high. An indicator is rated as 'low' if the indicator was not present; a 'moderate' rating is given if there was some evidence; and a 'high' rating was given for the clear presence of the risk indicator. Protective factors were rated as 'low' (0) representing no evidence of the protective factor being evident, 'moderate' (1) representing some indication or evidence of a protective factor being present, and 'high' (2) indicating clear evidence of a protective factor. The remaining 11 *Additional indicators* were scored on a dichotomous scale where (1) indicates the presence of factors associated with heightened vulnerability and (0) equates to the absence of these vulnerability factors.

Both authors attended an official three-day training course conducted by the Commonwealth Department of Home Affairs on the practical use and application of the VERA-2R risk assessment tool and were given copies of the manual authored by Pressman and colleagues and assessor forms used in the field. Scoring practices were informed by the VERA-2R manual and this training, and guidelines detailing specific instructions, parameters and explanations for scoring each of the VERA-2R indicators. To ensure that the assessments closely reflected how the VERA-2R is applied and used in practice, the authors liaised with VERA-2R trainers and other users to clarify the correct coding of indicators. There were often discrepancies between the information derived from the VERA-2R manual and the instructions provided in the official training. An example of this relates to the indicator designed to assess the presence of previous criminal violence (HAC 3). The manual instructs the user to report any criminal history but does not stipulate that this also includes the index offence (what offenders were charged with, ie terrorism offences), which was clarified by the VERA-2R training. It became clear that the official training was necessary to accurately interpret the risk domains and to clarify any ambiguity present in the manual. Both the first and second author each assessed the sample of 50 extremists separately and blind. Once indicator ratings for each profile were completed, a total risk judgement was given. Both authors gave either a *low*, *moderate* or *high* judgement.

It should be highlighted that these assessments were completed on retrospective data as captured in PIRA, which is based on past activities comprising violent and non-violent acts that were used as a proxy to determine risk levels. This approach has been adopted to assess other violent extremism risk assessment tools such as the TRAP-18 and also existing studies of the VERA-2R (eg Böckler et al. 2020; Corner and Taylor 2023b).

Forms of analysis

Testing the interrater reliability of the VERA-2R

The level of consistency among ratings of the VERA-2R was examined by measuring interrater reliability. To ensure comprehensiveness when testing for interrater reliability, this was analysed using four measures: percentage agreement, Cohen's kappa, Krippendorff's alpha and intraclass correlation coefficient (ICC). When calculating interrater reliability estimates, all missing values (originally coded as −99) were treated as low (coded as 0).

Percentage agreement gives a rough estimate of the level of agreement between assessors but is susceptible to limitations (Hallgren 2012) as it is unable to account for variance between coders (in this case the two assessors—the authors) or agreements that occur by chance. The use of additional interrater reliability estimates should be employed to supplement this approach. Average percentages for risk domains and overall percentage agreements are reported.

Cohen's kappa is best used with two coders and weighted kappas (κ) were used to account for ordinal scales of the VERA-2R risk indicators: low, moderate, high (Grooten et al. 2019). Weighted kappa statistics consider different levels of disagreement between categories (Tang et al. 2015), and linear-weighted coefficients are advised over quadratic weighted coefficients (see Vanbelle 2016). Interpretation of weighted kappa is similar to that of the unweighted kappa (Fleiss, Levin & Paik 2013). Using a sample of 50 cases for a two-rater study revealed statistically significant κ values ($p<0.05$), with a power of 90 percent (Sim & Wright 2005). According to Landis and Koch (1977), values <0.20 indicate slight agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 substantial agreement, and 0.80 to 1.0 almost perfect or perfect agreement. Kappa values were established for the VERA-2R indicators and risk judgement ratings. Mean kappa values for each domain were also established.

A Krippendorff's alpha (1970) can be applied to ordinal data for cases that are assessed by at least two raters and provides a more conservative interpretation than Cohen's kappa, where values less than 0.67 should be discounted and desired values are >0.80 (Krippendorff 2013). It also allows for different measurement levels (ie ordinal, categorical). A Krippendorff's alpha was calculated for each VERA-2R domain and risk judgement rating. Alpha values were calculated for ordinal risk ratings (low, moderate, high) for each of the five main risk VERA-2R domains and alpha values were calculated for nominal risk ratings (present or absent) for the additional indicators.

The ICC is considered a more robust measure of interrater reliability as it accounts for variance across raters (Shrout & Fleiss 1979). The number or cases coded (scored or rated) by the authors significantly influences the robustness of reliably estimates and the more cases coded increases the stability of reliability estimates (Walter et al. 1998). The minimal acceptable agreement level is 80 percent (Lacy & Riffe 1996). Following Bonett's (2002) formula to calculate the appropriate sampling size to achieve 95 percent confidence intervals across the two raters, the authors used 50 cases resulting in a precision of 0.1 (+/−) on an expected ICC reliability of 0.80 (95% confidence level). If an instrument has been designed and implemented in the correct manner, interrater values should be high (ie, it has high reliability). The ICC was examined using a two-way random effects model and absolute agreement type. The interpretation of ICCs was based on critical values for single measures. Interpretations follow Koo and Li (2016) based on 95 percent confidence intervals, where ICC values <0.50 are poor, 0.50 to 0.75 are considered moderate, 0.75 and 0.90 are good, and values >0.90 are judged as excellent. It has been pointed out that an ICC of 0.61 is desirable and commonly reported for established tools (Douglas & Reeves 2010), but ICC values above 0.75 are preferred (Vincent et al. 2012). ICC values were established for both the VERA-2R indicators and risk judgement ratings. Average ICC values are also provided for each of the VERA-2R risk domains.

Creating a final dataset for analysis

Following interrater reliability testing, a final dataset was created. That is, the two VERA-2R datasets of the same 50 cases which had been rated by both authors were combined into one dataset to allow for various forms of statistical analysis. This final dataset included the four major risk domains (BA, SCI, HAC and CM) and the additional indicators. To create a dataset suitable for analysis and to minimise bias, a defaulting system was employed in which there was a defaulting to the lowest rating level where there was a difference in the rating of high, medium and low. For example, if one assessor had rated an indicator as 'high' (2) and the second had rated the same indicator as 'moderate' (1), the 'moderate' rating (1) was taken as the default score. This was done for all indicator scores, from low, moderate to high. If an indicator was coded as missing, this was taken as the final code, regardless of what the second assessor rated. The same approach was employed to finalise risk rating judgements between coders (also rated as low, moderate or high). Using this protocol avoided over-inflation, which may have occurred using a consensus-based approach.

In the case of the additional indicators where disagreement was found, both assessors revisited case notes and, using the available data, followed a presence or absence approach. The same approach is outlined in the VERA-2R manual for this domain. For an indicator to be coded as present (1), there had to be clear evidence of this in the case notes. We deviated from the VERA-R2 manual when information for an indicator was not mentioned, in that we defaulted to a 'no' (0) rather than leaving it blank (ie missing), meaning there was less chance of assessor bias and no missing data would be present for analysis.

As mentioned, high instances of missing data are common when using open-source data. For the 50 cases, 40 percent (*n*=20) contained missing values. Due to the higher number of missing data for protective factors (12.3%), with some cases having missing data across all six protective factors, this domain was excluded from the final combined dataset, resulting in missing data in 18 percent (*n*=9) of cases. The exclusion of the protective factor domain from the final combined dataset for the purpose of trend analysis across the sample should not be seen as undermining our investigation. These indicators relate more to treatment and post-detention factors—of which we had little information on—than overall levels of risk, with the VERA-2R manual unclear how to recalibrate final risk judgements based on the presence of specific protective factors. The total missing values across the four main risk domains used for analysis were two percent (*n*=29), and over half of the missing values were in the *Beliefs, attitudes and ideology* domain. Due to the low occurrence of missing data and following the strategy taken for testing interrater reliability, all missing values were coded as 'low' (0).

Patterns and trends in risk indicators and extremist offenders

To observe any patterns in the final dataset, frequencies for each risk indicator were examined derived from the final combined dataset. A chi-square test was used to determine the association between risk indicators for the four major VERA-2R domains and the two groups of extremists. For cell frequencies fewer than five, Fisher's exact test was reported instead of chi-square. Phi coefficients (φ) were calculated as a measure of effect size and were interpreted as 0.10 small, 0.30 medium and 0.50 large (Cohen 1988). Due to the small sample size the original three-point scores for risk indicators (0=low, 1=moderate, 2=high) were converted into a dichotomous variable measuring the presence or absence of a risk indicator variable (0=absence, 1=presence), where presence equates to moderate (1) and high ratings (2), and absence equals low (0) ratings. This allowed us to measure the strength of associations between the presence of risk indicators for violent and non-violent cases.

Assessing predictive validity

Measures of predictive validity can be assessed by what is termed as either calibration or discrimination. Calibration examines how well the risk assessment agrees with observed risk and discrimination examines how well the instrument differentiates between those who went on to be violent and those who did not (Cook 2007). A combination of both performance indicators is recommended to measure different facets of predictive validity (Singh et al. 2013).

