

Testing the Reliability, Validity, and Equity of Terrorism Risk Assessment Instruments

Dr Emily Corner and Dr Helen Taylor
Centre for Social Research and Methods
The Australian National University



Australian
National
University

CENTRE FOR SOCIAL
RESEARCH & METHODS

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Executive Summary

Background

Recent policies to counter the terrorist threat in Australia have moved away from coercive actions and powers, and towards efforts founded in public health theories with the aim to identify, prevent and manage individuals at risk of radicalisation and violent extremism. Depending on the specification of risk posed by individuals, different risk assessment instruments should be used at different intervention points along the pathway to violent extremism and terrorism. However, a challenge for practitioners is knowing which instruments to use in order to accurately assess risk. Since 2015, the Australian Government has implemented at least two formal risk assessment instruments as part of their efforts to counter violent extremism, including the VERA-2R and Radar.

With the correct instrument, and with the correct implementation, risk assessment ensures that correctional programming is delivered in such a way that those at highest risk of offending receive the most intensive interventions, while those presenting as lower risk are kept from receiving too much intervention and therefore avoid the disruption of circumstances that were making them low risk in the first place. Thus, the proper allocation of scarce resources is contingent on the accurate identification of those who pose the highest risk, and of those who would benefit from various management strategies.

This research project undertakes a holistic and impartial analysis of the VERA-2R and Radar to demonstrate the extent to which these risk assessment instruments accurately classify offenders or overestimate or underestimate the risk they pose. In doing so, this research project also provides the most comprehensive overview of the state of the empirical knowledge of the causes of radicalisation and terrorism to date. The findings from this project should be used to inform the development of policy and practice in Australia's response to countering radicalisation and violent extremism.

This is the first piece of research to be performed on these instruments. Instruments in the general violence domain have been assessed multiple times using different population samples. We conducted an evaluation using the leading standard experimental methodology. Whilst the results of this method highlight the findings regarding both predictive validity and inter-rater reliability, these findings were drawn from this specific setting. Outcomes from the instruments in a non-experimental setting would likely differ. With this in mind, the following findings, implications and recommendations are now outlined.

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

Overview of Findings

This research was divided into four tasks. The first two tasks focused on the theoretical and empirical underpinnings of the instruments. The final task performed a series of experiments, scrutinising the risk factors within the instruments through assessment of the vignettes developed in the third task. The overall outcomes of this research have identified that both the VERA-2R and Radar lack a strong theoretical and empirical foundation, and have poor inter-rater reliability and questionable predictive validity.

More specifically, the findings have highlighted:

- There is much confusion as to the risk specification of the instruments, i.e. what setting the instruments should be applied in, and which individuals should be subject to assessments.
- The theoretical and empirical evidence base cited in the instruments' user manuals and supporting documentation is of poor quality, and is predominately composed of theoretical assertions, secondary citations of literature reviews, and media articles.¹
- Despite assertions by the authors of the instruments, neither instrument follows a true structured professional judgement (SPJ) approach, and it is more appropriate to describe them as SPJ 'lite' instruments.
- The wider empirical evidence base has identified over 1500 variables found to be statistically associated with movements towards radicalisation and violent extremism.
 - Of these variables, only three major themes were found to be consistent across the literature; the importance of social networks, being male, and being younger.

¹ However, it is likely that factors were also informed by the authors' communication with practitioners in law enforcement, intelligence and correctional agencies, and that these communications would not be appropriate to cite in the manuals. This is recognised in the thematic analysis of the VERA-2R, as the authors do note that practitioners were involved in the formulation of earlier iterations of the instrument. However, there is less information in the work underpinning Radar.

- Although both the VERA-2R and Radar include factors that contain social elements, these factors are not explicitly connected to the social networks of the individuals. Rather, these factors are thematically closer to beliefs and radical behaviours.
- The cited evidence base for both instruments is not reflective of the high-quality research that is freely available.
 - Both instruments lack the inclusion of the majority of variables identified as significantly associated with radicalisation and terrorism. The VERA-2R covers 14.2% of variables, and Radar only 5.9%.
- Despite assertions made by the authors of the instruments that predictive validity is not ascertainable, many studies evaluating a range of risk assessment instruments across multiple domains have identified that it is attainable. This research identified that for the sample employed, Radar offered good, and the VERA-2R poor, predictive validity.²³ Radar's better performance on predictive validity is the likely due to at least two factors:
 - First, qualitative evidence suggests that assessors did not rely on the instrument to make risk decisions, and instead used their own knowledge and experience.
 - Second, Radar's ambiguous risk specification (with seven different risk specifications identified) allowed for greater scope of cases for inclusion and thus a wider range of predictive outcomes.
- Neither instrument shows any acceptable standard of inter-rater reliability.

² It is also noted that an outcome of poor predictive validity does not necessarily negate the use of an instrument. Instruments with weaker predictive validity are still useful for the design of risk formulations, management strategies, and scenario planning. However, these uses were not assessed during this research. Further research should seek to assess the instruments' validity in these areas also.

³ It is recommended that this statement be supplemented with the discussion section in Task 4.

Implications

The lack of evidence underpinning both instruments has potentially serious implications for their validity and reliability. Without a strong theoretical and empirical basis for factor inclusion, it is not reasonable to anticipate that the instruments are able to predict their specified risk with anything other than chance. If an instrument with a weak evidence base is employed as a predictive instrument by practitioners, it is not possible to determine if individuals who pass through assessment processes would ever be suitable for the management plan as determined by the risk decision outcome made on the instrument.

The results regarding the predictive validity and inter-rater reliability of both instruments were identified during an experimental protocol, where conditions are not equivalent to the real-world situations that the instruments are employed in.⁴ The qualitative feedback from users casts doubt over the extent of use and thus the true value of the process and risk factors within the instruments during the assessment of risk of individuals. The confusion over the risk specifications of both instruments also has implications for how these instruments are used by assessors. Feedback implies that neither instrument is semantically or thematically coherent, and that they are both overly complex and induce fatigue. This affects the inter-rater reliability, which, when used in the field, has significant implications for the consistency in assessments across users.

Recommendations

This report puts forward four main recommendations that would improve the quality of the two instruments for future use:

- First, the research conducted here was an impact evaluation, where the effectiveness of risk assessment instruments was holistically assessed using an experimental protocol. Given the feedback from assessors, and the implications noted above, it is recommended that a process evaluation is undertaken to evaluate the actual implementation of these instruments in the field. This will allow for the determination of the true

⁴ This evaluation employed a random assignment of experimental and control conditions during the assessment process. This is recognised as the leading standard in evaluation design. However, this methodology is not able to assess how the instruments are actually employed in a real-world setting. Therefore, the results of Task 4 must be interpreted with this in mind.

user experience for professionals who employ the instruments in the field, aid the understanding of the reality of assessment processes outside of an experimental setting, help determine 'What Works', and identify best practice for risk assessment and management of those suspected of moving towards violent extremism.

- Second, it is recommended that the authors of both instruments enact far more thorough evaluations of the wider theoretical and empirical literature to help develop risk factors that accurately reflect behavioural trajectories towards radicalisation and terrorist violence. A strong theoretical and empirical foundation is critical for the deployment of these instruments, and forms the basis for any tests of validity, reliability and equity that they are subject to. This research identified over 1500 variables across the literature, which were not identified by the authors as pertinent. Introduction of these variables into the risk factors would aid in the formulation of holistic and consistent risk specifications, which would help support the validity, reliability, and equity of the instruments.
- Third, this research identified that both instruments' approach is better classified as SPJ 'lite'. This is likely to be a more appropriate approach due to the intended user audience. However, this should be reflected in the user manuals and training protocols for each instrument.
- Fourth, inter-rater reliability, which was found to be poor for both instruments, would be improved by refining the risk specification of each instrument. It is also expected that if the theoretical and empirical basis of each instrument, and risk factors within, were strengthened, this too would improve inter-rater reliability.

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Table of Contents

Testing the Reliability, Validity, and Equity of Terrorism Risk Assessment Instruments

Executive Summary.....	1
Background.....	1
Overview of Findings.....	2
Implications	4
Recommendations.....	4
List of Tables.....	9
List of Figures.....	11
Background	12

Task 1: Comparing the (Dis)Similarities Across Currently Published Terrorist Risk Assessment Instruments

Introduction.....	18
Thematic Analyses	18
Rationale	18
Evaluation Criteria	19
Protocol.....	36
Review of applications and evaluations	36
Results.....	42
Instrument 1 – The Violent Extremism Risk Assessment – Version 2 Revised (VERA-2R).....	42
Instrument 2 – Radar.....	61
Implications	81

Task 2: Synthesising the Existing Evidence Base

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Introduction.....	85
Rationale.....	85
Protocol.....	87
Secondary Systematic Review.....	87
Identification Stage.....	87
Screening Stage	91
Eligibility Stage.....	91
Analysis.....	92
Results.....	94
Study Characteristics	94
Thematic frequencies	95
Thematic Evaluation	96
Comparative Evaluation.....	101
Implications.....	105

Task 3: Developing the Case Test Library

Introduction.....	108
Case Vignettes.....	108
Rationale	108
Case Categories	108
Protocol.....	111
Case Identification.....	111

Task 4 - Testing the Reliability, Validity, and Equity of Instruments

Introduction.....	116
Instrument Testing.....	117
Rationale	117

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Protocol	124
Participant Recruitment	124
Participants	126
Vignette Assessment.....	127
Experimental Protocol	127
Data Collation	129
Results.....	131
Participant Bias.....	131
Instrument 1 – The Violent Extremism Risk Assessment – Version 2 Revised (VERA-2R).....	132
Instrument 2 – Radar.....	141
Implications.....	152
Conclusions	156
Bibliography	161

List of Tables

Table 1 Theoretical approaches underpinning psychological mechanisms of radicalisation	28
Table 2 Empirical weighting of theoretical psychological mechanisms of radicalisation.....	29
Table 3 Search Terms Employed during Review	37
Table 4 Categories within Thematic Evaluation	39
Table 5 Thematic Analysis of factors within the VERA-2R.....	47
Table 6 Descriptive analysis of included factors within VERA-2R.....	51
Table 7 Empirical Value of Cited Supporting Literature in the VERA-2R.....	51
Table 8 Thematic analysis of factors within the Radar Initial Risk Screening component.....	69
Table 9 Thematic analysis of factors within the In-Depth Risk Analysis component.....	70
Table 10 Descriptive analysis of included citations in the Risk Analysis protocol in Radar.....	73
Table 11 Empirical value of cited supporting literature in the Risk Analysis protocol in Radar.....	74
Table 12 Search terms used during SR.....	89
Table 13 Characteristics of studies analysed	95
Table 14 Top Variables Identified as Significantly Associated to Radicalisation or Terrorism	97
Table 15 Descriptive comparison of the SR evidence base with the VERA-2R.....	102
Table 16 Descriptive comparison of the SR evidence base within Radar's Risk Analysis protocol.....	103
Table 17 Reliability continuum of data sources	111
Table 18 Vignettes gathered by the research team	114
Table 19 Protocol to be used when undertaking reliability analyses	121
Table 20 Variations in inter-rater reliability studies	122
Table 21 Breakdown of Participant Experience and Expertise across Groups.	126

Table 22 Average Scores of Bias Vignettes across Participant Groups.....	131
Table 23 Factors deemed most useful during assessment process using the VERA-2R.....	132
Table 24 Factors deemed most useful during assessment process using Radar	142

List of Figures

Figure 1 Full SR process.....	90
Figure 2 Word cloud of 40 most frequently cited words within reviewed studies.....	96
Figure 3 VERA-2R risk scoring across case vignette categories.....	134
Figure 4 Outputs from Area Under the Curve Analysis for the VERA-2R.....	137
Figure 5 Confidence ratings across participant groups for the VERA-2R	140
Figure 6 Radar risk scoring across case vignette categories.....	143
Figure 7 Outputs from Area Under the Curve Analysis for Radar.....	147
Figure 8 Confidence ratings across participant groups for Radar Risk Screening.....	150
Figure 9 Confidence ratings across participant groups for Radar Risk Analysis.....	151

Background

Since September 11, 2001, over 50 pieces of anti-terror legislation have been introduced into the Australian Parliament (Williams, 2013). The majority of this legislation focuses on coercive measures to counter the terrorist threat including involvement in military action, expanding policing powers, and intelligence gathering activities. It is only in recent years in Australia, that attention has turned to non-coercive measures such as efforts to identify, prevent and manage individuals at risk of radicalisation and violent extremism (Harris-Hogan, Barrelle & Zammit, 2016). The Australian Government's Living Safe Together website states that since 2013, the Australian Government has invested over \$50 million through the countering violent extremism program with the aims of building the resilience of communities to violent extremism; reducing the spread of terrorist propaganda online; and diverting and deradicalising at-risk individuals.⁵

Applying a public health classification system to countering violent extremism (CVE), Harris-Hogan, Barrelle and Zammit (2016) argue that CVE work in Australia should be understood as a policy spectrum: At one end, primary CVE initiatives focus on preventing radicalisation in the first instance through; secondary level CVE initiatives focus efforts on diverting those displaying 'symptoms' of radicalisation and deemed to be 'at risk' of becoming 'significant members of extremist groups'; and at the other end of the spectrum are tertiary level interventions which deal with individuals already considered extremists.

Depending on the specification of risk, different risk assessment instruments should be used at different intervention points along the pathway to violent extremism. However, a challenge for CVE practitioners is having the right types of instruments to accurately assess risk and therefore make decisions for the type of CVE intervention required. Since 2015, the Australian Government has implemented at least two formal risk assessment instruments as part of its CVE efforts, including the VERA-2R and Radar, which are the subjects of analyses in this research project.

⁵ <https://www.livingsafetogether.gov.au/get-the-facts/what-is-the-australian-government-doing>

The concepts of risk and of risk assessment have informed research and practice in a variety of disciplines and for a range of applications (see Borum 2015 for an overview). Since the mid 1990s, there has been a marked increase in the development of violence risk assessment instruments (Douglas & Skeem, 2005; Monahan & Steadman, 1994; Otto, 1992). According to a systematic review conducted by Singh, Serper, Reinharth & Fazel (2011), over 150 risk assessment tools have specifically been developed to assess the risk of violence.

If implemented correctly, the risk principle ensures that correctional programming is delivered in such a way that the highest risk offenders receive the most intensive interventions while keeping lower risk offenders from receiving too much intervention and avoiding the disruption of circumstances that were making them low risk in the first place (Andrews, Bonta & Hoge, 1990; Lowenkamp, Latessa & Holsinger, 2006). Thus, the proper allocation of scarce resources is contingent on the accurate identification of those who pose the highest risk, and of those who would benefit from various management strategies (Kropp & Hart, 2000).

Yet current efforts to develop scientifically rigorous and operationally relevant research related to the risk assessment and management of the terrorist remains worryingly underdeveloped. Monahan (2012) outlines several conceptual challenges that need to be overcome if the risk assessment of terrorism is to make the same progress that in recent years has distinguished the risk assessment of other forms of violence.

The first conceptual challenge is that the substantive content for a risk assessment and management instrument required for a terrorist cannot mirror the risk assessment of other violent offenders, and decisions need to be made about the optimal degree of structuring within the risk assessment. Second, (and relatedly), robust individual risk factors need to be identified with regards to terrorism. Third, any developed instrument needs to be empirically validated (Monahan, 2012).

Finally, Monahan stresses that researchers need to be clear about what is being assessed. To date, academic approaches to understanding who becomes a 'terrorist' have largely tended toward generalist explanations. Such explanations, be they psychopathological or psychoanalytical, theoretical or descriptive analyses of large N datasets, tend to treat each individual group member equally. That is, they fail to effectively distinguish different member types not just across terrorist groups, but also within them. Analyses of 'the terrorist' frequently treat actors as monolithic in nature, differing merely in presumed personality traits and little else. Terrorist organisations, however, tend to possess some form of command and functional structure, be it hierarchical or linear. Within this

structure, a wide variety of roles, responsibilities and behaviours are delegated to individual members and sub-units. Such responsibilities may range from storing weapons to engaging in shooting attacks; from procuring vehicles for car-bombing attacks to being a bomb-maker; from being a suicide bomber to being a recruiter of suicide bombers; from being a foot soldier to being an executive leader. Because of this differentiation of focus and task, there are some important differences in both the nature and level of involvement by different members of terrorist groups. An individual may hold one or several distinct roles over time in their 'terrorist career'. These roles may be distinctive in the nature of the social, psychological and organisational demands required of the person in that role. The differences range from the level of violence the individual either directly engages in or merely facilitates others for; in expertise levels; individual risk to personal liberty and harm; and responsibility for overall strategy (Taylor & Horgan, 2006). Moreover, the security environment has evolved, and within Western countries the majority of terrorist attacks are now perpetrated by lone individuals who are inspired by, but are not part of, a larger terrorist group (Europol, 2017).

As Monahan (2012) notes, 'lumping' all forms of terrorism into one outcome variable diverges considerably from criminological studies where the norm is to typically 'split' the outcome variable. Rather than assessing general risk, studies that split the outcome variable tend to assess the risk of specific violent types. Examples include domestic violence, sexual violence, workplace violence, and so on. With the exception of the suicide bomber, rarely does the study of individual terrorist roles figure in the existing literature. Existing approaches therefore may miss the subtle psychological, behavioural, socio-demographic or organisational factors that may explain how and why some individuals are more likely to take part in particular terrorist roles than others. Plentiful evidence suggests that disaggregating our understanding of the terrorist is of value. Studies comparing male and female terrorists, lone and group actors, foreign and domestic fighters, and mentally ill and non-mentally ill lone actors (Jacques & Taylor, 2008; Gill, Horgan & Deckert, 2014; Gruenewald, Chermak & Freilich, 2013; Hegghammer, 2013; Corner & Gill, 2015) have identified distinct behavioural differences. Such endeavours remain predictive however and lack the specificity needed for individual risk assessment.

Gill (2015) added several additional considerations. First, the recent development of databases and larger datasets than have ever been available in terrorism studies, our knowledge of the characteristics of individuals involved in violent extremism, and of the factors associated with their involvement in extremist organisations, has evidently improved (LaFree, 2013). However, this knowledge is largely statistical in nature. Lists of variables (for example;

demographic, behavioural, historical, social, to a lesser extent, biological) that co-vary significantly with radicalisation, recruitment, involvement in, and to a lesser extent, disengagement from violent extremism has accumulated. As datasets continue to expand – involving an ever-larger number of variables across an ever-larger number of cases – the list of significant statistical associations is likely to grow. It is now time to take stock of the evidence rather than endlessly developing new (and largely untested) risk factors.

Second, base rates remain a continual problem. Quite simply, we have no grasp on what the societal prevalence of the vast majority of radicalisation indicators are. In some cases, like issues surrounding mental disorders, it is easier because of epidemiological studies (Corner, Gill & Mason, 2016). Other behaviours, such as leakage, are far more difficult to quantify. Without a sense of base rates, we are not able to measure with any certainty how reliable any one factor is, either in isolation or in combination with other factors. Instead, we can only sample on the dependent variable, which is not good practice.

Third, there is a distinct lack of research concerning protective factors. The literature just does not account for them. Instruments generally only look for ‘risk factors’, which may lead to a series of confirmation biases amongst intelligence analysts. Within those instruments that do include protective factors, they are sparse, rarely supported by any evidence, and often do not have a role in risk decisions. Protective factors may come in many forms and include individual factors (for example; attitudes, academic achievement, social orientation, self-control, personality factors), peer factors (e.g. close relationships with non-criminal peers, pro-social norms within peer group, number of affective relationships), and family factors (e.g. highly connected to family, involvement in social activities).

The final problem identified by Gill (2015) is the weighting of factors. In most studies, indicators of radicalisation are all treated equally. Very few risk assessment instruments discriminate between factors to any extent. One of the particular innovations within the TRAP-18 risk assessment instrument is that it splits factors into two categories – distal and proximal – and outlines how their relative presence should inform different responses with the presence of distal characteristics bringing attention to a subject for monitoring and the presence of proximal characteristics (warning behaviours) requiring active risk management of the subject (Meloy et al, 2015). Böckler, Hoffman and Zick (2015) go a step further, and state that the presence of one proximal factor necessitates instant disruption. This is an interesting, logical and yet rarely made argument. Of course, in reality not all factors are equal. A part of the problem is that the study of factors is almost trying to do too much – from highlighting factors of someone

adopting an extremist ideology to highlighting factors of someone planning an attack. These are very different processes, underpinned by very different behaviours and necessitating intervention by very different parts of the policing/intelligence/partner agency framework.

This research project therefore attempts to address these problems, and undertakes a holistic and impartial analysis of two risk assessment instruments currently used in the radicalisation and violent extremism domain in Australia - the VERA-2R and Radar. In particular, we demonstrate the lines of demarcation between these instruments, and outline their structures and goals, validation history and empirical underpinnings. We also provide a comprehensive overview of the state of the empirical knowledge of the causes of radicalisation and violent extremism, which these instruments purport to use to formulate their protocols. We also perform one of the world's first experimental testing of terrorist risk assessment instruments. In doing so, this research project demonstrates the extent to which the risk assessment instruments under scrutiny accurately classify offenders and which one's overestimate or underestimate risk.

This research is stratified across four research tasks that build upon one another. Task 1 compares the (dis)similarities across currently published terrorist risk assessment tools through qualitative thematic analyses. Task 2 synthesises the existing evidence base through a large-scale systematic review of the empirical literature on risk factors associated with terrorism. Task 3 builds a case test library of 60 anonymised test cases, exemplifying true positive (actual violent extremists), and true negatives. Finally Task 4 assesses these test cases to formally test the reliability, validity and equity of the tools under scrutiny.

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Task 1: Comparing the (Dis)Similarities Across Currently Published Terrorist Risk Assessment Instruments

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Introduction

This task charts the commonalities and differences in the risk assessment instruments and protocols under scrutiny. This task has been completed over 18 months of the project timeline. This involved the development of a set of standards through which to assess the quality, validity, reliability, and equity of the instruments. These standards were influenced by the wider risk assessment literature, particularly systematic reviews of other general risk assessment instruments (for example, Scarcella, Page, and Furtado (2016)). The task involved two interconnected qualitative thematic analyses that chart; (a) the content, (b) the method underlying instrument development, (c) the commonalities, (d) the points of departure, (e) the degree to which instruments are replicative of one another, (f) the level of 'structure' within the assessment components, (g) the ontological states of the risk factors, and (h) the intended use of the instruments. The purpose of this task is not to say which is the ideal instrument, but rather to provide a granular level analysis of the development of the instruments, how they are supposed to work in practice, and the quality of risk factors they include.

Thematic Analyses

Rationale

Instruments purporting to identify potential terrorists and extremists are increasing in volume and scope. Three factors have driven these developments. First, the quality of empirical research in this area has increased exponentially in the last 10 years (Schuurman, 2018). This increase in knowledge has led to improvements in capabilities for the identification of factors for inclusion in such instruments. Second, within this timeframe, there has also been an increase in the volume of individuals attracting assessment by security agencies due to the rise of the Islamic State and the extreme right (Koehler, 2019; Winter & Spaaij, 2018), which has led to an increase in need for valid assessment protocols for those whose remit is to provide policing and security. Third, there has been expansion in the scope of the type of risks being assessed, expanding from risk of terrorist violence, to risk of radicalisation (pre-violence), to 'vulnerabilities' to radicalisation (pre-pre-violence). This has widened the number of agencies tasked with the assessment and management of individuals under scrutiny, from policing and intelligence into other areas tasked with dealing with vulnerable populations (e.g. social work, healthcare, and education).

This increase in capability, need, and scope has led to the rapid development of a wide variety of terrorist and extremist risk assessment instruments,⁶ each with differing rationale, risk specifications, components, and factors. This development warrants an impartial comparative and factual analysis. Such analysis is necessary due to the potential serious costs associated with errors in risk assessment and subsequent management (Guy, 2008). Individuals who are incorrectly assessed as high risk may face imprisonment or involuntary hospitalisation, mandated treatment, forced medication, and a loss of multiple other civil liberties (Guy, 2008; Monahan et al., 2001). Conversely, individuals who are incorrectly assessed as low risk may go on to violate the safety and rights of others.

Given the above implications, this research seeks to undertake a holistic and impartial analysis of two risk assessment instruments currently used by those countering radicalisation and violent extremism in Australia (the VERA-2R and Radar). In this particular task we audit both instruments, and seek to demonstrate the lines of demarcation between these instruments, and outline their structures and goals, validation history and theoretical and empirical underpinning. The results presented below provide information concerning the background of each instrument, their genesis and characteristics, and all of the available examinations and evaluations of the instruments. This task also seeks to undertake comprehensive analyses of the empirical value of these instruments. This section begins by outlining the empirical underpinnings of the evaluation criteria, including the expected and required standards, the necessary considerations that should be undertaken during instrument adoption, and the areas which require immediate research focus to assist future instrument development.

Evaluation Criteria

Each of the instruments under scrutiny across this research⁷ were assessed across ten criteria: The approaches of the instruments, their stated risk specifications, the thematic classification of the included risk factors, the empirical quality of these factors, the weighting of the factors, the theoretical mechanisms underpinning the factors, the stated effect, validity, and reliability of the instruments, the moderators that will affect application of the

⁶ Mirroring the rapid development in the violent offending domain. Guy (2008) notes that over five decades, 457 tools were developed in this space.

⁷ Including those assessed in the project being undertaken by University College London, University of New Brunswick, and University of Waterloo.

instruments, and the implementation burdens that are specified as necessary for use. These criteria were chosen to ensure systematic and comprehensive evaluation of the theoretical, empirical, and practical underpinnings of each instrument. The following subsections outline the rationale for the inclusion of each specific criterion.

Approach

In the context of human behaviour, risk has traditionally been perceived as a static dispositional trait, often viewed as either present or absent (Borum, 2000). However, given the vast empirical advancements in the understanding of deviance and violence, risk is now understood to be contextual, dynamic, and continuous (Borum, 2000). This has affected the measurement and assessment of risk. The initial assessments of risk were based solely on the clinical judgement of a qualified practitioner. Research has long called into question the validity of such methods. In 1981, Monahan demonstrated that clinical judgement alone is no better than chance at predicting risk of violence. However, later studies concluded that clinical judgements of risk may have some predictive validity (Monahan & Steadman, 1994; Mossman, 1994). Conversely, more modern actuarial assessment measures, based solely on empirical evidence and statistical measures, have been found to have high predictive accuracy (Douglas, Cox, & Webster., 1999; Monahan et al., 2001; Quinsey, Harris, Rice, & Cormier, 1998), but are subject to their own weaknesses,⁸ and leave little room for human discretion. To balance these approaches, research moved towards the creation and implementation of instruments that incorporate the high statistical accuracy of actuarial instruments and the experience of practitioners inherent in clinical judgements, known as structured professional judgement (Skeem & Monahan, 2011). Each of the approaches considered for review in this project will now be described below.

(Unstructured) Clinical Judgement

In clinical judgement assessments, practitioners identify characteristics and risk factors based on clinical (non-empirical) experience, the practitioner's own knowledge of empirical research findings, and intuition (Hilton et al., 2006; Grove et al., 2000). The clinical judgement approach is commonly known as a "non-approach" or

⁸ See Borum (2000), Hart, Douglas and Guy (2016), Hilton, Harris, and Rice (2006), Melton, Petrila, Poythress, and Slobogin (1997), and Quinsey et al. (1998) for an in-depth discussion of the clinical versus actuarial debate. However, to summarise, this debate is based on the questioning of whether risk should be solely determined by clinical judgement or statistical measures.

anamnestic, as it is absent of any systematised structure or procedure (Hart, et al., 2016). Practitioners conducting assessments have complete discretion over the information they take forward for consideration in the final risk judgement during the assessment process, and there are no constraints on the information that can be used (Doyle & Dolan, 2002). This flexibility allows for an in-depth focus on case-specific influences. However, this approach has been widely criticised across the literature. Grove and Meehl (1996) describe clinical judgement as an 'unstructured', informal, subjective, and impressionistic approach to understanding and assessing risk, due to the lower levels of predictive accuracy as compared to assessment based on empirical evidence. Kemshall and Pritchard (1996) argue that such an approach is beset with bias due to the assessment process (interview, observation, and self-report). Hart, Michie, and Cooke (2007) note that alongside limitations in accuracy, clinical judgements are also affected by inter-clinician agreement (due to lack of specificity in the reporting of decision making).

Actuarial Measures

Actuarial instruments differ from clinical judgement assessments as final risk judgements are reached based on empirically valid data which is assessed according to fixed and explicit statistical rules. Such instruments use mathematical algorithms, and have strict instructions on how to formulate risk decisions. The data used to formulate the factors included for assessment have been empirically verified to be associated with the type of risk under scrutiny (Doyle & Dolan, 2002). Actuarial instruments conceptualise risk only in terms of the probability of the act occurring, and they neglect to include nature, severity, imminence, duration, and frequency of the possible act (Hart, et al., 2007). Many studies have demonstrated that instruments employing actuarial measures perform just as well as, or in many instances better than, clinical judgements (Egisdóttir et al., 2006; Borum et al., 1993; Dawes et al., 1989). However, much like clinical judgement, actuarial instruments are subject to critique on several levels, including, but not limited to; difficulty in generalisability beyond samples used in instrument development, their rigidity and propensity for excluding important risk factors, and errors in predicting an individual's personal risk of future violence (as compared to the prediction of groups) (Douglas et al., 2007; Dvoksin et al., 2001; Hart, 1998; Litwack, 2001).

Structured Professional Judgement

To counter the concerns present in unstructured clinical judgement and actuarial instruments, elements from both formats were combined, resulting in the formation of the structured professional judgement (SPJ) instrument (Douglas & Kropp, 2002). SPJ instruments are defined by the use of guidelines “that reflect current theoretical, clinical, and empirical knowledge about violence” (Douglas & Kropp, 2002, p. 626). SPJ instruments afford practitioners the opportunity to review all available information on the subject under scrutiny to identify potential risk factors as guided by a structured manual which is founded on empirical evidence. Hart and Logan (2011) note that in SPJ instruments, the practitioner gathers case information guided by recommendations within the instrument manual, identifies risk factors as specified in the manual, evaluates the relevance of the identified factors,⁹ undertakes scenario planning to determine likely outcomes given all gathered information and the relevance of this information, and conducts risk mitigation planning given the hypothetical scenarios. The practitioner then uses all of this information to formulate their own judgement of risk.¹⁰ SPJ instruments are not designed with fixed guidance on final risk decisions, but are designed to guide the decision-making process of the practitioner. The development of these instruments is supported by input from various stakeholders, such as researchers, clinicians, and administrators, and they include a list of risk factors, with factor inclusion also being dependent on the systematic review of empirical evidence. Unlike actuarial instruments, SPJ instruments afford the user the discretion to make the final decision of risk (as opposed to a decision made on fixed rules and statistical calculations) (Pedersen et al., 2010). There is agreement in the literature that SPJ instruments are both valid and reliable when assessing risk of violence (Pedersen et al., 2010). However, validity has not yet been assessed in the radicalisation and violent extremism domain.

Each assessment approach (clinical, actuarial, SPJ) has benefits and pitfalls. The choice between which approach type to apply should be determined based on which is the most appropriate in that instance. Research has shown that clinicians are often hesitant to use numerical probabilities to estimate risk (Heilbrun, Philipson, Berman, & Warren, 1999). Clinicians argue that numerical probabilities are too precise, and they instead prefer categorical

⁹ Relevant factors are those that the practitioner determines are functionally (causally) related to the potential risk.

¹⁰ Hart, Sturmey, Logan, and McMurran (2011, p. 118) define formulation as “the preparation of an evidence-based explanation of a person’s difficulties – their form, their origins, and their development and maintenance over time.”

estimates often found in SPJ instruments, with research showing that when clinicians are requested to use numerical rating scales, they give lower scores for risk of violence compared to other forms of response scales (Slovic & Monahan, 1995; Slovic, Monahan, & MacGregor, 2000).

In this project, each instrument is scrutinised to determine the professed (and actual) approach to risk assessment.

Risk Specification

Much like the aforementioned changes in assessment approaches, how we specify what risk is being measured has developed considerably. There is an agreed consensus in the theoretical and empirical literature that no single factor can explain the complexity of becoming radicalised and/or becoming involved in terrorist groups or violent behaviour. Research is consistent in its persistence that the movement towards terrorist violence should be viewed as a continuum, with discrete phases throughout (Borum, Fein, & Vossekuil, 2012; Corner, Bouhana, & Gill 2019; Gill & Corner, 2017). Translating this to risk assessment, the discrete phases each represent a different form of risk. Specificity matters. Individuals within each different discrete phase will present with different experiences, behaviours, and outcomes, and therefore, will produce different scores depending on the risk specification of the instruments used.¹¹

There are many theoretical models of the movements toward terrorist behaviour each citing multiple phases. These include, but are not limited to: Borum's (2003) terrorist mind-set model (including the phases of grievance, injustice, target attribution, and devaluation); Moghaddam's (2005) 'staircase to terrorism' model (psychological interpretations of material conditions, perceived options to fight unfair treatment, displacement of aggression, moral engagement, solidification of categorial thinking and the perceived legitimacy of the terrorist organisation, and terrorist act and sidestepping inhibitory mechanisms); Precht's (2007) conversion to terrorism model (pre-radicalisation, conversion and identification, conviction and indoctrination, and action); Silber and Bhatt's (2007)

¹¹ For example, assessing a radicalised individual using an instrument with a risk specification of "violent extremism" may produce a conclusion of low risk, whereas this same individual being assessed with an instrument with the risk specification of "radicalisation", would lead a user to conclude that the individual is of high risk. The implications for interventions for both assessments are different: In the first instance, an outcome of low risk may lead to a lack of intervention, which could lead to an escalation towards eventual violence. Whereas, in the second assessment, the individual may be referred towards the appropriate intervention, which would ensure that the individual of interest would not escalate in their behaviour.

radicalisation model (pre-radicalisation, self-identification, indoctrination, and jihadisation); and Wiktorowicz's (2004) radicalisation model (cognitive opening, religious seeking, frame alignment, and socialisation).

Other authors look to model distinct behavioural phases prior to, and resulting in radicalisation; how individuals form an attitudinal affinity with a terrorist ideology, cause, or movement. These scales encompass a range of non-violent behaviours and ideologies, and have been developed by Amjad and Wood (2009; acceptability of aggression), Bélanger et al. (2014; self-sacrifice), Bhui et al (2016; sympathies for violent protest and terrorism), Gerber et al. (2016; support for violence in the context of intergroup conflict), Liht et al. (2011; fundamentalism), Moskalenko and McCauley (2009; activism and radicalism intention), Nivette, Eisner, and Ribeaud (2017; violent extremist attitudes), Putra and Sukabdi (2014; Islamic fundamentalism), Saucier et al. and Stankov et al. (2009; 2010; militant extremist mindset), and Kruglanski et al. (2009; quest for personal significance).

For this research, risk assessment instruments that have been included for assessment¹² specify one of the following risks:

- Risk of vulnerability to radicalisation.
- Risk of radicalisation.
- Risk of terrorist violence.
- Risk of reoffending.

This research does not focus on instruments that have been developed based on research examining attitudinal affinity scales and phases.

Themes

Guy (2008) argues that risk assessment instruments must have a holistic and comprehensive coverage of factors for the phenomenon under scrutiny. Instruments that do not contain all empirically identified factors can result in inappropriately low or high estimates of risk. Guy also notes that these factors should be clearly operationalised

¹² Including those assessed by the London and Canadian teams.

and semantically organised. This categorisation is critical to promoting inter-rater reliability, and to inform users of the empirically verified associations between the included factors and the type of risk being measured.

Within this research, we assess the thematic categorisation of factors for two purposes; which factors are aggregated into which themes, and why the authors state this is the case. To ensure a consistent comparison of all included instruments across this research and the work being undertaken by the United Kingdom and Canada (to facilitate the future testing of face validity), each factor within each instrument under scrutiny is coded into one of the five themes listed below:

- Static Risk Factors.
 - Non-changeable historical factors of the individual.
- Dynamic Risk Factors.¹³
 - Situational and other factors that are amenable to change and occurred recently to problem behaviour.
- Beliefs, Attitudes, and Behaviour.
 - Facets of the individual typically associated with radicalism.
- Social Context and Capability.
 - Information regarding the individual's social network and capability for violence.
- Protective Factors.
 - Conditions or attributes that actively help mitigate or eliminate the risk posed by the individual.
They play an active role in modifying the risk.

¹³ Guy (2008) notes that instruments should be designed to facilitate risk reduction. To achieve this, the instrument must contain dynamic risk factors that can be changed through intervention. These factors provide information about individuals that is not accurately captured by static, historical factors (Hanson, 2005).

For instruments with a stated intention of examining behaviours close to violence, it is expected that they would contain more factors present within the Social Context and Capability theme. Conversely, for instruments that purport to measure the risk of radicalisation, the majority of factors should be within the Beliefs, Attitudes, and Capability theme. Finally, for instruments aiming to assess risk of vulnerability to radicalisation, we expect there to be more factors present in the Static and Dynamic themes.

Factors¹⁴

Quality

Risk assessment instruments are typically developed to measure a standard list of factors which are included based on empirical evidence (for example; abuse, age, history of behaviour, substance use), with each factor being interrogated during the assessment process. The use of empirical evidence to determine factor inclusion is critical across all approaches to assessment (Guy, 2008). Meta-analyses of risk factors across multiple offence types have exponentially strengthened understanding of risk, and have clarified the knowledge of major, moderate, and minor risk factors (Andrews, Bonta, & Wormith, 2006). It is therefore crucial to assess the empirical quality of the factors including in the instruments being reviewed.

Determining the quality of included factors can be achieved using two methods; interrogating the empirical evidence base underlying each factor within the instruments, and interrogating the wider empirical evidence base. Therefore, this research performs both a critical assessment of the stated literature underlying factors within the instruments, and a comprehensive systematic review of all existing literature on known causal factors of radicalisation and terrorism.¹⁵

¹⁴ Risk factors and indicators are two separate concepts, but are often treated interchangeably in research literature. For example, whilst smoking may be a risk factor for lung cancer, it is the cancerous lump which is an indicator of cancer. Often those items thematically coded as related to beliefs, attitudes and behaviours are indicators of radicalisation, in other words, by-products of radicalisation. The same is also true for many of the variables coded under social context and capability.

¹⁵ The systematic review was undertaken in task 2.

Weighting

Within risk assessment, the accurate determination of risk is highly dependent on the weighting of the individual importance of factors, and final risk judgements are determined based on the combinations of these weightings. Clinical judgement procedures typically use factors as guidance during assessment, with weighting dependent on the experience during assessment (Monahan & Skeem, 2014). Actuarial measures contain highly structured individualised guidance for identification measurement, and weighting of factors. They have a defined combination of factors, and a final score for the level of risk (Monahan & Skeem, 2014). Whereas SPJ instruments identify the presence of factors, and measure presence using a standardised scoring system of 0, 1, 2 (e.g. the VERA-2R uses the following scoring system; 0 for low, 1 for moderate, and 2 for high), which is applied across factors.¹⁶ Much like clinical procedures, the weighting of risk factors is often instructed to be conducted at the discretion of the user, as it is recognised that not all factors have equal weight for all individuals (Douglas, 2019).

Concerningly, within instruments related to radicalisation and violent extremism risk, factors are often weighted equally across the aforementioned risks themes. For example, the EU funded Safire Project outlined 21 factors, ranging from “lingering concerns with questions of meaning and identity” to “dependence on communication technology” to “associating with extremist groups” and “training travel” with no difference in measurement and weighting (cordis.europa.eu, 2020). Other risk assessment instruments discriminate between factors to a small extent. One of the particular innovations within the TRAP-18 is that it splits the included factors into two categories – distal and proximal – and outlines how their relative presence should inform different responses, with distal factors requiring monitoring, and proximal requiring management (Meloy et al., 2015). Böckler, Hoffman and Zick (2015) go further, stating that the presence of just one proximal factor necessitates instant disruption.

In this research we identify the cited protocol used for the weighting of factors.

Mechanisms

The development of risk assessment instruments also requires a theoretical justification for the inclusion of factors. Authors argue that the rationale behind inclusion of factors often is vague, or rests on untested assumptions

¹⁶ This scoring system is inverted for the Protective/Risk Mitigating factors.

regarding behaviour (Horgan, 2016; Gøtzsche-Astrup, 2018). However, the identification of factors or criminogenic needs that may predict risk requires a sound theoretical basis. Gøtzsche-Astrup (2018) conducted the first known review of the theoretical evidence base that underpins purported causal mechanisms of radicalisation. Table 1 presents the seven theoretical mechanisms that were investigated, and Table 2 presents the weighting of the evidential strength of the central elements within these theoretical mechanisms. These elements are categorised according to the amount of empirical evidence across the research literature. Within this research, each of the instruments under scrutiny will be assessed according to elements identified in Table 2.

Table 1 Theoretical approaches underpinning psychological mechanisms of radicalisation

Theory	Seminal Research	Central Tenets	Dependent Variable	Explanatory Variable(s)
Uncertainty-Identity Theory	Hogg & Adelman (2013)	Motivational and social identity theory	Joining and supporting radical groups Autocratic leadership Behavioural aggression and hostility	Self-uncertainty Group entitativity Social identity complexity Peripheral membership
Significance Quest/"3N"	Webber & Kruglanski (2018)	Social psychological motivational theory	Coming to see as socially normative violent behaviour that is deviant from the majority perspective	Need for significance Narratives legitimising violence Networks and groups
Devoted Actor Model	Atran (2016)	Ideology, value	Unconditional commitment, sacrifice and willingness to engage in extreme behaviour for a group	Sacred Values Identity fusion
Mindset and Worldview	Borum (2014)	Distinction between mindset factors and worldview factors	Psychological "climate" that increases the risk of involvement in violent extremism	Psychological mindset authoritarianism
Reactive Approach Motivation	McGregor, Hayes, & Prentice (2015)	Motivational framework encompassing personality and social dynamics	Aggressive religious radicalisation	Motivational process diverted through interaction between personality, threats and group affordances

Two-pyramid Approach	McCauley & Moskalenko (2017)	Distinguishing opinion and action	Radical opinion (support) and radical action (behavioural intentions)	Individual factors Group factors Mass factors
----------------------	------------------------------	-----------------------------------	---	---

Table 2 Empirical weighting of theoretical psychological mechanisms of radicalisation

Strong empirical evidence	<ul style="list-style-type: none"> - Normal psychological mechanisms rather than psychopathology - Motivational processes rather than instrumental calculations of risk and reward - Negative life experiences that put the individual in flux in terms of fundamental questions - Experience of fundamental uncertainty or loss of meaning or significance - Shift in social identity towards a single social group rather than many - Small group dynamics drive the process to behavioural extremes - Heightened dispositional anxiety, aggression and impulsivity - "Sacred values" as necessary in later stages of radicalisation
Moderate empirical evidence	<ul style="list-style-type: none"> - Psychological mindset of authoritarianism, dogmatism and fundamentalism - Negative emotions, particularly anger and contempt
More research needed	<ul style="list-style-type: none"> - Other individual difference factors such as ambiguity, intolerance and need for cognitive disclosure - The relative roles of uncertainty, significance, sacred values etc. - The causal role of ideology: justification or cause?