Sensitivity and specificity tests are regarded as the simplest performance indicators to measure the predictive validity of an instrument. Sensitivity is a high-risk discrimination index and specificity is a low-risk discrimination index (Singh et al. 2013). Sensitivity and specificity scores are calculated by using 2×2 contingency tables that organise assessment and outcomes into true positives, false positives, true negatives and false negatives. These scores were computed using cross-tabulations between dichotomous variables to calculate the proportion of individuals who engaged in violent acts and who were judged to be at high risk (sensitivity) and the proportion of individuals who engaged in non-violent acts and who were judged to be at low risk (specificity). They do, however, require a dichotomous outcome and therefore struggle with SPJ instruments that have multiple risk categories.

Two additional performance indicators with arguably greater relevance are positive predictive values (PPVs) and negative predictive values (NPVs), which measure high and low risk calibration. These estimates can measure calibrations and measure the accuracy of the classification outcomes, which also require dichotomous outcomes (Singh 2013). Individuals who were classified as high risk were assessed on their likelihood of being violent extremists (PPV) and individuals who were classified as low risk were assessed on their likelihood of being non-violent extremists (NPV). PPV and NPV are base rate dependant and should be used in conjunction with sensitivity and specificity (Singh 2013). Like sensitivity and specificity, positive and negative predictive values can only classify outcomes into dichotomous outcomes, limiting their usefulness in risk assessment tools with more than two risk categories.

To be able to complete the above validity tests a dichotomous risk judgement variable (low risk vs high risk) is required (Singh 2013). To meet requirements to conduct the above tests, we have taken what was a three-point judgement and converted it into a dichotomous measure, where 0=low risk and 1=high risk. In order to do this, we decided to combine low and moderate ratings into a low-risk category, and high ratings would represent the high-risk category. This allowed a base threshold for predictive validity to be produced. While categorising individuals who had moderate risk ratings as low (rather than moderate and high being combined), this possibly underestimates the level of risk for some offenders. However, it was decided that this was more appropriate than combining moderate and high-risk categories to ensure our approach was conservative.

The receiver operating characteristic analysis is another method of testing for predictive validity (Singh et al. 2013). It assesses accuracy by providing an area under the curve (AUC) statistic. To obtain the AUC a violent and non-violent dichotomous outcome is required (violent vs non-violent) and categorical risk judgements (low, moderate, high; Singh et al. 2013). The AUC index uses an instrument's cut-off thresholds to determine if the instrument can discriminate between types of offenders. It estimates the probability that a randomly selected violent individual received a higher risk classification (ie risk judgement) than a randomly selected non-violent individual (Altman & Bland 1994). AUC values of >0.80 are considered 'excellent', 0.70 to 0.79 are judged as 'acceptable' and 0.60 to 0.69 are seen as 'weak'. For violence risk assessment tools, AUC values over 0.70 are considered 'moderate' while values over 0.75 are 'good' (Geraghty & Woodhams 2015).

Unlike the above tests, AUC is resistant to changes in the base rates and can differentiate between more than two outcomes. However, solely testing predictive validity based on AUC fails to capture the calibration of instruments and therefore is not a complete picture of the effectiveness of an instrument, and it is recommended that researchers employ a range of tests to fully capture both calibration and discrimination (Singh 2013). However, it is important to highlight that small sample sizes (n<200) can also result in inaccurate estimates (Hanczar et al. 2010; Singh 2013). When interpreting validity tests, readers should keep in mind the low power of small sample sizes.

Results

In summary, the sample contained 50 extremists who were mostly male (94%) between 15 and 50 years old. A large portion were first or second-generation immigrants (78%). At the time of their offending, just under half were married, 42 percent had not completed their high school education, but 10 percent had enrolled in tertiary education. Individuals in the sample were suspected of or convicted for a variety of ideologically motivated activities. This included espousing extremist rhetoric online, compiling and disseminating extremist materials, financing a terrorist organisation, facilitating overseas travel, logistical support such as procuring weapons, and conspiring in or planning a domestic terrorist attack. Seventy percent (*n*=35) were formally charged and convicted of a terrorist-related offence.

Interrater reliability testing of the VERA-2R

The overall interrater reliability of the VERA-2R domains was calculated for the whole sample (*n*=50 subjects; 45 indicators, 2,250 separate codings/ratings). For each case, the authors indicated whether the VERA-2R risk indicator was low, moderate, high or missing (coded as 0=low, 1=moderate, 2=high and −99=missing). As mentioned, the authors both assessed the 50 cases separately and blind. Interrater reliability was found to be high between the assessors (authors), with percentage agreements for overall ratings for the six VERA-2R domains of *Beliefs, attitudes and ideology* (BA), *Social context and intention* (SCI), *History, action and capacity* (HAC), *Commitment and motivation* (CM), *Protective factors* (P) and *Additional indicators* being 88 percent, 84 percent, 84 percent, 84 percent, 93 percent and 95 percent respectively. Agreement for all 45 VERA-2R indicator domains was 89 percent, averaging five discrepancies across all indicators, ranging from zero to 10. The interrater percentage agreement for final risk judgements given by each assessor was 82 percent.

A Krippendorff's alpha was calculated for all the VERA-2R domains. Krippendorff's alpha scores for the VERA-2R domains were above the accepted values (>0.80), equating to good interrater agreement. Krippendorff alpha values for each domain were as follows: *Beliefs, attitudes and ideology* (α=0.89); *Social context and intention* (α=0.83); *History, action and capacity* (α=0.88); *Commitment and motivation* (α=0.87); *Protective factors* (α=0.80); and *Additional indicators* (α=0.86). The overall average Krippendorff's alpha for the VERA-2R (all domains) was 0.86, indicating well above desired standards (>0.80). Alpha scores for risk judgements also indicate excellent interrater reliability (α=0.83).

All kappa (κ) values ranged between 0.73 and 0.85, which are considered good to excellent agreement (Landis & Koch 1977). The average ICC values for indicators within the four main VERA-2R risk domains were between 0.80 and 0.87, showing good interrater reliability. For all the VERA-2R domains, including protective and additional indicators, the mean ICC value was 0.86, and the kappa value was 0.82, indicating good interrater reliability. Cohen's kappa for final risk judgements was found to have a strong level of agreement (κ=0.76, 95% CI [0.63, 0.90], p<0.001). The ICC value for risk judgement ratings between assessors indicated good reliability (0.83, 95% CI [0.72, 0.90], p<0.001).

Major risk VERA-2R domains

Here we consider the interrater reliability as it relates to specific indicators captured within the four main risk domains (see Tables 2 to 5). Kappa values are reported as a supplementary indicator of reliability and final conclusions on interrater agreement are derived from ICC values. Average ICC and kappa values for the four major VERA-2R risk domains were considered to have good interrater agreement (ICC=0.82, κ=0.77).

Out of the four major risk domains, *Beliefs, attitudes and ideology* (BA) received the highest level of interrater reliability. All indicators showed good to excellent interrater agreement (see Table 2).

Table 2: Kappa and ICC values for beliefs, attitudes and ideology VERA-2R indicators

Beliefs, attitudes and ideology (BA)	Kappa (95% CI)	ICC (95% CI)
Mean domain BA	**0.82**	**0.87**
BA 1 Commitment to ideology that justifies violence	0.74 (0.41, 0.80)	0.83 (0.48, 0.83)
BA 2 Perceived grievance and/or injustice	0.96 (0.88, 0.99)	0.98 (0.96, 0.98)
BA 3 Dehumanisation of designated targets associated with injustice	0.82 (0.69, 0.93)	0.88 (0.79, 0.93)
BA 4 Rejection of democratic society and values	0.81 (0.68, 0.93)	0.87 (0.78, 0.92)
BA 5 Expressed emotions in response to perceived injustice	0.84 (0.69, 0.98)	0.90 (0.82, 0.94)
BA 6 Hostility to national identity	0.82 (0.68, 0.95)	0.85 (0.74, 0.90)
BA 7 Lack of empathy and understanding for those outside one's own group	0.74 (0.58, 0.89)	0.78 (0.64, 0.87)

Note: N=50. 95% CI=95% confidence interval; all p-values were found to be significant at p<0.001. Kappa=weighted Cohen's kappa. ICC=intraclass correlation coefficient

The *Social context and intention* (SCI) domain had good overall reliability with three indicators showing excellent interrater agreement. However, this domain also demonstrated the poorest levels of interrater agreement for a single indicator. Reliability estimates for the indicator SCI 7 *Susceptibility to influence, control or indoctrination* showed poor agreement (ICC=0.45, κ=0.40).

When discussing the high interrater disagreement for this indicator, the authors concluded it was due to the ambiguity of the variable relating to what activities constituted influence and, if a person was a high-profile leader, whether they could still be categorised as being influenced by others. Interrater reliability for the indicator SCI 1 *Seeker, user or developer of violent extremist materials* presented as having moderate interrater agreement (see Table 3) which was also discussed as being due to the ambiguity of what represents a user verses a developer of extremist materials.

Table 3: Kappa and ICC values for social context and intention VERA-2R indicators

Social context and intention (SCI)	Kappa (95% CI)	ICC (95% CI)
Mean domain SCI	**0.76**	**0.81**
SCI 1 Seeker, user or developer of violent extremist materials	0.55 (0.35, 0.73)	0.63 (0.43, 0.78)
SCI 2 Target for attack identified (person, group, location)	0.91 (0.82, 0.99)	0.94 (0.90, 0.97)
SCI 3 Personal contact with violent extremists (informal or social context)	0.85 (0.71, 0.98)	0.89 (0.81, 0.94)
SCI 4 Expressed intention to commit acts of violent extremism	0.90 (0.79, 0.99)	0.94 (0.89, 0.96)
SCI 5 Expressed willingness and/or preparation to die for a cause or belief	0.81 (0.64, 0.96)	0.83 (0.72, 0.90)
SCI 6 Planning, preparation of acts of violent extremism	0.92 (0.84, 0.99)	0.96 (0.92, 0.97)
SCI 7 Susceptibility to influence, control, or indoctrination	0.40 (0.21, 0.58)	0.45 (0.08, 0.68)

Note: *N*=50. 95% CI=95% confidence interval; all *p*-values were found to be significant at $p<0.001$. Kappa=weighted Cohen's kappa. ICC=intraclass correlation coefficient

Indicators relating to *History, action and capacity* received the second highest level of overall interrater agreement (Table 4). The indicator HAC 5 *Training in extremist ideology in own country or abroad* had the lowest rater agreement but was still considered to have moderate reliability (ICC=0.71). For HAC 5 the term 'training' generated a lot of confusion as to what it actually involved. In the VERA-2R manual and as highlighted in the official VERA-2R training course, it appears the word 'training' can include formal face-to-face training, but also informal passive training including exposure to manuals online. The VERA-2R manual states it can include participation in at least one form of training and exposure to an extremist leader (which you are instructed in the VERA-2R manual to rate as moderate) or whether someone has acted as a trainer or an influential leader (which you are instructed in the VERA-2R manual to rate as high). The authors found that there was some ambiguity in rating someone 'moderate' compared with 'high' as the manual suggests. For example, if an individual receives one form of training in an extremist ideology (online or offline) or has been exposed to an influential leader, they are considered 'moderate', yet to be considered 'high' they had to be a leader or influential figure. The authors were unsure whether, if someone were to simply disseminate or share extremist propaganda among a network of individuals, that would even constitute training, or rather make them a 'trainer' or 'influential leader' in advancing an extremist ideology, requiring a rating of 'high'.

Table 4: Kappa and ICC values for history, action and capacity VERA-2R indicators		
History, action and capacity (HAC)	**Kappa (95% CI)**	**ICC (95% CI)**
Mean domain HAC	**0.75**	**0.82**
HAC 1 Early exposure to pro-violence, militant ideology	0.73 (0.48, 0.97)	0.80 (0.66, 0.88)
HAC 2 Network of family and friends involved in violent extremism	0.88 (0.74, 0.99)	0.92 (0.86, 0.95)
HAC 3 Previous criminal violence	0.78 (0.64, 0.92)	0.84 (0.72, 0.90)
HAC 4 Strategic, paramilitary and/or explosives training	0.70 (0.53, 0.87)	0.82 (0.67, 0.89)
HAC 5 Training in extremist ideology in own country or abroad	0.64 (0.46, 0.81)	0.71 (0.53, 0.82)
HAC 6 Organisational skills, access to funding and sources of help	0.74 (0.57, 0.89)	0.80 (0.67, 0.88)

Note: *N*=50. 95% CI=95% confidence interval; all *p*-values were found to be significant at *p*<0.001. Kappa=weighted Cohen's kappa. ICC=intraclass correlation coefficient

The eight indicators used to measure *Commitment and motivation* (CM) had the poorest average levels of agreement (see Table 5). Seven of these indicators were still considered to have acceptable levels of interrater reliability for established risk assessment tools (>0.61) (Douglas & Reeves 2010). The indicator CM 7 *Motivated by acquisition of status* indicated borderline moderate agreement (ICC=0.55, κ=0.48). Reasons for this are considered in the discussion and further commentary on this indicator and the CM domain is provided in that section.

Table 5: Kappa and ICC values for commitment and motivation VERA-2R indicators		
Commitment and motivation (CM)	**Kappa (95% CI)**	**ICC (95% CI)**
Mean domain CM	**0.74**	**0.79**
CM 1 Motivated by perceived religious obligation and/or glorification	0.87 (0.76, 0.98)	0.92 (0.86, 0.95)
CM 2 Motivated by criminal opportunism	0.81 (0.66, 0.96)	0.87 (0.78, 0.92)
CM 3 Motivated by camaraderie, group belonging	0.66 (0.49, 0.83)	0.73 (0.57, 0.84)
CM 4 Motivated by moral obligation, moral superiority	0.65 (0.47, 0.84)	0.75 (0.60, 0.85)
CM 5 Motivated by excitement and adventure	0.85 (0.65, 0.97)	0.85 (0.75, 0.91)
CM 6 Forced participation in violent extremisma		
CM 7 Motivated by acquisition of status	0.47 (0.26, 0.68)	0.54 (0.32, 0.71)
CM 8 Motivated by a search for meaning and significance in life	0.68 (0.48, 0.87)	0.68 (0.49, 0.80)

Note: *N*=50. 95% CI=95% confidence interval; all *p*-values were found to be significant at *p*<0.001. κ=weighted Cohen's kappa. ICC=intraclass correlation coefficient

Patterns across VERA-2R risk domains

The prevalence of risk indicators among the sample of 50 extremists was examined to identify any potential patterns (using the combined dataset). Risk judgements were also scrutinised across the overall sample and between violent and non-violent cases. General patterns were examined using descriptive statistics, *t*-tests and bivariate analysis.

Descriptive results

The most common risk factors among the sample were associated with *Beliefs, attitudes and ideology*, and specifically indicators that measured grievances (BA 1, BA 2 and BA 5) all received high ratings in over 66 percent of cases. Risk indicators relating to social networks, which in the VERA-2R are captured across two domains, were also high (HAC 2, 64%; SCI 3, 68%). Measures of intent were also high (see Figure 2). This suggests that the *Beliefs, attitudes and ideology* (BA) and *Social context and intention* (SCI) domains appear to dominate risk factors associated with radicalisation and violent extremism among a sample of Australian extremists. Results show some factors were less common and, in some cases, not present. Low prevalence of indicators seemed to cluster in the *Commitment and motivation* (CM) domain and were rarely reported as high and were not present (ie low) for the three indicators of motivated by criminal opportunism, motivated by excitement and adventure and forced participation in violent extremism (see Figure 4 and Table A1). The most common additional indicators were related to problematic upbringings or being placed in juvenile care and issues with school or work. Substance abuse and mental disorders was apparent (see Figure 5 and Table A2). This indicates unstable personal backgrounds, with mental disorders present in the sample.

Figure 1: Distribution of indicators across low, moderate and high risk ratings for beliefs, attitudes and ideology

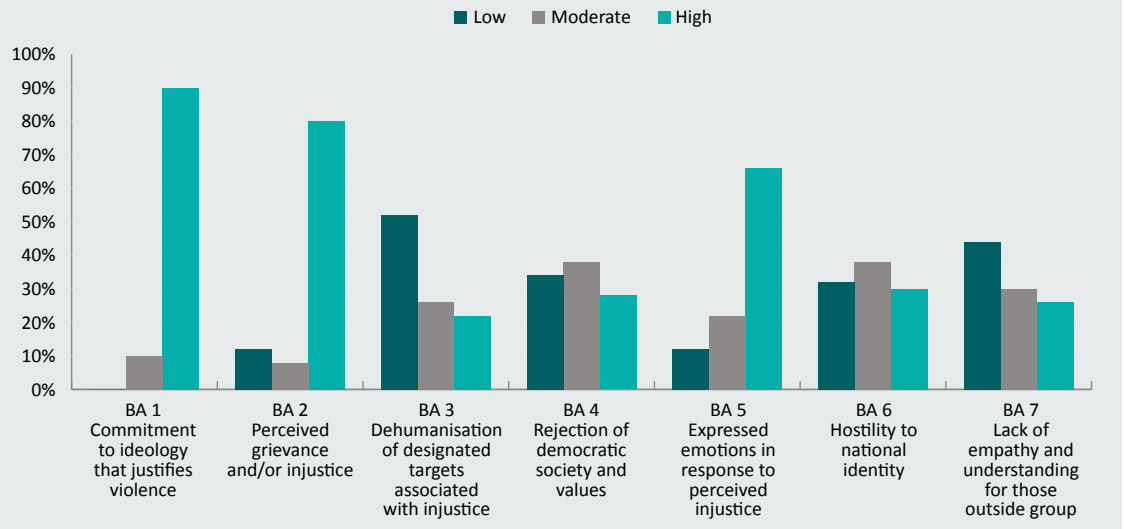


Figure 2: Distribution of indicators across low, moderate and high risk ratings for social context and intention