NB - Colouring system of Red, Amber, Green utilised to help the reader visualise where the mechanisms rest on a recognised scale of quality. This coding system is also used across this research to denote the quality of information.

Validity and Reliability

Due to the differing methodologies across the three approaches to risk assessment, there are differences in validity. High validity within an instrument demonstrates that the risk judgements (both individual factor and overall) represent the phenomenon they purport to. Clinical judgement is largely formulated through the individual experiences of the assessor, which results in low levels of overall predictive validity¹⁷ (Wong & Gordon, 2006).

¹⁷ The ability of the instrument to predict the outcome that we would expect it to predict (e.g. the ability of the VERA-OR to determine that an individual scoring as high risk will at some point engage in an act of violent extremism).

Monahan (2000) notes that this may be due, in part, to an underestimation of base rates of the phenomenon under scrutiny. Actuarial measures were developed to increase the consistency and therefore validity of the decision-making process. The structure and empirical evidence base in both actuarial and SPJ instruments promotes greater validity, with studies often demonstrating good predictive validity (Douglas, Yeomans, & Boer, 2005; Doyle, Dolan, & McGovern, 2002; Harris et al., 2003). However, as Singh, Desmarais, and Van Dorn (2013) note, although many studies have been conducted which purport to examine validity, there is a lack of consistency in its measurement and interpretation.

Effect

The overall effectiveness of risk assessment instruments can be evaluated by investigating their predictive validity. Predictive validity is defined as the degree to which outcomes (in the case of risk assessment instruments, scores) predict the performance of future behaviours (in this instance, radicalisation or extremist violence) (Shultz, Whitney, & Zickar, 2005). In order to assess predictive validity, it is critical to assess the instruments under scrutiny against performance indicators. Singh (2013) defines performance indicators as those that are used to measure levels of predictive validity, and therefore the ability of the instrument to correctly assess and predict risk. There are a range of performance indicators, and these measure global, high risk, and low risk accuracy. The use of performance indicators to measure the effectiveness of instruments provides comprehensive statistical descriptions of instrument performance. These descriptions are critical to understand whether the instruments are able to predict the risk that they purport to.

Appendix 1 is drawn from Singh (2013), and shows the equations and definitions of the performance indicators that are used to assess the predictive validity of risk assessment instruments. The earliest, and now most commonly reported factors are sensitivity and specificity. However, as Singh notes, in 1993, Hart, Walter, and Menzies questioned this approach (following movements from dichotomous to continuous constructs of risk). More recent work consistently seeks to determine predictive validity using the area under the curve (AUC).¹⁸ We seek to examine

¹⁸ Although, as authors note, solely testing predictive validity based on AUC fails to capture the calibration of instruments, and therefore is not a complete picture of the effect of the instrument. Therefore, it is recommended that authors employ a range of performance indicators to fully capture both calibration and discrimination (Singh, 2013).

if the literature associated with the instruments provide any information regarding predictive validity, and if any of the below performance indicators are used as evidence.

Validity

Due to the aforementioned problems that a lack of predictive validity can create, and the concerns over measurement processes, this research also seeks to examine if the instruments under scrutiny have been assessed for four other forms of validity, each of which affects the overall predictive validity of an instrument:

- Face validity concerns the stated aims of the instrument. Specifically, the degree to which the instrument appears effective in terms of its purported aims.
 - The weakest form of validity, as it results from a subjective assessment (Trochim, 2001). Face validity provides insight into how participants may interpret and respond to factors within instruments.
- Content validity is the extent to which the instrument represents all facets of the concept (e.g. radicalisation) it purports to assess.
 - It may be difficult to assess content validity when the construct of interest (risk specification of the instruments) is highly abstract and poorly defined.
- Criterion validity is the strength of the relationship between factors and the specified criterion (phenomenon being measured) of the instrument.¹⁹
 - Indicated when the correlation between individual factors supports the risk specification (as it relates to the definition of the phenomenon).
 - Criterion contamination (errors in measurement of the phenomenon) is a large threat to this form of validity (Schultz & Whitney, 2005).

¹⁹ Although this research seeks to test the other mentioned forms of validity, we will not be seeking to test criterion validity, as it is not appropriate for this research question. Criterion validity is most useful in determining the most suitable factors for the creation of new measurement procedures, or instruments.

- Construct validity is the degree to which the instrument measures the construct as defined by a theory (is the instrument measuring the theoretical phenomenon it purports to).

This research is also examining authorship effects. Recent works have expanded our understanding of validity, by also examining authorship effects. This research seeks to understand if authorship bias may exist in the validation of risk assessment measures (Singh et al., 2013). Bias may arise if authors research the predictive validity of their own measure. This can lead to the so-called authorship effect, with authors of the measure reporting higher predictive validity than independent researchers (Singh et al., 2013). Singh et al. (2013) identify two potential sources of the authorship effect: First, the bias may occur when administrators of instruments are trained with strict training protocol adherence (with actuarial instruments showing higher rates of authorship effects); and second, authors may be reluctant to publish results that show poor predictability of their instrument (Singh et al., 2013). Authorship effects have been identified in large scale analyses of risk assessment instruments. Blair, Marcus, and Boccaccini (2008) conducted a meta-analysis of the Violence Risk Appraisal Guide (VRAG), the Sex Offender Risk Appraisal Guide (SORAG), and the Static-99. They identified that validation studies conducted by the authors of these measures had higher predictive validity than independent studies. Singh et al. (2013) also conducted a meta-analysis in which 83 risk assessment validation studies published between 1966 and 2011 were assessed. They identified that authorship effect was highly evident in the studies that were published in peer-reviewed journals.

Reliability

Whereas tests of validity assess whether the instruments measure what they purport to (risk specification), tests of reliability assess the consistency of the results from the instruments across different situations. Standardisation of instruments, an empirical evidence base, structured information gathering, and education and training all improve the reliability of outcomes from risk instruments (Borum, 1996; Hilton, Harris, & Rice, 2006). In this research, we seek to examine several facets related to the consistency in ratings for replication and validation. We will report any investigations of the following reliability estimates:

- Internal Consistency (and item homogeneity) Reliability.
 - Internal consistency indicates how well the items in an instrument fit together conceptually.
- Inter-Rater Reliability.

- This is indicated when there is consistency across two or more assessors;
- Intra-Rater Reliability.
 - This is indicated when there is consistency of the one rater across different time intervals;
- Test-Retest Reliability.
 - This is measured by administering the same test to the same group of participants during different time periods. The correlation between the two outcome scores, or scores between individual indicators, indicates the test-retest reliability of the instruments.

Moderators

Risk assessment instruments are not developed for use across all conditions. The population under scrutiny can affect the observable outcomes (and thus the predictive validity of the instrument). Different moderators will affect the outcomes across different contexts. Moderators can include; The ideology under scrutiny (for example, a risk factor for one ideology may act as a protective factor for another (known as cross-cultural validity)), race, psychiatric status, and gender.²⁰ Questions on the suitability of assessing females using instruments built on evidence drawn from male samples are the most consistent. Multiple risk assessment instruments have been developed based on data from exclusively male samples, and research has shown that this affects the predictive validity of instruments when implemented in female samples (Holtfreter & Cupp, 2007; de Vogel & de Ruiter, 2005).²¹

In this research we seek to determine all stated moderators within the instruments under scrutiny.

Implementation

Alongside moderators within instruments, the conditions during implementation of assessments affect the integrity and fidelity of outcomes. It is critical for those tasked with instrument creation and implementation to understand

²⁰ Gender differences in criminal behaviour have been consistently shown within criminology (Farrington, 1986; Gottfredson & Hirschi, 1990; Moffitt, 1993; Odgers et al., 2005).

²¹ Although when examining instruments used in youth correctional settings, this effect is much less (Schwalbe, 2003).

the ideal conditions for instrument use, the difficulties that may arise during implementation, and the features that can facilitate or impede the assessment process.

A major implementation burden is training. Monahan (1993) produced a comprehensive list of guidelines which help facilitate effective risk assessment. Within this, his primary suggestion was that clinicians' understanding of risk and risk assessment research needs to be updated through continued education and training. Continued education and training has both a temporal and monetary impact on practitioners, and those responsible for assessment implementation.

Despite the burden, research has demonstrated the critical importance of training. McNeil, Fordwood, Weaver, Chamberlain, and Binder (2008a) evaluated training in suicide risk assessment. Using vignettes, McNeil et al. identified that participants who underwent training were better at identifying historical, clinical, and risk management variables in relation to risk of suicide, than those who did not undergo training. Training also improved communication skills (with regards to judgement reasoning) and self-confidence. McNeil et al. (2008b) replicated the methodology of their 2008a study, but examined violence risk assessment (the HCR-20 instrument). Replicating the results of their previous study, training in the HCR-20 improved identification of risk and protective factors, communication of reasoning, and self-confidence. Reynolds and Miles (2009) identified that systematic training in the HCR-20 can qualify non-clinicians to conduct risk assessment. This result has also been replicated across different instruments and practitioner types. Storey et al. (2011) evaluated the effect of training across multiple criminal justice professional groups.²² They found that practitioners showed improved knowledge, skills, attitude, and case management in violence risk assessment following systematic training. Given this evidence, this project seeks to evaluate the implementation burdens that are necessitated within each instrument.

In order to comprehensively thematically evaluate the instruments under scrutiny in this research, it was first necessary to source the origin documents for the instruments, and all accompanying literature that has evaluated or tested the instruments. The following section covers the methodology involved in the sourcing of this

²² Police officers, prosecutors, and civilian support staff.

documentation, the methodology for the thematic analysis of the factors within the VERA-2R and Radar, and the methodology underwriting the indicator quality assessment.

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Protocol

Review of applications and evaluations

Literature search

In order to supplement the instrument documentation supplied to the research team, we performed a comprehensive review of current literature to source all known publications which have either described, applied, or evaluated the VERA-2R²³ or Radar. The sourcing of literature employed a structured and systematic protocol. The search strategy was based on the Campbell Collaboration method, which is considered to be the standard-bearer for systematic reviews in the social sciences.²⁴

Search terms

To identify all studies relating to the description, application, or evaluation of the instruments, the research team employed a series of search terms. These terms are highlighted in Table 3.

Identifying studies: Databases and information sources

Studies for inclusion were identified using the following search methods:

- A keyword search of multiple electronic databases including grey literature and dissertation databases.²⁵
- Forward and backward citation searches of all eligible candidate studies.

Full text versions of identified studies were obtained through (in order of preference):

- Electronic copies via the university's e-journal service.

²³ Given the date of publication of the VERA-2R (Pressman et al., 2018), this search yielded no results. Therefore, evaluations and applications of the previous iterations of the tool (the VERA and VERA-2) were included for analyses.

²⁴ For more information, see www.campbellcollaboration.org

²⁵ PsychINFO, ProQuest Central Criminology Collection, ProQuest Central Social Science Database.

- Electronic copies of studies available elsewhere online.
- Paper copies.
- Electronic/paper copies requested through the University's inter-library loan system.
- Electronic/paper copies requested from the authors themselves.

In cases where the full text versions of the studies collated contained insufficient information to determine their eligibility for inclusion according to the coding strategy (described below), where possible the corresponding author of the study was contacted in an attempt to retrieve this information.

The protocol allowed for inclusion of published and unpublished (grey) studies.

Table 3 Search Terms Employed during Review

Instrument Name	Research Area
VERA	Terroris*
VERA 2	Radicali?ation
VERA-2	Risk
VERA 2R	Extremis*
VERA-2R	Violen*
Radar	
TRIM	
FIRE	

Data extraction and management

Primary level screening

Primary level screening involved the team examining the title and abstract of studies identified during the electronic and bibliographic searches. This initial search yielded 59215 (7679 regarding the VERA, VERA-2, or VERA-2R, 51536 regarding Radar, TRIM, or FIRE) study titles and abstracts which did not meet the inclusion criteria

(application or evaluation of the instruments of interest) and were subsequently excluded. A total of 22 studies were taken forward for secondary level screening.

Secondary level screening

Secondary level screening involved reading studies identified during primary screening and bibliographical cross checking to identify further studies for inclusion and exclusion. Following this secondary screening, nine studies were included for evaluation. This two-tier system ensured only articles which met all criteria were included.

Seven studies that were evaluated were regarding the VERA, VERA-2, or VERA-2R. It is unsurprising that the search yielded very few results for Radar, and that both studies are overviews of the process, as Radar is a classified risk assessment instrument and, as at the time of writing, has not been empirically validated.

Literature Assessment

All studies that were taken forward following screening were interrogated to discern information regarding the evaluation of the instruments. Information was drawn from the works if it was relevant to any items included in Table 4. This information was then used to supplement the thematic evaluation.

Thematic evaluation of Instruments

Primary Coding

The first stage of the thematic evaluation of the instruments was to discern if there were any further documentation that was crucial in the instrument development process, and the location of these documents.²⁶ The authors were supplied with two versions of the training manual,²⁷ and a range of supplementary material for the VERA-2R that is utilised during the training process, but no further documentation was identified. However, the development of Radar is unique, in that it's development was reliant on multiple theoretical models and thus all documentation relating to these models was also sourced.

²⁶ Also conducted in the screening protocol.

²⁷ Paper and pdf. Versions. There are differences in the literature included in these versions, which is highlighted in Appendix 3.

Once all documentation relating to the formation of instruments was sourced, these documents were then collated with the studies sourced during the literature search. Each document included was then thematically evaluated. This evaluation included the qualitative analysis of each document. This analysis was guided by the categories in Table 4. Coding included direct transcription from the source matched to the relevant category. Only information directly presented in the sourced documents was utilised.

Table 4 Categories within Thematic Evaluation

Category	Information to be sourced from Thematic Evaluation
Approach taken	e.g. structured professional judgement vs. actuarial.
Risk specification	e.g. risk of vulnerability to radicalisation, risk of radicalisation, risk of violence, risk of re-offending.
Themes within the instruments	Static risk factors, dynamic risk factors, beliefs attitudes and behaviour, social context and capability, protective factors.
Empirical grounding	The empirical research base supporting the tool.
Mechanisms for tool development	Theory of change underpinning the inclusion of individual factors or how factors produce their effects.
Validity	e.g. face, content, criterion, construct.
Effect	Performance indicators.
Reliability	e.g. inter-rater, intra-rater, test-retest.
Moderators	e.g. in what contexts and for whom, cultural translation.
Implementation burdens	e.g. readability and respondent burden.

Secondary Coding

To ensure reliability of the thematic evaluation, the documentation gathered on both instruments were then assessed by a secondary coder. This coding included the other member of the research team following the exact protocol undertaken as the primary coder.

Of particular concern with the thematic evaluation is the level of agreement between the two raters on the coding of the themes within the instruments (assignment of risk factors into themes). This is because this is the only category which relies on inductive reasoning, as opposed to deductive reasoning. To determine agreement between raters of the assignment of risk factors to the aforementioned themes the research team employed the The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

Cohen's Kappa agreement measure. The results indicated that for the factors within the VERA-2R, $K = 0.956$, $p < 0.001$. As per the categorisations set out by Landis and Koch (1977), there was almost perfect agreement between coders.

As Radar is the compilation of multiple assessment processes, the reliability of each component under analysis in this research was assessed separately. The results for the Risk Screening component were $K = 0.617$, $p > 0.001$. According to Landis and Koch (1977), there was substantial agreement between coders. The results for the In-Depth Risk Analysis component were $K=0.404$, $p < 0.001$. According to Landis and Koch (1977), this signifies moderate agreement between coders. The difference in coder reliability between the VERA-2R and Radar is likely due to the inherent complexity of the layout of Radar, the thematic differences between the explanations and case examples given for factors, and the thematic complexity of the factors themselves.²⁸

Further scrutiny of the differences in the thematic coding of the Radar components shows that the coders predominately disagreed on those factors classified as related to social relationships. This is likely due to the inclusion of ideological components of the social behaviours.²⁹

Indicator Quality Assessment

The first step in determining the theoretical and empirical underpinning of the factors within the instruments was to undertake systematic searches to source all texts cited within both the VERA-2R and Radar.³⁰ This included the sourcing of 181 documents cited in the VERA-2R documentation, and 142 cited in the Radar and FIRE documentation. The team were unable to source 10 texts cited in the VERA-2R, and 23 cited texts in Radar.³¹ Each sourced document was then read in its entirety during the systematic interrogation to determine:

²⁸ For example, the majority of the disagreement between coders was for those factors labelled as related to 'Social Relations'. However, many of these factors are actually describing changes in the beliefs and attitudes of individuals.

²⁹ This also has implications for users of Radar.

³⁰ The second stage is the systematic review within Task 2.

³¹ As mentioned, there are inconsistencies in the citations and literature review segments between the printed and digital versions of the VERA-2R provided to the team for assessment. The print version has four cited texts not in the digital version (Clarke, 2016 (NOTE this text is not included in the bibliography for the VERA-2R); Cornish, 2008; CT Morse, 2017 (NOTE this text is not included in the bibliography for the VERA-2R); ISD, 2011 (NOTE this text is not included in the bibliography for the VERA-2R)), the digital version has 31 extra citations (Aly, 2017; Atran, 2003; Atran, Sheikh & Gomez, 2014; Bandura, 1990; Basra et al., 2016; Bjørge, & Horgan, 2009; Böckler, Hoffman, & Meloy, 2017; Bora, 2015; Bronfenbrenner, 1979 (Note – no reference accompanies this citation); Conway, 2017; Cornish, Hughes, & Livingstone, 2017). The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

- Empirical value.
 - Here we evaluated the method of each document to determine the level of empirical work within. Documents were determined to be empirical if some form of empirical analysis was undertaken. Examples of non-empirical documents included newspaper reports, theoretical models, and literature reviews.
- Relationship to the cited statement in the instrument.
 - Here we evaluated the location of the cited statement from the source documents. This involved identifying any possible wording that caused the authors of the VERA-2R and Radar to cite these materials to support the statements within the instruments.
- Location of the cited statement within the sourced document.
 - The rationale here was to identify whether the cited statement had empirical value (beyond just being attributed to an empirical item).
 - This involved identifying which element of the source document the cited statement was attributed to. For example, if the statement was attributed to elements within the literature review, the findings, or the conclusion of documents.
 - If the cited statement was attributed to any area other than the findings, we then identified the source, and if any secondary citations should have been used by the authors of the instruments.³²

2009; Corte Ibanez, 2014 (NOTE this text is not included in the bibliography for the VERA-2R); Dalgaard-Nielsen, 2013; Francis, 2016; Hafez, 2007; Hofmann, 2016; Horgan, 2008; Horgan et al., 2017; Kohlberg, 1984; Kruglanski et al., 2007; Lankford, 2014a; Lankford, 2014b; Lennings et al., 2010; Ma, 2013; Ma, 2017; Merari, 2010; Piaget, 1926 (Note - no reference accompanies this citation); Reitano, Clarke, & Adal, 2017; Schwartz, Dunkel, & Waterman, 2009; von Behr et al., 2013; Weenink, 2015; Weimann, 2004). However, note that the reference lists within both documents have a combination of these extra citations across both print and digital versions.

³² For example, in indicator BA5 in the VERA-2R, the statement "Moral emotions are used to condemn others." (Pressman et al., 2018, p. 48), is attributed to Haidt (2003). However, upon scrutiny, Haidt actually attributed this statement to the work of Rozin, Lowery, Imada, and Haidt (1999). In this instance, all efforts from Pressman and colleagues should have been made to source the Rozin et al. document or provide a clear secondary citation.

Results

Instrument 1 – The Violent Extremism Risk Assessment – Version 2 Revised (VERA-2R)

The assessment of this instrument is based on the following reports and papers:

- Pressman, D. E., Duits, N., Rinne, T., & Flockton, J. S. (2018). *VERA-2R Violent Extremism Risk Assessment – Version 2 Revised: A structured professional judgement approach*. Utrecht, Netherlands: Netherlands Institute of Psychiatry and Psychology.
- Pressman, D. E. (2009). *Risk assessment decisions for violent political extremism*. Ottawa, Canada: Canadian Centre for Security and Intelligence Services.
- Pressman, D. E. (2016). The complex dynamic causality of violent extremism: Applications of the VERA-2 risk assessment methods to CVE initiatives. In A. J. Masys (Ed.) *Disaster Forensics: Understanding Root Cause and Complex Causality* (pp.249-269). Switzerland: Springer International Publishing.
- Pressman, D. E., & Flockton, J. (2012). Calibrating risk for violent political extremists and terrorists: The VERA 2 structured assessment. *The British Journal of Forensic Practice*, 14(4), 237-251.
- Beardsley, N. L., & Beech, A. R. (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.
- Scarcella, A., Page, R., & Furtado, V. (2016). Terrorism, radicalisation, extremism, authoritarianism and fundamentalism: A systematic review of the quality and psychometric properties of assessments. *PLoS ONE*, 11(12), e0166947.
- 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.

- Lloyd, M. (2019). *Extremism risk assessment: A Directory*. Lancaster, England: Centre for Research and Evidence on Security Threats.³³

The first listed 2018 publication is the current version of the VERA-2R which is used in the user training protocol. The 2009 publication is the academic publication (not the training manual) for the first version of the Violent Extremism Risk Assessment (VERA), which was produced for the Canadian Government, and is now publicly available. The 2012 and 2016 publications are publicly available published works which provide an overview of the second iteration of the VERA (VERA-2). The 2013 publication was conducted by independent researchers, and examines the validity and reliability of the VERA. The 2016 publication is a systematic review that assesses the quality of the risk factors and psychometric properties of a number of risk assessment instruments, including the VERA-2. The second 2018 publication was conducted by an independent researcher, and compares the VERA with the Extremism Risk Guidance (ERG) 22+. The 2019 publication includes an overview of the VERA-2R, and was compiled by an independent researcher with input from the authors of the VERA-2R.

As the VERA-2R has only recently been developed (Pressman et al., 2018), it was not possible to identify activity across all thematic categories directly from the instrument document. Therefore, the above cited evaluations and applications of previous iterations of the instrument were also scrutinised to identify missing information. The VERA-2R cites its development and strong reliability and validity based on evaluations of the earlier iterations, but to date, there is no specific evaluation of its own validity or reliability. This is concerning as the VERA-2R has 14 additional factors.³⁴ Rigorous evaluation of these elements is crucial when implementing a risk assessment instrument. The implications of this limitation are critical for the use of the VERA-2R, as the results of any tests of validity on the VERA (as completed by Beardsley & Beech, 2013) would not be transferrable to the VERA-2 or VERA-2R, as it is not possible to determine whether the additional factors would have an effect on the validity outcomes.

³³ Although Lloyd is the only stated author for this document, she notes "The originators of each framework completed their own sections" (Lloyd, 2019, p. 6)

³⁴ In Appendix 2, we offer a comprehensive estimation of the differences in factors across the VERA, VERA-2, and VERA-2R.

Approach

The VERA-2R builds on the previous iterations of the instrument, the VERA (Pressman, 2009) and the VERA-2 (Pressman & Flockton, 2012). The authors state that the original VERA instrument was “introduced as a consultative approach”, and developed “based on the existing knowledge base of violent extremists and terrorists and constructed into the structured professional judgement methodology”³⁵ (Pressman et al., 2018, p. 21). The VERA-2 was developed “following feedback from terrorism experts, national security analysts, law enforcement analysts, and professionals using the VERA with convicted terrorists in the high security prison setting” (Pressman et al., 2018, p. 21). The VERA-2R was developed from the original iterations, and empirical research on the subject of violent extremism and terrorism, from operational knowledge of law enforcement involved in terrorism and violent extremism, and from national security analysts.

The authors of the VERA-2R state that it is an SPJ instrument: “The VERA-2R supports professional judgement instead of replacing it” (Pressman et al. 2018, p. 22). The VERA-2R was not designed as a clinical instrument, as it can be implemented by psychiatrists, psychologists, analysts, and other professionals tasked with determining the risk of violent extremism in individuals. However, the authors state “it is standardized and in terms of structure and use, it is consistent with the methodology of the SPJ approach of psychological instruments” (Pressman et al., 2018, p. 23). This methodology consists of trained assessors examining a pre-established rating system (scoring factors with 0, 1, or 2) to identify the risk level of a defined set of risk and protective factors. According to the manual, the VERA-2R does not follow the SPJ approach as defined by Hart and Logan (2011). There is very little guidance on the gathering of case information (“All relevant background information should be compiled before the analysis is undertaken” and “The maximum number of available reports and supplementary information should be consulted for the background information” (Pressman et al., 2018, p. 25)), no mention of the process for

³⁵ Also note that Pressman (2009, p. 21) writes “Items on the VERA have been supported by the results of research undertaken in the area of radicalization and terrorism, are based on previous work undertaken in collaboration with RCMP personnel having operational experience with criminal violent extremists, have followed from discussions with professionals in the security and intelligence fields and have used relevant information obtained from interviews and self-report questionnaire data on radicalization.”

determining factor causality, and no scenario planning.³⁶ The process within the manual involves the gathering of case information, the rating of each factor, and the development of an overall risk judgment.³⁷

Although the VERA-2R lacks a systematic SPJ process, the authors of the VERA-2R do state that the instrument offers “a supplementary approach by psychologists and psychiatrists with *knowledge of violent extremism*” (emphasis added) (Pressman et al., 2018, p. 23), or other professionals who have undergone the specific training package from VERA trainers.

Following a traditional SPJ methodology, the VERA-2R uses a three-point scale in risk estimation. The instrument states “For reasons of standardization the user must carefully read and apply the definitions for each of the three risk levels (low, moderate, high). This distinguishes the VERA-2R from non-standardized assessments” (Pressman et al., 2018, p. 28). Pressman et al., (2018, p. 28) define the three risk levels as:

- ‘low’ if factor characteristics are “objectively *not* present”;
- ‘moderate’ if factor characteristics are “present to a *specified level*”;³⁸
- ‘high’ if factor characteristics are “*clearly* present or present to a *high level*”.

Factors within the Protective/Risk Mitigating category are coded within the three levels, however the calculation is inverted. Additional factors (Criminal History, Personal History, Mental Disorder) are coded as either present or not present.

During the assessment process, users are provided with a score sheet which allows them to keep track of their scoring (Pressman et al., 2018 p. 177-178). Once users have completed a score for each factor prior to the inclusion of the additional factors (note, there is no option for unknown, despite the authors noting that an

³⁶ Although the authors do state that the instrument does enable “the charting of risk pathways with the help of baseline measures and successive measurements repeated over the course of time to determine changes in each risk indicator” (Pressman et al., 2018, p. 22).

³⁷ The authors of the VERA-2R do use the term ‘formulation’, however, there is no mention of scenario planning or the analysis of past behaviour (Hart & Logan, 2011) within the assessment process, and without being privy to the training process for the VERA-2R, it is concluded that true formulation cannot occur in the use of the VERA-2R. This leads to the conclusion that the VERA-2R does not follow a true SPJ methodology.

³⁸ In this instance, specified refers to the wording within each indicator. For example, in BA1 – Commitment to Ideology that Justifies Use of Violence, s. 33(a)(i), s. 37(2)(b), s. 47E(d)

assessment may be completed with missing information), the score sheet provides an area for using the scoring to finalise their overall risk judgement (low, moderate, high). However, it is also possible for users to select an interval rating (low-moderate or moderate-high), giving flexibility in the overall risk assessment. This score sheet appears at odds with the earlier instructions that final risk judgments are arrived at following consideration of all available evidence, and not just the scoring of the factors.

Risk Specification

The VERA-2R states that it assesses the risk of violent extremism.³⁹ The authors state that “The development of an SPJ instrument specifically focused on violent extremism was essential due to analyses that revealed more than 70 percent of the factors used in other general violence risk-assessment instruments for judging the risk of general violence were not relevant to the characteristics of violent extremists” (Pressman et al., 2018, p. 20). However, the document also notes that “The VERA-2R... was specifically developed for the individual assessment of the risk of those persons inspired by any of the beliefs, principles, or philosophy that fall within the spectrum of ideologically motivated violence” (Pressman et al., 2018, p. 5), and “The VERA-2R uses mostly dynamic factors that correspond to the process of radicalization to extremist violence” (Pressman et al., 2018, p. 22). The inclusion of the terms ‘ideologically motivated violence’ and ‘radicalisation’ widens the remit of the originally specified risk. It also induces some confusion as to how a user should score individuals, and formulate an overall risk rating.

Themes

The VERA-2R groups the 45 included factors into six categories: Beliefs, attitudes, and ideology; social context and intention; history, action, and capability; commitment and motivation; protective/risk mitigating; and

³⁹ This differs from the specification of the VERA (“assess the risk of violence in ideologically motivated extremists” (Pressman, 2009), and the VERA-2 (“the risk assessment of terrorists and violent political extremists” and “intended for use with terrorists, it can be applied to the spectrum of violent extremists willing to use violence to further political, religious, or other ideological positions” (Pressman & Flockton, 2012, p. 244).

additional.^{40 41} The VERA-2 consists of 31 factors across six categories; Beliefs and Attitudes, Context and Intent, History and Capability, Commitment and Motivation, and Protective Items. The VERA includes 25 factors across four distinct categories; Attitudes/Mental Processes, Contextual/Social Factors, Historical Factors, and Protective Factors. The authors do not state their rationale for these categorisations, other than each category “includes” factors related to the categories (Pressman et al., 2018, p. 27). Appendix 2 highlights the similarities and differences in both factors and categories across all iterations of the VERA.

Given the VERA-2R's stated risk specification (risk of violent extremism), it would be expected that a large proportion of factors sit within the theme of Social Context and Capability. However, critical examination of where factors sit within the themes actually implies that the largest proportion of the factors within the VERA-2R are aligned more closely with assessing the risk of vulnerability to radicalisation (39.9% of factors are categorised as belonging to either Static or Dynamic risk factors). Further to this, 28.9% of factors are classified as appropriate for assessing the risk of radicalisation (Beliefs, Attitudes, and Behaviour). Whereas only 17.8% of factors are classified as indicative of a risk of terrorist or extremist violence (Social Context and Capability) (Table 5).

Table 5 Thematic Analysis of factors within the VERA-2R

Static Risk Factors	HAC1 – Early exposure to violence-promoting, militant ideology
	HAC2 – Network of family and friends involved in violent extremism
	HAC3 – Violent criminal history
	CH1 – Client of juvenile justice system/convicted for non-violent offence(s)*
	CH2 – Non-compliance with conditions or supervision*
	PH1 – History of family violence*
	PH2 – Problematic upbringing and/or placed in juvenile care*
Dynamic Risk Factors	CM2 – Motivated by criminal opportunism
	CM3 – Motivated by camaraderie, group belonging
	CM5 – Motivated by excitement and adventure

⁴⁰ Additional factors are intended to be used as supplementary to the overall risk decision determined by the assessment of all other factors. According to the document, “additional VERA – 2R indicators... may impact the risk of individuals engaging in violent extremism and terrorism... [additional indicators] may contribute to a vulnerability to future engagement in violent extremism activities when in combination with the presence of ideological, contextual, and motivational indicators identified” (Pressman et al., 2018, p. 121).

⁴¹ Additional factors are marked by an Asterix in Table 5.

	PH3 – Problems at school or work*
	SCI7 – Susceptibility to indoctrination
	MD1 – Personality disorder*
	MD2 – Depressive disorder*
	MD3 – Psychotic or schizophrenic disorder*
	MD4 – Autism spectrum disorder*
	MD5 – Post-traumatic stress disorder*
	MD6 – Substance use disorder*
Beliefs, Attitudes, and Behaviour	BA1 – Ideology that justifies the use of violence
	BA2 – Perceived grievances and/or injustice
	BA3 – Dehumanization of designated targets associated with injustice
	BA4 – Rejection of democratic society and values
	BA5 – Expressed emotions in response to perceived injustice
	BA6 – Hostility to national identity
	BA7 – Lack of empathy and understanding for those outside one's own group
	CM1 – Motivated by perceived religious obligation and/or glorification
	CM4 – Motivated by moral obligation, moral superiority
	CM7 – Motivated by acquisition of status
	CM8 – Motivated by search for meaning and significance in life
	SCI4 – Expressed intention to commit act of violent extremism
	SCI5 – Expressed willingness and/or preparation to die for a cause or belief
Social Context and Capability	HAC4 – Strategic, paramilitary and/or explosives training
	HAC5 – Training in extremist ideology in own country or abroad
	HAC6 – Organizational skills and access to funding and sources of help
	SCI1 – Seeker, user or developer of violent extremist materials
	SCI2 – Target for attack identified (person, group, location)
	SCI3 – Personal contact with violent extremists (informal or social context)
	SCI6 – Planning/preparation of acts of violent extremism
	CM6 – Forced participation in violent extremism
Protective Factors	P1 – Reinterpretation of the ideology
	P2 – Rejection of violence as a means to achieve goals
	P3 – Change in concept of the enemy

P4 – Participant in programs against violent extremism

P5 – Support from community for non-violence

P6 – Support from family and friends for non-violence

NB. * = Additional factors are to be considered to potentially "contribute to a vulnerability to future engagement in violent extremism *when in combination with* [emphasis added] the presence of ideological, contextual, and motivational indicators" (Pressman et al., 2018, p. 121).

Indicators

Quality

As the authors define the VERA-2R to be an SPJ instrument, there is an implication that each factor within the VERA-2R demonstrates "an empirical association with increased risk" (Guy, 2008, p. 15). The factors within the VERA-2R have been drawn from the two previous iterations of the instrument (the VERA and VERA-2). In the original VERA, the 25 factors were identified from a comparative item analysis of factors relevant for targeted violence identified by Fein and Vossekuil (1998), background factors identified by Borum et al. (1999), factors within the HCR-20 (Douglas et al., 2013), factors within SAVRY (Borum et al., 2006),^{42 43} and "personnel having operational experience with criminal violent extremists... discussions with professionals in the security and intelligence fields and have used relevant information obtained from interviews and self-report questionnaire data on radicalisation" (Pressman, 2009, p. 21). Given that the VERA was "intended for consultative purposes only" (Pressman, 2009, p. 21), in 2012, the VERA-2 was developed, revising the factors within the VERA based on operational knowledge and feedback from "terrorism experts, national security analysts, law enforcement analysts, and professionals" (Pressman et al., 2018, p. 21). The included risk factors developed through "empirical evidence obtained from applications of the VERA tool and feedback from experts"⁴⁴ (Pressman, 2016, p. 251).

⁴² None of these citations concern research on violent extremism or terrorism specifically. Fein and Vossekuil's (1998) study investigated characteristics of assassins and near lethal approachers of public figures, and is now considered seminal research in the area of fixated threat. Likewise, Borum et al.'s (1999) work is a state of knowledge and practice piece on risk and threat assessment in targeted violence (which encompasses a wide range of violent behaviours). The HCR-20 is a well-established and validated risk assessment tool for general violence, and the SAVRY is a risk assessment tool aimed specifically at violence in youth. It is noted that only a proportion of items in the HCR-20 and the SAVRY were relevant to violent extremism, and thus taken forward for inclusion.

⁴³ Items within the HCR-20 and the SAVRY that correspond with factors in the VERA are marked in Appendix 2.

⁴⁴ Applications included outcomes from use of the VERA on "convicted terrorists in the high security correctional setting" (Pressman, 2016, p. 251). Pressman and Flockton (2012) note that these settings were in Australia, but no further information as to the amount of evidence or number of cases is available.

The 14 new factors introduced into the VERA-2R were “identified as relevant to the radicalization to violence process”, and “identified as potential aggravating factors” (Pressman et al., 2018, p. 21). Pressman et al. (2018) do not specify the inclusion criteria or exact methodology behind the development of the 14 new factors. However, it is presumed that a literature review was included in this methodology, as each factor in the VERA-2R is preceded by a paragraph citing various works.

To examine the quality of the included factors, each cited source within the VERA-2R was scrutinised on three levels; empirical value of the cited works, relevance of the in-text statement in the VERA-2R to the cited works, and which element of the cited works the in-text statement refers to. This analysis highlighted that the supporting evidence base for the factors included in the VERA-2R may not truly fit within the defined elements necessary for a SPJ instruments.

Table 6 outlines the results of this analysis. Concerningly, a large proportion of the cited evidence base for factor development is not empirical (only 41.7% of cited works are empirical). Alongside this, only 48.1% of citations were accurately cited in the VERA-2R. That is, only 48.1% of cited statements in the VERA-2R were coded as accurately reflecting what was recorded in the cited texts.⁴⁵ Of the 51.9% of citations that did correspond to elements within the cited texts, only 48.5% of these (23.3% of the overall citations) were drawn from the empirical findings of the cited texts. Other areas of the texts that underwrote citations were predominately the literature reviews of the items, citing the works of other authors.⁴⁶ Additional to this, seven (15.5%) factors⁴⁷ have no citations included in the justification for inclusion.⁴⁸ Although the VERA-2R has 343 citations included in the literature reviews across all

⁴⁵ This is a concern as it indicates poor practice (attribution of statements to erroneous cited texts), which may indicate that the cited statement has no empirical basis.

⁴⁶ Proper practice would necessitate the authors of the VERA-2R citing these works instead. For example, for the statement “Foreign fighters perceive that they and the groups with which they identify, are more deprived, oppressed or persecuted than they should be, and that something should be done about it” was attributed in the VERA-2R to Dawson and Amarasingam (2017). However, Dawson and Amarasingam are citing from the concluding remarks of Weggemans, Bakker, and Grol (2014), and Bakker and Grol (2015). Proper practice would dictate that the authors of the VERA-2R source these documents and cite these works, or if unable to source the texts, provide a secondary citation, that accurately reflects that Dawson and Amarasingam (2017) did not identify this finding.

⁴⁷ HAC4 – Strategic, paramilitary and/or explosives training, P1 – Reinterpretation of ideology, P2 – Rejection of violence as a means to achieve goals, P3 – Change in concept of enemy, P4 – Participant in programs against violent extremism, P5 – Support from the community for non-violence, P6 – Support from family members, other important persons for non-violence.

⁴⁸ As the authors are not explicit as to the amount of evidence provided by feedback from practitioners, or tests conducted on prisoners, or how the evidence was used in the development of the VERA-2R, we are unable to assess where this evidence was utilised during instrument development.

factors, 203 (59.2%) of these citations were drawn from 57 published items. Appendix 3 presents the in-depth outcomes of this analysis using the red, amber, green colour coding system to show empirical quality of the factors within the instrument.

Table 6 Descriptive analysis of included factors within VERA-2R

Number of citations	343
Uncited statements (no support for statement)	95
Works relevant across multiple factors	57
Works relevant for one factor only	124
Empirical citations	143
Non-empirical citations	144
Citation relevant to text	165
Citation corresponds to empirical element of cited item	80

Table 7 provides an overview of the findings for the individual factor categories quality assessment. Each category of factors within the VERA-2R was critically assessed to determine its empirical value. The average cited empirical works is 40.8%, with only two categories (Commitment and Motivation, and Criminal History) having over 50% of the cited statements being attributed to empirical works. The average true empirical value of the categories (statements attributed to the empirical element of an empirical works) is only 26.5%, with the inclusion of factors related to personal history having the highest empirical value,⁴⁹ and those related to protective/mitigating the lowest, as there is no true empirical support for the inclusion of these factors.

Table 7 Empirical Value of Cited Supporting Literature in the VERA-2R

Indicator Category	Number of Cited Works	Empirical Works (%)	Cited Work Relevant to Statement (%)	Citation Relevant to Empirical Element of Works (%)
Beliefs, Attitudes, and Ideology	41	48.7	29.3	19.5

⁴⁹ Although this category has a very low number of cited works (20), which degrades the percentage value of 65%.

Social Context and Intention	71	38	50.7	16.9
History, Action and Capacity	54	48.1	51.9	29.6
Commitment and Motivation	48	52.1	60.4	22.9
Protective/Risk-mitigating	9	0	66.6	0
Supplementary Indicators				
Criminal History	11	54.5	9.1	36.4
Personal History	20	45	85	65
Mental Health	69	40.5	47.8	21.7

Weighting

Much like other risk assessment instruments focusing on extremism and terrorism, the authors of the VERA-2R state that although a scoring system is used, there is no fixed weight given to individual factors within the instrument, citing reasoning behind the SPJ methodology from Monahan (2012). Although there is an acknowledgement that some factors may be weighted by the assessor during assessment, there is no explicit mention of which factors should be weighted more heavily, only that the assessor has discretion during the process. To support the ambiguity, the authors state that “In line with the SPJ methodology, professional judgment is used to interpret the weighting of factors to arrive at an overall risk assessment based on evidence and professional judgment” (Pressman et al., 2018, p. 30-31). The authors of the VERA-2R specify that final judgements of risk are determined following the consideration of all available evidence (as guided by the scoring of indicators within the instrument), the consideration of the scoring of indicators, and the consideration of the relevance of the indicators as “related to the specified objective” (Pressman et al., 2018, p. 23).

Mechanisms

As noted above, the identification of factors that purportedly predict risk requires a sound theoretical basis. Although the authors of the VERA-2R do mention a range of “common elements or patterns” from a series of studies (including social bonds, shared identities, dogmatism, moral disengagement and rigidity, identity confusion, injustice, dehumanisation, enmity, hate, internalised martyrdom, frustration, rejection of Western values, and vengeance (Pressman et al., 2018, p. 16)), these findings are cited without any explicit mention to their exact role in movements towards violent extremism. Other than this list of findings from studies, no specific theory of change is provided. The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

(Table 2) is explicitly mentioned as fundamental for the development of the instrument across any of the originating documents of the VERA, the VERA-2, or the VERA-2R.

Lloyd (2019, p. 41) notes that theoretical underpinnings for the VERA-2R are drawn from “literature on violent extremism and terrorism” including; moral disengagement (Bandura, 1990; 2016), militant extremist mindset (Saucier et al., 2009), and quest for personal significance (Kruglanski et al., 2009). Following the findings of Gøtzsche-Astrup (2018), the work of Bandura (moral disengagement) is not included as providing empirical evidence, as this work remains theoretical, with no empirical work in the area of terrorism to date. The findings of Saucier et al. (elements within the militant extremist mindset) show moderate empirical evidence. The Kruglanski et al. piece is a theoretical model, which develops on strong empirical evidence (experience of fundamental uncertainty or loss of meaning or significance), but this theory of change (quest for significance) in and of itself currently requires an empirical evidence base.⁵⁰ These findings highlight that the VERA-2R requires deeper investment in its theoretical development.