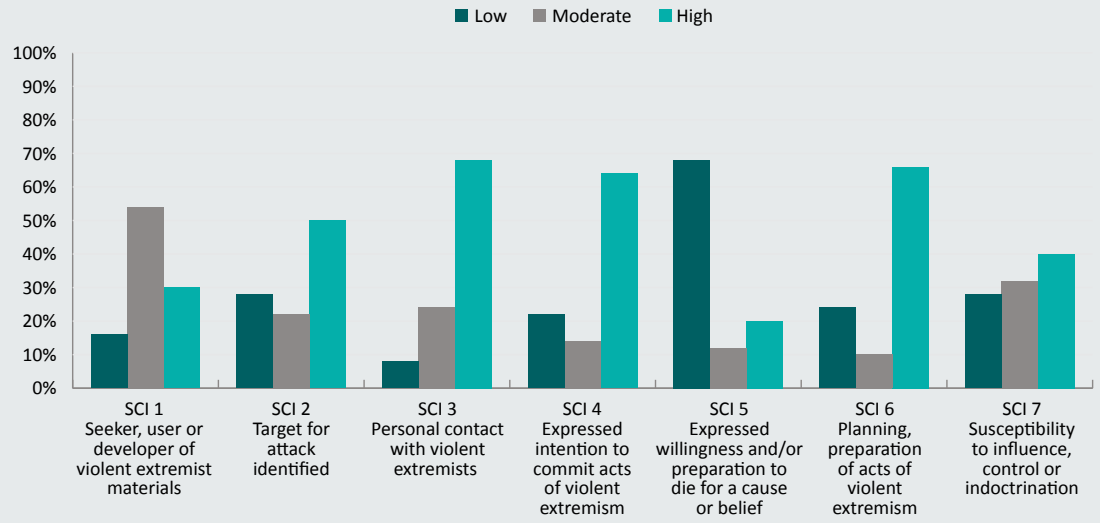


Figure 3: Distribution of indicators across low, moderate and high risk ratings for history, action and capacity

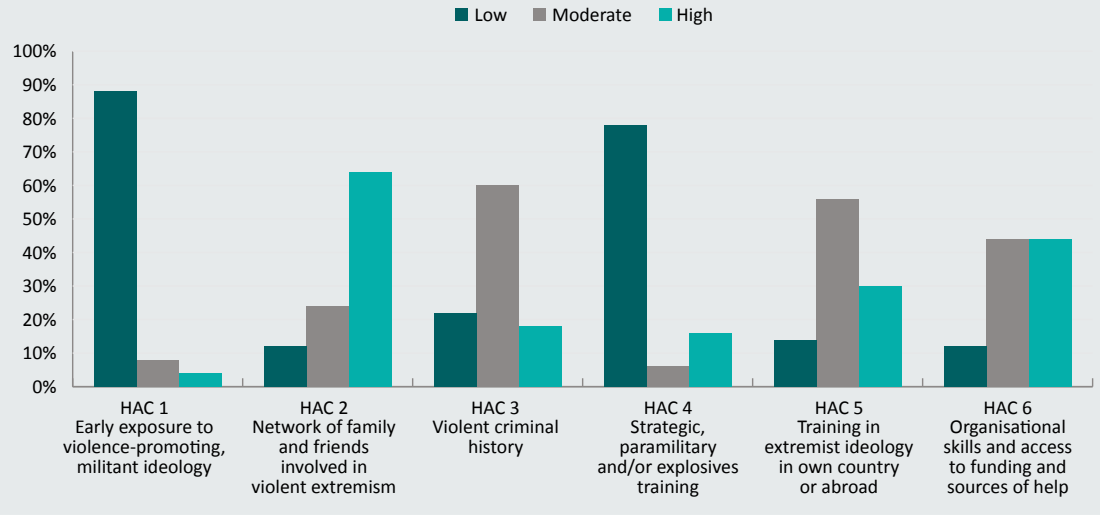


Figure 4: Distribution of indicators across low, moderate and high risk ratings for commitment and motivation

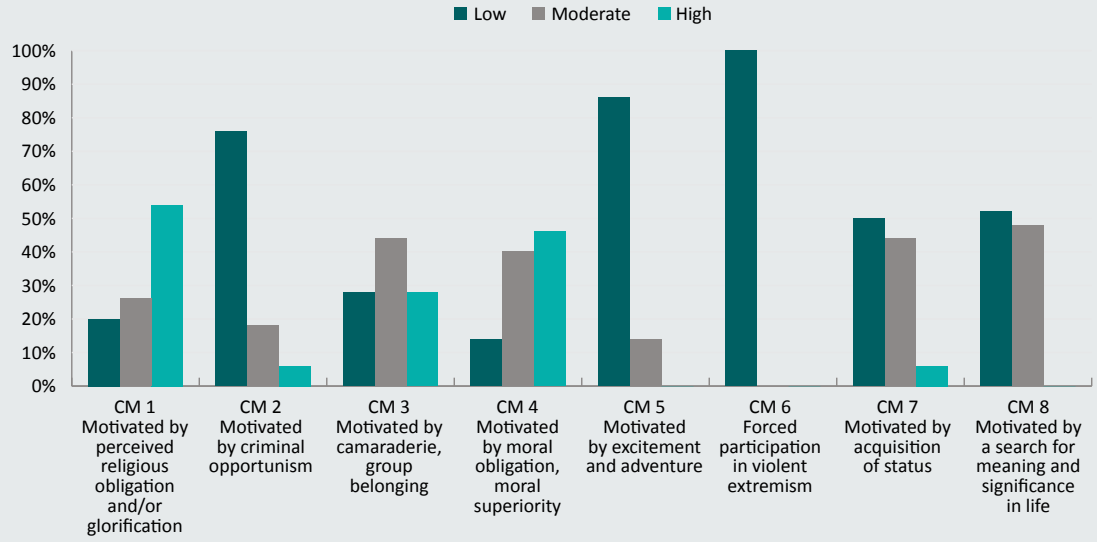
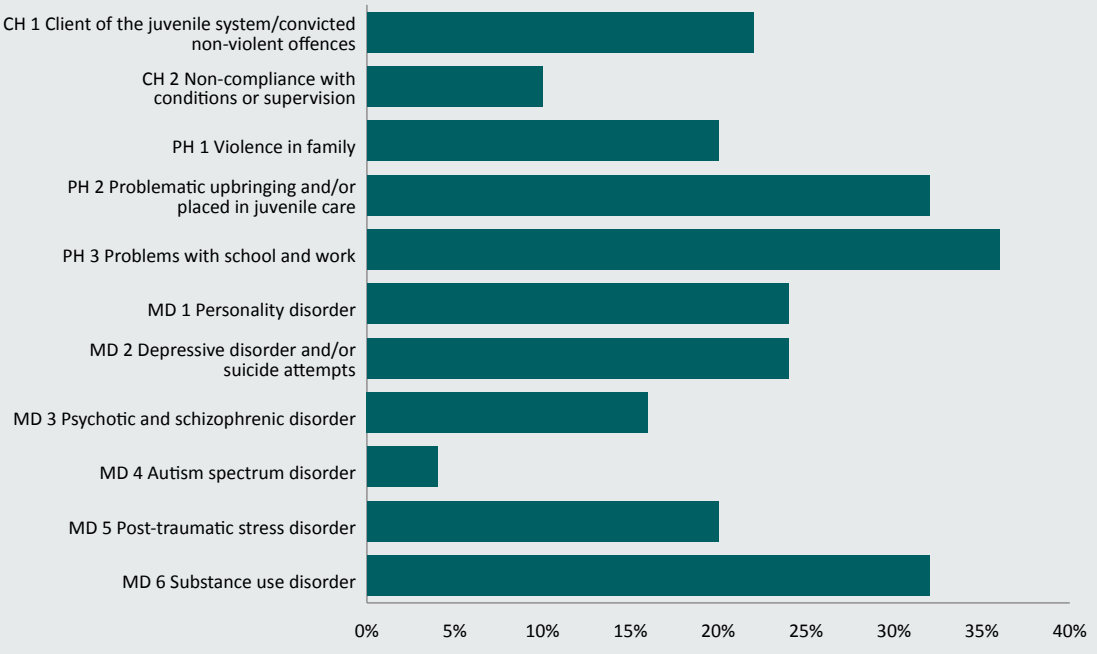


Figure 5: Distribution of additional indicators across the sample



VERA-2R indicators across violent compared with non-violent extremists

Results from chi-square analysis (Table 6) show four indicators across the four major VERA-2R risk domains to be associated with violent extremists, all of which were part of the *Beliefs, attitudes and ideology* category. Violent extremists were more likely to have a perceived grievance (p=0.033, φ=0.342), and express emotions in response to the grievance (p=0.033, φ=342), compared with those who were classified as non-violent. Violent extremists were also more likely to dehumanise designated targets associated with injustice (p=0.050, φ=0.296) compared with non-violent individuals. Compared with non-violent extremists, individuals who were categorised as violent demonstrated hostility towards a national identity (p=0.050, φ=0.278).