Validity and Reliability

Effect

No tests using performance factors have been conducted on the VERA, VERA-2, or the VERA-2R. Therefore, it is not possible to discern whether the VERA-2R, or the previous iterations have any predictive validity, and if the instrument is able to predict its risk specification. Pressman (2016, p. 260) also notes that “Predictive validity, although desirable, was not considered an appropriate or realistic goal based on the dynamic character of the process of radicalization to violence.” Pressman does not provide any supporting evidence for this statement.

Validity

No tests of any other form of validity have been conducted on the VERA-2R. This again limits our ability to assess the predictive validity of the instrument. The VERA and VERA-2 have also not been tested for criterion validity.

⁵⁰ It is worth noting that the addition of quest for significance was led by a series of professionals who were asked to test the content validity of the VERA-2, not from a comprehensive theoretical or empirical means.

However, the authors state that the VERA-2R has been developed following the testing of the following validity areas on previous iterations of the instrument.

Face validity

In 2013 an independent study of the validity and applicability of the VERA⁵¹ was published by researchers in the United Kingdom (Beardsley & Beech, 2013). The researchers rated five known male terrorist offenders, each with a different espoused ideology (single issue, right wing, left wing, Irish nationalist, religious cult) and rated most factors (although there were large differences in both historical and protective items) as ‘relevant and important to risk assessment’ (Beardsley and Beech, 2013, p. 12). These results are said to support the face validity of the VERA-2R (Pressman et al., 2018). Pressman (2016, p. 262) also purports that face validity was tested on the VERA-2: “The face validity of the VERA and VERA Version 2 has been demonstrated via empirical evidence obtained from expert operational users over the past 5 years.” No methodological details are included, so we are unable to assess the empirical value of the evidence. Pressman notes that a series of professionals across intelligence, police, and mental health services who use the VERA-2 have “reported that it is relevant to their analytical needs” (Pressman, 2016, p. 262). However, the author did not expand on, or provide information of, any formal assessment procedures that enabled this conclusion. Therefore, we are unable to conclude with any certainty that these conclusions meet the criteria for face validity, because it is very unclear as to how “relevant to their analytical needs” is related to the instrument appearing effective in terms of its purported aims (in the case of the VERA-2, “monitoring and managing individuals suspected or convicted of terrorist offences” (Pressman & Flockton, 2012)). It is worth repeating that the Beardsley and Beech study was conducted on the VERA, and the Pressman (2016) conclusions are regarding the VERA-2, and given the noted differences in risk specification and factor inclusion between all iterations of the VERA, it is not possible to conclude that these studies present a strong case for face validity of the VERA-2R.

⁵¹ Cases were developed from open source (newspaper) information.

Content validity

Pressman (2016, p. 260) contends that the Beardsley and Beech (2013) study also provides an assessment of the content validity of the VERA. They note that the results of this study “revealed that the risk factors identified in the VERA could be applied with equal efficacy to terrorists of different ages, lone-actors, individuals who were members of groups and those who supported the spectrum of ideological causes and objectives.” However, this is not indicative of content validity, which is defined as ‘the extent to which the instrument represents all facets of the concept it purports to assess’. Pressman (2016) also notes that content validity of the VERA-2 was tested by 60 analysts.⁵² The exact methodology for this is not known (as no citations are provided). The work sought opinions from analysts on the importance of the risk-increasing and risk-mitigating factors within the VERA-2. Pressman (2016, p. 261) concludes that “Raters reported that the majority of the ‘risk indicators’ used in the VERA-2 and the VERA were highly important or very important” This is worded differently in Pressman et al. (2018, p. 152), who note that the analysts “agreed that the VERA-2 risk factors were extremely important or very important, with an agreement of above 85 percent”. These results are said to support the content validity of the VERA-2R. Again, this is problematic due to the differences in the factors included in the VERA-2R. It also begs the question as to why 11 new factors were introduced into the VERA-2R if the VERA-2 had such high content validity.

Construct validity

The testing of construct validity is dependent on the degree to which inferences from the data outputs of the instrument can legitimately be made from the theoretical constructs on which the factors in the instrument are based. Therefore, any outcomes of tests of construct validity of the VERA-2R should demonstrate that the outcomes of the VERA-2R support the theoretical constructs. In the VERA-2R, the theoretical constructs (as noted by Lloyd, 2019) include those noted above; moral disengagement, militant extremist mindset, and significance quest. Together these theoretical constructs are what supports the case for the use of the VERA-2R for use with suspected violent extremists.

⁵² Note, in the Pressman et al. (2018) document, this figure is stated as 28 analysts. Given the lack of cited methodologies in both Pressman (2016) and Pressman et al. (2018), it is not possible to determine if these tests of content validity are from the same study.

The authors of the VERA-2R state that construct validity was tested on the VERA-2 (Pressman, 2016; Pressman et al., 2018). This work compared a group of convicted violent extremists (N = 11) to a group of violent, non-ideologically motivated offenders (N = 11), matched on the basis of sex and ethnic and religious background, in a western high-risk secure prison setting. This work demonstrated that the violent extremist sample scored an average of 26.5 on the VERA-2, which was significantly higher than other instruments used in the general violence domain.⁵³ The authors argue that these results demonstrate that the VERA-2 is superior in assessing convicted violent extremists (as compared to other general violence instruments). Although demonstrating construct validity of the VERA-2, they do not demonstrate construct validity of the VERA-2R, due to the addition of 11 new factors and the altered risk specification. The inclusion of these factors may well affect the results if this experiment was to be repeated with the VERA-2R.

Other forms of validity⁵⁴

'Deductive validity'

In philosophy, a deductive argument is made if we accept the evidence (premise) as true, then we also accept the conclusion as true. It is not a recognised form of validity, but a form of argument used in validity (the other form being inductive). Pressman (2016) and Pressman et al. (2018) argue that if an individual under scrutiny conducts behaviours that are recognised in law as being related to violent extremism, then they are a violent extremist. The authors argue that if outputs from a VERA-2R assessment conform to those within the legal definitions of a violent extremist, then the VERA-2R has deductive validity. There are issues here, as deduction is an argument, and therefore not testable.

'Impression' validity and 'user' validity

Pressman (2016) and Pressman et al. (2018) state that users of the VERA have reported that the instrument creates the impression that it encompasses the entire concept it claims to measure (noted as 'impression' validity).

⁵³ 8.6 for the PCL-SV ($p < 0.01$), 18.8 for the LSI-R ($p < 0.05$), 9.9 for the HCR-20 ($p < 0.001$), and 3.9 for the VRS-SV ($p < 0.001$).

⁵⁴ Deductive, impression, and user validity are not recognised forms of validity in the psychological literature. Therefore, there are no concrete definitions or recognised test outcomes that would support the claims made by the authors making the statements.

Pressman et al. also note that assessors have reported that using the instrument supported their professional judgment regarding the risks and threats. No citations are given to support how this evidence was gathered, or any statistical testing that occurred to reach these conclusions.

Reliability

Currently, reliability of VERA-2R has not been tested. Although, a reliability test was conducted on the VERA and the VERA-2. Inter-rater reliability of the VERA was investigated by Beardsley and Beech (2013). This study found the degree of conformity between persons providing ratings ($N = 2$ across 5 cases) of 85.7%, with a Kendall coefficient of 0.76 or higher across elements. Pressman et al. (2018) also cite a second study that investigated inter-rater reliability, with this study investigating the VERA-2. This research involved five analysts each assessing four case files. Pressman et al. report that the Kendall's coefficients between 0.55 to 0.78 (where 0 is no relationship and 1 is a perfect relationship). On the removal of one rater (who lacked training and experience), the coefficients ranged between 0.60 to 0.82. There is no further information on the authors, methodology, or outputs of this study as no attribution is given in the VERA-2R documentation.⁵⁵ The authors cite these tests as demonstrating reliability for this version of the instrument, despite the VERA-2R having additional factors as compared to the VERA and the VERA-2. These factors have not been assessed for their reliability independently, or for their global impact on the instrument.

Authorship effects

Many of the tests of validity on the VERA, VERA-2, and VERA-2R (face, content, and construct validity) have been reported by the primary author of all instruments, without any further citations which would provide insight into the researchers who conducted these tests (Pressman, 2016; Pressman et al., 2018). Only Beardsley and Beech (2013) were found to be independent authors. Without further information as to the processes underlying these attributions, we are unable to conclude that authorship effects have not affected the reporting of validity of the VERA-2R.

⁵⁵ Which, much like the espoused validity, may infer authorship effects.

Moderators

The VERA-2R has been “specifically developed for the individual assessment of the risk of those persons inspired by any of the beliefs, principles, or philosophy that fall within the spectrum of ideologically motivated violence” (Pressman et al., 2018, p. 5).⁵⁶ The instrument has been designed to be “applicable to all ideological types of violent extremism” (Pressman et al., 2018, p. 24). The authors note that the VERA-2R can be applied to cases across a spectrum of ideologies, including “violent environmental extremists, violent anti-abortionists, violent nationalists, violent right-wing or left-wing extremists, and individuals motivated by violent religious, cultural or social ideologies” (Pressman et al., 2018, p. 22). The initial supporting literature review covers nationalism, right-wing, left-wing, single issue (animal rights, eco-environmentalism, anti-abortion), and religiously inspired extremism and terrorism. The authors also state that the instrument is “applicable both to loners and members of extremist groups” (Pressman et al., 2018, p. 22).

This inclusion of all ideologies, and inclusion of both loners and group-based actors may limit the validity and reliability of the instrument, as research has shown that there are differences in the early life and pre-event experiences of terrorists who espouse different ideologies (Gill et al., 2014), and between loners and group members (Comer & Gill, 2015; Gruenewald et al., 2013).

The authors state that the VERA-2R can be employed repeatedly on persons of interest over time (without any supporting evidence from testing intra-rater reliability). The authors also state that the instrument can be utilised across multiple settings, including under supervision, arrest, conviction, post-conviction, and during rehabilitation.⁵⁷ The VERA-2R also specifies, without theoretical or empirical evidence, that it can be employed to support assessments concerned with risk reduction.⁵⁸ Authors note “It can also provide support for preventive programs

⁵⁶ This stated subject pool differs from the VERA, which aimed to “assess the risk of re-offending in those already convicted of terrorist offences” (Lloyd, 2019, p. 5), and the VERA-2R, which aimed to aid professionals “charged with the responsibility of monitoring and managing individuals suspected or convicted of terrorist offences” (Pressman & Flockton, 2012, p. 244).

⁵⁷ Although the VERA and the VERA-2 were designed to be used in assessing convicted offenders only. The VERA-2R documentation does not give any empirical or theoretical support for the inclusion of new settings on use on non-offending populations. This is a concern for all forms of validity, particularly predictive validity. Of particular note, in the cited validity and reliability tests for the VERA-2R, all have been conducted on the VERA and the VERA-2, within which all case studies included during these investigations have been those who had already convicted terrorist offences.

⁵⁸ Although the VERA-2R does contain risk mitigating/protective factors, there is no stated theoretical basis for their inclusion, and the results of the thematic analysis also highlight that there is no empirical grounding for the inclusion of. The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Comer and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

and decisions on priorities for supervision of individuals” (Pressman et al., 2018, p. 31). Lloyd (2019, p. 40) also notes that “The VERA-2R is appropriate for youths as well as for male and female adults. It has been used with aspiring and returning foreign terrorist fighters and their returning families”. This wide-ranging inclusion criteria remit allows for use by multiple practitioner groups across multiple settings, however, from a reliability and validity perspective it is concerning. Narrow risk specification and factor inclusion is essential during instrument development, as evidence has shown that instruments do not perform equally well across multiple forms of behaviour (Sarma, 2017).

Implementation Burdens

According to Pressman et al. (2018), the VERA-2R has been designed for use by: Forensic psychologists and psychiatrists; other specialists within the criminal justice system; forensic social workers; analysts at national security agencies; intelligence services; national and local police forces. Users must also hold a mandate from their organisation for violent extremism risk assessment, and be able to understand the VERA-2R rating definitions in the manner required for an SPJ methodology. Arguably a fundamental implementation burden of the VERA-2R is the requirement for specific training prior to use. All users of the VERA-2R must successfully complete a recognised VERA-2R training program. According to Pressman and Flockton (2012, p. 246) “This is to ensure familiarity and consistency in the methodology of the ratings and to ensure familiarity with the ratings and their correct interpretation. Training is intended to enhance the reliability of the approach and ensure that the ratings are interpreted in a consistent manner.”

This training program includes background information on the VERA-2R, the knowledge base related to violent extremism, terrorism and the radicalisation process, how to apply the VERA-2R factors, and the completion of assessments using case studies. Users must also have prior experience of interviews and/or structured assessment processes and knowledge of the violent extremism research literature. There is no information within any documentation as to the time or costs of this training program, but it is believed to be a multiple day program with substantial costs attached.

the protective factors (no factor has any reference to cited works, and the introduction section has no empirical citations; see Appendix 3).

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

Significant time is needed to administer the instrument, and users must obtain all relevant information from a wide range of sources. It is not necessary to draw information from self-report interview data, but this is not discouraged. The authors note that during this process, users of the VERA-2R must consider information according to: “1. The credibility of the source; 2. The correctness of the information; 3. The importance of the information (i.e. the weighting) in terms of general risk; 4. The appropriateness of the information for rating the specific risk factor; 5. The appropriateness of the general information for the provision of a substantiated and sound VERA-2R risk assessment at that moment” (Pressman et al., 2018, p. 29).

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Instrument 2 – Radar

The assessment of this instrument is based on the following reports and papers:

- Australian Government. (2017). *Radar: Countering violent extremism (CVE) intervention tools: User guidelines – version 2*. Canberra, Australia: Australian Government.
- Australian Government. (2018). *Radar: countering violent extremism (CVE) intervention tools training: Participant workbook – version 2*. Canberra, Australia: Australian Government.⁵⁹
- s. 47F(1) (2013). *Framework of factors of radicalisation and extremism (FIRE)*. Victoria: Global Terrorism Research Centre, Monash University.
- Barrelle, K. (2014). *Pro-Integration: Disengagement and life after extremism*. PhD Dissertation, Monash University.
- Barrelle, K. (2015). 'Pro-integration: Disengagement from and life after extremism', *Behavioral Sciences of Terrorism and Political Aggression*, 7(2), 129-142.
- Global Terrorism Research Centre. (2012). *Implementing the radicalisation indicators model (TRIM)*. Melbourne: Monash University.
- RTI International. (2017). *Countering violent extremism: The use of assessment tools for measuring violence risk*. Washington DC: Department of Homeland Security.
- Cherney, A. & Belton, E. (2019). Evaluating case-managed approaches to counter radicalization and violent extremism: An example of the Proactive Integrated Support Model (PRISM) intervention. *Studies in Conflict & Terrorism*, 0(00).

⁵⁹ It is noted that the research team also interrogated all available items that were referenced as further reading at the start of each module to determine if any were critical for the thematic analysis.

The 2017 Australian Government publication is the Radar risk assessment user guide. This guide is used during training of the Radar protocol. The 2018 publication is the participant workbook which is also used in the training protocol. This document includes case studies and worksheets that are used during the training. The s. 47F(1) (2013) document is the empirical and theoretical framework which forms the basis of the Risk Analysis used in Radar. This Framework (FIRE) presents the risk factors which are present in the Risk Screening and In-Depth Risk Analysis. s. 47F(1) (2013, p. 3), state in their report that “simplified versions of information contained within this report will be... disseminated throughout Australia by the Federal Attorney General’s department, in the form of a series of factsheets and a booklet”. It is unclear whether the booklet referred to is the 2017 Australia Government Radar user guide. None of these publications are available in the public domain. The 2014 publication is the PhD dissertation of the co-author of Radar and contains information on the Needs Analysis stage in Radar. This publication is publicly available. The 2015 publication is a journal article based on the 2014 PhD dissertation document. Both the 2013 and 2017 documents cite the model produced from the 2012 document. This document is a report to the Attorney General’s Department, and provides an overview of one of the theoretical models underpinning Radar. The 2017 RTI International document is a literature review which was prepared for the United States Department of Homeland Security. This document describes a series of risk assessment instruments, and includes a short section dedicated to Radar. The 2019 document is a peer-reviewed publication which evaluates the Proactive Integrated Support Model (PRISM) risk assessment. This document is included as it utilises assessment information which was drawn from the Needs Analysis protocol in Radar.

As the documentation for Radar has only recently been developed (Australian Government, 2017, 2018), and is made up of a number of unique instruments, each utilised at different stages during the assessment process, it was not possible to identify activity across all thematic categories directly from the instrument guidelines document. Therefore, the other above cited documents were also scrutinised to identify missing information.

[The Radar Assessment Process](#)

The Radar assessment process includes a series of five protocols which have been designed to be used at different stages along the countering violent extremism (CVE) intervention pathway. The first is referred to as the Client Intake and involves the completion of a Client Intake form which records the client’s personal details and the nature of the referral. This form also includes a section entitled “Early Flags” which includes a series of risk factors that have been

established as relevant to violent behaviour in other instruments (mental health, engagement with services, substance use, offending, and violence), and allows for the analyst to note “a range of other issues should they be overtly provided or immediately apparent” (Australian Government, 2017, p. 5). It is also presumed that this form is completed to help users establish a baseline to measure the presence of factors during the later protocols, as “radicalisation requires an increase in intensity over time” (S. 47F(1) [REDACTED], 2013, p. 22). The second protocol is the Client Contact Person (CCP) form, which identifies a suitable person to “provide the primary point of contact for the Intervention Panel and support services with regard to the client” (Australian Government, 2017, p. 9).

The third protocol is the Risk Analysis which is composed of two components: The Initial Risk Screening, made up of 18 factors,⁶⁰ and the In-Depth Risk Analysis, made up of 27 factors. The 15 risk factors within the Risk Screening are also within the In-Depth Risk Analysis.⁶¹ The rationale behind the requirement for the Risk Screening appears to be some form of triage.⁶² The authors of the 2017 document are careful to note that:

“The purpose of the risk screening is not to fully determine the level of risk an individual poses. Rather, the screening is designed to make an initial determination as to the suitability of a client for an Intervention programme, and whether to proceed for a full risk analysis. This approach was taken based on guidance provided by the CVESC EIWG” (Australian Government, 2017, p. 15).

Based on the outcomes of the Risk Screening,⁶³ an individual may be selected to proceed to the In-Depth Risk Analysis stage.⁶⁴

⁶⁰ Which includes 15 risk and 3 protective factors.

⁶¹ The authors do not provide any justification or explanation is provided for why these 15 indicators were selected for inclusion in the risk screening stage.

⁶² However, there is no theoretical or empirical justification given for the choice of the 15 risk factors (out of the final 27) for the Risk Screening.

⁶³ If there are only Notable factors (out of a possible six) identified, the individual only proceeds at the discretion of the intervention panel, if three or more (out of a possible six) Concerning factors identified, the individual should proceed (and if less than three, proceed at the discretion of the panel). If any Attention factors (out of a possible three) are identified, the individual should proceed.

⁶⁴ The authors note that for any individual who passes through the screening process and is determined to require examination using the In-Depth Risk Analysis, “it is first necessary to pass this individual through a de-confliction process.” (Australian Government, 2017, p. 17). It is noted that this process is a requirement of the EIWG. However, no further information of what the de-confliction process entails is given.

The fourth protocol is the Needs Analysis which was developed from the work undertaken by Barrelle (2014, 2015).⁶⁵ The purpose of the Needs Analysis is to assess a client's "rehabilitation needs" and consists of "five evidence-based domains relevant to disengagement from violent extremism and subsequent societal re-integration" (Australian Government, 2017, p. 43). The final protocol "maps the progress of clients involved in an intervention programme to ensure that services are meeting the programme's objectives and the individual's identified needs" (Australian Government, 2017, p. 3).

The protocol under scrutiny during this research is the Risk Analysis (both Risk Screening and In-Depth Risk Analysis components). The rationale being that the Risk Analysis focuses on the identification of suitable individuals for intervention, which matches the VERA-2R process. It would also not be possible to ascertain validity measures for the first two protocols (Client Intake, and Client Contact person) as these are administrative instruments. The Needs Analysis and Intervention process are specifically designed to be used as part of the disengagement and intervention process once at-risk individuals have been identified using the Risk Analysis. Therefore, these protocols are also not assessed in this research.

Underwriting Documentation

TRIM and FIRE

TRIM was the outcome of a four-year Australian Research Council (ARC) Linkage project undertaken by Victoria police, the Victorian Department of Premier and Cabinet, Corrections Victoria, and the AFP. TRIM is defined as "a coherent model" and "a systematic framework to assist in identifying whether a person is becoming radicalised towards political violence" (Global Terrorism Research Centre, 2012⁶⁶). Within the 2012 document it is stated that, "TRIM derives its content from both bottom-up data from actual cases of radicalisation as well as top down from

⁶⁵ The Needs Analysis appears to be based on the theoretical Pro-Integration Model (PIM) developed by Barrelle (2014; 2015). According to Barrelle (2015), "PIM is a new conceptualisation of disengagement from violent extremism and reintegration into society that combines the relevant empirical and theoretical literature with five emergent themes (p. 133). These themes (or domains) include social relations, ideology, action orientation, coping and identity (Australian Government, 2017; Barrelle, 2015). The Needs Analysis is accompanied by a Radar Chart developed to visually map the risk ratings across the five domains. According to the Radar user guide, "The Radar Chart is an evidence-based analytical tool" designed to measure disengagement from violent extremism (Australian Government, 2017, p. 71). Unfortunately, despite extensive searches and contact with the authors of Radar (undertaken by the Department), any further information on the theory or methodology underlying PIM was not sourced.

⁶⁶ No page number available.

theories of radicalisation” (Global Terrorism Research Centre, 2012⁶⁷). However, no further information is provided on where this data was sourced.⁶⁸

TRIM consists of two components, each with three elements. The first component consists of theoretical categories, which “have been identified as present in forms of serious radicalisation” (Global Terrorism Research Centre, 2012⁶⁹). These categories are; Social Relations, Ideology, and Action Orientation.⁷⁰ The second component focuses on three levels of “radicalisation intensity”, Notable (minor), Concerning (moderate), and Attention (major). (Global Terrorism Research Centre, 2012⁷¹). These three levels of intensity are employed as part of the assessment “to identify an escalation in radicalisation” s. 47F(1) 2013, p. 18). The components are compiled to form a matrix, with nine distinct elements.⁷²

The 27 risk factors contained within the Risk Screening and In-Depth Risk Analysis components of the Risk Analysis protocol are drawn from the Framework of Indicators of Radicalisation and Extremism (FIRE) with three indicators within each area of the matrix within TRIM. FIRE was developed in a collaborative project between the Australian Federal Police (AFP), Monash University, Victoria Police, and the State Government of Victoria. FIRE is underwritten by The Radicalisation Indicators Model (TRIM), and “over 200 real world (mostly Australian) examples” of radicalisation s. 47F(1) 2013, p. 14).

As both TRIM and FIRE are core theoretical components of the Risk Screening and In-Depth Risk Analysis, they are also subject to evaluation during this research.

⁶⁷ No page number available.

⁶⁸ The document does state “GTRC has drawn these indicators together into a coherent model called TRIM and we anticipate publication of this model in a prestigious peer-reviewed academic journal mid-2013” (Global Terrorism Research Centre, 2012). However, despite the systematic search protocol of this research, no document was sourced.

⁶⁹ No page number available.

⁷⁰ However, the FIRE documentation lists these categories as Social Relations, Ideology, and Criminal Orientation s. 47F(4) 2013), and the Radar documentations lists them as Social Relations, Ideology, and Criminal Orientation, and also Social Relations, Ideology, and Action Orientation (Australian Government, 2017, 2018).

⁷¹ No page number available.

⁷² These elements are; Social Relations_Notable, Social Relations_Concerning, Social Relations_Attention, Ideology_Notable, Ideology_Concerning, Ideology_Attention, and Action Orientation_Notable, Action Orientation_Concerning, Action Orientation_Attention.

Approach

Within the process of Radar, the protocol that is utilised to determine the risk specification is the Risk Analysis (which includes the Risk Screening and In-depth Risk Analysis). This protocol is used “to determine whether or not a potential client may be appropriate for involvement in an intervention programme” (Australian Government, 2017, p. 15). The documentation also states that the outcomes from this protocol, are “not suitable for use in judicial settings against the client as proof of ‘risk’” (Australian Government, 2017, p. 16)

The 2017 report states that the approach taken within both the Risk Screening and In-Depth Risk Analysis components align with the SPJ methodology, as “the purpose of the approach adopted here aligns with the Structured Professional Judgement tools in that it aims to identify a high risk group/cohort, to which risk management interventions can be applied, rather than predicting a low base rate event” (Australian Government, 2017, p. 16). The report also states that the protocol “utilises evidence-based behavioural factors of radicalisation within a structured methodology to assist assessors to make estimations of risk based on their own professional judgements” (Australian Government, 2017, p. 19), and “assessments should be made on the basis of the professional judgement of the analyst supported by the structured arrangement of the data” (Australian Government, 2017, p. 17). Similarly, the FIRE report states that when conducting the risk analysis, “decisions are made on the basis of the structured professional judgement of the assessor” s. 47F(1) 2013, p. 34,⁷³ and the TRIM documentation states “TRIM is based on the theory of structured professional judgement (SPJ), an approach commonly used in assessing and predicting complex human behaviour... The SPJ approach adopted in TRIM combines the best of research knowledge with the front-line professional or personal experience of those who have contact with people who may be radicalising towards violent extremism” (Global Terrorism Research Centre, 2012⁷⁴).

⁷³ The 2017 document also states that “Parallel to conducting an in-depth risk analysis, a threat assessment regarding a potential client should also be conducted.” (Australian Government, 2017, p. 19). The document goes on to state “Such an assessment should follow standard policing procedures in order to determine all risks posed, beyond those specifically related to violent extremism.” (Australian Government, 2017, p. 19). No further details are provided. It is also noted that this statement does not cover the definition of a threat assessment. Threat assessments are conducted following the identification of an individual who poses a threat to a specific target, even in the absence of identifiable risk factors.

⁷⁴ No page number available.

However, as shown within the assessment of the VERA-2R, these statements are not sufficient to formally define the Risk Analysis protocol in Radar as an SPJ instrument (Guy, 2008). The specified process for Radar includes assessors⁷⁵ examining a pre-established rating system of factors to establish a risk estimation. There is no guidance within the instrument of the structured protocol for the gathering of case information (no instructions for the sourcing of information or the acceptable forms of information to be used). The inclusion of the CCP form is promoted as a facilitator for the “information gathering process for a risk and needs analysis and to assist in the intervention design itself” (Australian Government, 2017, p. 9). However, the protocol for the CCP relates to the choice of individual as suitable for the CCP, not the information they are expected to provide. There is also no mention of the process used for the determination of factor causality. The TRIM protocol helps users determine the relative weighting of included factors, but not how such factors cause the outcome (radicalisation).⁷⁶ There is also no scenario planning included in the Radar process.⁷⁷

The Risk Screening and In-Depth Risk Analysis protocols also do not follow a traditional SPJ scoring scale for risk estimation. In both the Risk Screening and In-Depth Risk Analysis, the scoring includes;

- partially or historically reported (P),
- fully reported (F).

In the In-Depth Risk Analysis, users are also encouraged to score;

- not reported.

There is little rationale given for the inclusion of the need for noting the partial and historical reporting of factors, other than a presumption that it is used to compare to the baseline indicators noted in the Client Intake Form.⁷⁸

⁷⁵ It is noted that the TRIM documentation also states that the “27 indicators are intended to provide markers which will assist intelligence officers, law enforcement, community and religious leaders and even friends and family to identify individuals who may be at risk”. The inference here that any of these stated individuals may be assessors – which does not fit within the typical SPJ methodology, which requires experienced professionals.

⁷⁶ The FIRE protocol does dictate that “Observed behaviours must become more intense and extreme over time when compared to that person’s previous or ‘normal’ baseline of behaviour” (S. 47F(1) 2013, p. 22).

⁷⁷ The training manual includes case studies to aid in the training protocol, but this is not scenario planning.

⁷⁸ Despite the factors noted in the Client Intake Form being focused on mental health and violence, where mental health is only included in the Needs Analysis, and violence (as linked to an ideology) is found only within the attention indicators.

However, we are unable to find support for this, and in fact, the reporting of partial risk factors appears to be nullified in the final risk judgment, as users are instructed not to report these factors during finalisation of the Risk Analysis results form as “Only indicators which have been assessed as ‘reported’ are transferred onto the results form. ‘Partially reported’ and ‘not reported’ indicators from the risk analysis summary forms are not included in the risk analysis results.” (Australian Government, 2017, p. 38).

Users of Radar are provided with score sheets for both the Risk Screening and In-Depth Risk Analysis components, which allows them to keep track of their scoring. In the Risk Screening protocol, the score sheet includes individual areas for the scoring across the TRIM structure, and a space for the summing of factors within each area of TRIM. In the In-Depth Risk Analysis score sheet, there is space for users to mark factors as either reported, partially reported, or not reported. Users are then asked to sum the number of both reported and not reported factors for each element of the TRIM structure. Users are then provided with an In-Depth Analysis results form, which is designed as a matrix. Users are encouraged to sum the number of factors both across the rows and down the columns, but the figure taken forward for the risk judgment is only the sum from the row totals.

Risk Specification

The language used to describe the risk specification of Radar varies, and it is often unclear what the exact risk specification is. The Radar user guidelines documents states that “The aim of a CVE intervention is to connect at-risk individuals... that may be able to assist them to disengage from violent extremism” (Australian Government, 2017, p. 1), and also “The Risk Analysis Tool has been specifically designed to assist in the identification of behaviours which may result in someone threatening, advocating or participating in politically motivated violence” (Australian Government, 2017, p. 16). However, later, the document also states that the “in-depth Risk Analysis Tool assesses the risk level of those who have been identified as at risk of, or vulnerable to radicalisation towards violent extremism”, and also “applied to individuals undergoing the radicalisation process” (Australian Government, 2017, p. 19). The FIRE report states that the instrument is both “designed to facilitate identification of radicalization” s. 47F(1) 2013, p. 1), and “designed to facilitate recognition of radicalisation which results in people threatening, advocating or participating in politically motivated violence in Australia” s. 47F(1) 2013,

p. 3). The Global Terrorism Research Centre (2012⁷⁹) notes that the specification of TRIM is to “assist in identifying whether a person is becoming radicalised towards political violence” and that the “27 indicators are intended to provide markers.... to identify individuals who may be at risk of serious radicalisation.”

Themes

Given Radar’s stated risk specification; assessing the risk of those “vulnerable to radicalisation towards violent extremism” (Australian Government, 2017, p. 19), it is expected that a large proportion of the 27 factors in the Risk Screening and In-Depth Risk Analysis components will sit within the Static and Dynamic Risk Factors, and a small proportion may be classified as sitting within Beliefs, Attitudes and Behaviours.

However, the majority of the factors (44%) in the Risk Screening (Table 8) are classified as screening for facets of an individual typically associate with already being radicalised (Beliefs, Attitudes & Behaviours), and 39% are classified as indicative of a risk of terrorist or extremist violence (Social Context and Capability). There are no factors which fall under the category of Static or Dynamic Risk Factors. This is surprising, given that the screening process is designed to make “an initial determination as to the suitability of a client for an Intervention Programme” (Australian Government, 2017, p. 15).

This pattern is mirrored in the In-Depth Risk Analysis. Table 9 demonstrates that over half of the factors (55.6%) were thematically classified under Beliefs, Attitudes & Behaviours, which infers that these factors are associated with radicalism, and 44.4% of factors are associated with the individual’s social network and capability for violence, which infers that these factors indicate that the individual is at risk of conducting terrorist or extremist violence.

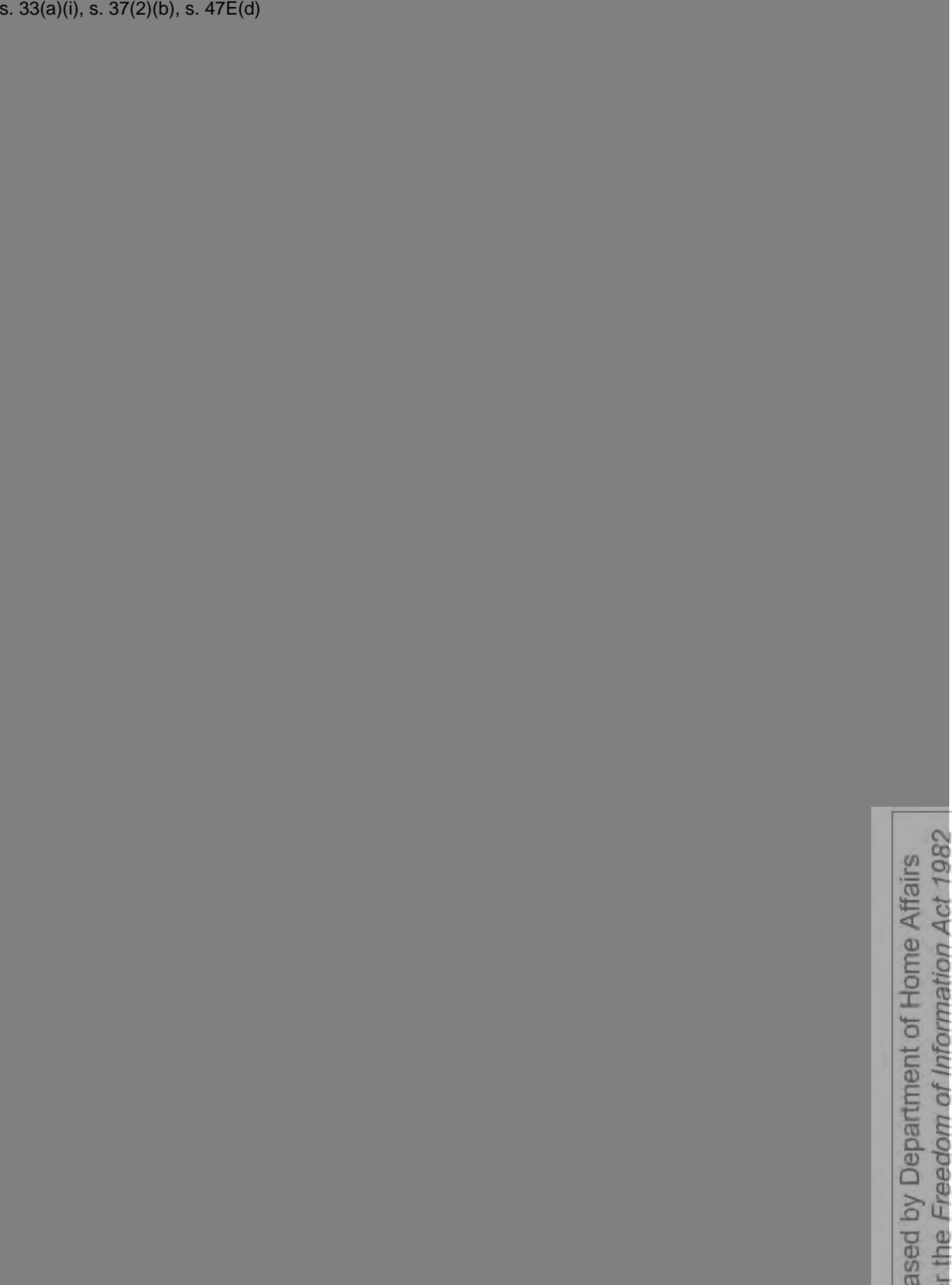
s. 33(a)(i), s. 37(2)(b), s. 47E(d)

⁷⁹ No page number included.

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Instruments.

The content of this report may be shared within agencies but is not to be reproduced or distributed further.

s. 33(a)(i), s. 37(2)(b), s. 47E(d)



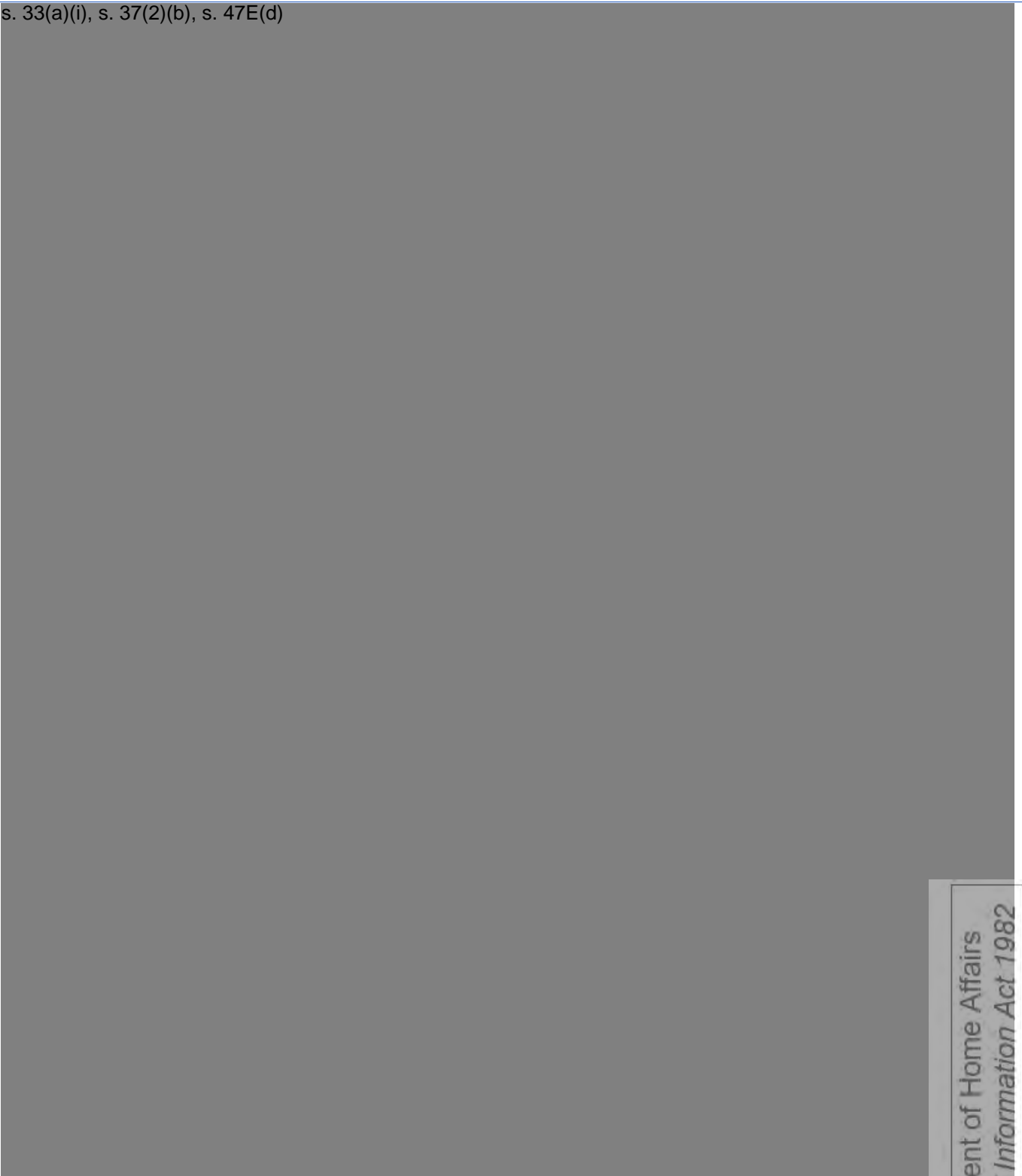
Released by Department of Home Affairs
under the Freedom of Information Act 1982

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Instruments.

The content of this report may be shared within agencies but is not to be reproduced or distributed further.

s. 33(a)(i), s. 37(2)(b), s. 47E(d)



Released by Department of Home Affairs
under the Freedom of Information Act 1982

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

Indicators

Quality

s. 47F(1) (2013) state that Risk Screening and In-Depth Risk Analysis elements of the Risk Assessment protocol follow an SPJ methodology. Much like the VERA-2R, this implies that each included factor has been developed through the investigation of the empirical evidence base.

The categories and levels of intensity of the TRIM were included based on an investigation of “the history of violent extremism in Australia, in order to better understand the unique elements of the specific geographical location” and through a “thorough literature review of all work into the study of radicalisation which is directly relevant or comparable to the Australian context” s. 47F(1) 2013, p. 10).⁸¹ Works that were reviewed included “theoretical academic works, government and academic empirical studies and individual open source case studies of radicalisation” s. 47F(1) 2013, p. 10). There is no information regarding the protocol for this review, or the number of studies included, which would have provided support for the espoused approach for this protocol.

The authors of the FIRE document also note that following the literature review, the researchers also gathered numerous forms of data. This data included: “over 50 interviews... with current or former violent extremists, drawn mostly from Australia or other directly comparable countries and from a diverse range of ideologies” s. 47F(1) 2013, p. 10); interviews “with 20 law enforcement officers from four countries (including former undercover operatives in terrorism investigations)”; “extensive practical knowledge of four senior Victoria Police Counter-Terrorism personnel”; more than “30 interviews... with religious and community leaders around Australia”; and 481 (4,000 pages) evidentiary Listening Device (LD) and Telephone Intercept (TI) transcripts from a joint State and Federal police investigation”⁸² s. 47F(1) 2013, p. 11). This data then underwent “research triangulation” with media sources and court documents to help develop the 27 indicators that are within FIRE.

⁸¹ The authors note that studies comparable to the Australian context include those from “countries with broadly comparable political and legal systems and with relatively similar modern populations. This included New Zealand, Canada, the United States, the United Kingdom and Scandinavia” s. 47F(1) 2013, p. 10).

⁸² This investigation commenced in 2004, and evolved over an 18-month period.

There is no more information on any of the methodology underpinning the collection or collation of this data. Worryingly, there is also a lack of information that would allow replication of this study, a necessity for testing of reliability and validity of the research underpinning the 27 included factors.

To examine the quality of each risk factor within FIRE, and the categories within TRIM, each source that is cited in the documentation regarding TRIM and FIRE was scrutinised on three levels; the empirical value of the cited works, the relevance of the in-text statement in documentation to the cited works, and to which element of the cited works the in-text statement refers to. This analysis highlighted that, much like the VERA-2R, the supporting evidence base for the 27 factors included in the Risk Analysis protocol in Radar does not fit within that necessary for an SPJ instrument.

Table 10 identifies the outcomes of the analysis, and thus the provenance of citations included in the TRIM, FIRE and Risk Screening and In-Depth Risk Analysis documentation. Despite the assertions that the TRIM and FIRE documentation is underwritten by a multitude of different sources, there are only 28 citations underwriting the three categories of TRIM, and 142 citations underwriting the 27 indicators of FIRE. Much like the VERA-2R, a large proportion of the cited evidence base for factor development was not empirical (71.6%), and a significant proportion (62%) of the cited documents were drawn from media sources. Unlike the VERA-2R, a large proportion of the citations in Radar were coded as accurately reflecting the cited texts. However, within these statements, only 15% were drawn from the empirical findings of the cited texts. Other areas of the texts that underwrote citations were predominately the opinion of the authors of the original texts, with no empirical source for these opinions identified. What is highly concerning is that 30.5% of all statements were not attributed to any evidence other than the opinion of the authors of the Radar documentation. Appendix 4 shows the in-depth outcomes of this analysis using the green, amber, red colour coding system to show empirical quality of the factors within the protocol.

Table 10 Descriptive analysis of included citations in the Risk Analysis protocol in Radar

Number of citations	232
FIRE Documentation	28
Indicators	204
Uncited statements (no support for statement)	102

Works relevant across multiple factors	34
Works relevant for one factor only	108
Empirical citations	66
Non-empirical citations	151
Media Sources	88
Interviews ^a	20
Transcripts ^a	6
Citation relevant to text	165
Citation corresponds to empirical element of cited item	35

^a - As these sources were unreferenced, it is not possible to discern if these figures represent independent elements, or are repetition (i.e. if there are 20 cited interviews, or one interview cited 20 times).

Table 11 offers an overview of the analysis examining the categories of TRIM, and the factors within them. There is an average of nine citations underpinning the categories, with an average of 21% of citations having empirical value. The Action Orientation category⁸³ has no empirical works underpinning its inclusion. The average true empirical value of the three categories is 8.1%. In relation to the factors within each of the TRIM categories, an average 28.7% were from empirical works, with 15.6% holding true empirical value. The Social Relations category has the highest empirical support underpinning its inclusion.

Table 11 Empirical value of cited supporting literature in the Risk Analysis protocol in Radar

Indicator Category	Number of Cited Works	Empirical Works (%)	Cited Work Relevant to Statement (%)	Citation Relevant to Empirical Element of Works (%)
Social Relations	14	42.9	64.3	14.3
Indicators	67	43.3	56.7	20.9
Ideology	10	20	80	10
Indicators	59	23.7	76.3	11.9

⁸³ This issue is returned to in task 4.

Criminal/Action Orientation	4	0	75	0
Indicators	78	19.2	80.8	14.1

Weighting

Within the elements of the Risk Analysis protocol, there is little information with regards to the weighting of factors. Both elements use the TRIM intensity structure (Notable, Concerning, Attention), however, and there is some inference that factors within the Notable section have little or no weighting, those within the Concerning section have some weighting, and those within the Attention section have heavy weighting applied.⁸⁴

Mechanisms

The FIRE report outlines “three social science theories” which “have together formulated a nuanced understanding of many of the elements present within the radicalisation process...These include new social movement theory, social psychology, and conversion theory” (s. 47F(1) 2013, p. 7).⁸⁵ Closer inspection of the cited literature underlying FIRE reveals that the support for the inclusion of ‘social movement theory’ is drawn from work citing shared identities, and Wiktorowicz’s (2005) radicalisation model. The support for ‘social psychology theory’ focuses on group influence, social identity, moral disengagement, and intergroup conflict, and a deeper examination of ‘conversion theory’ reveals the focus is on the “individual experience of a transformation in belief and ideology” (s. 47F(1) 2013, p. 7). Much like the VERA-2R, the authors draw on the work of Bandura, citing moral disengagement and dehumanisation, and they do touch on relative deprivation, but attribute no specific evidence to this⁸⁶ (s. 47F(1) 2013), but as mentioned above, the work of Bandura is not noted as having an empirical basis in the domain of terrorism studies. Other than this, no specific theory of change (matching those within Table 2) is explicitly mentioned as fundamental for the development of the instrument across

⁸⁴ The instructions in the Risk Screening protocol instruct users to use discretion (for proceeding) if only Notable factors are noted, to proceed to the In-Depth Risk Analysis if three or more Concerning factors are noted (and discretion if less than three), and to immediately proceed to In-Depth Risk Analysis if any Attention factors are noted (Australian Government, 2017, p. 18). However, on page 17 of the same document, it also states “any screening which does not identify an indicator in the red or orange [Attention or Concerning] may not proceed to an intervention”. In the Risk Analysis, similar weighting applies across the three sectors.

⁸⁵ It is worth noting that social psychology is an academic discipline, not a theory.

⁸⁶ So, we are unable to assess the empirical basis for this statement.

any of the originating documents of the Risk Analysis protocol in Radar. This highlights that the theoretical basis of the Risk Analysis protocol in Radar is particularly weak.

Validity and Reliability

Effect

No tests of performance factors have been conducted on any literature that has been scrutinised during the assessment of Radar. Therefore, we are unable to discern whether Radar, or any of the underlying protocols have any predictive validity, and if the instrument is able to predict its risk specification. In fact, the documentation underwriting the Risk Screening and In-Depth Risk Analysis components states “it is extremely difficult (potentially impossible) to predict with any level of accuracy who will become a violent extremist. In this way risk analysis does not provide statistically predictive results... the Risk Analysis Tool presented here has not been empirically validated for predicting risk in an individual” (Australian Government, 2017, p. 15).

Validity

No tests of any other form of validity have been conducted on Radar, or the underlying models or protocols that assist in the identification of risk. This again limits our ability to assess the predictive validity of the instrument. The 2013 s. 47F(1) document details an appended report written by Victoria Police provides some preliminary findings on the application of FIRE and TRIM (but not the 27 indicators within) in frontline police training programs. The outcomes of this training included “evaluation and feedback received following the Victorian sessions was extremely positive and indicated an improvement in the overall awareness of radicalisation indicators as a direct result of the training” s. 47F(1) 2013, p. 67), and “The evaluation and feedback received throughout the Queensland... [and] Tasmanian sessions was again extremely positive” s. 47F(1) 2013, p. 69).

These sessions may offer some (weak) evidence of face validity, as the assessors interpreted the factors, and provided feedback that they strengthened their understanding and practice, however, there is no explicit mention of the 'expertise' of these individuals, or their level of awareness of radicalisation prior to the training. The feedback is therefore more likely offering insight into the learning element of the training, rather than the validity of the instrument.

Reliability

Currently, reliability of any of the Radar protocols have not been tested. We are therefore unable to assess the reliability of the Risk Analysis protocol.

Authorship Effects

As no tests of validity or reliability have been conducted on Radar, no authorship effects have been detected.

Moderators

The FIRE documentation specifies that the theoretical and conceptual framework underpinning the Risk Analysis has been “specifically designed to better understand radicalisation... in Australia.” s. 47F(1) 2013, p. 12). The authors also note that “the indicators were developed within a liberal democratic society with a strong rule of law, non-authoritarian or corrupt policing and with a range of social services available to citizens” s. 47F(1) 2013, p. 12). The Radar user guide supports this, noting, “the analysis cannot be applied in the absence of such state structures as the factors and circumstances driving radicalisation are often vastly different” (Australian Government, 2017, p. 16). There is a caveat, however, that the Risk Analysis can be applied to “Australia and/or immediately comparable countries” s. 47F(1) 2013, p. 12).

This expansion to use outside of Australia is likely due to the data underlying the Risk Analysis. Although the authors do not present a comprehensive methodology which offers specific information regarding data collection or collation, they do offer the information that a proportion of the data was drawn from “other directly comparable countries” s. 47F(1) 2013, p. 10).⁸⁷ Closer scrutiny of this reveals that a large proportion of the citations within the Risk Screening and In-Depth Risk Analysis components are actually drawn from a range of Western countries. This limits the statements made by the authors of the documentation for Radar regarding its specific applicability to Australian cases.

The Radar user guide purports to be “ideologically neutral, meaning that it can be applied to individuals undergoing the radicalisation process regardless of the associated political or religious ideology” (Australian Government, 2017,

⁸⁷ No further details are presented.

p. 19). However, unlike the VERA-2R, there is very little detailed information regarding the specific ideologies underlying the data that were utilised to develop the protocol.^{88 89} However, more details can be inferred from examining the citations which underwrite the Risk Analysis protocol (Appendix 4). There is also an inference in the FIRE documentation that the protocol can also be applied to both individuals within and outside of social groups.

Despite the flexibility in ideological inclusion, the documentation dictates that only those over the age of 16 should be considered for the Radar process, with the additional caveat that users “applying these risk analysis tools to any intervention programme clients between the age of 16 and 18 exercise additional caution” (Australian Government, 2017, p. 16).⁹⁰

Implementation Burdens

Radar is designed for use by professionals from law enforcement, social services, health, multicultural and community relations and corrective services (Australian Government, 2017, p. iv).⁹¹ However, the FIRE documentation contradicts this slightly, noting that FIRE should be operationalised by “CVE personnel” s. 47F(1) 2013, p. 47). A large component of the Risk Analysis protocol is the involvement of a panel of individuals, so the Risk Screening and In-Depth Risk Analysis components require the input of multiple individuals, increasing the time required for an assessment of Risk, particularly as the Radar user guide notes: “Analysts within each Intervention Panel are encouraged to regularly communicate in order to recalibrate thresholds where necessary, and to ensure consistency of use” (Australian Government, 2017, p. 17).

Significant time is needed to administer Radar across the five protocols that a client may progress through (Intake and Managing Client Information, Initial Risk Screening,⁹² In-Depth Risk Analysis, Needs Analysis, and Mapping

⁸⁸ The authors of the FIRE document do note; “a diverse range of ideologies (including direct action environmentalists, left and right wing extremists, ethno nationalists and neojihadists” s. 47F(1) 2013, p. 11).

⁸⁹ It is also noted that the literature review within the FIRE document only has sections covering “Religiously/Politically Justified Violence... Issue Oriented Violence... Ethno-Nationalist Violence” s. 47F(1) 2013, p. 12-13). There is no literature covering any other form of ideology.

⁹⁰ This additional specification is based on statements regarding (uncited) evidence concerning Neo-jihadists only.

⁹¹ The TRIM documentation also infers that family and friends may be assisted by the indicators within the Risk Screening and In-Depth Risk Analysis.

⁹² And the accompanying “de-confliction process” (Australian Government, 2017, p. 17).

Client Progress and Discharge). This is partly due to the amount of information gathering and triangulation that is required to be conducted as, “it is recommended that information is sought from a number of individuals with reliable knowledge about the client... reducing potential bias and the scope for misinterpretation” (Australian Government, 2017, p. 41). Much like the VERA-2R, the Radar instrument warns against missing information, noting that “Overall judgements can still be made in the absence of information regarding one or more specific indicators, but only with appropriate caution. If many of the risk indicator cells have missing information, the risk analysis may need to be suspended pending the additional acquisition of reports or knowledge” (Australian Government, 2017, p. 41).

The Radar user guide is also accompanied by a participant training workbook containing 12 modules for completion. It is unclear whether it is a requirement that members of the intervention panel complete the workbook before conducting any analysis and whether there is an accreditation process involved.

The Risk Analysis components are completed by a nominated CVE/CT analyst.⁹³ The user guide states that, “it is recommended that a nominated analyst perform the analysis to increase consistency” (Australian Government 2017, p. 41). Analysts must complete the Initial Risk Screening (15 risk factors and three protective factors) before determining whether an individual is suitable for the In-Depth Risk Analysis (27 factors). The subsequent Needs Analysis Worksheet is intensive with 38 factors, most of which require in-depth knowledge of the client. Once the analyst has conducted the assessment, their findings are presented to an ‘intervention panel’, made up of stakeholders from the abovementioned professions, who make a final decision on the basis of their “professional judgement” (Australian Government, 2017, p. 41). The decision to be made depends on where along the CVE intervention process the client is (for example, Initial Risk Screening or In-Depth Risk Analysis stage). With regards to the completion of the Needs Analysis protocol, the user guide states, “it is strongly recommended that the person or persons conducting the analysis be skilled in interacting with members of the public in sensitive situations. Ideally this individual will also be familiar with the CVE context as well” and “at the very least they should be familiar with and have experience in conducting psychosocial assessments” (Australian Government, 2017, p. 43). Once the Needs Analysis and accompanying Radar Chart have been completed the “Intervention Panel officer”

⁹³ There is no specific mention that this analyst is required to undergo training in Radar prior to use.

responsible coordinates the identification of provisional intervention goals. The user guide recommends that goals be provisionally formulated by the Intervention Panel but that the responsibility for the ongoing management and reformulation of goals ultimately rests with the community-based programme that accepts the referral (Australian Government, 2017, p. 67).

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Implications

The outcomes of this task have highlighted that both the VERA-2R and Radar require further theoretical and empirical assessment. The determination of validity, reliability, and equity cannot be completed by scrutinising the purported outcomes of such tests alone. The outcomes for such measures of the VERA-2R are potentially contaminated by authorship effects, and Radar has yet to be assessed. The complexity of both instruments (more so Radar) meant that the methodology for assessment in this task was expanded on two occasions; to systematically search all known external literature to provide a more holistic assessment of each instrument; and to include an analysis of the empirical value of the cited evidence base.

The analyses demonstrated that both instruments have significant weaknesses across the assessment criteria. Much of these weaknesses span from the evidence base that underwrites the instruments. This has significant implications for the validity, reliability, and equity of the instruments.

Both instruments purport to follow the SPJ approach to risk assessment. However, when the protocols were scrutinised, this does not appear to be the case. SPJ instruments follow a strict six-stage format that includes a structured process for gathering case information, an evaluation of the relevance of factors, scenario planning, risk mitigation scenarios, and the use of all of these elements to formulate a final risk decision. These elements of the SPJ structure are not present in the VERA-2R or Radar. What the VERA-2R and Radar present is what has been termed in the literature as the SPJ 'lite' approach (Logan & Lloyd, 2019). In SPJ lite instruments, there is no detailed and supportive process for the full assessment approach, and due to the required timelines of assessments, no requirement for scenario planning and formulation. SPJ lite instruments are advantageous in some areas, such as the terrorism domain, as they are less time and resource consuming, and can be conducted by those who are not expert risk assessors (McEwan, Bateson, & Strand, 2017). However, due to their use by non-experts, a fundamental requirement of SPJ lite instruments is a strong evidence base for factor inclusion. This evidence base can provide the non-expert users with the most appropriate and valid information that will help guide their decision-making process. Guidance which is based on non-empirical evidence is likely to lead to errors in the identification of truly at-risk individuals. However, the above analyses have highlighted that the evidence bases underwriting both instruments are less concerned with scientifically validated information, as opposed to the opinion of the authors.

The professed risk specifications for the instruments differ significantly, however, both instruments suffer from a lack of clarity when information regarding the precise details of the specifications are sought. The VERA-2R cites two distinct specifications, and Radar, seven. This lack of clarity was confirmed when the research team attempted to thematically categorise the risk factors within each instrument. The findings highlighted that the factors within the VERA-2R are actually more suited to identifying those at risk of radicalisation (Radar's specification), and those within Radar's Risk Analysis protocol are better placed to identify individuals who are radicalised and at risk of committing violence (the VERA-2R's risk specification).

The relative (and in the case of Radar, complete) lack of static and dynamic risk factors across both instruments is surprising given the salience of such factors across general violence risk assessment instruments, and the empirical literature base in studies examining radicalisation and violent extremism. Typically, factors included within 'Beliefs, Attitudes and Behaviours' and 'Social Context and Capability' are a response to issues caused by the presence of static and dynamic factors. However, in these instruments (and in particular, Radar), greater prominence is afforded to the former categories with a lack of information or interest as to how these issues develop. These issues are likely driven by the lack of theoretical mechanisms of change, and the poor empirical evidence base underwriting both instruments. For the purposes of instrument development and improvement, there is a necessity to also scrutinise the wider non-terrorism literature on which interventions impact on static/dynamic risk factors.

Alongside the requirement for a strong empirical basis, a second fundamental requirement of the SPJ approach is the necessity of a theoretical underpinning which has been verified to demonstrate changes in perceptions and behaviours in the movement towards radicalisation and violent extremism. This analysis has demonstrated that the VERA-2R has, at best, weak theoretical support, as the theoretical work that it is cited has not yet been empirically validated. The outcomes regarding Radar paint a weaker picture still. Though the documentation underpinning the Risk Analysis protocol in Radar does touch on broad theoretical categories or disciplines, there is no evidence that these have been attended to in the citations that underwrite the development of the instrument.

Lastly, the analysis has highlighted that the VERA-2R has undergone more scrutiny in terms of validity and reliability. The authors cite numerous investigations that demonstrate a wide range of forms of validity and reliability. There are some concerns over some of the purported types of validity, as many examples are either not recognised forms of validity, or are miscited and incorrect. A second concern with these outcomes is the number of outcomes that

have no cited source and are only described in the VERA-2R documentation. Only Beardsley and Beech (2013) are cited as external investigators. Therefore, due to the concerns over authorship effects, we are only able to conclude with any certainty that the VERA-2R has demonstrated inter-rater reliability and face validity.⁹⁴

From the above analyses we can tentatively conclude that authors for both instruments should enact far more thorough evaluations of the wider literature to help develop better risk factors that match the stated risk specifications, which would help support the validity, reliability, and equity of the instruments. Therefore, the next task in this research seeks to expand on our understanding of the empirical evidence base that could be used to underpin the instruments. We then use this evidence base to further scrutinise the strength of the instruments

⁹⁴ Although the size of the sample in the Beardsley and Beech (2013) is concerning (N = 5)

Task 2: Synthesising the Existing Evidence Base

Introduction

This task employs a range of quantitative and qualitative methods in searching for and synthesising the empirical evidence base for causes of radical and extremist behaviour. This task has been conducted over 18 months of the project timeline. We have employed principles of information science (systematic searches, backward, forward searches; Rubin, 2017) and meticulous coding strategies to systematically source, collate, and synthesise all currently known empirical data examining the causes of terrorist behaviour. The results of this work also aid in the further assessment of the strength and quality of the factors within the risk assessment instruments under scrutiny. Overall, the results of this task; (1) organise the synthesis of what empirical evidence is available, (2) aid the identification of gaps in the evidence, and (3) supplement the findings of the thematic analysis of the instruments conducted in task 1. Key findings are summarised in graphical, qualitative, and quantitative form, and implications for the risk assessment instruments under scrutiny are discussed.

Rationale

The rise in the development and use of actuarial and SPJ risk assessment instruments as part of counter terrorism policy and practice has been facilitated by an increase in the empirical evidence base. In actuarial instruments, risk prediction is based on the statistical significance of specific factors. In SPJ instruments, overall predictions are founded on the factors, which are included based on their empirical validity. Therefore, to ensure validity and reliability, the instruments under scrutiny should be based on valid empirical evidence.

A wide range of empirical studies investigating the nature and determinants of radicalisation, and terrorist-related behaviour have been conducted. These, in turn, have influenced theory, policy and practice in areas concerned with prevention, disruption, and management. Some studies regularly cite the ambiguities and seemingly contrasting findings regularly uncovered within various empirical studies (e.g., Gill et al., 2014; Horgan, 2004; Victoroff, 2005). However, such differences may be a by-product of misunderstandings, methodological approaches, sampling, or interpretation. These differences are also evident by examining the wide range of risk factors that are included across risk assessment instruments in this area. This incentivises a need for a rigorous synthesis of the existing evidence base.

We have used a two-staged process to complete this synthesis and determine the strength of the empirical evidence base underlying the risk assessment instruments under scrutiny. The first stage was undertaken in task 1, and involved investigating the strength and quality of the cited theoretical and empirical evidence base for each included factor. The second stage was carried out in this task. This task involved a systematic literature review (SR) of the wider empirical evidence base related to causes of radicalisation and terrorist behaviour. This SR provides a comprehensive empirical foundation to use as a baseline for further assessing the instruments for the strength and quality of the evidence for each included factor.

SRs have emerged as an important means of synthesising research findings. Whilst literature reviews can be conducted relatively quickly,⁹⁵ they are subject to considerable bias, likely to be incomplete, and do not require a formal process of rating the evidence on which they are based. In contrast, SRs are substantial pieces of research requiring the use of reproducible, comprehensive literature searches (the search terms, inclusion criteria and methods used are proposed a priori in an independently reviewed protocol) and formal synthesis methods.

The objective of this SR is to assess the influence of multiple factors upon individuals' attitudes, intentions, and behaviours in the context of radicalisation and violent extremism; synthesising the existing evidence base. The SR specifically investigates the existing empirical literature on causes of radicalisation and terrorist behaviour, including its coverage, common themes, methodological strengths and weaknesses, and implications. The purpose of this is threefold: First the results will aid the identification of gaps in the empirical evidence base; second, the factors included in the instruments under scrutiny will be compared with the results, and, third the analysis will draw out key findings that can be used to inform operational activity and strengthen the development of terrorism risk assessment instruments.

The below sections first provide a detailed summary of the SR protocol. Following this, we provide a quantitative overview and in-depth qualitative analysis of the results of the SR conducted by the research team. Then we present

⁹⁵ Which both the VERA-2R and Radar cite as a component of the methodology for factor development.

the results of a comparative analysis between the SR and the outcomes of the thematic evaluation of the evidence base for the instruments from task 1.

Protocol

Secondary Systematic Review

The initial SR undertaken by the London team (which was originally anticipated to be the entire SR for this research) identified 191 studies of contributory causes of radicalisation and terrorist behaviour. On scrutiny of these studies, both the London and ANU research teams noted that there was a paucity of known empirical studies taken forward for review. Second to this, a significant proportion of studies that were included for final review were identified during the citation search process (120 out of 191 studies). This discrepancy implied that the databases that were utilised did not hold a large proportion of literature which investigates the criteria under scrutiny. Therefore, a secondary SR was undertaken by the ANU team. This protocol matched that of the London team, but examined a different set of databases. The rationale here was to expand the range of empirical studies that would be considered for review, and thus strengthen the findings of this research.

Figure 1 highlights the overall SR process⁹⁶ and the resulting output. However, for completeness, the following sections detail the full protocol of the SR, including details of both the primary and secondary stages.

Identification Stage

Databases and Information Sources

Following a similar protocol to that undertaken in task 1, the search strategy for the SR was based on the Campbell Collaboration method. Studies were identified using the following search methods:

- A keyword search of multiple electronic databases.

⁹⁶ Carried out by both the London and ANU research teams.

- PsychINFO, ProQuest Central Criminology Collection, ProQuest Central Social Science Database (Primary).
- International Bibliography of Social Sciences, Sociological Abstracts, and Scopus (Secondary).
- Forward and backward citation searches of all eligible candidate studies.

Full text versions of identified studies were obtained through (in order of preference):

- Electronic copies via the university's e-journal service.
- Electronic copies of studies available elsewhere online.
- Paper copies.
- Electronic/paper copies requested through the University's inter-library loan system.
- Electronic/paper copies requested from the authors themselves.

In cases where the full text versions of the works collated contained insufficient information to determine their eligibility for inclusion according to the coding strategy (described below), where possible the corresponding author was contacted in an attempt to retrieve this information.

More generally, the review considered published and unpublished (grey) studies. No date restrictions were applied.

Studies however had to be available in English, French or German since available resources limited the research team's ability to search and translate studies in other languages.

Search Terms

In order to identify the relevant studies for the review, a number of search terms were used in the above databases (Table 12). These include terms relevant to radicalisation, extremism, terrorism, and causation.

Table 12 Search terms used during SR

Terrorism/Radicalisation	Causation		
	Factor	Risk	Pathway
Terroris*			
Insurgen*	Mechanism	Vulnerability	Process
Rebel	Caus*	Context	Profile
Radicalisation	Motive	Stressor	Indicator
Radical	Motivat*	Behaviour	Predictor
Extremis*	Determinant	Behavior	Reward
Militant	Propensity	Influence	Attitude
	Trigger	Personality	Root
	Antecedent	Opportunity	Explanation
	Susceptib*	Reward	

Selection Criteria

The selection of appropriate studies was conducted in a number of stages. The first stage involved the research teams screening all identified studies (45,217) based on their title and abstract. Studies were screened against the following criteria:

- The study must report an explicit goal of understanding the determinants of radicalisation or behaviour associated with a terrorist offence.
- The study must report at least one measure in a quantitative or qualitative sense. Outcome data can comprise official measures (such as police recorded data) or unofficial measures (such as self-reported experiences). These measures could relate to causal mechanisms activated in the context of radicalisation, substantive information relating to the environmental conditions that impact upon radicalisation, or substantive information relating to the offender that impact upon radicalisation.

Studies failing to meet the inclusion criteria for the full review were excluded with reason for exclusion and rates of attrition noted. At this stage, 833 studies were deemed appropriate for inclusion based on title and abstract.

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

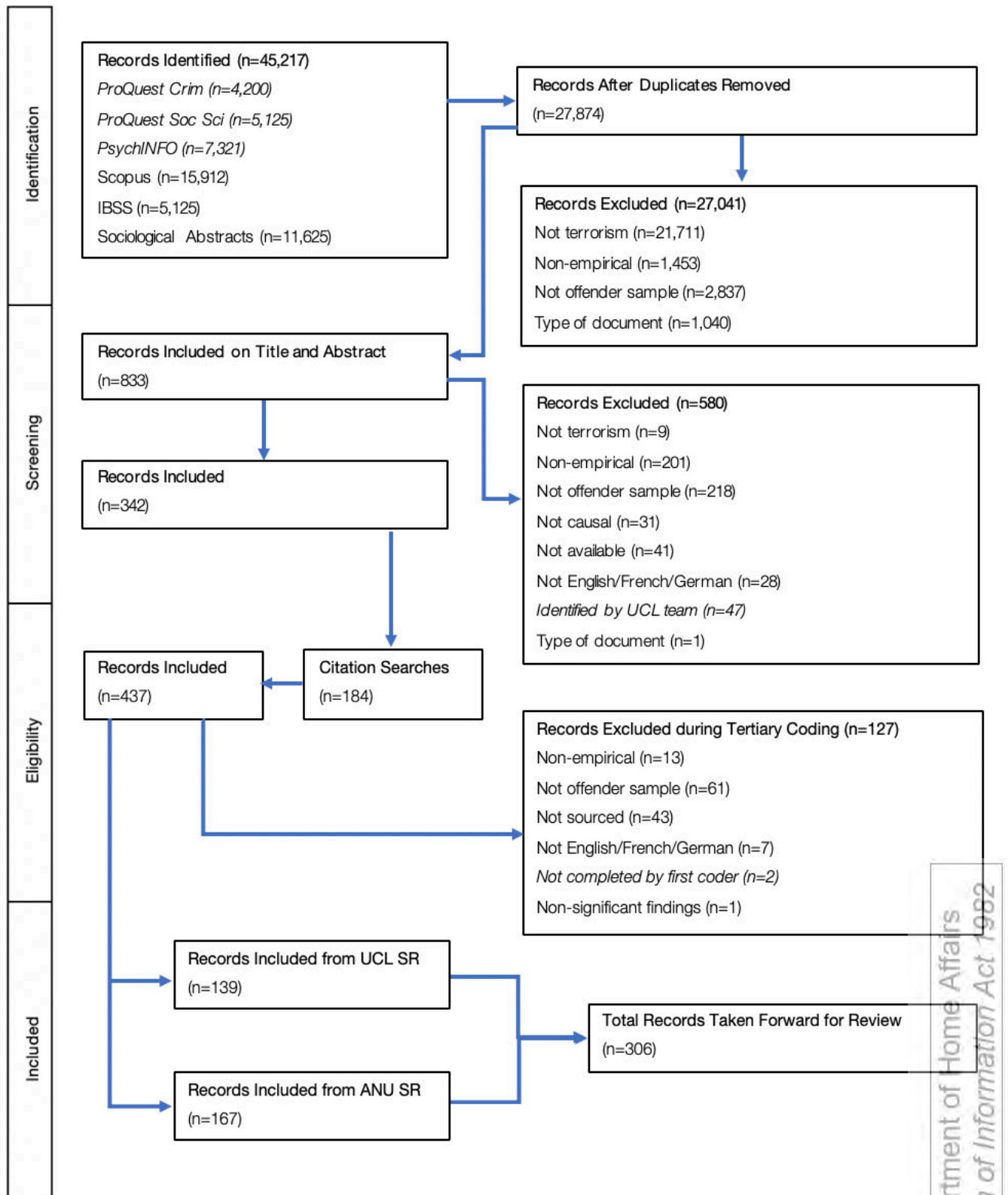


Figure 2 Full SR process.

NB, UCL-related processes are italicised for emphasis

Screening Stage

During the screening stage, all 833 studies carried forward were read in their entirety to determine their eligibility using the same inclusion and exclusion criteria as above. A further 580 studies were excluded from the final analysis. During this stage, each study was also used to conduct backwards and forwards citation searches to identify further candidate studies. This process involved first reviewing titles of cited studies and also subsequent citations that each candidate study accrued up to July 2019. Each appropriate title was then examined and judged based on the previously mentioned selection criteria. For each study identified in the backwards and forwards searches, additional searches were conducted until all citations had been fully identified. As depicted in Figure 1, 437 studies were brought forward for final review. This included 184 studies identified through the backwards and forwards citation searches

Eligibility Stage

Study Coding

The coding protocol for the SR required an in-depth critical examination of each of the 437 studies captured during the eligibility phase. This involved two independent coders reading each of the included studies in their totality, extracting information on the source of the data, sample size, participants, and variables of interest. Variables of interest included those indicated by authors of the studies as significantly related to radicalisation and violent extremism. For studies employing a quantitative methodology, significance of variables was determined by examination of the significance values and coefficients of each variable within the models presented in the study.⁹⁷ This was a straightforward method of determining which variables to include in the SR. For studies employing qualitative methods (for example participant observation, case studies and small n interviews) variables were selected for inclusion based on a reading of the authors' analyses and argument. This was a more complicated way of determining significance as the nature of qualitative results is also influenced by the reader's interpretation.

⁹⁷ Studies that did not measure significance, but reported other outcomes, such as effect size were assessed using the discretion of the coder. This predominately relied on assessing the core value of the statistics measured during the study and following appropriate guidelines regarding individual statistical tests.

During this process, each coder also highlighted studies that were deemed inappropriate for inclusion in the SR if it became apparent that they did not match the criteria for the synthesis component. Excluded studies were flagged as inappropriate for several reasons. At the end of this process, the two coders came together to discuss the studies that each coder had highlighted for exclusion. Where the coders could not agree on exclusion, these studies were sent to a tertiary coder for review and final decision on exclusion.⁹⁸ A further 127 studies were excluded as a result of this process. This left 306 (167 from the ANU team and 139 from the UCL team) studies taken forward for final review.

During the coding discussion, the coders also jointly critically re-assessed each of the included studies to ensure consistency across the terminology of variables of each study.⁹⁹ This was predominately due to the proportion of qualitative studies included for assessment. During this process, all variables that were identified by both coders were carried forward for analysis, and where there were inconsistencies in variable identification, both coders interrogated each study to reconcile differences in variable inclusion.

Analysis

Thematic Analysis

All 306 studies were then analysed using Nvivo (QSR International, nd.). Nvivo is a qualitative data analysis software program used by practitioners and researchers engaged in qualitative and mixed methods research. Nvivo is able to sort, categorise and analyse unstructured text data providing meaningful qualitative insights. Nvivo was used by the research team to identify the most frequently cited words, variables, and themes in the data.

Comparative Analysis

Once the research team had sorted the qualitative data using Nvivo, all identified variables were scanned for completeness, duplicates were removed (but noted), and the final list of variables (and their accompanying studies)

⁹⁸ Within the London team.

⁹⁹ For example, the reviewer identified all use of the term 'radical peers', 'radical friends' and 'social bonds' and, after checking the source document to ensure accuracy in the meaning of the term, changed this to 'social networks' for greater consistency across the variables.

was thematically compared to the factors within the instruments. This process involved both coders independently sorting through the variables to assign all those deemed thematically appropriate to each factor in the instruments.

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Results

Study Characteristics

Table 13 highlights the descriptive characteristics of the studies within the final sample for review. A significant majority of research conducted on factors associated with either the process of radicalisation towards violent extremism, or behaviour associated with a terrorist act, has been published. The predominant outlet for publication is peer reviewed academic journal articles (79.1%). The remainder of the studies have been published as reports (4.2%), books (6.9%), or book chapters (3.3%). Only 6.5% of studies are classified as unpublished or grey literature, suggesting that research examining casual factors of radicalisation and terrorist behaviour is relatively accessible to researchers and practitioners.

As mentioned above, a criterion for study selection was that the study be empirical, requiring a sample population as the unit of inquiry. More than half (63.4%) of the studies analysed terrorists as their sample population. The next most popular sample choice was general populations (26.1%), including students and residents of a particular country, state or region. The remaining sample populations included mixed samples of terrorists with general populations (5.9%), those who had radicalised but remained non-violent (2.6%) and family members of terrorists (2.0%). Sample sizes varied between 1 and over 16000 with a majority of studies (52.6%) employing large N samples of 100 or above. The most common data collection techniques employed were interviews (29.4%), the collection of open source information (27.8%) and surveys or questionnaires (27.8%). Only four studies utilised closed source or classified data.¹⁰⁰

Finally, the SR examined the ideological focus of each of the studies. A wide range of terrorist groups were assessed in the literature as well specific ideologies. Some studies were less specific referring only to 'right wing' or 'Islamic'. To ensure consistency and comparability, therefore, this section of the analysis focussed solely on the general ideology referred to in the study. More than half (53%) of all studies examined individuals espousing a Jihadist ideology. Almost a fifth (19%) of studies focused on individuals holding right wing ideologies, including proponents

¹⁰⁰ For example, Schuurman and Eijkman (2015) performed a qualitative analysis of police files, court documents, and interviews. The direct results of this analysis were not published, but the authors were able to report the themes that were identified in the data.

of white supremacy, neo-Nazism and anti-immigration beliefs. Ethno-national terrorism, including individuals involved in political violence in Northern Ireland and the Basque region, received coverage in a small proportion (11%) of studies. Left wing ideologies, including socialism, communism, and anti-Vietnam war extremism were examined in only 7% of studies. Finally, individuals espousing single issue ideologies, such as anti-abortion extremists and animal rights extremists, received 4% of the coverage.

Table 13 Characteristics of studies analysed

Characteristics of Studies (N = 306)									
Publication Type	(N)	Sample Size	(N)	Population Sample	(N)	Ideologies Examined ^a	(N)	Data Collection Technique	(N)
Article	(242)	500 +	(89)	Terrorists	(194)	Jihadist	(214)	Interviews	(90)
Book	(21)	100 – 500	(72)	General Population	(80)	Right Wing	(77)	Open Source	(85)
Dissertation	(15)	50 -100	(56)	Mixed Sample	(18)	Ethno-Nationalist	(43)	Surveys/ Questionnaires	(85)
Report	(13)	10 – 50	(36)	Radicalised (non-violent)	(8)	Left Wing	(27)	Existing Dataset	(21)
Book Chapter	(10)	1 - 10	(32)	Family Members	(6)	Single Issue	(18)	Case Studies	(13)
Working/ Conference Paper	(5)					Other	(26)	Ethnographic/ Fieldwork	(4)
								Experiment(s)	(4)
								Mixed Source	(4)

^a A number of studies examined multiple ideologies. Therefore, the total N reported for this element is 405

Thematic frequencies

An effective and user-friendly method of visually representing results of qualitative data analysis is through the use of word clouds. Word clouds display the frequencies of text data, allowing for easy interpretation of the broader themes within the text. Figure 2 presents the top 40 most frequently cited empirical findings within the 306 studies in the SR. The colours follow the red, amber, green scale that has been used in Task 1. Green depicts the most

[illegible]

Thematic Evaluation

There is growing consensus among terrorism researchers that individuals who carry out terrorist acts do not fit a single profile (Borum, 2011; Gill et al., 2014; Horgan, 2008). The findings of this SR further validate this argument, as the outcomes of the analysed studies were marked by their diversity rather than their homogeneity. The findings highlight that there are numerous factors that contribute to pathways into terrorism. This is evident by the sheer number of variables deemed to be significantly associated with radicalisation and terrorist behaviour that were identified in the review (N = 1,532). It is therefore not practical to thematically assess all variables that have been

¹⁰¹ For example, the word 'social' (depicted in green) was verified empirically across 180 studies, compared to 'religion' (depicted in amber) which was verified in 49 studies, and 'poverty' (depicted in red), which was only empirically verified 17 times.

identified as associated with radicalisation or terrorist behaviour. Table 14 highlights the most frequently identified empirically verified variables, which make up 33.2% of all identified variables. Of these, only 3 variables were empirically verified in more than 60 studies; social networks, younger age, and being male. The frequency of empirical support for all other variables declines rapidly after these three variables as demonstrated by the amber and red categories.

Table 14 Top Variables Identified as Significantly Associated to Radicalisation or Terrorism

Empirical Evidence	Variable	Number of Verifications	Percent of all Identified Variables
Strong empirical evidence	Social networks	131	8.5
	Age (younger)	68	4.4
	Gender (male)	66	4.3
Moderate empirical evidence	Criminal history	43	2.8
	Religion/religious beliefs	42	2.7
	Education (higher)	32	2.1
More research needed	Unemployment	27	1.8
	Ideological beliefs	22	1.4
	Trigger event	21	1.4
	Poverty	17	1.1
	Perceived injustice	17	1.1
	Military experience	16	1.0

Social Networks

An individual's social networks, either online or offline, stand out as being the most commonly cited variable found to be significantly associated with radicalisation and terrorist behaviour. This variable was found to be significant across the ideological spectrum. Within the majority of studies examining the role of social networks in radicalisation and terrorism, an individual's friendship network, as opposed to other social relationships, is of high importance. The work of Sageman (2004, 2008), for example, concluded that a growing proportion of violent extremist attacks have been perpetrated by individuals within small, informal non-hierarchical social networks. These radical social

networks developed out of 'cliques' of similar friends who, through the development of a shared collective identity, can "literally transform lives and...change the meaning and impact of friendship bonds that pave the way to joining the jihad" (Sageman, 2004, p. 154). Another highly influential study, on Middle Eastern terrorists, conducted by Post, Sprinzak & Denny (2003, p. 173) argued that peer groups are very influential in the recruitment process, finding that "...in many cases it was a friend or acquaintance in the group who recruited the subject."

More recent studies have expanded on the findings of these seminal pieces, identifying the explicit behavioural functions of the friendship networks. Ahmad (2016) examined the radicalisation of young male extremists in Pakistan. They noted that it was only after becoming socially embedded within networks inside the organisation (Islami Jamiat Talba) that these individuals underwent a change in personal perceptions, preferences and worldview. Another recent study, examining 99 German foreign fighters, Reynolds and Hafez (2019) argued that local networks among the radical milieu is a critical for driving foreign fighter recruitment. The majority of fighters were mobilised within a single interconnected network, and 71.7% had a close personal connection with at least one other German foreign fighter before departing Germany. Similarly, from interviews with twenty Jihadist foreign fighters, Dawson and Amarasingam (2017) identified that the process of radicalisation began in early adolescence and involved small groups of friends, who facilitated the process. Analyses of interview records of over 12,000 Revolutionary Armed Forces of Colombia (FARC) militants highlighted that personal connections between individuals and serving FARC members facilitated the radicalisation, mobilisation, and recruitment of potential recruits (Rosenau, Espach, Ortiz, & Herrera, 2014). And in interviews with 141 members of al-Shabaab, the role of friends in recruitment to the group was emphasised by the majority of respondents. These interviews revealed that they were recruited by their friends, and on entry to the group, would then go on to recruit other friends (Botha, 2014).

Another particular type of social network that is often hypothesised to be associated with radicalisation and terrorism is an individual's family connections. The importance of family ties was highlighted in several studies during the review. Robert White's oral history of the Provisional Irish Republican Army (PIRA) reveals that many Republicans who joined the PIRA grew up in households that espoused strong political beliefs, as stated by one of the interviewees: "I'm from a Republican background. My parents were members of the Republican movement in the 1950s and that's how I came to know what the movement was about" (White 1993, p. 39). Results from a survey examining separatist rebel group the Moro Islamic Liberation Front (MILF) on the island of Mindanao in the

Philippines, concluded that having a family member in the MILF was the main driving force for voluntary recruitment (Özerdem, Podder, & Quitariano, 2010). In an examination of 491 terrorists who engaged in over 30 groups across 13 conflicts, just over 30% were raised in a family that supported and engaged in terrorism. Of these individuals, 61.2% cited their familial connection to terrorism as a motivating factor for their own engagement (Jacques & Taylor, 2013).

Further examination of the precise role of family in radicalisation and terrorism shows that there are several different scenarios that may push an individual towards engagement, including tradition (Bosi, 2012), the death of family members (Florez-Morris, 2007) and the desire to avenge their death (Özerdem & Podder, 2012; Podder, 2011), pressures from family members to become part of the organisation (Amble & Meleagrou-Hitchens, 2014) and the financial rewards offered for the families of militants (Kimhi & Even, 2004).

Withdrawing from previous social networks in the process of establishing new ones was also cited in several studies as being a significant causal factor in radicalisation and terrorist behaviour. Analysing the radicalisation pathway of convicted Australian terrorist Jack Roche, Aly and Striegheer (2012) concluded that Roche withdrew from his 'mainstream' Muslim friends and deepened his relationship with an individual linked to the terrorist organisation Jemaah Islamiyah. In Roche's own words, "if one were to look at my life as a process of becoming radicalised" it was at this point when he withdrew from old social networks and created new ones that was the tipping point (Aly & Striegheer, 2012, p. 855). This process of withdrawing from one social network and deepening connection with another was also evidenced in a study of Chechen suicide bombers. Out of the 34 suicide bombers studies, 32 were found to have withdrawn from family members prior to increasing their visits to mosques where militant Wahhabi beliefs were being preached (Speckhard & Ahkmedova, 2006).