An unexpected finding was that non-violent extremists were more likely to have received paramilitary or explosives training (p=0.023, φ= −0.346) compared with those who were classified as violent. Upon closer inspection of the data, it was found that for those who received training, they were involved with facilitating overseas travel, producing and disseminating terrorist manuals and were leaders of terrorist organisations (all of whom are considered non-violent according to our classification).

Table 6: Major risk VERA-2R indicators association with violent and non-violent extremists

VERA-2R domain risk indicators	Violent (n=37)		Non-violent (n=13)		Sig (p)
	n	%	*n*	%	
Beliefs, attitudes and ideology indicators (BA)					
BA 1 Commitment to ideology that justifies violence	37	100.0	13	100.0	
BA 2 Perceived grievance and/or injustice	35	94.6	9	69.2	*
BA 3 Dehumanisation of designated targets associated with injustice	21	56.8	3	23.1	*
BA 4 Rejection of democratic society and values	25	67.6	8	61.5	
BA 5 Expressed emotions in response to perceived injustice	35	94.6	9	69.2	*
BA 6 Hostility to national identity	28	75.7	6	46.2	*
BA 7 Lack of empathy and understanding for those outside one's own group	22	59.5	6	46.2	
Social context and intention indicators (SCI)					
SCI 1 Seeker, user or developer of violent extremist materials	30	81.1	12	92.3	
SCI 2 Target for attack identified	28	75.7	8	61.5	
SCI 3 Personal contact with violent extremists	34	91.9	12	92.3	
SCI 4 Expressed intention to commit acts of violent extremism	30	81.1	9	69.2	
SCI 5 Expressed willingness and/or preparation to die for a cause or belief	12	32.4	4	30.8	
SCI 6 Planning, preparation of acts of violent extremism	29	78.4	9	69.2	
SCI 7 Susceptibility to influence, control or indoctrination	29	78.4	7	53.8	

Table 6: Major risk VERA-2R indicators association with violent and non-violent extremists (cont.)

VERA-2R domain risk indicators	Violent (*n*=37)		Non-violent (*n*=13)		Sig (*p*)
	n	%	*n*	%	
History, action and capacity indicators (HAC)					
HAC 1 Early exposure to violence-promoting, militant ideology	5	13.5	1	7.7	
HAC 2 Network of family and friends involved in violent extremism	33	89.2	11	84.6	
HAC 3 Violent criminal history	29	78.4	10	76.9	
HAC 4 Strategic, paramilitary and/or explosives training	5	13.5	6	46.2	*
HAC 5 Training in extremist ideology in own country or abroad	32	86.5	11	84.6	
HAC 6 Organisational skills and access to funding and sources of help	32	86.5	12	92.3	
Commitment and motivation indicators (CM)					
CM 1 Motivated by perceived religious obligation and/or glorification	30	81.1	10	76.9	
CM 2 Motivated by criminal opportunism	8	21.6	4	30.8	
CM 3 Motivated by camaraderie, group belonging	28	75.7	8	61.5	
CM 4 Motivated by moral obligation, moral superiority	32	86.5	11	84.6	
CM 5 Motivated by excitement and adventure	7	18.9	0	0.0	
CM 6 Forced participation in violent extremism	0	0.0	0	0.0	
CM 7 Motivated by acquisition of status	20	54.1	5	38.5	
CM 8 Motivated by a search for meaning and significance	20	54.1	4	30.8	

*statistically significant at $p<0.05$

Predictive validity of the VERA-2R

To test the predictive validity of the VERA-2R instrument, final risk judgements were subjected to five performance indicators: sensitivity and specificity, positive predictive values (PPV), negative predictive values (NPV) and area under the curve (AUC), to assess how well the VERA-2R tool performed in identifying high-risk and low-risk individuals. The validity of the VERA-2R instrument using these final risk judgements was first explored using descriptive statistics.

Final risk judgements

Each case was given a risk rating of either *low*, *moderate* or *high*—their final risk judgement. According to descriptive patterns (see Table 7), the two assessors rated violent extremists as high in 57 percent of cases, and as moderate risk in another 30 percent. However, five violent extremist cases were categorised as low risk. Risk judgements for non-violent extremist cases were categorised equally across low, moderate and high risk levels.

Table 7: Risk judgement distribution across violent extremists, non-violent extremists, and total sample

Risk judgement rating	Violent (*n*=37)		Non-violent (*n*=13)		Total sample (*N*=50)	
	n	%	*n*	%	*n*	%
Low	5	13.5	4	30.8	9	18.0
Moderate	11	29.7	5	38.4	16	32.0
High	21	56.8	4	30.8	25	50.0

To perform the four validity tests—sensitivity and specificity, positive predictive values (PPV), negative predictive values (NPV)—the final risk judgements were converted into a dichotomous risk outcome variable (high risk vs low risk). This was because these types of tests require a dichotomous indicator. There were equal proportions between risk categories: high risk (*n*=25, 50%) and low risk (*n*=25, 50%).

Sensitivity and specificity

The results for tests of sensitivity show the VERA-2R has a sensitivity value of 57 percent, indicating that 57 percent of cases classified as violent extremists were correctly judged to be high risk. The tests for specificity found a specificity value of 69 percent, indicating that 69 percent of cases classified as non-violent extremists were correctly identified as low risk during assessments. The use of a dichotomous low- and high-risk variable should be considered when interpreting sensitivity and specificity values because in practice risk levels fall on a continuum outside of being considered simply either low risk or high risk.

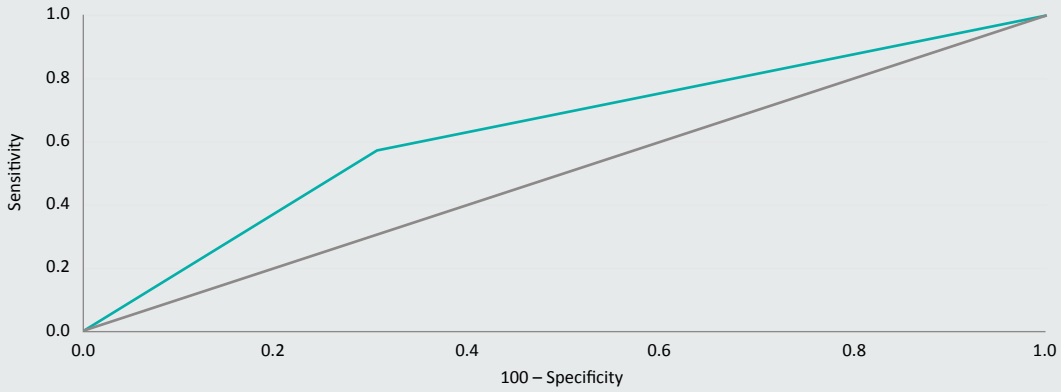
Positive and negative predictive values

Positive predictive values (PPV) demonstrated that, for those assessed as high risk using the VERA-2R, 84 percent of cases were violent extremists. This measure includes preparatory acts of violence (see above distinction between violent and non-violent extremists as captured in the sample). Negative predictive values (NPV) show that, among those assessed as low risk, 36 percent of cases were non-violent. That is, of the individuals who were judged as low risk, 36 percent did not commit an act of violent extremism. It should be noted that these values are influenced by base rates. PPV and NPV rely heavily on the distribution of the outcome variable, which in this study was whether extremists were violent or non-violent. In our study, violent offending base rates were higher than non-violent offending base rates.

Area under the curve

The AUC captures the probability that individuals classified as violent extremists compared with non-violent extremists received a higher risk classification. Risk classifications were measured using dichotomous classification of final risk judgements (where low and moderate risk ratings were judged as low risk and high ratings were judged as high risk; see Figure 6). The AUC value was 0.630 (95% CI [0.455, 0.805]), and the outcome was not significant (p>0.1), indicating poor predictive validity. It is important to note that sample size affects the accuracy of the result.

Figure 6: Area under the curve analysis for VERA-2R risk ratings (low vs high)



Discussion and conclusion

In this project 50 individual cases drawn from the Profiles of Individual Radicalisation in Australia database were assessed by both authors according to the VERA-2R risk assessment tool. This was informed by both the official VERA-2R training, engagements with practitioners and the academic background of the authors.

The results show that the VERA-2R overall had good interrater reliability. The tests for interrater reliability indicated consistency in risk judgements and in the assessment of available information. This provides evidence that the VERA-2R, despite being an SPJ tool that is underpinned by assessor discretion, can lead to consistent judgements. This appears to be the case when assessors are appropriately trained. However, it should be highlighted that a minority of specific indicators did demonstrate moderate to low interrater reliability.

The following needs to be kept in mind regarding the observed result for reliability. The outcome could be the result of the fact that the majority of cases captured within the sample were convicted for terrorism-related offences, with such cases easier to assess given the availability of information sources. There is the potential for bias to creep into assessments when knowing from the data sources that someone acted violently or non-violently. The academic background of the authors and knowledge about risk assessment, including familiarity with a number of the cases in the sample, also may have influenced the result. While not risk assessment practitioners, both authors received the VERA-2R training, and its application reflected as closely as possible how the tool is applied in the field. However, a different result for reliability may have been achieved with a more varied number of assessors.