The role that online social networks play in radicalisation pathways has received increasing attention in the scientific literature. Schafer, Mullins and Box (2014, p. 192) undertook analyses of online activities and biographical and autobiographical accounts of white supremacists. They identified that peers were portrayed as more influential than other actors in helping drive a person to their 'awakening' to white supremacist ideology. Gill's (2015) analysis of 111 lone actor terrorists highlighted a growing trend amongst lone actors making use of virtual connections in their pre-attack behaviours. These individuals ventured online to reinforce their beliefs, seek legitimisation for their proposed actions, disseminate propaganda and attempt to recruit others. The use of the internet to develop social

networks and radical intentions is not distinct to lone actors, Gill, Corner, Thornton, and Conway (2015) analysed 227 cases of individuals who radicalised online, across both Islamic and right-wing ideologies. They identified that individuals use their social networks to both access resources and ideological content online. They also noted that individuals who do communicate with a radical social network online, are just as likely to also affiliate with this network offline. With the online space offering different functions for radicalisation and planning behaviours than the offline space. For example, terrorists may engage in face to face interactions which reinforce their ideological convictions while simultaneously communicating with online networks regarding the specifics of bomb-making. They conclude that the process of radicalisation should be seen as 'cyber enabled' rather than cyber-dependent (Gill, Corner, Thornton, & Conway, 2015, p. 37).

Age and Gender

Being young and male were also commonly cited as significant variables of radicalisation and terrorist behaviour across a diverse range of studies. In fact, only 12 studies highlighted older individuals (compared to 68 for younger individuals) and 5 highlighted females (compared to 66 for males) as more likely to become radical or commit a terrorist act. These findings are unsurprising, and conform to common belief. Bakker (2006, p. 36) notes that his research on Jihadi terrorism in Europe confirms "conventional wisdom" that terrorists are generally male, with 237 males and only five females in their sample. The results from a study of a random sample of adult Americans further support this, revealing that men were 1.76 times more likely than women to produce online hate material (Costello & Hawdon, 2018). Based on quantitative analysis from a sample of ETA (*Euskadi ta Askatasuna* or Basque Homeland and Freedom) militants, Reinares (2004) finds the overwhelming majority of members to be male and to have joined the organisation at a young age. In a study of deceased Hezbollah militants, Krueger and Malečková (2003, p. 132) found that "Hezbollah fighters tended to be in their late teens and early 20s when they died".

Interviews with 16 paramilitaries and former paramilitaries (Republican IRA and Loyalist UVF) in Northern Ireland revealed that majority of those interviewed who joined as young men, did so in response to experiencing a 'critical incident' (Ferguson, Burgess & Hollywood, 2008). This was a common finding across studies. Blazak (2001) concluded that terrorist recruiters take advantage of critical incidents, or "red flags of strain", as opportunities to intensify their recruitment activities. In a seven-year ethnographic study of skinhead violence, Blazak (2001) found that racist skinhead groups used flyers distributed at high schools and informal youth networks to seek out

'strained' individuals who might be easily recruited into the Aryan skinhead movement. Blazak also noted that the selection of schools as recruitment areas was based on two factors; the presence of racist skinheads in the school or younger siblings of older skinheads who could be used as contacts, and whether there were any perceived threats (strains) to young white male students that could be manipulated to increase the chances of recruiting these individuals. These findings also demonstrate the importance of social networks in terrorist recruitment drives.

Comparative Evaluation

The final stage of the SR was to compare the evidence base for the factors included in the VERA-2R and Radar's Risk Analysis protocol with the variables that have been empirically validated in the wider literature. In task 1, we presented a thematic evaluation of the strength of the evidence base underwriting the instruments. The following sections expand on this analysis, comparing the evidence base underwriting the instruments with the findings of the SR, highlighting both where the instruments correspond with the empirically verified variables and the areas of departure.

Instrument 1 – The Violent Extremism Risk Assessment – Version 2 Revised (VERA-2R)

As shown in task 1, the cited evidence base for the factors within the VERA-2R is weak. Only 48.1% of cited statements in the VERA-2R were coded as accurately reflecting what was recorded in the cited texts. Of the 51.9% of citations that did correspond to elements within the cited texts, only 48.5% of these (23.3% of the overall citations) were drawn from the empirical findings of the cited texts. Seven (15.5%) factors have no citations included in the justification for inclusion. Of the 343 citations included in the literature reviews across all factors, 203 (59.2%) are drawn from 57 published items. Further to this, the average cited empirical works across factor categories is 40.8%.

The SR identified 157 studies (50.8% of the total studies identified) that provide empirical evidence for the factors within the VERA-2R, across these studies, 220 variables were thematically identified to be directly attributable to the VERA-2R. This represents 14.2% of the variables across the evidence base identified in the SR. Table 15 presents the amount of available supporting empirical evidence for the factors across the instrument, the number of studies that identified this evidence, and the number of these studies that are cited by the authors of the VERA-

2R as justification for factor inclusion. As noted, only seven studies were accurately identified as relevant for the evidence base for the VERA-2R.

For each factor within the VERA-2R, there was an average of 8.7 variables identified. The SR revealed the highest empirical support was for 'BA2 – Perceived Grievances and Injustice', with 40 studies identifying 20 empirical variables that are associated with these behaviours that have been empirically validated as associated with radicalisation and terrorist behaviour. The second highest evidence base was found for 'HAC2 – Network of Family and Friends Involved in Violent Extremism', with 28 studies identifying 13 variables.¹⁰² All other factors had less than 20 studies supporting them. The factors 'CH2 - Non-Compliance with Conditions or Supervision' and 'MD4 - Autism Spectrum Disorder' were not supported by any empirical evidence that was drawn from the SR.

Table 15 Descriptive comparison of the SR evidence base with the VERA-2R

Factor Categories in the VERA-2R ^a	Number of Variables Identified in the SR	Number of Studies Identified in SR	Number of SR Studies Correctly Attributed to Factors in the VERA-2R
Beliefs, Attitudes, and Ideology	46	80	1
Social Context and Intention	40	50	1
History, Action, and Capability	41	70	1
Commitment and Motivation	47	69	2
Additional Indicators	46	73	2

^a NB – Protective/Risk Mitigating factors not examined here, as the VERA-2R does not provide any evidence for the justification of each factor.

Instrument 2 – Radar

Much like the VERA-2R, task 1 demonstrated that a large proportion of the cited evidence base for factor development in the Radar Risk Analysis protocol was not empirical (71.6%), and a significant proportion (62%) of the cited documents were drawn from media sources. However, unlike the VERA-2R, a large proportion of the citations in Radar were coded as accurately reflecting the cited texts. However, within these statements, only 15%

¹⁰² Note, as we followed the rationale given by the authors of the VERA-2R the thematically organise variables, 'social networks' is not included, as this term is not specified in the VERA-2R documentation.

were drawn from the empirical findings of the cited texts. What was particularly concerning was that, in Radar, 30.5% of all statements that underpinned the factors were not attributed to any evidence other than the opinion of the authors of the Radar documentation, and the “Action Orientation” category had no empirical work underpinning its justification for inclusion.

The SR identified 55 studies (17.8% of the total studies identified) that provide empirical evidence for the factors within Radar’s Risk Analysis protocol. Across these studies, 91 variables were thematically identified to be directly attributable to Radar. This represents 5.9% of the variables across the evidence base identified in the SR. Table 16 compares the amount of available supporting empirical evidence for the factors across the protocol, the number of studies that identified this evidence, and the number of these studies that are cited by the authors of the Radar documentation as justification for factor inclusion. As noted, only one study was accurately identified as relevant for the evidence base for the Radar Risk Analysis protocol.

For each factor within the Risk Analysis protocol, there was an average of 3.8 variables identified. The SR revealed the highest empirical support was for s. 33(a)(i), s. 37(2)(b), s. 47E(d)

with 10 studies identifying 8 empirical variables associated with these behaviours that have been empirically validated as associated with radicalisation and terrorist behaviour. No other factors were supported by 10 more studies. The factors s. 33(a)(i), s. 37(2)(b), s. 47E(d)

were not supported by any empirical evidence that was drawn from the SR. This is likely due to these factors being indicators of someone already being radicalised or embedded as a terrorist, and not risk factors for such behaviours.

Table 16 Descriptive comparison of the SR evidence base within Radar's Risk Analysis protocol

Factor Categories in Radar Risk Analysis Protocol ^a	Number of Variables Identified in the SR	Number of Studies Identified in SR	Number of SR Studies Correctly Attributed to Factors in Radar
Social Relations	20	29	0
Ideology	53	59	1

Action Orientation	18	12	0
--------------------	----	----	---

^a NB – The three protective factors in the Risk Screening are not examined here, as Radar documentation does not provide any evidence for their justification.

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Implications

The objective of this SR was to synthesise the existing knowledge and evidence base of all relevant literature examining the influence of various factors upon individuals' attitudes, intentions and behaviours in the context of radicalisation and terrorism. This involved utilising a range of methods in searching for and synthesising the wide evidence base, drawing from the principles of information science (for example, systematic searches, backward, forward searches.) and employing meticulous coding strategies. The search of the published and unpublished literature across six different databases identified 306 empirical studies produced between 1952 and 2019.

The brief qualitative synthesis of the studies provide evidence for the range of empirical research that should be taken forward as evidence of factors to be included within risk assessment instruments. The vast number of findings that were identified during the SR (1532 variables across 306 studies) mean that it is not feasible to report all findings in this report. However, the results of the SR do provide an empirically rigorous base for the examination of radical and terrorist behaviour, as well as the ability to compare this base with the instruments under scrutiny.

A few observations from the SR are worth reinforcing here. First, there are multiple pathways into radicalisation and terrorism as demonstrated by the sheer number of variables identified in the studies that were reviewed. However, the SR did reveal key themes regarding social networks, gender and age which should help to inform the development of future risk assessment instruments. None of these themes should be surprising. The wider criminological literature is consistent in its agreement that social networks, age, and gender are all key elements that lead to delinquent and criminological behaviour. What is surprising is the lack of attention to these themes across the instruments under scrutiny. Although both the VERA-2R and Radar do include categories that include social elements, none of the elements within are linked to the social networks of the individuals. In fact, the risk factors are more closely aligned to beliefs and radical behaviours.¹⁰³

Second, the SR revealed that the empirical literature contains very rich data, with 63% of studies analysing terrorist populations and a majority of studies employing samples of 100 or more. This replicates the work of Schuurman (2018), who performed a systematic review of journal articles in this field. Thirdly, the findings of the SR show that

¹⁰³ Which goes some way to explain the differences in the thematic coding of the factors in Radar.

the preponderance of the scientific literature on individuals' pathways to radicalisation and terrorism has focused on those espousing a Jihadist ideology. This has implications for risk assessment instruments that purport to be ideologically neutral but which draw from an empirical literature which is skewed towards a particular ideology.

The findings of the comparative analysis supplement the findings of the thematic analysis in task 1. In task 1 it was identified that the empirical evidence base for the factors within the instruments was at best moderate. However, this task has revealed that this conclusion should be reassessed to poor. The studies identified in the SR span 67 years, and although a proportion of those were published after the development of the instruments, it is concerning that neither instrument cites the majority of these studies as evidence for inclusion of the factors within. In the case of the VERA-2R, there is a reasonable proportion of studies that support the risk factors, and yet these are not included in the justifications for the risk factors. This suggests that the literature review that underpins the instrument was not particularly rigorous, and this lack of rigour has the potential to impact the validity, reliability, and equity of the instrument.

In the case of the Radar Risk Analysis protocol, different concerns have been brought to the fore. There is a distinct lack of empirical evidence which could possibly underpin the risk factors. Particularly with regards to the factors in the Action Orientation category. This is concerning on two fronts: Firstly, in order to qualify as a recognised SPJ risk assessment instrument, the inclusion of an empirical evidence base is paramount. Secondly, without an empirical justification for inclusion, it is not possible to accurately determine whether any outcomes regarding predictive validity or reliability are due to anything other than chance. Therefore, any outcomes from validity analysis on this instrument will have to be evaluated alongside the qualitative responses of participants with regard to the usefulness of the protocol in the determination of risk.

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Task 3: Developing the Case Test Library

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Introduction

This task is a methodological task that was conducted to provide a foundation for instrument testing in task 4. Work within this task spanned six months of the project timeline, and involved the research team building a case library of 60 anonymised vignettes which included true positives, true negatives, false positives, and false negatives. The vignettes are separated into eight¹⁰⁴ different categories, and are based on the life histories of actual individuals, and range from radical to non-radical, violent to non-violent. The case vignettes were drawn from Australia, Europe, and Canada from both open and closed sources.

Case Vignettes

Rationale

The purpose of this research project is to demonstrate which risk assessment instruments accurately classify those at high risk of movements towards terrorism and which one's overestimate or underestimate risk. A fundamental stage in this process is to subject the instruments to the process for which they were developed; evaluation of cases. In order to achieve this, it was necessary to develop a series of cases that the research team were able to utilise during the testing phase of the project.

Case Categories

The case vignettes are comprised of eight different categories of individuals. The categories have been designed by the research team to draw a representative sample from the potential offender populations, in order to test the predictive validity of the VERA-2R and Radar. The categories differentiate between political and non-political, violent and non-violent, and lone and group-based actors. The cases include a full range of motivating political ideologies (religious, nationalist/separatist, left-wing, right-wing, and single-issue), and also personally aggrieved individuals. The categorisation of cases covers true positives (actual violent extremists), true negatives (non-violent non-extremists), false positives (non-violent radicals (VERA-2R), non-violent political (non-radical) activists (Radar),

¹⁰⁴ Following a teleconference held on 1 August 2019 facilitated by the Department, a request was made for the inclusion of an eighth category to capture individuals deemed to be 'at risk' of radicalisation.

and false negatives.¹⁰⁵ The case categories have also been designed so that it is possible to follow the movements individuals take on the pathways towards committing extremist violence (apolitical, political interest, at risk, radical, involvement in non-violence, involvement in violence). This will allow for the assessment of the risk specifications of both the VERA-2R and Radar. The case categories are as follows:

- Violent Extremist
 - Includes individuals who have become radicalised, identify with a terrorist group, and who have committed an act of violence.
 - Individuals in this category will have committed their violence together with other individuals.
- Non-Violent Extremist
 - Includes any individuals who have radicalised, and who identify with a terrorist group, but have not yet committed an act of violence.
 - Individuals in this category may have contemplated violence that was not carried out, or may have facilitated others to commit violence.
- Violent Group Member Non-Political
 - Includes any individuals who have committed violent acts, but are not driven by political or religious motives.
 - These individuals include gang or organised crime members.
- Non-Violent Group Member Political
 - Includes any individuals who have clear political motives and who actively represent their political agenda, but do not act violently.
 - May include members of activist organisations.
- Lone Extremist Violent
 - Includes individuals who have been radicalised and committed a violent act, but who have no clear connection to a terrorist group.
- Lone Extremist Non-Violent

¹⁰⁵ False negatives are defined as error in which a result improperly indicates no presence (the result is *negative*), when in reality it is present.

- Includes individuals who have been radicalised, but have not yet committed acts of violence, and who have no clear connection to a terrorist group.
- Individuals may have contemplated, or openly expressed a wish, to act violently, but have not engaged in these plans.
- At Risk
 - Includes individuals who present as at risk of radicalisation.
 - These individuals have not yet committed an act of violence for the furtherance of an ideological cause, and they may or may not hold associations with a radical group.
- Control Group (Neutral)
 - Includes individuals who fit within social norms.
 - These individuals are not radicalised, have no strong political motive, are not associated with a radical group, and have not committed any violent acts.

Released by Department of Home Affairs
under the Freedom of Information Act 1982



Protocol

Case Identification

The research team performed a comprehensive search of legal and non-legal databases (Australasian Legal Information Institute (AustLII), Google News, Lexis Advance, and Westlaw) to source all known cases in Australia which have involved acts that are either non-politically violent, politically violent or non-violent which fit within the case categories used in this research. The sourcing of the cases employed a structured and systematic protocol, and involved the rating of source reliability to determine the strength of the information drawn. Table 17 highlights this source reliability rating scale, which has been employed by the primary researcher across a number of the research projects they have conducted.

During this coding process, court transcripts and associated documents were deemed most reliable, as these documents recorded finalisations of judicial decisions. Competency evaluations, sworn affidavits and indictments were deemed reliable, as these were carried out post arrest and prior to trial, when initial investigations had been made. Manifestos were deemed somewhat reliable, as the individual may not have been honest (or have insight; Erickson & Erickson, 2008). Warrants and Expert Witness reports were also reasoned to be somewhat reliable, as warrants are produced prior to arrest, and like Expert Witness reports are subject to unreliability and bias (Gutheil & Simon, 2004; Murrie, Boccaccini, Guamera, & Rufino, 2013). Media articles were then placed on a separate continuum within the less reliable end of the spectrum; with personal opinion blogs at the lower end, and broadsheet newspapers at the upper.

Table 17 Reliability continuum of data sources

Least Reliable		Partially Reliable		Somewhat Reliable		Reliable		Very Reliable	
									
Media Articles				Manifestos		Competency Evaluations		Trial Transcripts	
									
Internet Blogs		Tabloids	Broadsheets	Warrants		Sworn Affidavits		Trial Memorandums	
				Expert Witness Reports		Indictments			

Case Building

Cases Sourced by the Research Team

Once all potential cases were identified (N = 153), the research team then used the same search engines to identify all known case documentation (again using Table ## to guide decisions regarding reliability). This search protocol involved the use of the name of the actor (and all known aliases) within the case and Boolean and wildcard search terms (the same as those presented in Table 4, under the heading 'Research Area'). Once all appropriate documentation was sourced, the research team followed a systematic search process to identify relevant information for vignette building. The quality of documentation across the 153 cases varied widely, and this led to a reduction in the number of cases taken forward for the final list of vignettes. The final number of included vignettes across categories is presented in Table 18.

The information used in the vignette building for each case represents that which could reasonably be found in an investigational file. Information spanned a range of static, dynamic, risk and protective factors that the individuals possessed (following the categories laid out in the thematic analysis in task 1, and supported by the empirical findings of the SR undertaken in task 2). This included; information regarding the individual's childhood, family, academic achievement, religious upbringing, prior criminal convictions, drug use, travel, ideology, social networks, social activities, social norms, affective relationships, attitudes, personality factors, self-control, self-efficacy, and mental health. Information regarding any violent or terrorist act was excluded from vignettes.¹⁰⁶

Once all information was gathered for each case, it was transformed into vignettes between one and four pages in length. Following this, each was de-identified and personal pronouns were made gender neutral.

Cases Sourced from London and Canada

Given that Radar specifies that it has been designed to be used specifically in Australia, and was developed from Australian case data, and part of the data underpinning the development of the VERA-2R is drawn from convicted Australian offenders, the research team gathered non-Australian cases to test the validity of using both instruments

¹⁰⁶ To ensure that the outcome was not known during testing, as this would affect the rating of cases by participants

in the assessment of both Australian and non-Australian cases. Research teams in London and Canada provided cases from both Europe and Canada. Once the ANU research team received these cases, each was scrutinised for quality and similarities in factor inclusion and length, to match those gathered by the Australian research team and were anonymised and made gender neutral.

Closed Source Cases

The final source for cases was the Department. These vignettes were provided by participating state and territory agencies (sourced through the Department's contacts and affiliates), and were built using investigative files and information held during case assessments by these agencies. Agencies who agreed to provide vignettes were given an instructional pack which covered the vignette writing protocol used by the ANU (and London and Canada) research team. This information pack covered the types of required information (as stated above), style guidance, and anonymisation instructions. Vignettes were sent to the research team via a secure email system.¹⁰⁷ Once received by the research team, vignettes were given new identifying markers, and styles altered to match the cases gathered by the research team.

This closed source data offers the opportunity to evaluate if the standard of information that is often used during the assessment process of the VERA-2R and Radar is of sufficient quality to perform assessments as stated by the VERA-2R and Radar.¹⁰⁸

Ideological Manipulation

Of the 46 cases taken forward for use in task 4, 14 underwent ideological manipulation. These cases were chosen due to the richness of available information, and the lack of ideological specific content. This manipulation allows for the testing of the equity of the instruments.

¹⁰⁷ This included a secure email address, and this inbox was only accessed in a secure room within the research team's building at the University.

¹⁰⁸ Although researchers often criticise the outcomes of analyses from data that is gathered from open source outlets, research has shown that, when handled correctly, and researchers use strict contingency procedures during collection, coding, reconciliation, and analysis, open source data can be highly reliable, and comparable with closed source data (Gill, Corner, McKee, Hitchen, & Betley, 2019).

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Instruments.
The content of this report may be shared within agencies but is not to be reproduced or distributed further.

Table 18 Vignettes gathered by the research team

Case Category	Number of Vignettes			Total
	ANU	London, Canada	Closed Source	
Violent Extremist (VE)	10	5	2	17
Non-Violent Extremist (NVE)	9	2	3	14
Violent Group Member Non-Political (VGMNP)	5	0	0	5
Non-Violent Group Member Non-Political (NVGMP)	3	1	0	4
Lone Extremist Violent (LEV)	4	3	1	8
Lone Extremist Non-violent (LENV)	1	1	0	2
At Risk (AR)	0	0	5	5
Control Group (CG)	3	2	0	5
Total	35	14	11	60

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Task 4 - Testing the Reliability, Validity, and Equity of Instruments

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

Introduction

This task is the final critical examination of the instruments under scrutiny. The report below presents quantitative and qualitative outcomes of an experimental protocol. In this protocol, a series of assessors followed the risk assessment protocol of both the VERA-2R and Radar to evaluate the vignettes that were developed in Task 3. This task was completed over a four-month period.¹⁰⁹ In particular, in this task, we critically assess the validity and reliability of the instruments. To do so, the team employed multiple measures. The inclusion of false positives, false negatives, and counter-factual cases allowed for the ability to test for validity (sensitivity and specificity, predictive values, and the area under the curve),¹¹⁰ and inter-rater and intra-rater reliability (percentage agreement, Krippendorff's alpha). As the vignettes cover a range of ideologies and terrorist types, we also assessed equity (how well the instruments apply to different categories of individuals). Finally, written and verbal feedback from the participants was assessed to better understand the true extent of the implementation burdens of the instruments.

These assessments allow for answers to the following questions: (1) Is each risk assessment instrument sufficiently reliable to ensure consistency in decision-making regarding the level of risk posed? (2) What specific factors within each risk assessment instrument are less/more reliable? (3) Is each risk assessment instrument valid and to what degree does it discriminate between levels of risk? (4) Could any instrument be improved by adding/deleting specific factors within them or by adopting factors mentioned in other instruments? (5) Is each instrument valid for various subgroups: e.g. across ideologies, loners vs. group actors (6) What costs are associated with each instrument in terms of workload and speed? (7) Which instruments consistently do well across issues surrounding motivation, intent, capability and preparation and do some instruments score better on particular issues? In sum, the outcomes of this task have the potential to demonstrate whether the VERA-2R and Radar accurately classify offenders or overestimate or underestimate risk inaccurately. The outcomes also have the potential to provide international standards for reliability, validity, and equity for future instruments.

¹⁰⁹ The anticipated completion of the experiments was delayed due to the extreme weather conditions that Canberra faced between December 2019 and March, 2020. This also affected the size of the cohort of the participants.

¹¹⁰ Diagnostic and Logistic Odds ratio were not used due to instability arising from small sample sizes, and point biserial not used as it is used to measure direction and strength of association between continuous and dichotomous variables.

Instrument Testing

Rationale

In tasks 1 and 2, the research team performed thematic and comparative assessments of the VERA-2R and the Risk Analysis protocol in Radar using specific standards. The outcomes of task 1 offered understanding of the stated approaches and risk specifications of the instruments. The authors of the VERA-2R and Radar assert that the instruments both adopt the SPJ approach to risk assessment, and that the risk specifications are distinct from one another. Given this, it was necessary to critically examine the quality of the evidence base underwriting the instruments, the theoretical mechanisms, and the weighting of the factors. Task 1 only evaluated the author's statements regarding each of these categories.¹¹¹ In order to present conclusions with any certainty, we must look to the instruments themselves. Therefore, in task 2, we detailed the outcomes of a systematic analysis of both the content and construct validity of the VERA-2R and Radar. The outputs of this task identified the weaknesses in the empirical underpinnings of the factors within the instruments. However, the other categories assessed in task 1; validity and reliability, moderators¹¹², and implementation burdens cannot be fully assessed solely by thematic or comparative analyses.

Both the VERA-2R and Radar purport to assess dynamic risk factors to predict the likelihood that specific individuals will go on to either radicalise (in the case of Radar) or conduct violent extremist acts (in the case of the VERA-2R). The premise of predicting the most 'at-risk' individuals is underwritten by the assumption that by identifying and targeting such individuals, resources can be directed toward those who would most benefit from some form of intervention and management (Andrews & Bonta, 2010; Singh, 2013). Both the VERA-2R and Radar are employed across a range of settings to help facilitate interventions for those of highest risk. In order to support

¹¹¹ The authors of the VERA-2R state that the instrument has face, content, construct, 'deductive', 'impression', and 'user' validity, and inter-rater reliability. The authors of Radar do not make any assertions regarding either validity or reliability, as no tests have been conducted on the instrument. With regards to moderators, both instruments claim to be 'ideologically neutral'. The literature underwriting the instruments supports these claims (see task 2). However, there is confusion over the purported supporting data due to poor methodological reporting. The VERA-2R also claims to be appropriate for use on all age groups, however, closer scrutiny of the supporting literature does not appear to support this claim. Both instruments purport to be suitable for both lone and group-based actors, despite a lack of consideration of the distinct psychological differences between these actor types in both the supporting literature and the included factors. The VERA-2R documentation is steadfast in the need for training, with the Radar documentation being less clear.

¹¹² Which offers understanding of the equity of the instruments.

the use of the VERA-2R and Radar for their stated risk specifications, it is necessary to determine if the instruments are able to differentiate between individuals in terms of their individual risk factors and overall risk in a valid and reliable manner.

A wide range of research has systematically evaluated the validity and reliability of risk assessment instruments. These efforts predominately lie in the violence domain, with studies employing a variety of performance indicators to statistically measure validity and reliability (Singh, 2013). Despite the widespread use of instruments used in the countering violent extremism and counter terrorism domain, this practice has not yet been implemented for those instruments. Employing instruments that have not statistically demonstrated validity has major implications for the security practice. Without assessing validity, it is not known whether instruments accurately identify those at risk of conducting terrorism offences.

Validity

The accurate identification of risk of future offending is critical for offender management and public safety (Geraghty & Woodhams, 2015). Identification accuracy is most often measured by assessing the predictive validity of the instrument (Singh, 2013). Employing an instrument with no proven predictive validity has serious implications, particularly in the terrorism domain, where the consequences for the inaccurate identification of offenders are potentially deadly. Inaccurate instruments may identify individuals who are subsequently subjected to intervention procedures or deprived of their liberty, despite not being of any risk. The other consequence of a lack of predictive validity is that an instrument may fail to capture those individuals who do move to carry out an act of mass violence, which has serious societal and security consequences.

Sensitivity and specificity are the simplest and most user-friendly performance indicators than can be used to measure the predictive validity of an instrument. They are risk discrimination indexes that identify the proportion of offenders who were judged by the instrument to be of high-risk and the proportion of non-offenders who were judged by the instrument to be of low risk. These analyses are calculated using the information regarding true positives, false positives, true negatives, and false negatives (Singh, 2013). They require only simple calculations and can be performed without statistical software. Both sensitivity and specificity are influenced by base rates however, and unable to indicate predictive accuracy across populations. They also are affected by the cut off-

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

threshold for risk. These indicators are only able to assess the accuracy of prediction of dichotomous outcomes (low-risk vs. high-risk), and therefore struggle with outcomes with more than two conditions. This is problematic for SPJ instruments with scoring thresholds that include three categories,

A more statistically complex performance indicator of predictive validity is the Receiver Operating Characteristic (ROC) analysis, which was first used in the violence risk assessment domain in the 1990s (Douglas, Cox, & Webster, 1999; Geraghty & Woodhams, 2015; Mossman, 1994). The ROC is now viewed as the de facto standard in the identification of predictive validity (Singh et al., 2013). The outcome of this analysis is the Area Under the Curve (AUC) statistic (Singh, 2013). This statistic offers insight into the probability of a randomly selected offender scoring higher on the instrument than a randomly selected non-offender.¹¹³ This offers insight into whether the instrument is able to discriminate between types of individuals (the global discrimination index). An AUC value of 0.00 represents perfect negative prediction, a value of 0.50 indicates chance prediction, and value of 1.0 indicates perfect positive prediction. Generally, AUC values over 0.70 are considered 'moderate' and values over 0.75 'good' (Geraghty & Woodhams, 2015). AUC estimates are often sought during the evaluation of risk assessment instruments as they are independent of base rates and selection ratios (Rice & Harris, 1995), and are able to differentiate between more than dichotomous outcomes.

Despite the wide application of sensitivity and specificity and the ROC and resulting AUC, they are fundamentally measures of an instrument's ability to discriminate between those who go on to commit an offence and those who do not. Predictive validity is also dependent on an instrument having calibration. Calibration is defined as how well the instrument's prediction of risk agrees with the actual observed risk (Singh, 2013). Thus, other performance indicators are often drawn on to supplement sensitivity and specificity of the AUC, and provide calibration estimates.

Positive and Negative Predictive Values (PPV and NPV) are two such estimates that are able to measure calibration, and therefore provide greater insight into the usefulness of instruments in practice, as they emphasise the prospective prediction of offending outcomes (Singh, 2013). The PPV offers insight into the proportion of individuals judged to be of high-risk who do go on to offend, and the NPV gives information of the proportion of individuals

¹¹³ So, for the VERA-2R, the 'offender' would be a violent extremist as compared to a non-violent, non-extremist.

who are judged to be of low-risk who do not go on to offend (Altman & Bland, 1994; Singh, 2013). However, much like sensitivity and specificity, both the PPV and NPV are limited to predicting dichotomous outcomes only in the sample under assessment, and are affected by base rates.

However, authors argue that despite the issues with the PPV and NPV, they can be used in conjunction with global discrimination indices, such as sensitivity and specificity or the AUC (Singh, 2013). Despite the benefits of the AUC, its lack of ability to measure calibration means that any outcomes are unable to offer insight into the prospective predictive accuracy of an instrument. Therefore, the results from calibration indices are able to supplant the findings of the AUC. In this task we employ all of the above-mentioned indicators; sensitivity and specificity will provide a descriptive overview, the AUC can offer results that are not bound by the sample under scrutiny, and the PPV and NPV will offer insight into the prospective predictive validity of the instruments.¹¹⁴

Reliability

As well as validity, risk assessment instruments should also hold high reliability. The reliability of any risk assessment is central to its continued use by professionals. If users are regularly unable to agree on the presence and or relevance of certain risk factors in an individual case,¹¹⁵ then the method of risk assessment may be contentious. For SPJ approaches, much like any other risk assessment approach, it is important to formally test for the agreement between those using the instruments.

The literature on reliability studies for general violence risk assessment tools (see Table 20 for illustrative examples) follows a standard approach to designing reliability studies (points 1-8 in Table 19 below). It should be understood that, unlike the general violence tools, there are far fewer empirical studies to draw on to establish best practice approaches to measuring reliability within instruments employed in the terrorism domain (Lloyd, 2019).

The standard approach to critically assessing reliability is as follows:

¹¹⁴ The other performance indicators detailed in Appendix 1 are not suitable for use due to the format of the data that was gathered during the experiments.

¹¹⁵ As both the VERA-2R and Radar encourage the use of panels and teams of assessors, this issue is particularly pertinent.

Table 19 Protocol to be used when undertaking reliability analyses

- 1 Decide if reliability is to be measured in the 'field' or a 'laboratory' setting.
- 2 For raters - decide their number, experience and training required.
- 3 For cases to be assessed – decide on the number and whether they are 'real world' cases or constructed vignettes. Decide on the format and environment in which case information is presented. For example, this could be roundtable presentations or inspection of digital/written case files.
- 4 Decide on the levels of analysis. Depending on the instrument structure, reliability might be measured at the single item level, scale level (items congregated together) or summary assessment level (all items together).
- 5 Likewise, consideration is needed of how the reliability of scenarios or formulations will be assessed (qualitatively or possibly quantitatively).
- 6 Pragmatic considerations (modulated by power calculations) may often come into play when deciding on the availability of cases and raters. There is considerable variation in numbers of cases and raters in published studies which impacts on the robustness of the design.¹¹⁶
- 7 Decide on the statistic to be used to measure reliability. Depending on size and quality of the data available, the complexity of the method may be relatively simple (e.g. Cohen's kappa) or more complex (Krippendorff's Alpha or Intraclass Correlation Coefficient (ICC)). Percentage agreement amongst raters is helpful but it does not account for variance between users or control for agreement that happens by chance (Cohen, 1960). ICC produces an index of agreement while accounting for variance across users.¹¹⁷
- 8 Compare the statistic to commonly accepted 'good practice' values. Fleiss (1981) proposed that for single measure ICCs, values .75 and above be considered "excellent," those from .60 to .74 "good", between .40 and .59 "moderate" and under .40 "poor".¹¹⁸ For well-established tools, ICCs above 0.6 are commonly reported. For example, Douglas and Reeves (2010) summarised interrater reliability research on the HCR-20 Version 2 scale scores across 36 studies and reported that the inter-rater reliability to be in the good to excellent range (ICC = .67 to .95), with a median reliability coefficient of .85. Vincent et al (2012) noted "a tool should strive for ICCs of at least .61 and preferably above .75".

¹¹⁶ For example, Douglas & Belfrage (2014) commented "10 ratings on 50 patients would have provided an extremely robust design that would not be influenced by minor perturbations or outliers in the data". Pragmatically, however, this study actually used 3 raters and 32 cases.

¹¹⁷ Shrout and Fleiss (1979).

¹¹⁸ However, Landis and Koch (1977) suggest the following guidelines: Below 0 is "poor", .00 to .20 is "slight", .21 to .40 is "fair", .41 to .60 is "moderate", .61 to .80 is "substantial", and .81 to 1.00 is "almost perfect."

The number of raters and cases significantly influences the robustness of estimates of reliability. For example, Table 20 shows four reliability studies of different and widely-used SPJ instruments, illustrating the variations in number and type of cases and number of raters. Douglas and Belfrage (2014, p. 138) noted the use of three raters “reduces somewhat the number of cases necessary to achieve relatively stable estimates of reliability.” Formal power calculations may be undertaken for ICCs based on the work of Bonett (2002). Software tools¹¹⁹ are available including for estimates of the number of raters and cases at varying levels of reliability and precision. For example, using 3 raters and 10 cases will allow for a precision of 0.2 (+/- expected) on an expected ICC reliability of 0.80 (95% confidence level).

Table 20 Variations in inter-rater reliability studies

Author	Instrument	Cases	Raters	Statistical test
Douglas & Belfrage (2014)	HCR20 v3	32 real cases	3 from 6 (3 chosen at random) experienced	ICC
Ryan (2016)	SARA v3	30 real cases	2 Experienced	ICC
Timmins, Evans, & Tully (2018)	START	2 test cases	15 experienced	ICC
Vincent, Guy, Fusco, & Gerschenson (2012)	SAVRY	80 real cases	2 experienced	ICC

NB – Examination of these studies will allow the teams to measure (a) how consistent scoring is over a short time interval, (b) how individual users change their scoring over time, and (c) possible drift in tool use over longer time periods.

Activities within this task will be focused on conducting reliability testing to determine the level of consensus among item and scale ratings. If an instrument has been designed and implemented in the correct manner, then similar outcomes will be expected across users when examining the same case. That is, formal inter-rater reliability (IRR) values should be high.

In order to test the validity and reliability of the instruments under scrutiny in this research, the vignettes produced during task 3 were employed as cases for assessment to a range of users. Of particular note, in order to determine

¹¹⁹ For example, <https://wnariffin.github.io/ssc/ssicc.html>

the most effective test of IRR, the guidelines in Table 19 were all followed in the design of the experimental methodology. The following section outlines this methodology, including the identification and engagement of participants, the experimental process, and the data collation and analytical procedures.

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Protocol

Participant Recruitment

*Participant Categories*¹²⁰

Given that both the VERA-2R and Radar involve a training protocol for potential users, participants for this task were drawn from three sources:

- University students who are both untrained and non-experts.¹²¹ These individuals were approached because they had successfully completed an undergraduate-level course focusing on risk and threat assessment, and had achieved either a Distinction or High Distinction grade.
- Experts. These individuals were drawn from three sources; the academic network of the primary researcher on this project, from the Department of Home Affairs, and from across the State and Territory agencies who supplied case vignettes during task 3.¹²²
- Trained assessors. These individuals were also drawn from the primary researcher's network, the Department of Home Affairs, and the agencies. Individuals were classified as trained if they had received training in either the VERA-2R or Radar.

In order to adequately undertake reliability analyses, the employment of power calculations were undertaken by the team to determine the minimum required number of participants. If we estimate the reliability of the VERA-2R to mirror the findings of Beardsley and Beech (2013) who identified a Kappa statistic of 0.76, whilst using five case

¹²⁰ Given that this research employed human participants during the data collection, and a proportion of the case vignettes (18.3%) that were assessed during this task were drawn from closed source information, university protocol required that an ethics agreement was granted, and confidentiality forms were signed by all participants prior to the start of any experimental protocol. Ethics approval was granted prior to any activity on this task.

¹²¹ For the purposes of this project, experts are defined as those individuals who work within the countering-terrorism and countering violent extremism areas, and have experience dealing with such cases, but who have not received specific training in either the VERA-2R or Radar. These individuals may be trained in other forms of risk assessment, such as general violence or other terrorism-based instruments, or they may have no formal training in any form of risk assessment, but they have sufficient expertise from their work.

¹²² To ensure there were no errors in assessment bias, after randomisation (see below for more details) Case vignettes that had been allocated to these individuals were cross-checked to ensure that they were not the same vignettes that had been provided during task 3.

studies and two raters, for each vignette to be assessed by the same three participants, we were required to have a sample of at least 67 participants, each assessing all 60 case vignettes.

However, these power calculation estimations do not consider two significant issues that were identified during the experimental protocol; timing and fatigue. These issues had major implications for the number of vignettes that could be reasonably assessed during the protocol. The number of factors within both the VERA-2R and Radar assessment protocols (45 across both instruments)¹²³ meant that significant time was needed for participants to assess each case vignette. Due to the issue of fatigue, the research team set a time limit of three and a half hours per experiment for each assessor. Within this timeframe, the maximum number of assessments that could be successfully conducted without significant fatigue was identified to be four. This meant that the calculations of reliability could not be calculated with any statistical confidence using ICC. Therefore, the research team opted to use Krippendorff's Alpha (Krippendorff, 2011), which does not require three assessments of each case vignette, and is able to calculate ordinal, non-dichotomous data, which is not possible with Cohens Kappa.

Recruitment Process

All potential participants were approached by email. In order to follow the ethics and security protocol required, a secure functional email account was used for this process. The email account was only accessed in a secure room on a computer that was not accessible to any other persons. All correspondence were securely destroyed following the experimental process. Upon first contact, all participants were provided with an information pack, which included a participant information sheet and the ethics agreement information. Participants were also informed of the time the assessment process would take. All potential participants were requested to respond with their willingness to participate. Participants who indicated their willingness were assigned dates for assessment based on their availability.

Given the original power calculation estimations, the research team did attempt to recruit 67 participants, but despite these efforts, only 30 participants completed the full assessment protocol. There were a number of practical

¹²³ Including both the Risk Screening and Risk Analysis components in Radar.

constraints that influenced non-completion, including; non-response, timing clashes, reluctance to participate for the required time for assessment, technological issues, access restrictions, illness, role termination, and extreme weather conditions. Despite these issues, it is worth noting that the completion of assessments by 30 participants, and particularly expert and trained participants, is highly substantial for a project in this academic discipline.

Participants

Upon recruitment, all 30 participants were assigned a randomly allocated unique identifier that they used during the assessment process.¹²⁴ Overall the gender split of participants was 60-40 in favour of females. The majority of assessors had graduated with some form of higher education qualification (86.7%). This is not surprising given the requirements of two of the categories. To protect identities, no further demographic or personal information was gathered on the participants. However, as it was critical for the categorisation process, the breakdown of participant experience and expertise is detailed in Table 21 below. On average trained participants had more years of experience in violence risk assessment ($M = 11.60$, $SE = 1.89$) than experts ($M = 9.78$, $SE = 2.11$), although t-tests revealed that this difference was not significant, $t(12) = -.571$, $p > 0.1$; with a low effect size, $r = 0.16$. However, on average, trained participants did have significantly more experience in risk assessment instruments ($M = 9.00$, $SE = 2.48$) than experts ($M = 3.50$, $SE = 1.01$), $t(12) = -2.496$, $p < 0.05$; with $r = .58$, indicating a high effect. These findings are expected, as a requirement of inclusion in the 'trained' participant group was training in either the VERA-2R or Radar. However, the results do indicate that some individuals in the expert group had received training in other terrorism risk assessment tools, which may have given them an advantage during the assessment process.¹²⁵

Table 21 Breakdown of Participant Experience and Expertise across Groups.

¹²⁴ On completion of the assessment process, these unique identifiers were then randomly changed using software to ensure that the identity of assessors was anonymous.

¹²⁵ Although the VERA-2R documentation notes that irrespective of experience, all users must undergo specific VERA training, indicating that there are elements within this process which are advantageous for the assessment of this instrument.

Participant Group	N	Gender	Law Enforcement Experience (average years)	Violence Risk Assessment Experience (average years)	Terrorism Risk Assessment Experience (average years)
Novice	10	8 Female, 2 Male	0	0	0
Expert	9	5 Female, 4 Male	1 (30)	5 (11.6)	3 (10.3)
Trained	11	7 Female, 4 Male	7 (15.1)	9 (9.8)	11 (3.2)

Vignette Assessment

Allocation of Cases

All cases were initially allocated randomly. However, as some participants had also provided case vignette information for task 3, all cases were cross checked to ensure that no individual received a vignette that they had supplied to the research team. Further to this, familiarity became an issue during case allocation. As two groups of participants either directly worked in the security space, or had previous access, and also were knowledgeable about terrorism cases, during testing, multiple assessors reported that they were too familiar with some cases to perform the assessment without bias. In these instances, participants were provided with a new randomly allocated vignette. This affected the allocation of cases, and further prevented the aim to undertake three assessments of all vignettes.

Experimental Protocol

All participants were offered three alternative experimental settings, in order of preference, these conditions were; face to face, video conference, and phone conference. All novice participants completed their assessments in the face to face setting. Due to the geographical locations of many expert and trained participants, the number of participants who undertook their experiments in the video conference setting was equal to those who undertook face to face testing. Only two participants undertook their experiments over phone conference setting. The overall protocol across these setting did not differ; however, we note the differences in footnotes.