When using the VERA-2R domains to map trends across the sample, the results indicated that some indicators do not seem to be associated with an Australian extremist cohort, which raises questions about their relevance. This included, for example, early exposure to militant ideologies (HAC 1), motivation related to criminal opportunism (CM 2) or forced participation in extremism (CM 6). Among the sample, certain indicators were more prevalent pertaining to the presence of grievances and the influence of social networks. Risk factors relating to *Beliefs, attitudes and ideology* and *Social context and intention* were more often associated with radicalisation and violent extremism among a sample of Australian extremists. In relation to the prevalence of grievances and radicalised social networks, this aligns with key risk factors identified in the broader literature (Belton, Cherney & Zahnow 2023; Cherney et al. 2022; Corner & Taylor 2023a; Thijssen et al. 2022, 2023). Based on whether someone is violent or non-violent, particular risk factors can vary according to the VERA-2R, with violent individuals in the sample having a clustering of factors in the *Beliefs, attitudes and ideology* (BA) domain, in particular relating to grievances and a sense of injustice, dehumanisation, emotional expressions in response to perceived injustice and hostility to national identity. It is perhaps the presence of these risk factors that helps these individuals overcome inhibitions towards the use of violence and provides justifications for violent action, and hence these factors are more pronounced among individuals who have acted violently. The prominence of indicators within the BA domain draws attention to the importance of addressing specific cognitions, thought processes and beliefs captured by VERA-2R risk indicators among violent extremist offenders, particularly given their relationship to violence.

When it came to testing the predictive validity of the VERA-2R, the analysis indicated the tool has some capacity to delineate between levels of risk but that this classification did not meet the required threshold. That is, our analysis showed differences in VERA-2R risk indicators associated with violent and non-violent individuals, with violent individuals demonstrating specific behavioural and motivational traits. However, the predictive validity of the VERA-2R was found to be low. This result must be interpreted with caution because of our sample size.

When it comes to the VERA-2R the findings are both positive and negative. The adopted methodology, as well as the results, do provide lessons both for violent extremism risk assessment and the application of the VERA-2R. We now canvass some of these issues.

Observations and lessons when applying specific indicators

The ambiguity of some VERA-2R indicators was raised in the reliability section (eg relating to the meaning of training and what this comprised). This is further compounded by the lack of clarity surrounding the explanation of some indicators and their theoretical basis (Corner & Taylor 2023b). The authors also observed that when rating specific indicators, the possibility of double counting could occur—hence leading to elevated risk judgements—because of the overlap across indicators or what might be assumed to be logical connections between certain factors.

For example, the first VERA-2R domain captures the role of *Beliefs, attitudes and ideology* that may promote violent extremism, with two specific indicators measuring elements of grievance and injustice: the presence of perceived grievances and/or injustice (BA 2) and the expression of emotions in response to perceived injustice (BA 5). The VERA-2R manual refers to emotional reactions such as anger, hate and frustration, but the authors, based on assessing their cases, could not work out under what circumstance you would have the presence of grievances and/or a sense of injustice and have no corresponding emotional response. Also, a sense of injustice is a form of grievance about something being unfair, with grievances part of an emotional reaction to that perceived injustice (Mikula 1986; Mikula, Scherer & Athenstaedt 1998; Nivette, Eisner & Ribeaud 2017; Van den Bos 2018). For a majority of the cases when the authors coded BA 2 as high, BA 5 was also high. The argument could be made this was an outcome of sample characteristics, but it could also be the fact that BA 2 and BA 5 capture the same process and theoretical constructs, leading to double counting.

Similar issues were revealed across other risk domains. For example, social networks are captured in SCI 3 and HAC 2 and both measure personal contact with other violent extremists. The main distinction is that SCI 3 measures the *frequency* of contact with either informal or formal networks and HAC 2 measures the *presence* of close family and friendship networks. However, again, how would frequency not equate to presence, with family and friendship networks being formal and informal group structures? Ambiguity and overlap were also encountered when measuring indicators in the *Commitment and motivation* (CM) domain. This is where the highest level of disagreement was found when it came to interrater reliability. This could be the result of the data sources used to undertake assessments, with this domain relating to psychological factors that might not be well captured in open sources, such as psychological reports. It is also potentially a highly subjective domain given its focus on motivations. Also, CM 8 *Motivated by a search for meaning and significance in life* derives itself from Kruglanski et al.'s (2014, 2017) significance quest theory, in which overcoming a loss of significance and seeking opportunities for gaining significance is a key motivator for violence extremist individuals (Kruglanski et al. 2009). The problem is that pursuing a sense of status is an important aspect of that theory, with its acquisition a motivator for seeking significance (Kruglanski et al. 2009; Schuurman & Carthy 2023). This is referred to, and separately captured, in CM 7 (*Motivated by acquisition of status*). While practitioners that apply the VERA-2R might not be aware of these various theoretical subtleties, it does raise the issue about whether certain indicators are capturing similar processes, the possibility of double counting when applying the VERA-2R in practice and the redundancy of some indicators.

The issue of measuring predictive validity

In this project we set out to test the validity of the VERA-2R, undertaking common tests that are used in risk assessment research and that also have been applied in various studies. However, there are limitations in the design and research sample that have a bearing on any conclusion that the VERA-2R does or does not have predictive validity. Also, a true test of predictive validity requires certain conditions, which as outlined by Hart and Logan (2011: 90) are very difficult to obtain:

> It is necessary to recruit a cohort of patients and offenders, assess them, follow them up over a long period of time, and then detect violence that occurs in institutional or community settings. The sample must be sufficiently large and the follow-up sufficiently long to yield a base rate of violence amenable to statistical analysis.

Meeting these basic conditions when it comes to terrorist offending and reoffending is somewhat challenging. We know terrorist offending has a low base rate, with reoffending rates also being relatively low (Whelan, Bright & Fletcher 2021). Moreover, in some circumstances when terrorist offenders are released from prison, they are subject to what can be categorised as an artificial experience of reintegration. This is because terrorist offenders can be placed on extended supervision and control orders, have restrictions imposed on their movements and associations, including with the use of electronic devices, and be subject to surveillance and periodic visits and compliance checks by many different agencies from corrections to state and federal police. This can either artificially suppress the risk of reoffending because normal conditions that might lead to recidivism (exposure to previous associates) are constrained, or might increase reoffending risk because of the frustrations these conditions cause, pushing an individual to re-engage. Hence in the terrorism space there can be a range of factors that impact on offending and reoffending risk. It should be acknowledged that no authors of the most commonly used extremism assessment tools, whether that be the VERA-2R, TRAP-18, ERG 22+ or Radar, claim they are predictive. Any debate about the predictive validity of the VERA-2R needs to be qualified by a recognition of the nuances surrounding measures of validity as they relate to the application of a risk assessment tool to certain cohorts, and the various factors that may or may not influence reoffending risk. It also needs to be prefaced with a consideration of factors that influence engagement in, and disengagement from, violent extremism. This requires tools and approaches that not only focus on risks, but also individual needs, and that aim to promote behavioural change relating to disengagement from violent extremism.

Future research

This project has contributed to the emerging research on the VERA-2R. Future projects employing larger sample sizes with a wider variety of cohorts would help to further assess the application of the VERA-2R and its reliability and validity. Reliability testing should ideally involve practitioners trained in the VERA-2R and who apply it as part of their roles. Finally, examining how the VERA-2R tool is applied in the field by practitioners is recommended, with studies on risk assessment implementation helping to provide insights on the application of the tool and the relevance of its various indicators.

References

Alberda D, 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. https://doi.org/10.3389/fpsyg.2022.1000186

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

Allely CS & Wicks SJ 2022. The feasibility and utility of the Terrorist Radicalization Assessment Protocol (TRAP-18): A review and recommendations. *Journal of Threat Assessment and Management* 9(4): 218–259

Altman DG & Bland JM 1994. Statistics notes: Diagnostic tests 2: Predictive values. *British Medical Journal* 309: 102

Beardsley NL & Beech AR 2013. Applying the violent extremist risk assessment (VERA) to a sample of terrorist case studies. *Journal of Aggression, Conflict and Peace Research* 5(1): 4–15. https://doi.org/10.1108/17596591311290713

Becker MH 2019. When extremists become violent: Examining the association between social control, social learning, and engagement in violent extremism. *Studies in Conflict and Terrorism* 44(12): 1104–1124. https://doi.org/10.1080/1057610X.2019.1626093

Belton E & Cherney A 2023. *Profiles of Individual Radicalisation in Australia (PIRA) Codebook (draft)*. University of Queensland

Belton E, Cherney A & Zahnow R 2023. Profiles of Individual Radicalisation in Australia (PIRA): Introducing an Australian Open-Source Extremist Database. *Perspectives on Terrorism* 17(1): 18–35

Böckler N, Allwinn M, Metwaly C, Wypych B, Hoffmann J & Zick A 2020. Islamist terrorists in Germany and their warning behaviors: A comparative assessment of attackers and other convicts using the TRAP-18. *Journal of Threat Assessment and Management* 7(3–4): 157

Bonett DG 2002. Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics* 27(4): 335–340. https://doi.org/10.3102/10769986027004335

Borum R 2015. Assessing risk for terrorism involvement. *Journal of Threat Assessment and Management* 2(2): 63–87

Challacombe DJ & Lucas PA 2019. Postdicting violence with sovereign citizen actors: An exploratory test of the TRAP-18. *Journal of Threat Assessment and Management* 6(1): 51

Chermak SM, Freilich JD, Parkin WS & Lynch JP 2012. American terrorism and extremist crime data sources and selectivity bias: An investigation focusing on homicide events committed by far-right extremists. *Journal of Quantitative Criminology* 28(1): 191–218

Cherney A, Belton E, Norham SAB & Milts J 2022. Understanding youth radicalisation: An analysis of Australian data. *Behavioral Sciences of Terrorism and Political Aggression* 14(2): 97–119

Cherney A, Grossman M & Khalil L 2022. Guest editorial: Special issue on violent extremist risk assessment. *Journal of Policing, Intelligence and Counter Terrorism* 17(3): 235–245

Cohen J 1988. *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge Academic

Cook AN, Hart SD & Kropp PR 2014. Multi-Level Guidelines for the assessment and management of group-based violence, Version 2. Burnaby, Canada: Mental Health, Law & Policy Institute, Simon Fraser University

Cook NR 2007. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 115(7): 928–935. https://doi.org/10.1161/circulationaha.106.672402

Corner E & Pyszora N 2022. The Terrorist Radicalization Assessment Protocol-18 (TRAP-18) in Australia: Face validity, content validity, and utility in the Australian context. *Journal of Policing, Intelligence and Counter Terrorism* 17(3): 246–268

Corner E & Taylor H 2023a. *Grievance-fuelled violence: Modelling the process of grievance development*. Research Report no. 27. Canberra: Australian Institute of Criminology. https://doi.org/10.52922/rr78917

Corner E & Taylor H 2023b. *Testing the reliability, validity and equity of terrorism risk assessment instruments.* Centre for Social Research and Methods, The Australian National University (FOI released document version). https://www.homeaffairs.gov.au/foi/files/2023/fa-230400097-document-released-part-1.PDF

Cubitt T & Wolbers H 2023. *Review of violent extremism risk assessment tools in Division 104 control orders and Division 105A post-sentence orders*. Special reports. Canberra: Australian Institute of Criminology. https://www.aic.gov.au/publications/special/special-14

de Bruin A, Duits N, Kempes M & Prinsen M 2022. *Interrater and intrarater reliability of the Violent Extremism Risk Assessment tool.* Custodial Institutions Agency, Ministry of Justice, Science and Education, NIFP, Netherlands

Douglas K & Kropp P 2002. A prevention-based paradigm for violence risk assessment. *Criminal Justice and Behavior* 29(5): 617–658

Douglas K & Reeves KA 2010. Historical-Clinical-Risk Management—20 (HCR-20) Violence Risk Assessment Scheme: Rationale, application, and empirical overview. In RK Otto & KS Douglas (eds), *Handbook of violence risk assessment*. Routledge/Taylor & Francis Group: 147–185

Duits N, Alberda DL & Kempes M 2022. Psychopathology of young terrorist offenders, and the interaction with ideology and grievances*. Frontiers in Psychiatry* 13: 801751. https://doi.org/10.3389/fpsyt.2022.801751

Fleiss JL, Levin B & Paik MC 2013. *Statistical methods for rates and proportions*. Hoboken, NJ: John Wiley & Sons

Geraghty K & Woodhams J 2015. The predictive validity of risk assessment tools for female offenders: A systematic review. *Aggression and Violent Behavior* 21. https://doi.org/10.1016/j.avb.2015.01.002

Gill P 2015. *Lone-actor terrorists: A behavioural analysis*. Routledge

Gill P, Corner E, McKee A, Hitchen P & Betley P 2019. What do closed source data tell us about lone actor terrorist behavior? A research note. *Terrorism and Political Violence*. https://doi.org/10.1080/09546553.2019.1668781

Gill P, Marchment Z, Zolghadriha S, Salman N, Rottweiler B, Clemmow C & Vegt IVD 2020. Advances in violent extremist risk analysis. In D Silva & M Deflem (ed), *Radicalization and counter-radicalization* (*sociology of crime, law and deviance*, Vol 25). Emerald Publishing: 55–74

Grooten WJA, Tseli E, Äng BO, Boersma K, Stålnacke BM, Gerdle B & Enthoven P 2019. Elaborating on the assessment of the risk of bias in prognostic studies in pain rehabilitation using QUIPS-aspects of interrater agreement. *Diagnostic and Prognostic Research* 3(5). https://doi.org/10.1186/s41512-019-0050-0

Hallgren KA 2012. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology* 8(1): 23–34

Hanczar B, Hua J, Sima C, Weinstein J, Bittner M & Dougherty ER 2010. Small-sample precision of ROC-related estimates. *Bioinformatics* 26(6): 822–830. https://doi.org/10.1093/bioinformatics/btq037

Hart SD, Cook AN, Pressman DE, Strang S & Lim YL 2017. *A concurrent evaluation of threat assessment tools for the individual assessment of terrorism*. Waterloo, Ontario: Canadian Network for Research on Terrorism, Security, and Society

Hart SD, Douglas KS & Guy LS 2016. The structured professional judgement approach to violence risk assessment: Origins, nature, and advances. In DP Boer et al. (eds), *The Wiley handbook on the theories, assessment, and treatment of sexual offending*. Wiley-Blackwell

Hart SD & Logan C 2011. Formulation of violence risk using evidence-based assessments: The structured professional judgment approach. In P Sturmey & M McMurran (eds), *Forensic case formulation*. Chichester, UK: Wiley-Blackwell: 83–106

Hassan G et al. 2022. PROTOCOL: Are tools that assess risk of violent radicalization fit for purpose? A systematic review. *Campbell Systematic Reviews* 18: e1279. https://doi.org/10.1002/cl2.1279

Hayes AF & Krippendorff K 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1(1): 77–89. https://doi.org/10.1080/19312450709336664

Herzog-Evans M 2018. A comparison of two structured professional judgment tools for violent extremism and their relevance in the French context. *European Journal of Probation* 10(1): 3–27

Independent National Security Legislation Monitor 2023. *Review into Division 105A (and related provisions) of the Criminal Code Act 1995* (Cth). Commonwealth of Australia

Koo TK & Li MY 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15(2): 155–163

Krippendorff K 2013. *Content analysis: An introduction to its methodology*. Sage Publications

Krippendorff K 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement* 30(1): 61–70

Kruglanski AW, Chen X, Dechesne M, Fishman S & Orehek E 2009. Fully committed: Suicide bombers' motivation and the quest for personal significance. *Political Psychology* 30(3): 331–357

Kruglanski AW, Gelfand MJ, Bélanger JJ, Sheveland A, Hetiarachchi M & Gunaratna R 2014. The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Advances in Political Psychology* 35(Supplement S1): 69–93

Kruglanski AW, Jasko K, Chernikova M, Dugas M & Webber D 2017. To the fringe and back: Violent extremism and the psychology of deviance. *American Psychologist* 72(3): 217

Lacy S & Riffe D 1996. Sampling error and selecting intercoder reliability samples for nominal content categories. *Journalism & Mass Communication Quarterly* 73(4): 963–973. https://doi.org/10.1177/107769909607300414

LaFree G, Jensen MA, James PA & Safer-Lichtenstein A 2018. Correlates of violent political extremism in the United States. *Criminology* 56(2): 233–268

Landis JR & Koch GG 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159–174

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.