Prior to the start of the experiment, participants were assigned their instruments for the experiment. In the novice and expert groups, this process used random allocation. In the novice group, the instrument allocation was split. The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

evenly, with five participants assigned the VERA-2R, and five Radar. In the expert group, five participants were assigned the VERA-2R, and five were assigned Radar. In the trained group, those participants who were trained in both the VERA-2R and Radar were randomly allocated an instrument for testing. With those participants who were only trained in one instrument, all efforts were made to ensure they were assigned that instrument during testing.¹²⁶

On arrival into the experimental setting, participants were again supplied with the participant information booklet, and a confidentiality agreement.¹²⁷ Participants signed the confidentiality agreement prior to the start of the experiment, and before any further information was given to them. Following this, participants were given the opportunity to ask questions that arose from the information sheet, and then were provided with, and signed a written consent form.¹²⁸

Following this process, all experiments began with a training session. For novice participants, this involved an introduction to the underlying principles of the instrument that was assigned to the participant, followed by an in-depth coverage of the assessment process and the risk factors within the instrument. Expert and trained participants were offered a less comprehensive overview of the introduction section, but were provided with the same training as novices regarding the specific instruments. All participants were afforded the opportunity for further questions at this stage. Following the training, all participants completed an assessor survey, where details of their experience were gathered, and tests for bias were conducted.¹²⁹

Upon completion of the assessor survey, all participants were provided with a copy of the factors within their assigned instrument (categorised as per the instrument documentation), their case vignettes, and an assessment questionnaire for each case vignette. During the assessment of vignettes, participants were required to provide an

¹²⁶ However, this was not possible for two participants.

¹²⁷ During both video and phone conference experiments, all paperwork was provided via the secure email address, and experiments were conducted in a secure room with restricted access.

¹²⁸ For remote participants, these forms were scanned and emailed back to the research team prior to the commencement of any testing.

¹²⁹ Results reported below.

overall assessment (based on the guidance of the instrument documentation),¹³⁰ as well as their overall confidence in the risk decision based on the instrument (using a Likert scale where 1 indicated no confidence, and 5 indicated high confidence), an inclusion of up to five factors that were most pertinent in their decision making process, and their own opinion of the most pertinent items within the case vignette that influenced their decision.

Given that both the VERA-2R and Radar encourage collaboration during assessment protocols, all participants were afforded the opportunity to talk to fellow participants,¹³¹ or the research team, regarding any items in the assessments that were ambiguous. Participants also used this opportunity to bounce ideas around to help them arrive at decisions that they felt were supportive in their decision-making process.

Following the completion of all assessments, participants were debriefed, and informed of the gathering of data to assess their level of bias. Participants were also informed of the ideological manipulation of a proportion of vignettes to help supplement the investigation of equity of the instruments. No participants elected to withdraw their data at any time.

All assessments were conducted within the three-and-a-half-hour time period, with 28 participants completing four vignettes, one completing three vignettes, and one completing five.

Data Collation

Once all information was collected, paper copies were redacted, provided with a newly randomly allocated participant id to ensure that the research team were blind to who each participant was during testing, and scanned, and stored on a secure hard drive in a secure location that was only accessible to the primary researcher.

¹³⁰ For Radar, participants completed both the Risk Screening and In-Depth Risk Analysis components. Here participants were required to make a judgement as to whether each vignette should proceed to the Analysis following the Screening (as per the Radar protocol).

¹³¹ Given time and resource constraints, if participants were recruited from the same agency, or were geographically close, they were encouraged to complete the assessments at the same time and in the same location. This may have affected bias; however, it is more likely that this form of environment was more similar to a real assessment process, and more conducive to better outcomes.

All information was then transformed into both quantitative and qualitative data in preparation for analysis. Analytical procedures were all computed using appropriate statistical software, and assumptions of testing were all met prior to analysis.

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Results

Participant Bias

To critically assess the underlying bias of participants, all were assessed for their underlying opinions regarding which political ideologies are indicative of radical and terrorist behaviour. To do this, participants were each presented with a series of five short vignettes, each with a different political ideology. Participants were requested to provide a score on a Likert scale of their own personal opinion of the risk of the case committing an act of violent extremism. Likert scales were ranked 0-10 with 0 being 'No Risk' and 10, 'High Risk'. The average scores are presented in Table 22.

Table 22 Average Scores of Bias Vignettes across Participant Groups

Participant Group	Average Score				
	Right Wing Vignette	Environmental Vignette	Islamist Vignette	Pro-Life Vignette	Nationalist Vignette
Novice	4.8	3.7	2	2	4.4
Expert	3	3.6	2.9	2.8	3
Trained	4.2	4.8	2.7	4.1	3

Overall, participant bias did not appear to differ across vignette categories. To identify any significant differences in bias between groups, a series of Kruskal-Wallis tests were conducted. The only vignette that yielded significant differences was the pro-life vignette, $H(2) = 6.02$, $p < .05$. Dunn's pairwise comparisons revealed that this effect was due to significant differences between the novice and trained groups ($p < .05$, adjusted using the Bonferroni correction). The results of this analysis indicate that we can expect participant bias did not differ significantly across groups during the full assessment process, and therefore the ideology of the case vignettes did not affect the outcomes of risk judgements between groups.

Instrument 1 – The Violent Extremism Risk Assessment – Version 2 Revised (VERA-2R)

Validity

Face validity

To assess the face validity (the degree to which the instrument appears effective in terms of its purported aims) of the VERA-2R, we requested all participants provide information on the five risk factors that they found most pertinent during assessment. Table 23 below details the factors that were noted as most appropriate for an assessment of the risk specification (violent extremism) with their cited importance ranked accordingly. As shown, the most commonly cited useful factors were similar across all three participant groups. ^{s. 33(a)(i), s. 37(2)(b), s. 47E(d)}


That only nine of out a possible 45 risk factors were most frequently cited as useful (here, 'useful' was determined if a factor was cited five or more times across assessments) has implications for the global face validity of the instrument.

Also of note is the difference between groups. The novice and expert groups cited an equal number of factors that they felt useful for their risk decisions, showing an equal amount of reliance on the instrument during assessments. However, the trained group were twice as likely to not cite any factors as useful during their decision-making process. Examination of the espoused rationale given by participants for this phenomenon highlighted that 'trained' participants were more likely to cite a range of other behaviours and risk factors not present within the VERA-2R, and it was note that there was difficulty assessing the case using the instrument. This was predominately due to the brevity of the case vignettes. However, when asked if such brevity is not representative of true cases that would realistically be assessed, trained participants noted that often cases have even less information at assessment.

Table 23 Factors deemed most useful during assessment process using the VERA-2R.

¹³² Of note is the result that BA2 – Perceived Grievances and/or Injustices was also cited by two groups as critical in their decision making. Task 2 demonstrated that this factor has the greatest underlying empirical support.

s. 33(a)(i), s. 37(2)(b), s. 47E(d)



Predictive validity

Descriptive Results

Scrutiny of the case vignette categories reveals that the participants who used the VERA-2R were significantly more likely to discriminate between categories when assigning a final risk decision. $X^2 (14) = 24.552, p < 0.05$. This outcome however, was skewed by participants correctly assessing control group cases as low risk, and incorrectly assessing violent extremist cases as having moderate risk. Figure 3 below highlights the pattern of risk scoring across categories.

To further interrogate the predictive validity of the VERA-2R, we followed the work of Singh (2013), and used a range of performance indicators. The following sections detail the outputs from analyses of five performance indicators; sensitivity, specificity, positive predictive value, negative predictive value, and the AUC.¹³³

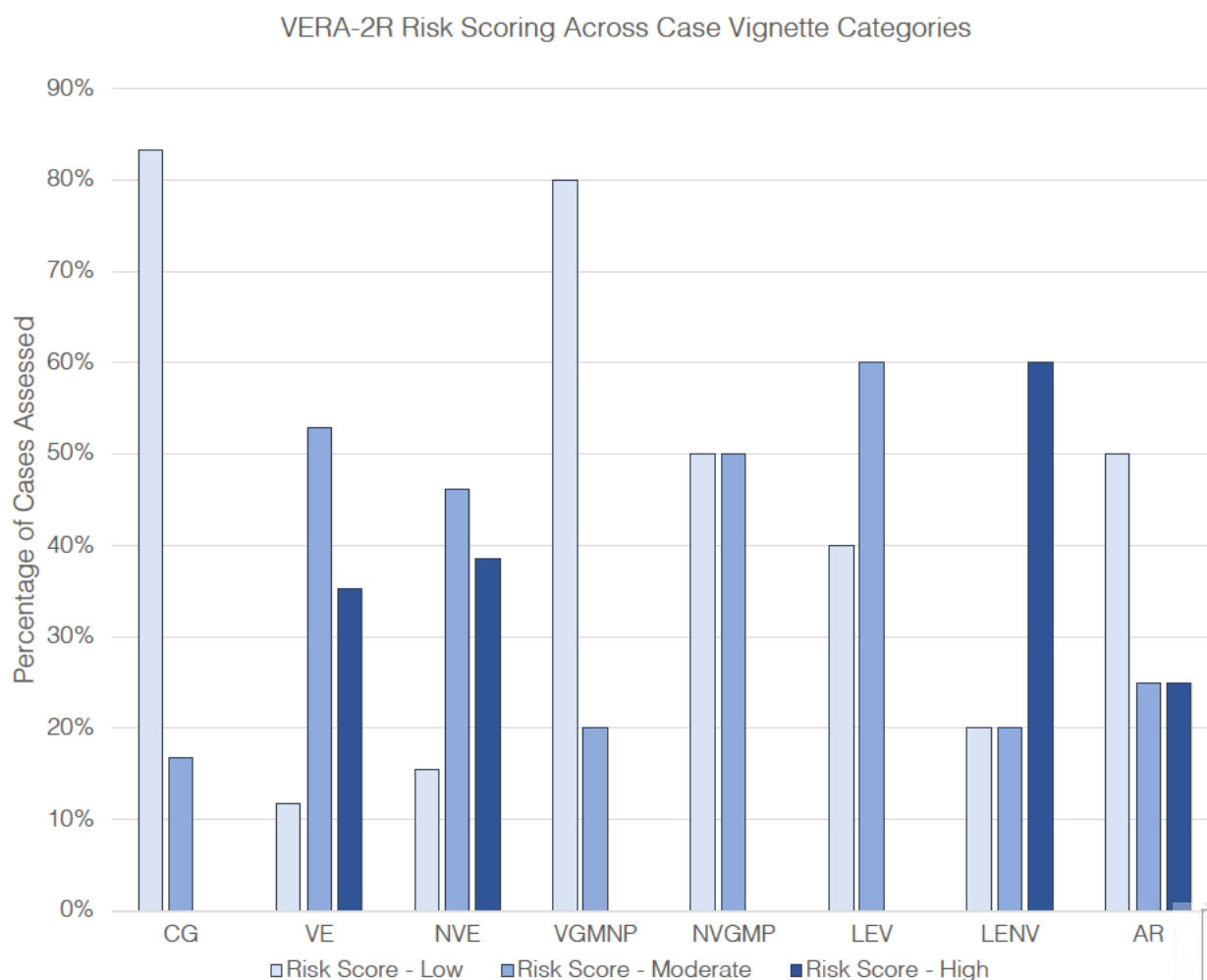


Figure 4 VERA-2R risk scoring across case vignette categories

¹³³ Statistical calculations for each of these performance indicators against both the VERA-2R and Radar are detailed in Appendix 5.

Sensitivity and Specificity

The results for tests of sensitivity (the proportion of individuals who engage in specified risk who are correctly identified by the instrument as high risk) show that the VERA-2R has a sensitivity value of 60%. That is, 60% of case vignettes were correctly identified as high-risk during assessment. The results for tests of specificity (the proportion of individuals who do not engage in the specified risk who are correctly identified as low risk) determine that the VERA-2R has a specificity value of 64%. That is, 64% of case vignettes were correctly identified as low risk during assessment.

Alongside the risk specification, the VERA-2R's ability to predict violent outcomes was also examined using these performance indicators. The results demonstrated that, despite the inclusion of risk factors drawn from the HCR-20, only 30% of case vignettes were identified as high risk for violence, and 64% were correctly identified as non-violent.

However, as noted above both sensitivity and specificity are affected by base rates (and the decision categories used in the experiment mirrored that within the VERA-2R, and included low, moderate, and high). The greater the amount of violent extremist vignettes in the sample, the higher the specificity (and lower the sensitivity). In the sample of case vignettes assessed using the VERA-2R, although there were eight case categories, 28.8% the total cases were violent extremists, which may have skewed the resulting sensitivity and specificity values.

Positive and Negative Predictive Values

As noted above, PPV and NPV both move beyond sensitivity and specificity, emphasise the prospective prediction on outcomes (as opposed to sensitivity and specificity, which are more retrospective indicators). Examination of the predictive values for the VERA-2R show amongst those cases assessed as high risk, 64% went on to perform an act of violent extremism, and in those cases assessed as low risk, 80% did not go on to commit an act of violent extremism.

Much like the outcomes for using sensitivity and specificity indicators, the VERA-2R does not have good prospective prediction for violent outcomes. Within case vignettes judged to be high risk, only 43% went on to perform an act of violence, and in those judged to be low risk, 57% did not go on to commit violence.

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

However, authors warn of using positive and negative predictive values for instruments with more than two risk outcomes (e.g. low, medium, and high), as both indicators rely on the use of single cut off thresholds. Therefore, these outcomes may not be truly representative of the VERA-2R's predictive ability.

The Area Under the Receiver Operating Characteristic Curve

The AUC is used to determine that probability that a randomly selected (in the case of the VERA-2R) violent extremist received a higher risk score than a randomly selected non-violent, non-extremist. Unlike the aforementioned performance indicators, the AUC is resistant to changes in the base rate.

The outputs for the AUC calculations are in Figure 4 below. The results show that the VERA-2R lacks both sensitivity and specificity, as the blue curve is closer to the red diagonal line across all three risk outcomes measured (1 = the risk specification for the VERA-2R; violent extremism, 2 = the risk specification for Radar; radicalisation, and 3 = violence). AUC values of closer to 1 indicate that the instrument is able to reliably distinguish between outcomes. The AUC value for violent extremism is 0.603. Accordingly, this indicates a poor predictive validity (values of 0.5 to 0.6 indicate that the instrument is worthless as it is unable to predict outcomes). This outcome is not significant ($p > 0.1$ (CI: 0.457, 0.749)), however, the strength of the value does indicate a pattern of predictive value. The AUC value for the VERA-2R's ability to predict the risk specification for Radar (radicalisation) is stronger than its own risk specification. Here, the value is 0.699 ($p < 0.05$ (CI: 0.552, 0.849)), which indicates that its predictive validity is on the threshold of fair. Much like the outcomes for the other performance indicators, the VERA-2R's predictive validity for violent outcomes is extremely low. With a value of 0.510 ($p > 0.1$ (CI: 0.361, 0.658)), the VERA-2R's ability to predict violence borders on worthless.

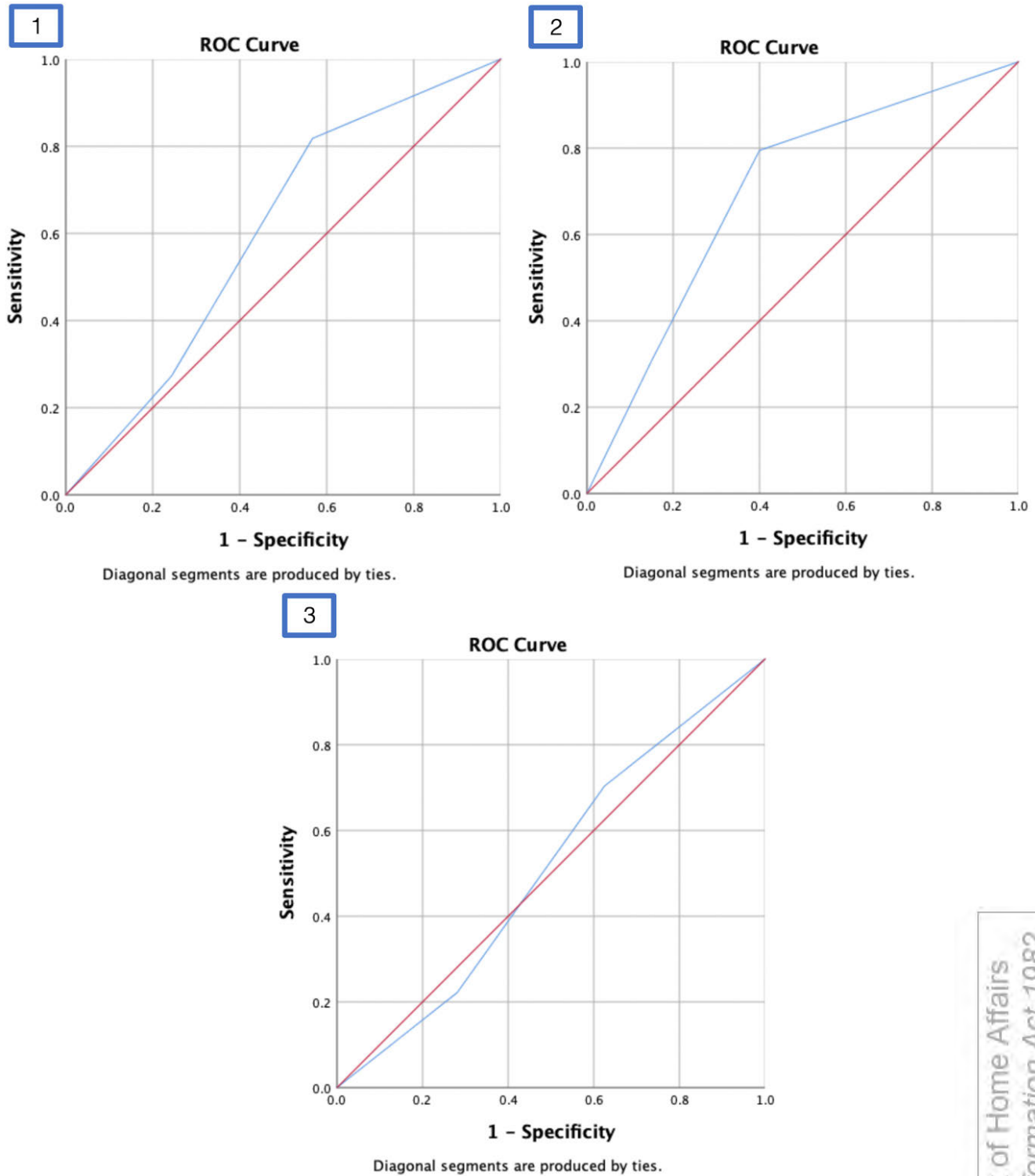


Figure 5 Outputs from Area Under the Curve Analysis for the VERA-2R

NB – 1 = Output for VERA-2R risk specification (violent extremism), 2 = Output for Radar risk specification (radicalisation), 3 = Output for violence specification .

Reliability

Inter-Rater Reliability

We used the Krippendorff's alpha test (Krippendorff, 2011) to assess 13 (30.2% of the total cases assessed using the VERA-2R) cases for inter-rater reliability (IRR) of the factors within the VERA-2R. Chosen cases were assessed by either two or three coders. Individual alpha (α) values are reported in Appendix 5. The average α value for all assessed cases was 0.242. This indicates that inter-rater reliability of the VERA-2R is extremely low (values below 0.67 are considered worthless, values between 0.68 and 0.8 are considered poor, and values above 0.8 are considered good). The results also demonstrate, that if we were to conduct IRR testing across a population sample of cases, there is an 88.2% probability that the α would be below 0.500 (indicating no inter-rater reliability).

Equity

Ideology

To determine whether the VERA-2R is suitable for determining risk of violent extremism across different ideologies, and thus measure its equity, the research team performed a three-way log-linear analysis. The likelihood ratio of this model was $X^2(0) = 0, p = 1$. This indicated that the highest-order interaction (assigned risk score x violent extremism x ideology) was significant. However, the effect was influenced by the association between violent extremism and ideology ($X^2(7) = 13.308, p < 0.1$). This outcome shows that the assignment of risk was not significantly associated with any one ideology, or whether the individual was a violent extremist or not. This result both supports the assertions of the authors of the VERA-2R, that the instrument is ideologically neutral, and also the results from the examinations of predictive validity, as the scoring of risk was not associated with the rating of extremism.

Implementation Burdens

Training

The authors of the VERA-2R are specific in their assertions that individuals who undertake assessments using the instrument should have expertise in the area of countering violent extremism or terrorism and have undergone a

recognised VERA-2R training program. To determine if expertise and training significantly affects the decision-making protocol during the assessment process, the research team undertook a three-way loglinear analysis to compare the likelihood of making correct risk determinations across all three categories of participants.

The analysis revealed that the likelihood ratio of this model was $X^2(0) = 0, p = 1$. This indicated that the highest-order interaction (participant group x case category x risk score) was significant. However, this effect was influenced by the association between the case category and the final risk score ($X^2(14) = 27.724, p < 0.05$). This outcome shows that the assignment of correct level of risk was not significantly associated with training of participants, which does not support the assertions of the authors of the VERA-2R.

It is also worth noting that training and experience does not appear to affect the IRR of the VERA-2R. Of those cases where coders had the same level of experience, the α values did not significantly differ from the cases assessed by coders with differing levels of experience, $t(11) = -0.565, p > 0.1$. This implies that training and expertise does not affect the interpretation of the presence of factors within the VERA-2R, further supporting the results of the log linear analysis.

Confidence

To further assess whether expertise or training may be necessary for the use of the VERA-2R, we examined the confidence of participants during the assessment process. Overall, participants showed moderate confidence in the VERA-2R with regards to how it helped them formulate their final risk decision. Figure 5 below shows the confidence ratings across participant groups. Further chi-square analysis did not reveal any significant differences in confidence ratings between participant groups, further supporting the above results regarding the effect of expertise and training.

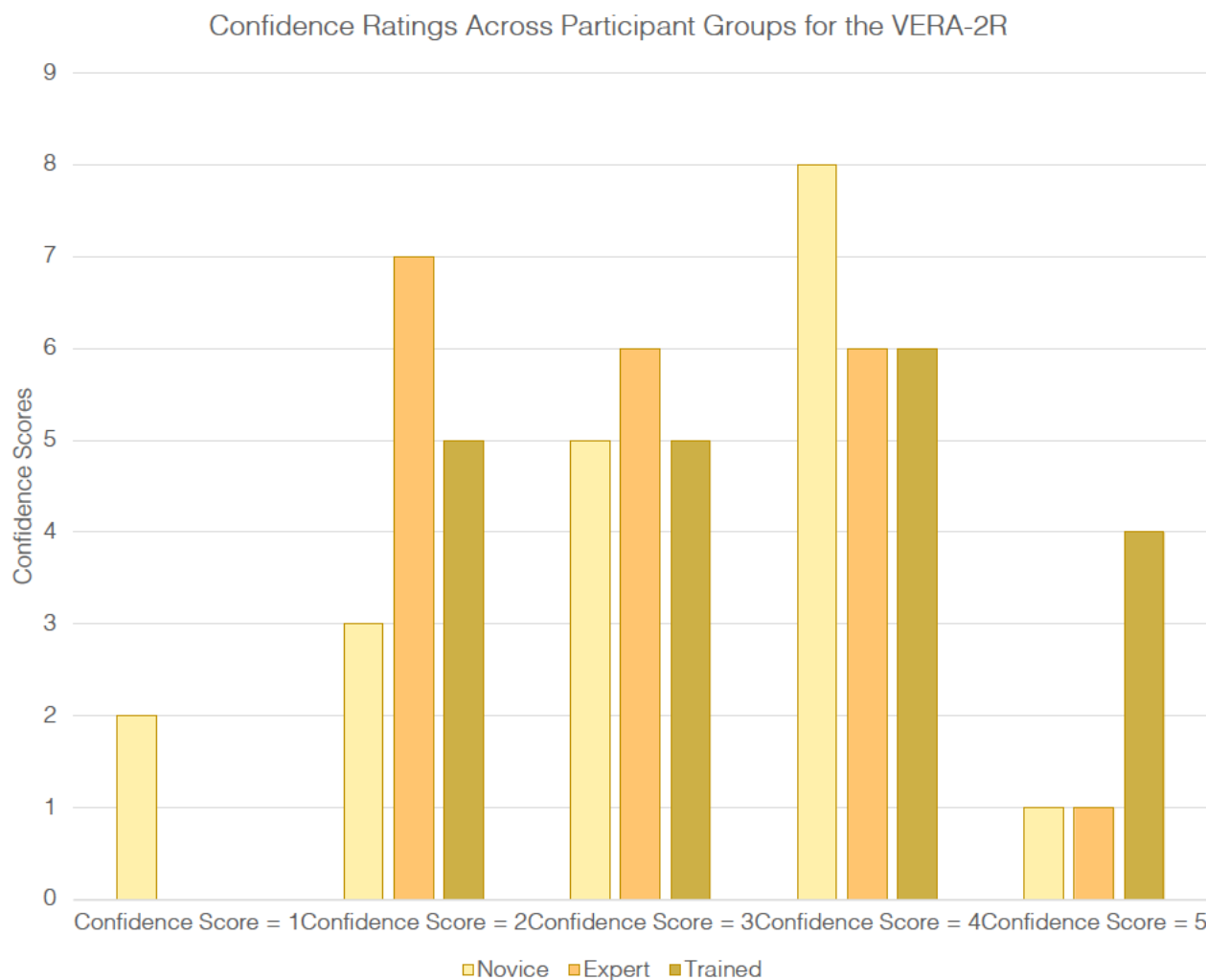


Figure 6 Confidence ratings across participant groups for the VERA-2R

Instrument 2 – Radar

Validity


Face validity

To assess the face validity of the Risk Analysis protocol in Radar, we requested all participants provide information on the five risk factors that they found most pertinent during assessment of the vignettes. Table 24 below details the factors that were noted as most appropriate for an assessment of the risk specification (radicalisation). As shown, the most commonly cited useful factors were similar across all three participant groups. The one consistent factor identified across all three groups was s. 33(a)(i), s. 37(2)(b), s. 47E(d)

. But its ranked importance differed. This difference in opinion is also matched in the amount of assessments which cited no factors as useful in the decision-making process (14 cited no factors as useful, and 15 cited three or less factors they deemed pertinent). Matching the assessment of the VERA-2R, only factors that were cited five or more times for their usefulness were included. The trained group cited seven factors five or more times that were useful in their determination of risk, so all are included. However, the expert group did not cite any factors five or more times, so the two that are included in the table were both cited four times.

Unlike in the VERA-2R, where the trained group were less likely to cite factors, for Radar, this group were more likely to cite five factors. This indicates that those individuals trained in Radar relied on it more heavily during assessment than the other two groups of participants. However, this finding masks a more convoluted outcome: when we examine the qualitative expansions of the decision-making process, the trained group were also more likely to give far more detail of their decision-making process than the other two groups, and much like with the VERA-2R, cite a range of other behaviours and risk factors that are not within Radar, and note that there was difficulty assessing the case using the instrument. There was no mention of vignette brevity impacting the process using Radar, but themes did include; too greater specificity of factors, the number of protective factors (not drawn from Radar, but own personal knowledge), and the inability to attach ideological information in vignettes to risk factors in Radar.

s. 33(a)(i), s. 37(2)(b), s. 47E(d)



Predictive validity

Scrutiny of the case vignette categories reveals that the participants who used Radar were significantly more likely to distinguish between categories when assigning a final risk decision. ($\chi^2(12) = 22.893, p < 0.05$). This outcome however, was skewed by participants correctly assessing control group and violent non-political group members as low risk of radicalisation. Raters were not as successful in their determinations of risk scores across the categories that contained radicalised individuals (violent and non-violent extremists (group members and loners)), or those at risk of radicalisation. Figure 6 below highlights the pattern of risk scoring across categories.

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

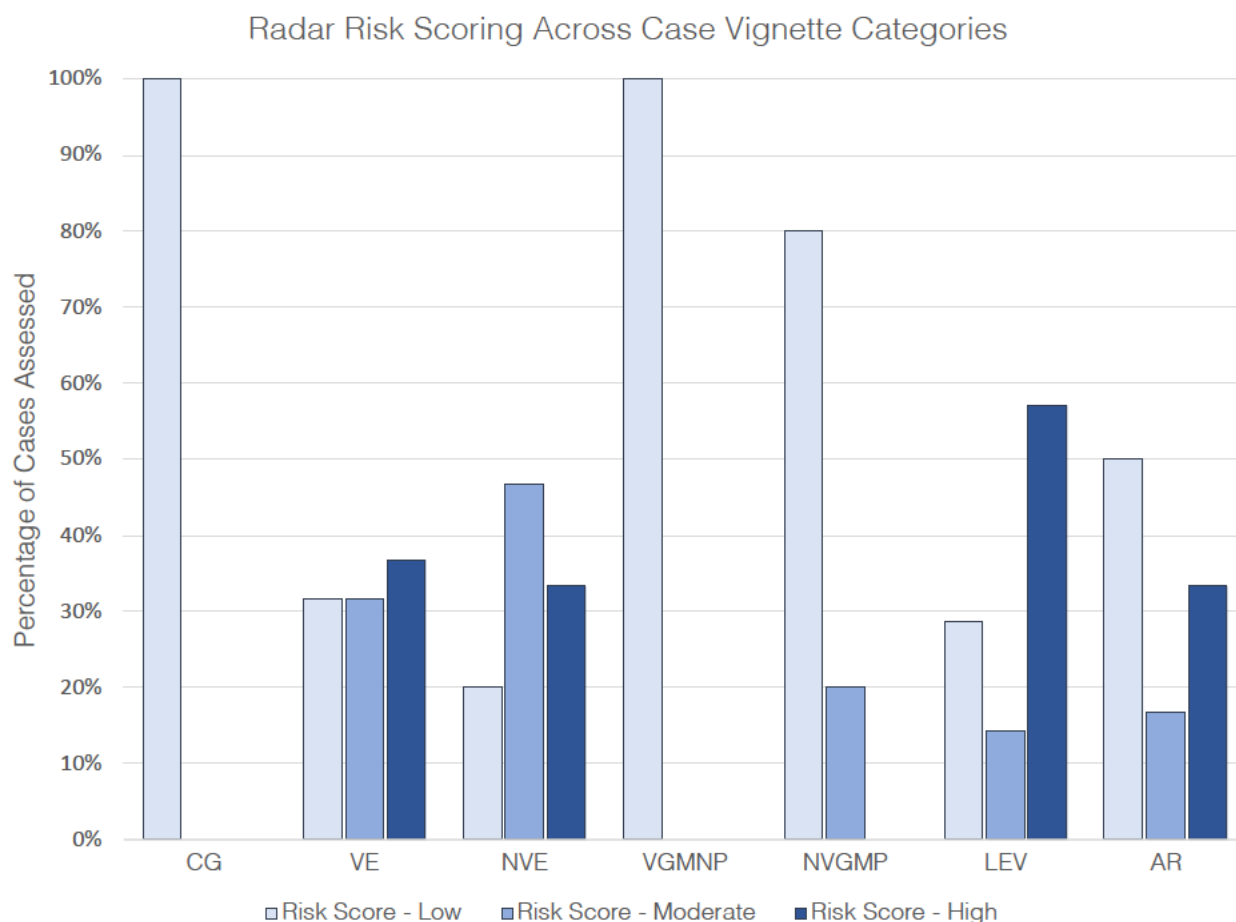


Figure 7 Radar risk scoring across case vignette categories

To further interrogate the predictive validity of Radar, we followed the protocol used to assess the predictive validity of the VERA-2R; using five performance indicators; sensitivity, specificity, positive predictive value, negative predictive value, and the AUC.¹³⁴

Sensitivity and Specificity

As the Radar Risk Analysis protocol is made up of both the Risk Screening and In-Depth Risk Analysis components, both were assessed for their validity. The results for tests of sensitivity show that the Risk Screening component

¹³⁴ Statistical calculations for each of these performance indicators against both the VERA-2R and Radar are detailed in Appendix 5.

has a sensitivity value of 68%. That is, 68% of case vignettes were correctly identified as high enough risk to proceed for further analysis during assessment. The specificity of the In-Depth Risk Analysis was only 58%. This infers that a proportion of the additional factors in the In-Depth Risk Analysis will lead to errors in the overall identification of the risk specification. The results for tests of specificity determine that the Risk Screening component has a specificity value of 93%. That is, 93% of case vignettes were correctly identified as not posing a high enough risk for proceeding for further analysis. The specificity value for the In-Depth Risk analysis is 100%, meaning that in 100% of the case vignettes, there was correct identification of individuals who did not pose a high risk of radicalisation.¹³⁵

Alongside the risk specification, the two components were assessed for their ability to predict violent outcomes. The results supported the documentation of Radar, demonstrating that, the Risk Screening and Risk Analysis protocol had poor sensitivity (53% and 40% respectively) and specificity (55% and 55% respectively) as compared to the outcomes for these indicators in the assessment of Radar's risk specification.¹³⁶

Positive and Negative Predictive Values

Examination of the positive and negative predictive values for the Risk Analysis protocol of Radar show amongst those cases assessed as needing to proceed during the Risk Screening, 97% went on to radicalise, and in those cases assessed as not requiring closer scrutiny, 52% did not go on to radicalise.

Examination of the In-Depth Risk Analysis component shows that for those individuals who are classified as high risk of radicalisation, 100% go on to radicalise, and much like the Risk Screening, for those determined to be of low risk, 52% do not go on to radicalise.

¹³⁵ It is likely this figure is skewed by the requirement of measuring dichotomous outcomes, and does not consider those individuals classified as presenting with moderate risk.

¹³⁶ Of note, both the Risk Screening and In-Depth Risk Analysis show greater specificity than the VERA-2R for the prediction of violent outcomes, which is concerning for the VERA-2R, as it contains risk factors drawn from the HCR 20.

Although the positive predictive values for both components show that Radar has excellent prospective prediction, concern lies in the outcomes for the negative predictive values. This is due to the high number of false negative outcomes in both the Risk Screening and Risk Analysis components.

In the case of the Risk Screening, where the outcome is to proceed or not proceed (to the In-Depth Risk Analysis) the predictive values are able to support the assertions, as this is a dichotomous outcome. However, the outcome of the risk decision measured was not dichotomous, which will have affected the 'excellent' prospective predictive ability.

The Area Under the Receiver Operating Characteristic Curve

In the case of Radar, the AUC is used to determine that probability that a randomly selected radical individual proceeded to an In-Depth Risk Analysis (in the case of the Risk Screening) and received a higher risk score (in the case of the In-Depth Risk Analysis) than a randomly selected non-radical individual. Unlike the aforementioned performance indicators, the AUC is resistant to changes in the base rate.

The outputs for the AUC calculations are in Figure 7 below. The results show that the Risk Screening and In-Depth Risk Analysis show some promise in their ability to predict radicalisation, as the blue curve is distinct from the red diagonal line (items 1 and 2 in the Figure). The AUC value for processing forward in the Risk Analysis protocol in the Risk Screening is 0.822 ($p < 0.01$ (CI: 0.707, 0.937)). Accordingly, this indicates a good predictive validity. The AUC value for the In-Depth Risk Analysis' ability to predict radicalisation is also good (0.838, $p < 0.01$ (CI: 0.740, 0.937).

When we examine the predictive validity of the Risk Screening and In-Depth Risk Analysis with regards to the risk specification of the VERA-2R (violent extremism), we find that both components have better predictive validity than the VERA-2R (Risk Screening; 0.672, $p < 0.05$ (CI: 0.530, 0.814), In-Depth Risk Analysis; 0.644, $p < 0.1$ (CI: 0.501, 0.786)). Both outcomes are still classified as poor, but the value is higher than that of the VERA-2R.

To further examine whether the predictive validity of Radar is due to items regarding the action/orientation component of the FIRE factors (which may also explain the results regarding its predictive validity of violent extremism), we examined whether the Risk Screening and In-Depth Risk Analysis have any predictive validity for

violence. The outcomes suggest that this is not the case. The AUC values for the Risk Screening are 0.573, $p > 0.05$ (CI: 0.445, 0.701), and for the In-Depth Risk Analysis are 0.573, $p > 0.05$ (CI: 0.445, 0.701). The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

0.1 (CI: 0.427, 0.719), and for the In-Depth Risk Analysis, 0.546, $p > 0.1$ (CI: 0.400, 0.692). These results imply that Radar shows poor predictive validity for violence, and supports the work in task 2 that concluded the distinct lack of empirical evidence underpinning these factors. This also supports the outcomes regarding face validity, and the qualitative information that suggests that factors within the action/orientation sector do not provide much input in the decision-making process for Radar.

Released by Department of Home Affairs
under the Freedom of Information Act 1982

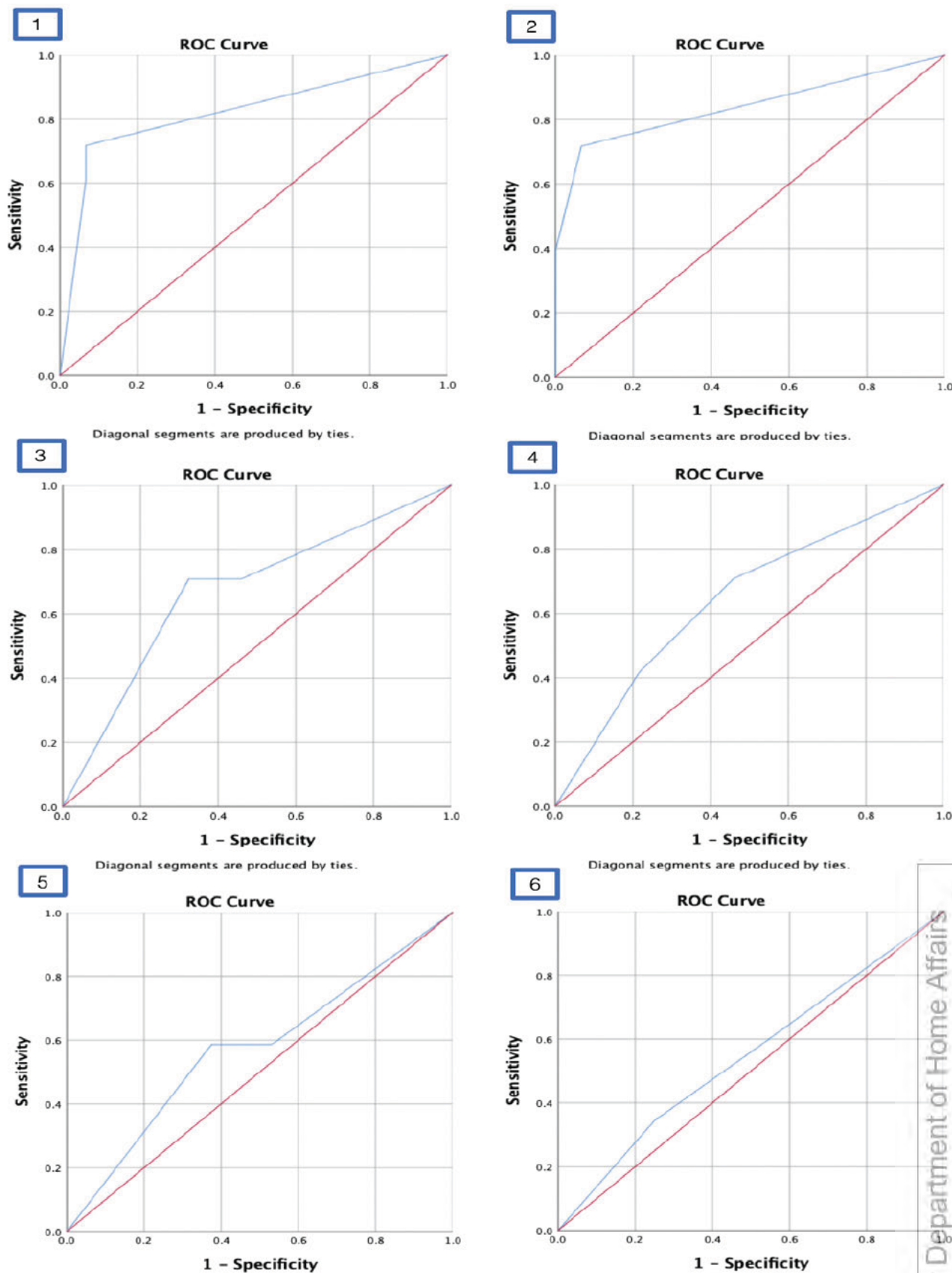


Figure 8 Outputs from Area Under the Curve Analysis for Radar.

The NB - 1 = Output for Risk Screening, 2 = Output for In-Depth Risk Analysis, 3 = Output for Risk Screening with the VERA-2R risk specification, 4 = Output for In-Depth Risk Analysis with the VERA-2R risk specification, 5 = Output for Risk Screening with violence specification, 6 = Output for Risk Screening with violence specification

Aus

Reliability

Inter-Rater Reliability

We used the Krippendorff's alpha test (Krippendorff, 2011) to assess 13 (31.7% of total cases assessed using Radar) cases for IRR of the factors within the Risk Screening and In-Depth Risk Analysis components of Radar. Chosen cases were assessed by either two or three coders. Individual case alpha (α) values are reported in Appendix 5. The average α value for all assessed cases for the Risk Screening was 0.469. This indicates that, much like the VERA-2R, IRR of the Radar Risk Screening is extremely low. The results also demonstrate, that if we were to conduct IRR testing of the Risk Screening across a population sample of cases, there is a 53.8% probability that there would be no IRR,

In the seven cases where raters agreed that the outcome was to proceed to the In-Depth Risk Analysis were then analysed to assess the IRR of this component also. Individual case alpha (α) values are reported in Appendix 5. The average α value for all assessed cases for the In-Depth Risk Analysis was 0.354. If we were to conduct further testing across a population sample of cases, there is a 72.8% probability there would be no IRR (α of below 0.500).

Equity

Ideology

To investigate the equity of Radar, we conducted a three-way log-linear analysis. The likelihood ratio of this model was $X^2(0) = 0, p = 1$. This indicated that the highest-order interaction (assigned risk score x ideology x radicalisation) was significant. However, the effect was influenced by the associations between the assigned risk score and radicalisation ($X^2(2) = 6.088, p < 0.05$) and between radicalisation and ideology ($X^2(7) = 15.756, p < 0.05$). This outcome shows that the assignment of risk was not significantly associated with any one ideology, but it was associated with whether the case under scrutiny was radicalised, supporting the assertions of the authors of Radar.

Due to the issues in the sampling methodology, the complications during case assignment, and the poor IRR of the Risk Analysis protocol, the cases with the manipulated ideologies were not assessed using statistical testing.

Therefore, despite the effort to formulate these assessments, they are only able to offer a supplementary insight.

In those cases where both the non-manipulated and manipulated cases were assessed using Radar, in two

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

instances the ideology affected the overall risk decision. In particular, these outcomes were; a higher risk score for Islamist as compared to Christian Fundamentalist, and a higher risk score for Right-Wing compared to Grievance.

Implementation Burdens

Training

Although the authors of the Radar documentation are not clear in whether training is a requisite to undertake the Radar process, the inclusion of training workbooks in the instrument literature infers that training is preferred. The authors do note the range of individuals who are fit to undertake assessments, citing the requirement of expertise. Therefore, to determine if expertise and training significantly affects the decision-making protocol during the assessment process, the research team undertook a three-way loglinear analysis to compare the likelihood of making correct risk determinations across all three categories of participants.

The analysis revealed that the likelihood ratio of this model was $X^2(0) = 0, p = 1$. This indicated that the highest-order interaction (participant group x case category x risk score) was significant. However, this effect was influenced by the association between the case category and the final risk score ($X^2(14) = 25.210, p < 0.05$), and between the participant group and the case category ($X^2(14) = 27.025, p < 0.05$). This outcome shows that the assignment of correct level of risk was significantly associated with the case categories and not the expertise or training of participants.

Also of note, training and experience does not appear to affect the IRR of Radar. In the Risk Screening, of those cases where coders had the same level of experience, T-Tests revealed that the average α values did not significantly differ from the cases assessed by coders with differing levels of experience, $t(11) = 0.373, p > 0.1$. We were not able to compute this outcome for the In-Depth Risk Analysis due to the lack of cases carried forward for the In-Depth assessment.

Confidence

To further assess whether expertise or training may be necessary for the use of Radar, we examined the confidence of participants during the assessment process. Overall, participants showed moderate to high confidence in the Risk Screening and In-Depth Risk Analysis protocol with regards to how it helped them formulate their final risk

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

decision. Figures 8 and 9 below show the confidence ratings across participant groups. In the Risk Screening component, chi-square analysis revealed that novice participants were less likely¹³⁷ to espouse confidence in the component than experts or trained assessors ($\chi^2(8) = 15.138, p < 0.1$). However, this effect disappeared during the In-Depth Risk Analysis (possibly due to the repetition of the Risk Screening factors, and the learning process that this induces). Overall, across both the Risk Screening and the In-Depth Risk Analysis, trained participants showed the greatest levels of confidence in their scoring, which does indicate that training may be useful for the use of Radar.

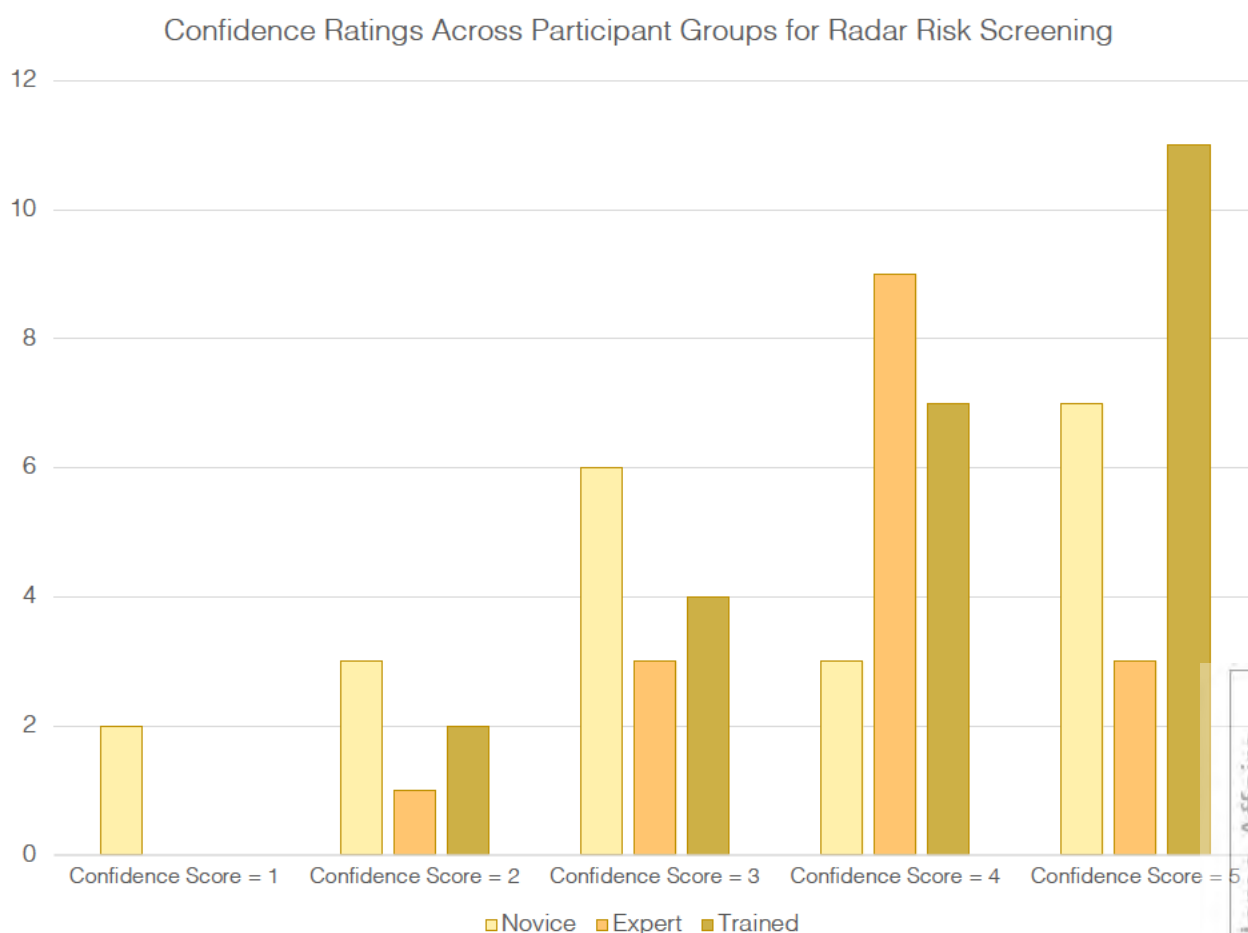


Figure 9 Confidence ratings across participant groups for Radar Risk Screening

¹³⁷ The p value for this analysis bordered on significant (0.057) so it was deemed worth including.

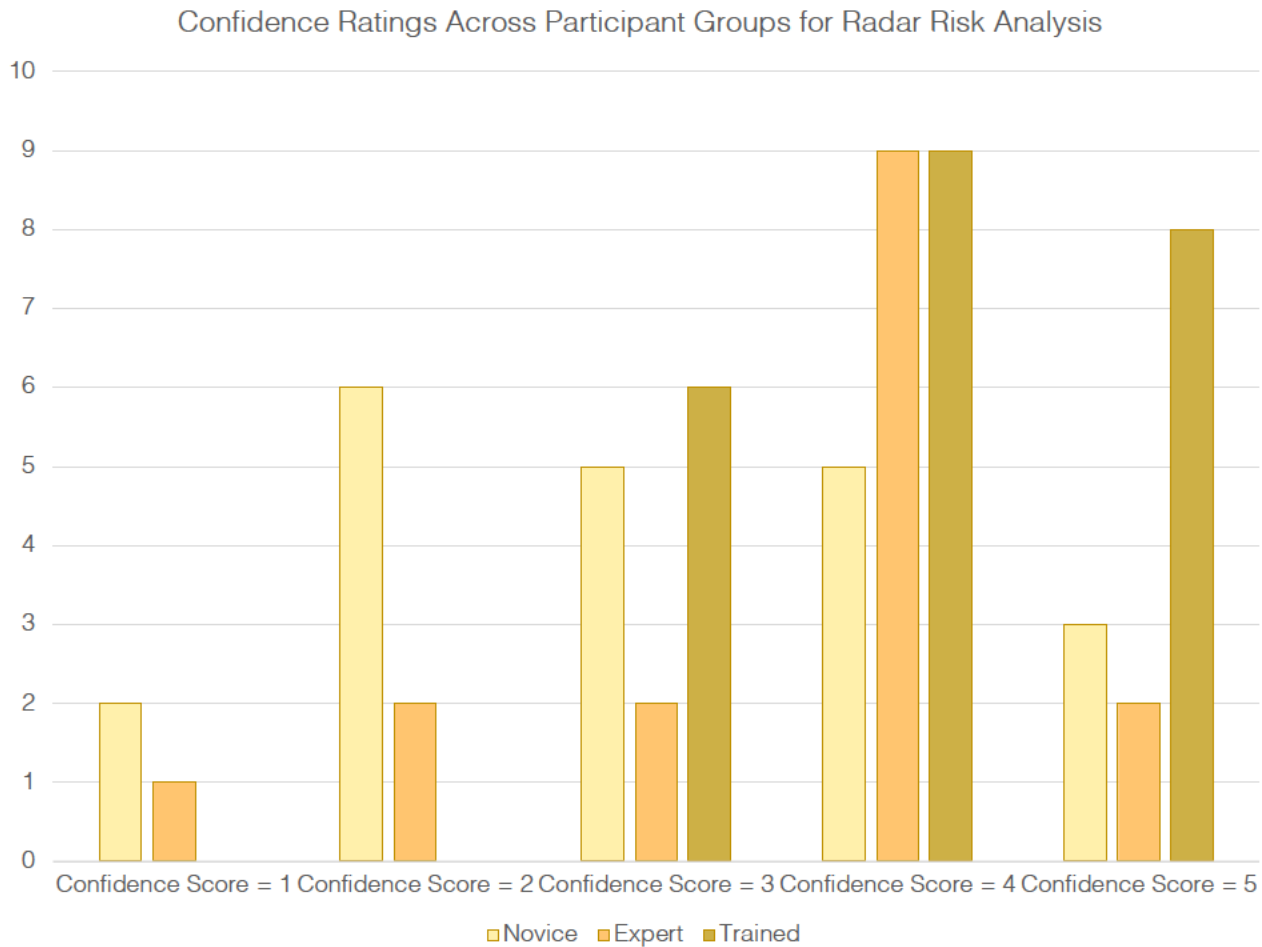


Figure 10 Confidence ratings across participant groups for Radar Risk Analysis

Implications

The research that has been undertaken in this task has been the first to empirically assess the predictive validity of both the VERA-2R and Radar. Authors of both instruments state that ascertaining predictive validity is not feasible or even possible. However, this is not the case. Multiple risk assessment instruments employed in the general violence domain are assessed for their predictive validity, and it has now become an expected process; there are even teams of researchers who are recognised experts in this specific area (Singh, Yang, & Mulvey, 2015). If instruments in the terrorism domain are not assessed for their ability to predict what they specify to predict, any outcomes from such instruments cannot be reasonably expected to represent what an assessor is seeking to identify. This has two major implications: Incorrectly classifying an individual as high-risk and going on to deprive them of their liberty and rights; or incorrectly classifying an individual as low-risk, who subsequently goes on to commit an act of terrorist violence. It is possible, however, that instruments with poor predictive validity are still of use to practitioners. These instruments still have value for helping design risk formulations, management strategies, and scenario planning.

It is worth noting, that although both desirable and possible, predictive validity in instruments in the domain of terrorism risk assessment requires much greater interrogation and assessment. The findings of this research are preliminary. This is the first piece of research that has been performed on these instruments. Instruments in the general violence domain have been assessed multiple times using different population samples. We conducted an evaluation using the leading standard experimental methodology. Whilst the results of this method highlight the findings regarding both predictive validity and inter-rater reliability, these findings were drawn from this specific setting. Outcomes from the instruments in a non-experimental setting would likely differ.

The lack of assessment of reliability of these instruments is also of concern. If an instrument has not been assessed for reliability, we are unable to say with any confidence that assessors are able to consistently come to any firm conclusions in their risk decisions. More specifically, for this analysis, if assessors are not able to reach similar conclusions on the presence of indicators, or on final risk decisions on cases, then the overall process of using these instruments is flawed. In an instrument with poor inter-rater reliability, the interpretation of a factor that is deemed to indicate high-risk by one assessor may be interpreted to be of low-risk by another. These two decisions would lead to different outcomes for the case under scrutiny. Although this research was conducted in an experimental setting, which has implications for how the specific factors in the instruments may have been

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

interpreted, the results highlighted that, particularly for the VERA-2R, training had very little effect on decision-making. This is likely a true representation as to how the training would work in practice. Although assessors must undergo training, the length of time between the training process and the assessment processes will have an impact on the users of the assessors.

The outcomes of the above analyses paint both instruments in a poor light. The outcomes suggest that the VERA-2R has both extremely poor predictive validity and IRR. The Risk Analysis protocol of Radar fares better in its ability to predict outcomes, but has comparably poor IRR. However, one potential starting point for improvement lies in the confidence of participants. Overall, participants were more confident when using the VERA-2R as compared to Radar. This may be due to the comparative lack of complexity in the assessment process of the VERA-2R. This is supported by the finding that participants who had received prior training in Radar reported higher confidence when using it.

The identified predictive validity of both instruments must also be clarified on two points: The first is the outcomes of the first two tests of predictive validity (Sensitivity and Specificity and PPV and NPV) are reliant only on the dichotomous outcomes of the risk decisions (low risk and high risk). However, the measurement of risk in SPJ instruments is often ordinal (low, moderate, high). Therefore, the outcomes from these tests do not capture the range of individuals rated as presenting with moderate risk. If we examined the raw numbers for these tests, the Radar Risk Analysis protocol presents with only 1 false positive, which skewed the outcomes highly. This is particularly pertinent when we consider the finding that the novice group displayed lower confidence in the instrument, and this affected their risk ratings (this group submitted more decisions of moderate risk). This lack of confidence was also found in the VERA-2R. Participants who tested this instrument were more likely to report a lower confidence in their decision-making than those who used Radar (we anticipated that this result is skewed by the trained participants who rated confidence as higher in Radar). Therefore, the outcomes of the AUC are a more accurate representation of predictive validity of both instruments.

The second element that must be considered is the qualitative information that participants gave during the assessment process. Participants were asked to provide information on any other aspects of the case vignette that they personally saw as relevant to their decision-making. Many participants used this opportunity to note a number of other variables that they saw as pertinent to an individual's risk score. Participants were also permitted to use

these elements in their formulation of risk (as is practice in the SPJ approach), which indicates that we are not able to fully accept that the predictive validity of the Risk Analysis protocol in Radar is due to the factors within. This was particularly evident when participants would escalate cases through the Risk Screening stage despite the outcome (according to as directed by Radar) to ensure that they fully captured all factors in the case that the Risk Screening element did not capture,

One limitation of the methodology which has also potentially affected the validity of all analytical outcomes for this task was the final sample size of the participants. Due to the anticipated fatigue and access restrictions,¹³⁸ it was not feasible to expect all participants to agree to each complete an assessment of all 60 case vignettes (which would have allowed for a smaller sample of participants). This was taken into consideration, and as noted, following power calculations, all efforts were made to source a sufficient number of participants to allow for the assessment of each vignette by three separate participants. However, due to unforeseen extreme weather conditions that affected the South East coast of Australia, alongside the anticipated rates of attrition, only 30 of the 67 approached participants were able to complete the assessment process. This smaller sample size then affected our ability to run the initially anticipated reliability analysis. It also affected the validity analysis, as a larger participant group would have allowed for a larger pool of assessed cases for analysis. Further work should rectify this by employing a larger sample of participants.

A second issue that came to light during testing was the familiarity that participants had with cases. In a number of instances, participants informed the research team that they were familiar with one or more of the cases they had been randomly assigned. This led to the reallocation of cases to these participants, which affected the overall number of assessments that were conducted across cases, further degrading the sample that could be assessed for validity and reliability. A number of validity analyses were run, to both ensure calibration and discrimination of the instruments was captured, but also to help support outcomes from each analysis. Like all statistical procedures, discrimination and calibration indicators are affected by sample size. As found, a range of the outcomes were not statistically significant, which would normally negate the reporting of such outcomes, however, the patterns in the

¹³⁸ Given that two thirds of participants were employed across a range of roles within CVE and counter terrorism areas, and that 80% of the novice participants also were in full time employment.

validity analyses (the extremely low values of the statistics) present a holistic overview of the predictive validity of these instruments. The further inferential statistical analyses that were run (t-tests, loglinear analyses, chi-square tests) were also employed to supplement the outcomes of the performance indicators. However, it is necessary to note that, despite the patterns in the data, the sample size may have inhibited the outcomes, and further testing is required to paint a more comprehensive picture of the predictive validity of the instruments.

To help counter the sampling issue when testing for reliability, the team were able to perform Krippendorff's alpha testing on a range of cases that were assessed. Krippendorff's alpha is advantageous to the more commonly used Cohen's Kappa statistic, as it allows for the assessment of both dichotomous (for the Radar Risk Screening) and ordinal data, (for the VERA-2R and Radar Risk Analysis component), and also can assess the IRR between two or more assessors (as some cases were assessed by three assessors). These outcomes are not dependent on statistical significance, and they are able to help us infer IRR. However, again, a larger sample size would have enabled the research team to employ the more comprehensive, and more widely accepted ICC measure.

A final issue that is worth noting when interpreting the data is the specification that was used to measure the predictive validity of Radar against. In the documentation, seven different risk specifications are given. This meant that during the design of the risk specification for analysis (the designation of cases fitting the specification), the team faced some challenges assigning cases. The cases that were captured under the risk specification therefore spanned numerous categories, and included violent, non-violent, and at-risk individuals. This may have affected the outcomes for the predictive validity, as more cases were captured in the risk specification than the authors of Radar intend.

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Conclusions

The outcomes of each task in this research project have supplemented the findings of the previous. Task 1 helped the research team identify that the theoretical and empirical evidence underwriting the instruments is of poor quality. Given that both instruments purport to be SPJ instruments, their foundations should be built on a theoretically and empirically strong evidence base. This alone has serious implications for the validity, reliability, and equity of the instruments. As other authors have shown, the quality of empirical research in this field has increased exponentially in the last decade (Schuurman, 2018). Both instruments evaluated here have been published in this timeframe; in particular, the VERA-2R has undergone three iterative developments. However, the cited evidence base for both instruments is not reflective of the high-quality research that is freely available. Both instruments are heavily reliant on theoretical suppositions, secondary citations of statements made in literature reviews, and media articles. What can be inferred from this analysis is that the reviews of the evidence base that were conducted in the development of the instruments were not systematic.

Both instruments also purport to be supplemented by data. The evidence base underwriting Radar is supplemented with a range of data, however, as no methodology is included in the sourced documentation, we are unable to comment on the strength of this data, or how much of it was used in instrument development. The authors of the VERA-2R also make note of a supplementary cohort of data, but again, we are unable to comment on the strength of this, or how much influence it has on the factors.¹³⁹ What is clear however, is that a large proportion of the evidence that underpins these instruments is neither theoretically or empirically valid.

The lack of theoretical and empirical validity also feeds into the conclusion that neither instrument truly follows the recognised and practiced structure of an SPJ instrument. As noted by a range of authors (Hart & Logan, 2011; Logan & Lloyd, 2019; Monahan, 2012, 2018), many of whom are developers of a range of risk assessment instruments that have been empirically validated,¹⁴⁰ SPJ instruments require an empirical foundation. SPJ instruments also follow a strict format that includes a structured process for gathering case information, and

¹³⁹ Despite the inclusion of data in the VERA, and VERA-2, there is no mention of supplementary data to support the 41 extra factors in the VERA-2R.

¹⁴⁰ And who are also practitioners that conduct assessments and are therefore fully aware of the consequences of a lack of predictive validity and reliability.

evaluation of the relevance of factors, scenario planning, risk mitigation scenarios, and the use of all of these elements to formulate a final risk decision. This structure is not present in the VERA-2R or Radar. What the VERA-2R and Radar present is what has been termed in the literature as the SPJ 'lite' approach (Logan and Lloyd, 2019). In these instruments, there is no detailed and supportive process for the full assessment approach, and due to the required timelines of assessments, no requirement for scenario planning and formulation.

SPJ lite instruments are advantageous in some areas, such as the terrorism domain, as they are less time and resource consuming, and can be conducted by those who are not expert risk assessors (McEwan, Bateson, & Strand, 2017). It is noted that reclassifying both instruments' approach as SPJ 'lite' is likely to help clarify the issues with their approach, and is likely more appropriate given their intended user audience. Future versions of the instruments should reflect this change.

However, due to their use by non-experts, a fundamental requirement of SPJ lite instruments is a strong evidence base for factor inclusion. This evidence base can provide the non-expert users with the most appropriate and valid information that will help guide their decision-making process. This is crucial as users are not expected to be experts in either risk assessment or the discipline that they are conducting the assessment in, and are thus expected to be guided by the instrument in their assessments. Guidance which is based on non-empirical evidence is likely to lead to errors in the identification of truly at-risk individuals.

Task 2 further cemented the findings of task 1. The research team undertook the most comprehensive systematic review that has been conducted on the drivers of radicalisation and terrorist behaviour and violence to date. This is a substantial piece of research in its own right. The outcomes of this analysis showed exactly how complex human behaviour is. There is no single pathway towards radicalisation and terrorist behaviour. A finding that some academics (Corner et al., 2019; Gill, 2015; Horgan, 2003; Victoroff, 2005) have long purported, and recently been able to empirically verify. Over 1500 variables were found to be significantly related to a wide range of terrorist ideologies across numerous countries. The systematic review yielded research from all inhabited continents, and the final studies that were included for coding represent the highest quality empirical research in the field. Due to the complexity of human behaviour, only three themes were found to be consistent across the literature; the importance of social networks, being male, and being younger. These findings are not surprising, as the wider criminological research has never wavered on asserting the prevalence of these conditions across delinquency and

criminality. However, neither instrument truly focuses on any of these themes. Both instruments do have sections for social elements, but in both instances, the descriptive guidance for these factors is not explicit and focuses more on the role of ideology and radical behaviour than the importance of social networks.

A further, and much deeper concern is the finding that across both instruments, there is an extant lack of empirical evidence drawn from the wider literature to support the assertions of the authors: in the case of the VERA-2R, seven studies from the systematic review were correctly attributed to factors; and for Radar, only one study. Further to this, scrutiny of the empirically verified variables identified during the systematic review shows that both instruments lack the inclusion of the majority of variables identified as significantly associated with radicalisation and terrorism. The VERA-2R covers 14.2% of variables, and Radar only 5.9%. What we can infer from this finding is that not only was there no systematic process during the literature searches in the development of both instruments, but also that the factors that are included in the instruments are not supported by the empirical evidence base.

The lack of empirical evidence underpinning the instruments has potential serious implications for their validity and reliability. To assess the impact of this issue, the research team designed a series of case vignettes, drawn from both open and closed sources. The inclusion of closed sources is novel to the assessment of risk assessment instruments in this field, and the collaboration between academic research and security agencies was a major factor in the methodological strength of this research.

The statistical analyses that were conducted on the outcomes of the experiments were complex and systematic. They have allowed us to make relatively strong conclusions regarding the validity, reliability, and equity of the instruments. Despite the assertions by the authors of the instruments that predictive validity is not ascertainable, and in the case of the VERA-2R, the purported validity and reliability of the instrument,^{141 142} this research was able to identify that there is, at best, good predictive validity for Radar, and poor predictive validity for the VERA-2R. The results regarding IRR are surprisingly poor. Neither instrument performs well in the consistency of the rating of the factors within. Both of these issues are intrinsically connected to the lack of an empirically valid evidence base. So,

¹⁴¹ That should only be attributed to the iterations of the instrument that have been tested.

¹⁴² Also, note the concerns regarding authorship effects.

although these results may have been degraded by the interruptions to the assessment process, the pattern in outcomes is supported by evidence that was not affected by this problems.

A further area of support for the conclusions of task 4 can be drawn from the qualitative feedback that was received during the experiments. To maintain confidentiality, these outcomes have not been reported in this research, but some major themes are worth reporting. Overall, the assessors of the VERA-2R were far more satisfied with both the process of the assessment, and the factors that are included. Feedback that was received indicated that factors were more intuitive and it was fairly easy to understand the relevance of their inclusion. However, these participants did tire of the number of factors that were included, and often felt that the additional factors were not useful in their overall decision-making process. Assessors of Radar were much less satisfied with the process. Assessors were only required to undertake the Risk Analysis protocol, but the feedback that was received was that it was confusing, overly complicated by the inclusion of a screening component, with no rationale as to why those particular factors were more pertinent, and there was little coherence in the factors that were included. Many Radar assessors did report that they often found the screening component pointless across many case vignettes (particularly the violent cases), and they were highly likely to proceed a case through screening despite a lack of evidence of factors (which is not recommended in the Radar process). An overwhelming majority of Radar assessors also queried the utility of the 'partially' reported option found within both the Risk Screening and Risk Analysis components given that the Risk Analysis Results Form does not require the inclusion of any 'partially' reported indicators. Assessors also expressed confusion as to what constituted a 'partially' reported indicator.

Common feedback across both assessment groups was that the assessors would often use their own knowledge in their final risk decisions, and this was due to the lack of semantic organisation of the instruments, confusion over what behaviours constituted examples of the risk factors and lack of guidance that they felt both instruments gave. This outcome is concerning, as if assessors are likely to not find either instrument integral to guiding their decision-making, then it calls into question the usefulness of the instruments.

As mentioned earlier, the research team was faced with some interruptions to the experiments and endeavoured to overcome the interruptions by ensuring that the analytical processes were altered to reflect the issues arising in data collection. However, the impact of these problems cannot be understated. The conclusions that have been made regarding the findings in task 4, and the feedback that was received requires supplementing by subsequent

research. To move beyond the provisional conclusions from task 4, both instruments will require further assessment by undertaking a second wave of experiments with a larger sample size.

The employment of further assessments of the case vignettes will also afford researchers the ability to accurately assess the feedback from participants, as well as assess the full protocols of these instruments. In the assessments carried out in this research, timeframes for assessment were limited, fatigue was a major factor, and participants often drew attention to the lack of information in case vignettes. When participants were asked what the typical process for such an issue would be, it was repeatedly stated that an assessor would endeavour to gather more information prior to conducting a second assessment. In this research the team were only assessing the factors within the instruments. Further research should move to assess the process of the instruments, particularly in light of the finding that both the VERA-2R and Radar are not the format of instruments (SPJ) that they purport to be.

In evaluation research, there are two main forms of evaluation: Impact evaluations, such as this research, where the focus of the research is to determine what occurs after the implementation of an intervention or program; and process evaluations, which seek to evaluate the implementation of an intervention or program, to help determine the best practices for effective implementation (Lab, 2020). To truly understand 'What Works' in the risk assessment protocols of those suspected of moving towards violent extremism in Australia, this research should be supplemented by a process evaluation. An ideal process evaluation in this space would seek to understand the settings, implementation, and nuances in the assessment processes of the risk assessment instruments.

Released by Department of Home Affairs
under the Freedom of Information Act 1982

Bibliography

- Ægisdóttir, S., White, M.J., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook, R. S., & Rush, J.D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, 34(3), 341-382. DOI: 10.1177/0011000005285875
- Ahmad, A. (2016). The ties that bind and blind: Embeddedness and radicalisation of youth in one Islamist organisation in Pakistan. *The Journal of Development Studies*, 52(1), 5-21. DOI:10.1080/00220388.2015.1075976
- Altman, D., & Bland, J. (1994) Diagnostic tests 3: Receiver operating characteristic plots. *BMJ*, 309(6948), 188.
- Aly, A. (2017). Brothers, believers, brave mujahideen: Focusing attention on the audience of violent jihadist preachers. *Studies in Conflict & Terrorism*, 40(1), 62-76. DOI:10.1080/1057610X.2016.1157407
- Aly, A., & Striegheer, J.L. (2012). Examining the role of religion in radicalization to violent Islamist extremism. *Studies in Conflict & Terrorism*, 35(12), 849-862. DOI:10.1080/1057610X.2012.720243
- Amble, J.C., & Meleagrou-Hitchens, A. (2014) Jihadist radicalization in East Africa: Two case studies. *Studies in Conflict & Terrorism*, 37(6), 523-540.
- Amjad, N., & Wood, A.M. (2009). Identifying and changing the normative beliefs about aggression which lead young Muslim adults to join extremist anti-Semitic groups in Pakistan. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, 35(6), 514-519. DOI:10.1002/ab.20325
- Andrews, D.A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law*, 16(1), 39-55.
- Andrews, D.A., Bonta, J., & Hoge, R. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behaviour*, 17, 19-52.
- Andrews, D.A., Bonta, J., & Wormith, J.S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency*, 52(1), 7-27. DOI:10.1177/0011128705281756
- Atran, S. (2003). Genesis of suicide terrorism. *Science*, 299(5612), 1534-1539. DOI:10.1126/science.1078854

Atran, S. (2016). The devoted actor: unconditional commitment and intractable conflict across cultures. *Current Anthropology*, 57(S13), S192-S203. DOI:10.1086/685495

Atran, S., Sheikh, H., & Gomez, A. (2014). Devoted actors sacrifice for close comrades and sacred cause. *Proceedings of the National Academy of Sciences of the United States of America*, 111(50), 17702-17703. DOI:10.1073/pnas.1420474111

Australian Government. (2015). *Review of the Commonwealth's Counter-Terrorism Machinery*. Canberra, Australia: Commonwealth of Australia.

Australian Government. (2017). *Radar: Countering Violent Extremism (CVE) Intervention Tools: User Guidelines – Version 2*. Canberra, Australia: Commonwealth of Australia.

Australian Government. (2018). *Radar: Countering Violent Extremism (CVE) Intervention Tools training: Participant Workbook – Version 2*. Canberra: Commonwealth of Australia.

Bakker, E. (2006). *Jihadi Terrorists in Europe: Their Characteristics and the Circumstances in which they Joined the Jihad: An Exploratory Study*. Den Haag, Netherlands: Netherlands Institute of International Relations Clingendael.

Bakker, E., & Grol, P. (2015). Motives and considerations of potential foreign fighters from the Netherlands, *ICCT Policy Brief*. The Hague, Netherlands: International Centre for Counter Terrorism, ICCT.

Bandura, A. (1990). Mechanisms of moral disengagement in terrorism. In W. Reich (Ed.), *Origins of Terrorism: Psychologies, Ideologies, Theologies, States of Mind* (pp. 161–191). Cambridge, England: Cambridge University Press.

Bandura, A. (2016). *Moral Disengagement. How People Do Harm and Live with Themselves*. New York, NY: Worth Publishers.

Barrelle, K. (2014). *Pro-Integration: Disengagement and Life after Extremism*. PhD Dissertation, Monash University.

Barrelle, K. (2015). Pro-integration: Disengagement from and life after extremism. *Behavioral Sciences of Terrorism and Political Aggression*, 7(2), 129-142. DOI:10.1080/19434472.2014.988165

s. 47F(1) (2013). *Framework of Factors of Radicalisation and Extremism (FIRE)*. Victoria, Australia: Global Terrorism Research Centre, Monash University.

Basra R., Neumann, P., & Brunner C. (2016). *Criminal Pasts, Terrorist Futures: European Jihadists and the New Crime-Terror Nexus*. London, England: International Centre for the Study of Radicalisation.

Beardsley, N.L., & Beech, A.R. (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. DOI:10.1108/17596591311290713

Bélanger, J.J., Caouette, J., Sharvit, K., & Dugas, M. (2014). The psychology of martyrdom: Making the ultimate sacrifice in the name of a cause. *Journal of Personality and Social Psychology*, 107(3), 494. DOI:10.1037/a0036855

Bhui, K., Silva, M.J., Topciu, R.A., & Jones, E. (2016). Pathways to sympathies for violent protest and terrorism. *The British Journal of Psychiatry*, 209(6), 483-490. DOI:10.1192/bjp.bp.116.185173

Björge, T., & Horgan, J. (2009). *Leaving Terrorism Behind: Individual and Collective Disengagement*. New York, NY: Routledge.

Blair, P.R., Marcus, D.K., & Boccaccini, M. T. (2008). Is there an allegiance effect for assessment instruments? Actuarial risk assessment as an exemplar. *Clinical Psychology: Science and Practice*, 15(4), 346-360. DOI:10.1111/j.1468-2850.2008.00147.x

Blazak, R. (2001). White boys to terrorist men: Target recruitment of Nazi skinheads. *American Behavioral Scientist*, 44(6), 982-1000. DOI:10.1177/00027640121956629

Böckler, N., Hoffmann, J., & Zick, A. (2015). The Frankfurt airport attack: A case study on the radicalization of a lone-actor terrorist. *Journal of Threat Assessment and Management*, 2(3-4), 153–163. DOI:10.1037/tam0000045

Böckler, N., Hoffmann, J., & Meloy, J.R. (2017). “Jihad against the enemies of Allah”: The Berlin Christmas market attack from a threat assessment perspective. *Violence and Gender*, 4(3), 73–80. DOI:10.1089/vio.2017.0040

Bonett, D.G. (2002) Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine*, 21(9), 1331-1335.

Bora, K. (2015). ISIS uses mentally challenged children as suicide bombers; crucifies, buries them alive: UN. *International Business Times*. Retrieved from <http://www.ibtimes.com>

Borum, R. (1996). Improving the clinical practice of violence risk assessment: Technology, guidelines, and training. *American Psychologist*, 51(9), 945–956. DOI:10.1037/0003-066X.51.9.945

Borum, R. (2000). Assessing violence risk among youth. *Journal of Clinical Psychology*, 56(10), 1263-1288. DOI: 10.1002/1097-4679(200010)56:10<1263::AID-JCLP3>3.0.CO;2-D

Borum, R. (2003). Understanding the terrorist mindset. *FBI Law Enforcement Bulletin*, 72(7), 7-10.

Borum, R. (2011). Radicalization into violent extremism I: A review of social science theories. *Journal of Strategic Security*, 4(4), 7-36. DOI:10.5038/1944-0472.4.4.1

Borum, R. (2014). Psychological vulnerabilities and propensities for involvement in violent extremism. *Behavioral Sciences and the Law*, 32(3), 286-305 DOI:10.1002/bsl.2110

Borum, R. (2015). Assessing risk for terrorism involvement. *Journal of Threat Assessment and Management*, 2(2), 63–87. DOI:10.1037/tam0000043

Borum, R., Bartel, P., & Forth, A. (2006). *SAVRY: Professional Manual for Structured Assessment of Violence Risk in Youth*. Lutz, FL: Psychological Assessment Resources Inc.

Borum, R., Fein, R.A., Vossekuil, B., & Berglund, J. (1999). Threat assessment: Defining an approach for evaluating risk of targeted violence. *Behavioral Sciences and the Law*, 17(3), 323-337. DOI:10.1002/(SICI)1099-0798(199907/09)17:3<323:AID-BSL349>3.0.CO;2-G

Borum, R., Fein, R., & Vossekuil, B. (2012). A dimensional approach to analyzing lone offender terrorism. *Aggression and Violent Behavior*, 17(5), 389-396. DOI:10.1016/j.avb.2012.04.003

Borum, R., Otto, R., & Golding, S. (1993). Improving clinical judgment and decision making in forensic evaluation. *Journal of Psychiatry and Law*, 21(1), 35-76. DOI:10.1177/009318539302100104

Bosi, L. (2012). Explaining pathways to armed activism in the Provisional Irish Republican Army, 1969-1972. *Social Science History*, 36(3), 347-390.

Botha, A. (2014). Political socialization and terrorist radicalization among individuals who joined al-Shabaab in Kenya. *Studies in Conflict & Terrorism*, 37(11), 895-919. DOI:10.1080/1057610X.2014.952511

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*, Cambridge, MA: Harvard University Press.

The Campbell Collaboration. (n.d.). *The Campbell Collaboration*. Retrieved from www.campbellcollaboration.org

Cherney, A. & Belton, E. (2019). Evaluating case-managed approaches to counter radicalization and violent extremism: An example of the Proactive Integrated Support Model (PRISM) intervention. *Studies in Conflict & Terrorism*, 0(00) DOI:10.1080/1057610X.2019.1577016

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Conway, M. (2017). Determining the role of the internet in violent extremism and terrorism: Six suggestions for progressing research, *Studies in Conflict & Terrorism*, 40(1), 77-98. DOI:10.1080/1057610X.2016.1157408

cordis.europa.eu. (2020). Safire Project. www.cordis.europa.eu

Corner, E., Bouhana, N., & Gill, P. (2019). The multifinality of vulnerability indicators in lone-actor terrorism. *Psychology, Crime & Law*, 25(2), 111-132. DOI:10.1080/1068316X.2018.1503664

Corner, E., & Gill, P. (2015). A false dichotomy? Mental illness and lone-actor terrorism. *Law and Human Behavior*, 39(1), 23-34. DOI:10.1037/lhb0000102

Corner, E., Gill, P., & Mason, O. (2016). Mental health disorders and the terrorist: A research note probing selection effects and disorder prevalence. *Studies in Conflict & Terrorism*, 39(6), 560-568.

Cornish, P. (2008), *Terrorism, radicalisation and the internet*. England: Chatham House.

Cornish, P., Hughes, R., & Livingstone, D. (2009). *Cyberspace and the National Security of the United Kingdom: Threats and Responses*. London, United Kingdom: Chatham House.

Costello, M., & Hawdon, J. (2018). Who are the online extremists among us? Sociodemographic characteristics, social networking, and online experiences of those who produce online hate materials. *Violence and Gender*, 5(1), 55-60. DOI:10.1089/vio.2017.0048

Dalgaard-Nielsen, A. (2013). Promoting exit from violent extremism: Themes and approaches. *Studies in Conflict & Terrorism*, 36(2), 99-115. DOI:10.1080/1057610X.2013.747073

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899), 1668-1674. DOI: 10.1126/science.2648573

Dawson, L., & Amarasingam, A. (2017). Talking to foreign fighters: Insights into the motivations for Hijrah to Syria and Iraq. *Studies in Conflict & Terrorism*, 40(3), 191-210. DOI:10.1080/1057610X.2016.1274216

Douglas, K.S. (2019). Evaluating and managing risk for violence using structured professional judgment. In A. Day, C. Hollin, & D. Polaschek (Eds.), *The Wiley International Handbook of Correctional Psychology* (pp. 427-445). Hoboken, NJ: Wiley & Sons Ltd.

Douglas, K.S., & Belfrage, H. (2014). Interrater reliability and concurrent validity of the HCR-20 Version 3. *International Journal of Forensic Mental Health*, 13(2), 130-139. DOI: [10.1080/14999013.2014.908429](https://doi.org/10.1080/14999013.2014.908429)

Douglas, K.S., Cox, D.N., & Webster, C.D. (1999). Violence risk assessment: Science and practice. *Legal and Criminological Psychology*, 4(2), 149-184. DOI: 10.1348/135532599167824

Douglas, K.S., Guy, L.S., Edens, J.F., Boer, D.P., & Hamilton, J. (2007). The personality assessment inventory as a proxy for the psychopathy checklist-revisited: Testing the incremental validity and cross-sample robustness of the antisocial features scale. *Assessment*, 14, 255-269.

Douglas, K.S., Hart, S.D., Webster, C.D., & Belfrage, H. (2013). *HCR-20V3: Assessing Risk for Violence: User Guide*. Burnaby, British Columbia: Mental Health, Law, and Policy Institute, Simon Fraser University.

Douglas, K.S., & Kropp, P.R. (2002). A prevention-based paradigm for violence risk assessment: Clinical and research implications. *Criminal Justice and Behavior*, 29(5), 617-658. DOI: [10.1177/009385402236735](https://doi.org/10.1177/009385402236735)

Douglas, K.S., & Reeves, K.A. (2010). Historical-Clinical-Risk Management-20 (HCR-20) Violence Risk Assessment Scheme: Rationale, application, and empirical overview. In R.K. Otto & K.S. Douglas (Eds.), *International*

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

Perspectives on Forensic Mental Health: Handbook of Violence Risk Assessment (pp.147–185). New York, NY: Routledge/Taylor & Francis Group

Douglas, K.S., & Skeem, J.L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and the Law*, 11(3), 347-383.

Douglas, K.S., Yeomans, M., & Boer, D.P. (2005). Comparative validity analysis of multiple measures of violence risk in a sample of criminal offenders. *Criminal Justice and Behavior*, 32(5), 479–510. DOI: [10.1177/0093854805278411](https://doi.org/10.1177/0093854805278411)

Doyle, M., & Dolan, M. (2002). Violence risk assessment: combining actuarial and clinical information to structure clinical judgements for the formulation and management of risk. *Journal of Psychiatric and Mental Health Nursing*, 9(6), 649-657. DOI: 10.1046/j.1365-2850.2002.00535.x

Doyle, M., Dolan, M., & McGovern, J. (2002). The validity of North American risk assessment tools in predicting inpatient violent behaviour in England. *Legal & Criminological Psychology*, 7, 141-154.

Dvoksin, J. A., & Heilbrun, K. (2001). Risk assessment and release decision-making: Toward resolving the great debate. *Journal of the American Academy of Psychiatry and the Law*, 29, 6-10.

Europol. (2017). *European Union Terrorism Situation and Trend Report (TE-SAT)*. The Hague, Netherlands: European Union Agency for Law Enforcement Cooperation.

Fair, C. (2015). Explaining support for sectarian terrorism in Pakistan: Piety, maslak and sharia. *Religions* 6(4), 1137-1167. DOI:10.3390/rel6041137

Farrington, D. P. (1986). Age and crime. *Crime and Justice*, 7, 189-250. DOI:10.1086/449114

Fein, R. A., & Vossekuil, B. (1998). *Protective Intelligence and Threat Assessment Investigations: A Guide for State and Local Law Enforcement Officials*. Washington, DC: US Department of Justice, Office of Justice Programs, National Institute of Justice.

Ferguson, N., Burgess, M., & Hollywood, I. (2008). Crossing the rubicon: Deciding to become a paramilitary in Northern Ireland. *International Journal of Conflict and Violence*, 2(1), 130-137. DOI: 10.4119/ijcv-2779

- Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*. 2nd ed. New York, NY: Wiley.
- Florez-Morris, M. (2007). Joining guerrilla groups in Colombia: Individual motivations and processes for entering a violent organization. *Studies in Conflict & Terrorism*, 30(7), 615-634.
- Francis, M.D.M. (2016). Why the sacred is a better resource than religion for understanding terrorism. *Terrorism and Political Violence*, 28(5), 912-927. DOI:10.1080/09546553.2014.976625
- Geraghty, K.A., & Woodhams, J. (2015). The predictive validity of risk assessment tools for female offenders: A systematic review. *Aggression and Violent Behavior*, 21, 25-38.
- Gerber, M.M., Carvacho, H., & González, R. (2016). Development and validation of a scale of support for violence in the context of intergroup conflict (SVIC): The case of violence perpetrated by Mapuche people and the police in Chile. *International Journal of Intercultural Relations*, 51, 61–68.
- Gill, P. (2015). *Lone Actor Terrorists: A Behavioral Analysis*. Abingdon, Oxon: Routledge.
- Gill, P., & Comer, E. (2017). There and back again: The study of mental disorder and terrorist involvement. *American Psychologist*, 72(3), 231–241. DOI:10.1037/amp0000090
- Gill, P., Comer, 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*, 0(0), 1-18.
- Gill, P., Comer, E., Thornton, A., & Conway, M. (2015). *What are the Roles of the Internet in Terrorism? Measuring Online Behaviours of Convicted UK Terrorists*. [online]: VOXPol Network of Excellence.
- Gill, P., Horgan, J., & Deckert, P. (2014). Bombing alone: Tracing the motivations and antecedent behaviors of lone-actor terrorists. *Journal of Forensic Sciences*, 59(2), 425-435. DOI:10.1111/1556-4029.12312
- Global Terrorism Research Centre. (2012). *Implementing the Radicalisation Indicators Model (TRIM)*. Melbourne: Monash University
- Gottfredson, M. R., & Hirschi, T. (1990). *A General Theory of Crime*. Stanford: Stanford University Press.