Logan C & Lloyd M 2019. Violent extremism: A comparison of approaches to assessing and managing risk. *Legal and Criminological Psychology* 24: 141–161

Meloy R 2017. Terrorist Radicalization Assessment Protocol (TRAP-18)

Meloy JR, Goodwill AM, Meloy MJ, Amat G, Martinez M & Morgan M 2019. Some TRAP-18 indicators discriminate between terrorist attackers and other subjects of national security concern. *Journal of Threat Assessment and Management* 6(2): 93

Meloy JR, Roshdi K, Glaz-Ocik J & Hoffmann J 2015. Investigating the individual terrorist in Europe. *Journal of Threat Assessment and Management* 2(3–4): 140

Mikula G 1986. The experience of injustice: Toward a better understanding of its phenomenology. In HW Bierhoff, RL Cohen & J Greenberg (eds), *Justice in social relations*. Critical Issues in Social Justice. Boston, MA: Springer: 103–123. https://doi.org/10.1007/978-1-4684-5059-0_6

Mikula G, Scherer KR & Athenstaedt U 1998. The role of injustice in the elicitation of differential emotional reactions. *Personality and Social Psychology Bulletin* 24(7): 769–783

Monahan J 2012. The individual risk assessment of terrorism. *Psychology, Public Policy, and Law* 18: 167–205

Nivette A, Eisner M & Ribeaud D 2017. Developmental predictors of violent extremist attitudes: A test of general strain theory. *Journal of Research in Crime and Delinquency 54*(6): 755–790

Olver M, Stockdale K & Wormith J 2009. Risk assessment with young offenders: A meta-analysis of three assessment measures. *Criminal Justice and Behavior* 36(4): 329–353

Pressman E 2016. The complex dynamic causality of violence extremism: Applications of the VERA-2R risk assessment method to CVE initiatives. In AJ Masys (ed), *Disaster forensics: Advanced sciences and technologies for security applications*. International Publishing. https://doi.org/10.1007/978-3-319-41849-0_10

Pressman E, Duits N, Rinne T & Flockton J 2018. *Violent Extremism Risk Assessment—Version 2 Revised*. Utrecht, Netherlands: Netherlands Institute of Forensic Psychiatry and Psychology

Pressman E & Flockton J 2014. Violent extremist risk assessment: Issues and applications of the VERA-2 in a high-security correctional setting. In A Silke (ed), *Prisons, terrorism and extremism: Critical issues in management, radicalisation and reform*. Routledge

Pressman E & Flockton J 2012. Calibrating risk for violent political extremists and terrorists: The VERA 2 structured assessment. *British Journal of Forensic Practice* 14(4): 237–251

Renwick J 2018. *Report to the Prime Minister: The prosecution and sentencing of children for terrorism*. Australian Government—Independent National Security Legislation Monitor

Ripperger B 2021. The use of terrorism risk assessment tools in Australia to manage residual risk. In *Terrorism risk assessment instruments*. IOS Press: 165–192

Sarma KM 2017. Risk assessment and the prevention of radicalization from nonviolence into terrorism. *American Psychologist* 72(3): 278

Schuurman B & Carthy SL 2023. Understanding (non) involvement in terrorist violence: What sets extremists who use terrorist violence apart from those who do not? *Criminology & Public Policy.* https://doi.org/10.1111/1745-9133.12626

Shepherd S & Lewis-Fernandez R 2016. Forensic risk assessment and cultural diversity: Contemporary challenges and future directions. *Psychology and Public Policy* 22(4): 427–438

Shrout PE & Fleiss JL 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86(2): 420–428

Sim J & Wright CC 2005. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy* 85(3): 257–268

Singh JP 2013. Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law* 31(1): 8–22

Singh JP, Desmarais SL & Van Dorn RA 2013. Measurement of predictive validity in violence risk assessment studies: A second-order systematic review. *Behavioral Sciences & the Law* 31(1): 55–73

Singh JP, Grann M & Fazel S 2011. A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review* 31(3): 499–513

Tang W, Hu J, Zhang H, Wu P & He H 2015. Kappa coefficient: A popular measure of rater agreement. *Shanghai Archives of Psychiatry* 27(1): 62–67. https://doi.org/10.11919/j.issn.1002-0829.215010

Thijssen G, Masthoff E, Sijtsema JJ & Bogaerts S 2023. Understanding violent extremism: Identifying motivational classes in male jihadist detainees. *International Journal of Offender Therapy and Comparative Criminology* 67(15): 1455–1473. https://doi.org/10.1177/0306624X221144295

Thijssen G, Masthoff E, Sijtsema JJ & Bogaerts S 2022. Understanding violent extremism: Risk and protective factors in a jihadi male detainee population in the Netherlands. *European Journal of Criminology* 20(3): 973–995. https://doi.org/10.1177/14773708221132887

Van den Bos K 2018. *Why people radicalize: How unfairness judgments are used to fuel radical beliefs, extremist behaviors, and terrorism*. Oxford University Press

Vanbelle S 2016. A new interpretation of the weighted kappa coefficients. *Psychometrika* 81(2): 399–410. https://doi.org/10.1007/s11336-014-9439-4

Vincent GM, Guy LS, Fusco SL & Gershenson BG 2012. Field reliability of the SAVRY with juvenile probation officers: Implications for training. *Law and Human Behavior* 36(3): 225–236

Walter SD, Eliasziw M & Donner A 1998. Sample size and optimal designs for reliability studies. *Statistics in Medicine* 17(1): 101–110

Whelan C, Bright D & Fletcher P 2021. *Rapid evidence assessment: An international review of terrorist recidivism*. Deakin University, Addressing Violent Extremism and Radicalisation to Terrorism (AVERT) Research Network. https://www.avert.net.au/publications

Whitehead PR, Ward T & Collie RM 2007. Time for a change: Applying the Good Lives Model of rehabilitation to a high-risk violent offender. *International Journal of Offender Therapy and Comparative Criminology* 51(5): 578–598

Appendix: VERA-2R domains

Table A1: VERA-2R risk across the sample (*N*=50)						
	Low		Moderate		High	
	n	%	*n*	%	*n*	%
Beliefs, attitudes and ideology (BA)						
BA 1 Commitment to ideology that justifies violence	0	0.0	5	10.0	45	90.0
BA 2 Perceived grievance and/or injustice	6	12.0	4	8.0	40	80.0
BA 3 Dehumanisation of designated targets associated with injustice	26	52.0	13	26.0	11	22.0
BA 4 Rejection of democratic society and values	17	34.0	19	38.0	14	28.0
BA 5 Expressed emotions in response to perceived injustice	6	12.0	11	22.0	33	66.0
BA 6 Hostility to national identity	16	32.0	19	38.0	15	30.0
BA 7 Lack of empathy and understanding for those outside one's own group	22	44.0	15	30.0	13	26.0
Social context and intention (SCI)						
SCI 1 Seeker, user or developer of violent extremist materials	8	16.0	27	54.0	15	30.0
SCI 2 Target for attack identified (person, group, location)	14	28.0	11	22.0	25	50.0
SCI 3 Personal contact with violent extremists (informal or social context)	4	8.0	12	24.0	34	68.0
SCI 4 Expressed intention to commit acts of violent extremism	11	22.0	7	14.0	32	64.0
SCI 5 Expressed willingness and/or preparation to die for a cause or belief	34	68.0	6	12.0	10	20.0
SCI 6 Planning, preparation of acts of violent extremism	12	24.0	5	10.0	33	66.0
SCI 7 Susceptibility to influence, control or indoctrination	14	28.0	16	32.0	20	40.0

Table A1: VERA-2R risk across the sample (*N*=50) (cont.)						
	Low		Moderate		High	
	n	%	*n*	%	*n*	%
History, action and capacity (HAC)						
HAC 1 Early exposure to violence-promoting, militant ideology	44	88.0	4	8.0	2	4.0
HAC 2 Network of family and friends involved in violent extremism	6	12.0	12	24.0	32	64.0
HAC 3 Violent criminal history	11	22.0	30	60.0	9	18.0
HAC 4 Strategic, paramilitary and/or explosives training	39	78.0	3	6.0	8	16.0
HAC 5 Training in extremist ideology in own country or abroad	7	14.0	28	56.0	15	30.0
HAC 6 Organisational skills and access to funding and sources of help	6	12.0	22	44.0	22	44.0
Commitment and motivation (CM)						
CM 1 Motivated by perceived religious obligation and/or glorification	10	20.0	13	26.0	27	54.0
CM 2 Motivated by criminal opportunism	38	76.0	9	18.0	3	6.0
CM 3 Motivated by camaraderie, group belonging	14	28.0	22	44.0	14	28.0
CM 4 Motivated by moral obligation, moral superiority	7	14.0	20	40.0	23	46.0
CM 5 Motivated by excitement and adventure	43	86.0	7	14.0	0	0.0
CM 6 Forced participation in violent extremism	50	100.0	0	0.0	0	0.0
CM 7 Motivated by acquisition of status	25	50.0	22	44.0	3	6.0
CM 8 Motivated by a search for meaning and significance in life	26	52.0	24	48.0	0	0.0

Table A2: VERA-2R additional indicators across the sample (*N*=50)				
	Yes		No	
	n	%	*n*	%
CH Criminal history				
CH 1 Client of the juvenile system/convicted non-violent offences	11	22.0	39	78.0
CH 2 Non-compliance with conditions or supervision	5	10.0	45	90.0
PH Personal history				
PH 1 Violence in family	10	20.0	40	80.0
PH 2 Problematic upbringing and/or placed in juvenile care	16	32.0	34	68.0
PH 3 Problems with school and work	18	36.0	32	64.0
MD Mental disorder				
MD 1 Personality disorder	12	24.0	38	76.0
MD 2 Depressive disorder and/or suicide attempts	12	24.0	38	76.0
MD 3 Psychotic and schizophrenic disorder	8	16.0	42	84.0
MD 4 Autism spectrum disorder	2	4.0	48	96.0
MD 5 Post-traumatic stress disorder	10	20.0	40	80.0
MD 6 Substance use disorder	16	32.0	34	68.0

CRG reports

CRG 40/21–22

Adrian Cherney is a Professor in the School of Social Science at the University of Queensland.

Emma Belton is a Research Fellow in the Griffith Criminology Institute at Griffith University.

www.aic.gov.au/crg