Gøtzsche-Astrup, O. (2018). The time for causal designs: Review and evaluation of empirical support for mechanisms of political radicalisation. *Aggression and Violent Behavior*, 39, 90–99. DOI:10.1016/j.avb.2018.02.003

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, 2, 293-323. DOI: 10.1037/1076-8971.2.2.293

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1), 19-30. DOI:10.1037/1040-3590.12.1.19

Gruenewald, J., Chermak, S., & Freilich, J.D. (2013). Distinguishing “loner” attacks from other domestic extremist violence: A comparison of far-right homicide incident and offender characteristics. *Criminology & Public Policy*, 12(1), 65-91. DOI:10.1111/1745-9133.12008

Gutheil, T., & Simon, R. (2004). Avoiding bias in expert testimony. *Psychiatric Annals*, 34, 260-270. DOI: 10.3928/0048-5713-20040401-06

Guy, L.S. (2008). *Performance Indicators of the Structured Professional Judgement Approach for Assessing Risk for Violence to Others: A Meta-Analytic Survey*. Unpublished doctoral dissertation. Burnaby, Canada: Department of Psychology, Simon Fraser University.

Hafez, M.M. (2006). Moral agents, immoral violence: Mechanisms of moral disengagement in Palestinian suicide terrorism. In J. Victoroff (Ed), *Tangled Roots: Social and Psychological Factors in the Genesis of Terrorism* (pp. 292-307). Amsterdam, Netherlands: IOS Press.

Haidt, J. (2003) The moral emotions. In R. J. Davidson, K.R. Scherer & H.H. Goldsmith (Eds.) *Handbook of Affective Sciences* (pp. 852-870). Oxford, United Kingdom: Oxford University Press.

Hanson, R.K. (2005). Twenty years of progress in violence risk assessment. *Journal of Interpersonal Violence*, 20(2), 212-217.

Harris, G.T., Rice, M.E., Quinsey, V.L., Lalumière, M.L., Boer, D., & Lang, C. (2003). A multisite comparison of actuarial risk instruments for sex offenders. *Psychological Assessment*, 15(3), 413–425.

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

Harris-Hogan, S., Barrelle, K., & Zammit, A. (2016). What is countering violent extremism? Exploring CVE policy and practice in Australia. *Behavioral Sciences of Terrorism & Political Aggression*, 8(1), 6-24.

Hart, S.D. (1998). The role of psychopathy in assessing risk for violence: Conceptual and methodological issues. *Legal and Criminological Psychology*, 3, 121–137.

Hart, S.D., Douglas, K.S., & Guy, L.S. (2016). The structured professional judgement approach to violence risk assessment: Origins, nature, and advances. In D.P. Boer, A.R. Beech, T. Ward,

L.A. Craig, M. Rettenberger, L.E. Marshall, & W.L. Marshall (Eds.), *The Wiley Handbook on the Theories, Assessment, and Treatment of Sexual Offending* (p. 643–666). Chichester, United Kingdom: Wiley-Blackwell.

Hart, S. D., & Logan, C. (2011). Formulation of violence risk using evidence-based assessments: The structured professional judgment approach. In P. Sturmey & M. McMurrin (Eds.), *Forensic Case Formulation*. Chichester, United Kingdom: Wiley-Blackwell

Hart, S.D., Michie, C., & Cooke, D.J. (2007). Precision of actuarial risk assessment instruments: Evaluating the “margins of error” of group v. individual predictions of violence. *The British Journal of Psychiatry*, 190(S49), s60-s65. DOI:10.1192/bjp.190.5.s60

Hart, S., Sturmey, P., Logan, C., & McMurrin, M. (2011). Forensic case formulation. *International Journal of Forensic Mental Health*, 10(2), 118–126, DOI: [10.1080/14999013.2011.577137](https://doi.org/10.1080/14999013.2011.577137)

Hart, S.D., Webster, C.D., & Menzies, R.J. (1993). A note on portraying the accuracy of violence predictions. *Law & Human Behavior*, 17, 695–700. DOI: [10.1007/BF01044690](https://doi.org/10.1007/BF01044690).

Heilbrun, K., Philipson, J., Berman, L., & Warren, J. (1999). Risk communication: Clinicians' reported approaches and perceived values. *Journal of the American Academy of Psychiatry and the Law*, 27(3), 397-406.

Hegghammer, T. (2013). Should I stay or should I go? Explaining variation in western jihadists' choice between domestic and foreign fighting. *American Political Science Review*, 107(1), 1-15.

DOI:10.1017/S0003055412000615

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.

DOI:10.1177/2066220317749140

Hilton, N.Z., Harris, G.T., & Rice, M.E. (2006). Sixty-six years of research on the clinical versus actuarial prediction of violence. *The Counseling Psychologist*, 34(3), 400-409. DOI: 10.1177/0011000005285877

Hofmann, D.C. (2016). The influence of charismatic authority on operational strategies and attack outcomes of terrorist groups. *Journal of Strategic Security*, 9(2), 14-44. DOI:10.5038/1944-0472.9.2.1486

Hogg, M.A., & Adelman, J. (2013). Uncertainty–identity theory: Extreme groups, radical behavior, and authoritarian leadership. *Journal of Social Issues*, 69(3), 436–454. DOI:10.1111/josi.12023

Holtfreter, K., & Cupp, R. (2007). Gender and risk assessment: The empirical status of the LSI-R for women. *Journal of Contemporary Criminal Justice*, 23(4), 363–382. DOI: 10.1177/1043986207309436

Horgan, J. (2003). The search for the terrorist personality. In A. Silke (Ed.), *Terrorists, victims and society: Psychological perspectives on terrorism and its consequences* (pp. 3-27). Chichester, England: Wiley.

Horgan, J.G. (2004). The case for firsthand research. In A. Silke (Ed.) *Research on Terrorism: Trends, Achievements, and Failures*. London: Routledge.

Horgan, J.G. (2008). From profiles to pathways and roots to routes: Perspectives from psychology on radicalization into terrorism. *The ANNALS of the American Academy of Political and Social Science*, 618(1), 80-94.

DOI:10.1177/0002716208317539

Horgan, J.G. (2016). A call to arms: The need for more psychological research on terrorism, *Social Psychological Review*, 18(1), 25-28.

Horgan, J.G., Taylor, M., Bloom, M., & Winter, C. (2017). From cubs to lions: A six stage model of child socialization into the Islamic state. *Studies in Conflict & Terrorism*, 40(7), 645-664, DOI:10.1080/1057610X.2016.1221252

Jacques, K., & Taylor, P.J. (2008). Male and female suicide bombers: Different sexes, different reasons? *Studies in Conflict & Terrorism*, 31(4), 304-326.

Kemshall, H., & Pritchard, J. (1996). *Good Practice in Risk Assessment and Risk Management* (Vol. 3). London, United Kingdom: Jessica Kingsley Publishers.

Kimhi, S., & Even, S. (2004). Who are the Palestinian suicide bombers? *Terrorism and Political Violence*, 16(4), 815–840.

Koehler, A. (2019). Violence and terrorism from the far-right: Policy option to counter an elusive threat. ICCT Policy Brief. The Hague, The Netherlands: International Centre for Counter-Terrorism. DOI: 10.19165/2019.2.02

Kohlberg, L. (1984). *The Psychology of Moral Development: The Nature and Validity of Moral Stages*. New York, NY: Harper and Row.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. Retrieved from <https://repository.upenn.edu>

Kropp, P.R., & Hart, S.D. (2000). The Spousal Assault Risk Assessment (SARA) guide: Reliability and validity in adult male offenders. *Law and Human Behavior*, 24, 101–118.

Krueger, A. B., & Malečková, J. (2003). Education, poverty and terrorism: Is there a causal connection? *Journal of Economic Perspectives*, 17(4), 119-144. DOI:10.1257/089533003772034925

Kruglanski, A.W., 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. DOI: 10.1111/j.1467-9221.2009.00698.x

Kruglanski, A.W., 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), 217–230. DOI:10.1037/amp0000091

Lab, S. P. (2020). *Crime prevention: Approaches, practices, and evaluations*. New York, NY: Routledge

LaFree, G. (2013). Lone-offender terrorists. *Criminology & Public Policy*, 12, 59–62.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Lankford, A. (2014a). Evidence that suicide terrorists are suicidal: Challenges and empirical predictions. *Behavioral and Brain Sciences*, 37(4), 380-393. DOI:10.1017/S0140525X13003609

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

Lankford, A. (2014b). Précis of the myth of martyrdom: What really drives suicide bombers, rampage shooters, and other self-destructive killers. *Behavioral and Brain Sciences*, 37(4), 351-362. DOI:10.1017/S0140525X13001581

Lennings, C.J., Amon, K.L., Brummert, H., & Lennings, N.J. (2010). Grooming for terror: The internet and young people. *Psychiatry, Psychology, and Law*, 17(3), 424-437. DOI:10.1080/13218710903566979

Liht, J., Conway, L.G. III., Savage, S., White, W., & O'Neill, K.A. (2011). Religious fundamentalism: An empirically derived construct and measurement scale. *Archive for the Psychology of Religion*, 33, 1-25.

Litwack, T. R. (2001). Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy and Law*, 7(2), 409-443. DOI:10.1037/1076-8971.7.2.409

Lloyd, M. (2019). *Extremism Risk Assessment: A Directory*. Lancaster, United Kingdom: Centre for Research and Evidence on Security Threats.

Logan, C. & Lloyd, M. (2019). Violent extremism: A comparison of approaches to assessing and managing risk. *Legal and Criminological Psychology*, 24(1), 141-161. DOI: 10.1111/lcrp.12140

Lowenkamp, C.T., Latessa, E.J., & Holsinger, A.M. (2006) The risk principle in action: What have we learned from 13,676 offenders and 97 correctional programs? *Crime & Delinquency*, 52(1), 77-93.

Ma, H.K. (2013). The moral development of the child: An integrated model. *Frontiers in Public Health*, 1(57), 1-18. DOI:10.3389/fpubh.2013.00057

Ma, H.K. (2017). The development of altruism with special reference to human relationships: A 10-stage theory. *Frontiers in Public Health*, 5(271), 1-14. DOI:10.3389/fpubh.2017.00271

McCauley, C., & Moskalenko, S. (2017). Understanding political radicalization: The two-pyramids model. *American Psychologist*, 72(3), 205-216. DOI:10.1037/amp0000062

McEwan, T., Bateson, S., & Strand, S. (2017). Improving police risk assessment and management of family violence through a collaboration between law enforcement, forensic mental health and academia. *Journal of Criminological Research, Policy and Practice*, 3(2), 119-131. DOI: 10.1108/JCRPP-01-2017-0004

- McGregor, I., Hayes, J., & Prentice, M. (2015). Motivation for aggressive religious radicalization: Goal regulation theory and a personality \times threat \times affordance hypothesis. *Frontiers in Psychology*, 6, 1-18. DOI: 10.3389/fpsyg.2015.01325
- McNiel, D.E., Fordwood, S.R., Weaver, C.M., Chaimberlain, J.R., Hall, S.E., & Binder, R.L. (2008a). Effects of training on suicide risk assessment. *Psychiatric Services*, 59, 1462-1465.
- McNiel, D.E., Fordwood, S.R., Weaver, C.M., Chaimberlain, J.R., Hall, S.E., & Binder, R.L. (2008b). Impact of clinical training on violence risk assessment. *The American Journal of Psychiatry*, 165(2), 195-200.
- Meloy, J.R., Roshdi, K., Glaz-Ocik, J., & Hoffmann, J. (2015). Investigating the individual terrorist in Europe. *Journal of Threat Assessment and Management*, 2(3-4), 140–152. DOI:10.1037/tam0000036
- Melton, G.B., Petrila, J., Poythress, N.G., & Slobogin, C. (1997). *Psychological Evaluations for the Courts*, 2nd ed. New York, NY: Guildford Press.
- Merari, A. (2010). *Driven to Death: Psychological and Social Aspects of Suicide Terrorism*. Oxford, United Kingdom: Oxford University Press.
- Moffit, T.E. (1993). Adolescence-limited and life-course persistent antisocial behaviour: A developmental taxonomy. *Psychological Review*, 100, 674 –701. DOI: 10.1037/0033-295X.100.4.674
- Moghaddam, F.M. (2005). The staircase to terrorism: A psychological exploration. *American Psychologist*, 60(2), 161–169. DOI:10.1037/0003-066X.60.2.161
- Monahan, J. (1981). The clinical prediction of violent behavior. *Crime & Delinquency Issues: A Monograph Series*, ADM 81-921, 134.
- Monahan, J. (1993). Limiting therapist exposure to Tarasoff liability: Guidelines for risk containment. *American Psychologist*, 48, 242-250. DOI: [10.1037/0003-066X.48.3.242](https://doi.org/10.1037/0003-066X.48.3.242)
- Monahan, J. (2000). Violence risk assessment: Scientific validity and evidentiary admissibility. *Washington and Lee Law Review*, 57(3), 901-918.

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

Monahan, J. (2018). Predictions of violence. In T. Grisso & S. L. Brodsky (Eds.), *The roots of modern psychology and law: A narrative history* (pp. 143-157). New York, NY: Oxford University Press.

Monahan, J., & Skeem, J. (2014). The evolution of violence risk assessment. *CNS Spectrums*, 19(5), 419-424. DOI:10.1017/S1092852914000145

Monahan, J., & Steadman, H.J. (1994). Toward a rejuvenation of risk assessment research. In J. Monahan & H.J. Steadman (Eds.) *Violence and Mental Disorder: Developments in Risk Assessment*. Chicago, IL: University of Chicago Press.

Monahan, J., Steadman, H.J., Silver, E., Appelbaum, P.S., Robbins, P.C., Mulvey, E.P., & Banks, S. (2001). *Rethinking Risk Assessment: The MacArthur Study of Mental Disorder and Violence*. New York, NY: Oxford University Press.

Moskalenko, S., & McCauley, C. (2009). Measuring political mobilization: The distinction between activism and radicalism. *Terrorism and Political Violence*, 21(2), 239-260. DOI:10.1080/09546550902765508

Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, 62(4), 783. DOI:10.1037/0022-006X.62.4.783

Murrie, D.C., Boccaccini, M.T., Guamera, L.A., & Rufino, K.A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science*, 24(10), 1889–1897.

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.

DOI:10.1177/0022427817699035

Odgers, C.L., Moretti, M.M., & Reppucci, N.D. (2005). Examining the science and practice of violence risk assessment with female adolescents. *Law and Human Behavior*, 29(1), 7-27. DOI:10.1007/s10979-005-1397-2

Otto, R. (1992). The prediction of dangerous behavior: A review and analysis of “second generation” research. *Forensic Reports*, 5, 103-133.

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

- Özderdem, A. & Podder, S. (2012). Grassroots and rebellion: A study on the future of the Moro struggle in Mindanao, Philippines. *Civil Wars*, 14(4), 521-545. DOI:10.1080/13698249.2012.740202
- Pedersen, L., Rasmussen, K., & Elsass, P. (2010). Risk assessment: The value of structured professional judgments. *International Journal of Forensic Mental Health*, 9(2), 74-81. DOI: 10.1080/14999013.2010.499556
- Piaget, J. (1926). *The Language and Thought of the Child*. New York, NY: Harcourt Brace.
- Podder, S. (2011). Child soldier recruitment in the Liberian civil wars: Individual motivations and rebel group tactics. In A. Özderdem and S. Podder (Eds.) *Child Soldiers: From Recruitment to Reintegration* (pp. 50-77). Basingstoke, Hampshire: Palgrave Macmillan.
- Post, J., Sprinzak, E., & Denny, L. (2003). The terrorists in their own words: Interviews with 35 incarcerated Middle Eastern terrorists. *Terrorism and Political Violence*, 15(1), 171-184. DOI:10.1080/09546550312331293007
- Precht, T. (2007). *Home Grown Terrorism and Islamist Radicalisation in Europe: From Conversion to Terrorism*. Copenhagen, Denmark: Danish Ministry of Justice.
- Pressman, D. E. (2009). *Risk Assessment Decisions for Violent Political Extremism*. Ottawa, Canada: Canadian Centre for Security and Intelligence Services
- Pressman, D. E. (2016). The complex dynamic causality of violent extremism: Applications of the VERA-2 risk assessment methods to CVE initiatives. In A. J. Masys (Ed.) *Disaster Forensics: Understanding Root Cause and Complex Causality* (pp. 249-269). Switzerland: Springer International Publishing.
- Pressman, D.E., Duits, N., Rinne, T., & Flockton, J.S. (2018). *VERA-2R Violent Extremism Risk Assessment – Version 2 Revised: A Structured Professional Judgement Approach*. Utrecht, Netherlands: Netherlands Institute of Psychiatry and Psychology.
- Pressman, D.E., & Flockton, J. (2012). Calibrating risk for violent political extremists and terrorists: The VERA 2 structured assessment. *The British Journal of Forensic Practice*, 14(4), 237-251.
DOI:10.1108/14636641211283057

Putra, I.E., & Sukabdi, Z.A. (2014). Can Islamic fundamentalism relate to nonviolent support? The role of certain conditions in moderating the effect of Islamic fundamentalism on supporting acts of terrorism. *Peace and Conflict: Journal of Peace Psychology*, 20(4), 583-589. DOI:10.1037/pac0000060

QRS International. (nd.) Retrieved from <https://www.qsrinternational.com/>.

Quinsey, V., Harris, G., Rice, M., & Cormier, C. (1998). *Violent Offenders: Appraising and Managing Risk*. Washington, DC: American Psychological Association.

Reinares, F. (2004). Who are the terrorists? Analyzing changes in sociological profile among members of ETA. *Studies in Conflict & Terrorism*, 27(6), 465-488. DOI:10.1080/10576100490519741

Reitano, T., Clarke, C., & Adal, L. (2017). *Examining the Nexus between Organised Crime and Terrorism and its Implications for EU Programming*. CT-MORSE Consortium.

Reynolds, K., & Miles, H.L. (2009). The effect of training on the quality of HCR-20 violence risk assessments in forensic secure services. *The Journal of Forensic Psychiatry & Psychology*, 20(3), 473-480. DOI:10.1080/14789940802638366

Reynolds, S.C., & Hafez, M.M. (2019). Social network analysis of German foreign fighters in Syria and Iraq. *Terrorism and Political Violence*, 31(4), 661-686. DOI:10.1080/09546553.2016.1272456

Rice, M.E., & Harris, G.T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology*, 63(5), 737-748.

Rosenau, W., Espach, R., Ortiz, R.D., & Herrera, N. (2014). Why they join, why they fight, and why they leave: Learning from Colombia's database of demobilized militants. *Terrorism and Political Violence*, 26(2), 277-285. DOI: 10.1080/09546553.2012.700658

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of personality and social psychology*, 76(4), 574-586. DOI:10.1037/0022-3514.76.4.574

RTI International. (2017). *Countering Violent Extremism: The Use of Assessment Tools for Measuring Violence Risk*. Washington DC: Department of Homeland Security.

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

- Rubin, R.E. (2017). *Foundations of Library and Information Science*. American Library Association.
- Ryan, T. (2016) *An Examination of the Interrater Reliability and Concurrent Validity of the Spousal Assault Risk Assessment Guide – Version 3 (SARA-V3)*. Burnaby, Canada: Department of Psychology, Simon Fraser University.
- Sageman, M. (2004). *Understanding Terror Networks*. Philadelphia, PA: University of Pennsylvania Press.
- Sageman, M. (2008). *Leaderless Jihad: Terror Networks in the Twenty-First Century*. Philadelphia, PA: University of Pennsylvania Press.
- Sarma, K.M. (2017). Risk assessment and the prevention of radicalization from nonviolence into terrorism. *American Psychologist*, 72(3), 278. DOI: 10.1037/amp0000121
- Saucier, G., Akers, L. G., Shen-Miller, S., Knežević, G., & Stankov, L. (2009). Patterns of thinking in militant extremism. *Perspectives on psychological science*, 4(3), 256-271. DOI:10.1111/j.1745-6924.2009.01123.x
- Scarcella, A., Page, R., & Furtado, V. (2016). Terrorism, radicalisation, extremism, authoritarianism and fundamentalism: A systematic review of the quality and psychometric properties of assessments. *PLoS ONE*, 11(12), e1066947. DOI:10.1371/journal.pone.0166947
- Schafer, J.A., Mullins, C.W., & Box, S. (2014). Awakenings: The emergence of white supremacist ideologies. *Deviant Behavior*, 35(3), 173-196. DOI:10.1080/01639625.2013.834755
- Schuurman, B. (2018). Research on terrorism, 2007-2016: A review of data, methods, and authorship. *Terrorism and Political Violence*, 1-16. DOI:10.1080/09546553.2018.1439023
- Schuurman, B., & Eijkman, Q. (2015). Indicators of terrorist intent and capability: Tools for threat assessment. *Dynamics of Asymmetric Conflict*, 8(3), 215-231. DOI:10.1080/17467586.2015.1040426
- Schwalbe, C.S. (2008). A meta-analysis of juvenile justice risk assessment instruments: Predictive validity by gender. *Criminal Justice and Behavior*, 35(11), 1367-1381. DOI:10.1177/0093854808324377
- Schwartz, S.J., Dunkel, C.S., & Waterman, A.S. (2009). Terrorism: An identity theory perspective. *Studies in Conflict & Terrorism*, 32(6), 537-559, DOI:10.1080/10576100902888453
- Schultz, K.S., & Whitney, D.J. (2005). *Measurement Theory in Action*. Thousand Oaks, CA: Sage.

Shultz, K.S., Whitney, D.J., & Zickar, M. J. (2013). *Measurement Theory in Action: Case Studies and Exercises*. New York, NY: Routledge.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428.

Silber, M.D., & Bhatt, A. (2007). *Radicalization in the West: The Homegrown Threat*. New York, NY: New York Police Department.

Singh, J.P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law*, 31(1), 8-22. DOI:10.1002/bsl.2052

Singh, J.P., Desmarais, S.L., & Van Dorn, R.A. (2013). Measurement of predictive validity in violence risk assessment studies: A second-order systematic review. *Behavioral Sciences & the Law*, 31(1), 55-73. DOI:10.1002/bsl.2053

Singh, J.P., Serper, M., Reinharth, J., & Fazel, S. (2011). Structured assessment of violence risk in schizophrenia and other psychiatric disorders: A systematic review of the validity, reliability, and item content of 10 available instruments. *Schizophrenia Bulletin*, 37, 899–912. DOI: 10.1093/schbul/sbr093

Singh, J. P., Yang, S., & Mulvey, E. P. (2015). Reporting guidance for violence risk assessment predictive validity studies: the RAGEE Statement. *Law and Human Behavior*, 39(1), 15-22.

Skeem, J.L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science*, 20(1), 38-42. DOI: 10.1177/0963721410397271

Slovic, P., & Monahan, J. (1995). Probability, danger, and coercion: A study of risk perception and decision making in mental health law. *Law and Human Behavior*, 19, 49-65.

Slovic, P., Monahan, J., & MacGregor, D. G. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior*, 24(3), 271-296.

Speckhard, A., & Ahkmedova, K. (2006). The making of a martyr: Chechen suicide terrorism. *Studies in Conflict & Terrorism*, 29(5), 429-492. DOI:10.1080/10576100600698550

The content enclosed in this document reflects the results of a comprehensive analytical assessment of the validity, reliability, and equity of Terrorism based Risk Assessment Instruments used in Australia. This research was conducted by Dr Emily Corner and Dr Helen Taylor at the Australian National University and does not necessarily reflect the views of the Australian Government or the Department of Home Affairs.

- Stankov, L., Saucier, G., & Knežević, G. (2010). Militant extremist mind-set: Provivence, vile world, and divine power. *Psychological Assessment*, 22(1), 70. DOI:10.1037/a0016925
- Storey, J. E., Gibas, A. L., Reeves, K. A., & Hart, S. D. (2011). Evaluation of a violence risk (threat) assessment training program for police and other criminal justice professionals. *Criminal Justice and Behavior*, 38(6), 554-564. DOI:10.1177/0093854811403123
- Taylor, M., & Horgan, J.G. (2006). A conceptual framework for addressing psychological process in the development of the terrorist. *Terrorism and Political Violence*, 18(4), 585-601. DOI:10.1080/09546550600897413
- Timmins, K. L., Evans, L., & Tully, R. J. (2018). Inter-rater reliability of the Short-Term Assessment of Risk and Treatability (START). *The Journal of Forensic Psychiatry & Psychology*, 29(6), 968-988.
- Trochim, W. M. (2001). *Research Methods Knowledge Base* (Vol. 2). Cincinnati, OH: Atomic Dog Publishing.
- Victoroff, J. (2005). The mind of a terrorist: A review and critique of psychological approaches. *Journal of Conflict Resolution*, 49(1), 3-42. DOI:10.1177/0022002704272040
- Vincent, G.M., Guy, L.S., Fusco, S.L., & Gershenson, B.G. (2012). Field reliability of the SAVRY with juvenile probation officers: Implications for training. *Law and Human Behavior*, 36(3):225-236.
- de Vogel, V., & de Ruiter, C. (2005). The HCR-20 in personality disordered female offenders: A comparison with a matched sample of males. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, 12(3), 226-240. DOI:10.1002/cpp.452
- Von Behr, I., Reding, A., Edwards, C., & Gribbon, L. (2013). *Radicalisation in the Digital Era: The Use of the Internet in 15 Cases of Terrorism and Extremism*. Belgium, Brussels: RAND Europe.
- Webber, D., & Kruglanski, A.W. (2018). The social psychological makings of a terrorist. *Current Opinion in Psychology*, 19, 131-134. DOI:10.1016/j.copsyc.2017.03.024
- Weenink, A. (2015). Behavioral problems and disorders among radicals in police files. *Perspectives on Terrorism*, 9(2), 17-33.

Weggemans, D., Bakker, E., & Grol, P. (2014). Who are they and why do they go? The radicalisation and preparatory processes of Dutch jihadist foreign fighters. *Perspectives on Terrorism*, 8(4), 100-110.

Weimann, G. (2004). *www.terror.net: How Modern Terrorism Uses the Internet*. United States Institute of Peace.

White, R. W. (1993). *Provisional Irish Republicans: An Oral and Interpretive History*. London, United Kingdom: Greenwood Press.

Wiktorowicz, Q. (2005). *Radical Islam Rising: Muslim Extremism in the West*. Oxford, United Kingdom: Rowman & Littlefield.

Williams, G. (2011). A decade of Australian anti-terror laws. *Melbourne University Law Review*, 35(3), 1136-1176.

Winter, C., & Spaaij, R. (2018). The evolving threat of lone-actor terrorism. Inside Story. Retrieved from www.insidestory.org.au

Wong, S.C.P., & Gordon, A. (2006). The validity and reliability of the Violence Risk Scale: A treatment-friendly violence risk assessment tool. *Psychology, Public Policy, and Law*, 12(3), 279-309. DOI:10.1037/1076-8971.12.3.279

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 1 – Performance Indicators Commonly used to Assess Predictive Validity

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

Sensitivity (Sens)

$$\frac{TP}{TP + FN}$$

The proportion of those who engaged in the act who were judged to be at high risk

Specificity (Spec)

$$\frac{TN}{TN + FP}$$

The proportion of those who did not engage in the act who were judged to be at low risk

Positive predictive value (PPV)

$$\frac{TP}{TP + FP}$$

The proportion of those judged to be at high risk who did go on to engage in the act

Negative predictive value (NPV)

$$\frac{TN}{TN + FN}$$

The proportion of those judged to be at low risk who did not go on to engage in the act

Number needed to detain (NND)

$$\frac{1}{PPV}$$

The number of individuals judged to be at high risk who would need to be detained to prevent a single act

Number safely discharged (NSD)

$$\left(\frac{1}{1 - NPV} \right) - 1$$

The number of individuals judged to be at low risk who could be discharged before a single act

Diagnostic odds ratio (DOR)

$$\frac{TP \times TN}{FP \times FN}$$

The ratio of the odds of a high-risk classification in those who engaged in act to the odds of a high-risk classification in those who did not

Logistic odds ratio (DOR)

$$\frac{e^{a+bx}}{e^{a+bx} + 1}$$

The ratio of the odds of a lower risk classification in those who did not engage in act to the odds of a higher classification in those who did

Point-biserial correlation coefficient (r_{pb})

APPENDIX 1 – Performance Indicators Commonly used to Assess Predictive Validity

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

$$\frac{\bar{x}_{RiskClassification} - \bar{x}_{NotRiskClassification}}{\sqrt{\frac{\sum_{x=Min}^{x=Max} (x - \bar{x})^2}{n_{Total}}}} \times \sqrt{\frac{n_{Risk} \times n_{NotRisk}}{n_{Total}^2 - n_{Total}}}$$

The direction and strength of the association between risk classification and having engaged in act or not

Area under the curve (AUC)

$$\frac{1}{2} \sum_{T_{Min}}^{T_{Max}} (Sens_{T_{i-1}} + Sens_{T_i}) \times (Spec_{T_{i-1}} - Spec_{T_i})$$

The probability that a randomly selected individual who engaged in act received a higher risk classification than a randomly selected individual who did not

NB: TP, number of true positives; FN, number of false negatives; TN, number of true negatives; FP, number of false positives; T, cut-off threshold; n, number of participants; \bar{x} , mean score; x, individual score; X, independent variable; a, constant; b,

APPENDIX 2 – Comparison of factors across the iterations of the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

VERA	VERA – 2	VERA – 2R
Attitude	Beliefs and Attitudes	Belief, attitudes, and Ideology
A1 – Attachment to ideology justifying violence	BA1 – Commitment to ideology justifying violence	BA1 – Commitment to ideology <i>that justifies</i> violence
A2 – Significant view of injustice and grievances	BA2 – Perceived victim of injustice and grievances	BA2 – <i>Perceived grievances</i> and/or injustice
A3 – Identified target of injustice	NOT TAKEN FORWARD	
A4 – Dehumanization of identified target	BA3 – Dehumanization/demonization of identified targets of injustice	BA3 – Dehumanization of <i>designated</i> targets associated with injustice
A5 – Internalized martyrdom to die for cause	NOT TAKEN FORWARD	
A6 – Rejection of society and values/Alienation	BA4 – Rejection of democratic society and values	BA4 – Rejection of democratic society and values
A7 – High level anger, frustration, persecution*	BA5 – Feelings of hate, frustration, persecution, alienation	BA5 – <i>Expressed emotions in response to perceived injustice</i>
A8 – Need for group bonding and belonging	NOT TAKEN FORWARD	
A9 – Identity problems	BA6 – Hostility to national collective identity	BA6 – Hostility to <i>national identity</i>
A10 – Low empathy for those outside own group*	BA7 – Lack of empathy, understanding outside own group	BA7 – Lack of empathy <i>and</i> understanding <i>for those</i> outside one's own group
Context	Context and Intent	Social Context and Intention
C1 – User of extremist websites	CI1 – Seeker, consumer, developer of violent extremist materials	SCI1 – Seeker, <i>user</i> or developer of violent extremist materials
C2 – Community of support for violent action	CI2 – Identification of target (person, place, group) for attack	SCI2 – <i>Target for attack identified (person, group, location)</i>

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 2 – Comparison of factors across the iterations of the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

C3 – Direct contact with violent extremists	CI3 – Personal contact with violent extremists	SCI3 – Personal contact with violent extremists (<i>informal or social context</i>)
C4 – Anger at political decisions, actions of country	CI4 – Anger and expressed intent to act violently	SCI4 – Expressed <i>intention to commit act of violent extremism</i>
NOT IN INSTRUMENT	CI5 – Willingness to die for a cause	SCI5 – <i>Expressed</i> willingness <i>and/or preparation</i> to die for a cause <i>or belief</i>
NOT IN INSTRUMENT	CI6 – Expressed intent to plan, prepare violent action	SCI6 – <i>Planning/preparation of acts of violent extremism</i>
NOT IN INSTRUMENT	CI7 – Susceptible to influence, authority, indoctrination (If a leader, specify, and rate as high risk)	SCI7 – <i>Susceptibility</i> to influence, control or indoctrination
Historical	History and Capacity	History, Action, and Capacity
H1 – Early exposure to violence in home*	HC1 – Early exposure to pro-violence militant ideology	HAC1 – Early exposure to violence-promoting , militant ideology
H2 – Family involvement in violent action*	HC2 – Network (family, friends) involved in violent action	HAC2 – Network <i>of</i> family and friends involved in violent extremism
H3 – Prior criminal violence*	HC3 – Prior criminal history of violence	HAC3 – Violent criminal history
H4 – State-sponsored military, paramilitary training	HC4 – Tactical, paramilitary, explosives training	HAC4 – Strategic , paramilitary and/or explosives training
H5 – Travel abroad: non-state sponsored training/fighting	HC5 – Extremist ideological training	HAC5 – Training in extremist ideology <i>in own country or abroad</i>
NOT IN INSTRUMENT	HC6 – Access to funds, resources, organizational skills	HAC6 – Organizational skills and access to funding and sources of help
H6 – Glorification of violent action	MOVED INTO COMMITMENT AND MOTIVATION	
	Commitment and Motivation	Commitment and Motivation
IN HISTORICAL	CM1 – Glorification of violent action	CM1 – <i>Motivated by perceived religious obligation and/or glorification</i>

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 2 – Comparison of factors across the iterations of the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

NOT IN INSTRUMENT	CM2 – Driven by criminal opportunism	CM2 – <i>Motivated</i> by criminal opportunism
NOT IN INSTRUMENT	CM3 – Commitment to group, group ideology	CM3 – <i>Motivated by camaraderie, group belonging</i>
NOT IN INSTRUMENT	CM4 – Driven by moral imperative, moral superiority	CM4 – <i>Motivated</i> by moral <i>obligation</i> , moral superiority
NOT IN INSTRUMENT	CM5 – Driven by excitement, adventure	CM5 – <i>Motivated</i> by excitement and adventure
NOT IN INSTRUMENT	<i>Additional considerations such as being coerced to act, search for significance or status</i>	CM6 – <i>Forced participation in violent extremism</i>
NOT IN INSTRUMENT	NOT IN INSTRUMENT	CM7 – <i>Motivated by acquisition of status</i>
NOT IN INSTRUMENT	NOT IN INSTRUMENT	CM8 – <i>Motivated by search for meaning and significance in life</i>
Protective	Protective	Protective/Risk Mitigating
P1 – Shift in ideology	P1 – Re-interpretation of ideology less rigid, absolute	P1 – <i>Reinterpretation of the ideology</i>
P2 – Rejection of violence to obtain goals ^{&*}	P2 – Rejection of violence to obtain goals	P2 – Rejection of violence <i>as a means to achieve goals</i>
P3 – Change of vision of enemy	P3 – Change of vision of enemy	P3 – Change in <i>concept</i> of the enemy
P4 – Interest in constructive political involvement	P4 – Involvement with non-violent, de-radicalization, offence related programs	P4 – <i>Participant in programs against violent extremism</i>
P5 – Significant other/community support ^{&*}	P5 – Community support for non-violence	P5 – <i>Support from community</i> for non-violence
NOT IN INSTRUMENT	P6 – Family support for non-violence	P6 – <i>Support from family</i> and friends for non-violence

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 2 – Comparison of factors across the iterations of the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

Additional		
D1 – Sex Gender)	NOT IN INSTRUMENT	NOT IN INSTRUMENT
D2 – Married	NOT IN INSTRUMENT	NOT IN INSTRUMENT
D3 – Age	NOT IN INSTRUMENT	NOT IN INSTRUMENT
NOT IN INSTRUMENT	NOT IN INSTRUMENT	CH1 – <i>Client of juvenile justice system/ convicted for non-violent offence(s)</i>
NOT IN INSTRUMENT	NOT IN INSTRUMENT	CH2 – <i>Non-compliance with conditions or supervision</i>
NOT IN INSTRUMENT	NOT IN INSTRUMENT	PH1 – <i>History of family violence</i>
NOT IN INSTRUMENT	NOT IN INSTRUMENT	PH2 – <i>Problematic upbringing and/or placed in juvenile care</i>
NOT IN INSTRUMENT	NOT IN INSTRUMENT	PH3 – <i>Problems at school or work</i>
NOT IN INSTRUMENT	NOT IN INSTRUMENT	MD1 – <i>Personality disorder</i>
NOT IN INSTRUMENT	NOT IN INSTRUMENT	MD2 – <i>Depressive disorder</i>
NOT IN INSTRUMENT	NOT IN INSTRUMENT	MD3 – <i>Psychotic or schizophrenic disorder</i>
NOT IN INSTRUMENT	NOT IN INSTRUMENT	MD4 – <i>Autism spectrum disorder</i>
NOT IN INSTRUMENT	NOT IN INSTRUMENT	MD5 – <i>Post-traumatic stress disorder</i>
NOT IN INSTRUMENT	NOT IN INSTRUMENT	MD6 – <i>Substance use disorder</i>

NB: Bolded and italicised text indicates a change from the previous iteration of the instrument, & - Indicator drawn from the HCR-20, * - Indicator drawn from SAVRY.

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

Attached Statement	Indicator	Citation	Empirical	Relevant	Element Cited
BA1 - Commitment to Ideology that Justifies the use of violence					
“Ideology, beliefs and attitudes are generally cited as reasons for terrorist attacks and suicide terrorism”	ONLY CITED ONCE	Kruglanski et al. (2009)	Yes – analysis of media extracts (ONLY SUICIDE ATTACKS)	Fair – No mention of attitudes	Findings of study are specifically religion, less ideology/beliefs
“Ideology is the most frequently cited reason for terrorist attacks and suicide attacks in video messages left behind by so-called martyrs to describe their motives”				Fair - More specifically religion, less ideology	Findings of study are specifically religion, less ideology/beliefs
“They are sometimes also called the ‘emotional fuel’ for extremist attacks. Furthermore, they provide the justification for the use of violence to achieve the desired goals.”	NO CITATION				
“An analysis of written declarations by militant-extremist terrorist groups active in seven different regions of the world (Europe, Middle East, Sub-Saharan Africa, East Asia, South Asia, Latin America and North America) led to the identification of sixteen common themes”	IN BA1, BA2, BA3, BA4, BA5, SCI4, SCI5	Saucier et al. (2009)	Yes – Open source data	Yes	In Overview of Findings of Terrorism
“Analysts at intelligence services and academic researchers both acknowledge the fundamental	IN BA1, HAC3	Taher (2011)	No – Daily Mail article	No – Does not mention ideology as fundamental for justifying violence	N/A – Not applicable to statement

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

role that ideology plays in justifying terrorist violence”	IN BA1, HAC3	Mullins (2011)	Fair – Blurb of wider project	Fair – Does not explicitly mention ‘fundamental’ role of ideology in justifying violence.	Citing Porter & Kebbell (2010) in Literature Review
“conducted a study on lone actor terrorists and found that 83 percent of the perpetrators attracted the attention of their social environment in the run up to their deed. In 79 percent of the cases, people around the perpetrator recognized that the future attacker was devoted to an extreme ideology.”	IN BA1, SCI6, MDInt, MD3	Gill (2015)	Yes – 111 cases	74% aware of grievance, 69% aware of extremist ideology. No – Statistics are actually drawn from Gill, Horgan, & Deckert (2014) (119 as opposed to 111) as reads: “In 82.4% of the cases, other people were aware of the individual’s grievance that spurred the terrorist plot, and in 79%, other individuals were aware of the individual’s commitment to a specific extremist ideology.”	N/A – Not applicable to statement
BA2 – Perceived Grievances and/or Injustice					
“Grievances are generally seen as risk factors for terrorist acts and form an important element in the assessment of the risk of terrorism”	ONLY CITED ONCE	Monahan (2012)	No – Literature Review	Yes	Citing Atran (2003), McCauley & Moskaleiko (2011). Nibbett & Cohen (1996), Pressman & Flockton (2010)

Released by Department of Home Affairs under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

"The perception of injustice and dominating grievances play a role in radicalization"	IN BA2, HAC3	Silber & Bhatt (2007)	Yes – Five case studies	Fair – No mention of injustice, grievance mentioned in terms of 'solving global political grievance'	N/A – Not applicable to statement
"Grievances may be specific to an individual or applicable to the group or country with which the individual identifies"	IN BA2, SCI5	Atran (2003)	No – Literature Review	Fair – "Massive retaliation further increases people's sense of victimization and readiness to behave according to organizational doctrines"	Evidence from a Reuter's news story
"[The perception of injustice and dominating grievances...] support the final steps an individual takes towards a terrorist act"	IN BA2, SCI3, SCI6, HAC2	McCauley & Moskalenko (2011)	Yes – Case studies from Russian Terrorist Group in 1880s.	No – No mention of grievance specificity	N/A – Not applicable to statement
"and can arise from actual or imagined circumstances."	NO CITATION				
"Foreign fighters perceive that they and the groups with which they identify, are more deprived, oppressed or persecuted than they should be, and that something should be done about it"	IN BA2, SCI5, CM1, CM3	Dawson & Amarasingam (2017)	Yes – 25 Interviews with Foreign Fighters	Fair – Exact quote is "perceive that they, and the groups they identify with, are more deprived than they should be, and that something should be done about it" No mention of oppression or persecution.	Concluding remarks from Review of Weggemans, Bakker, and Grol (2014), and Bakker and Grol (2015)

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

“In their messages to the outside world, terrorists generally cite these grievances and perceived injustices.”	NO CITATION				
“This corresponds to an overarching theme for terrorists, namely that “one’s own group is tragically obstructed”	IN BA1, BA2, BA3, BA4, BA5, SCI4, SCI5	Saucier et al. (2009)	Yes – Open source data	Yes	Yes
BA3 - Dehumanization of Designated Targets Associated with Injustice					
“Dehumanization is classified as one of the eight phases in the process of genocide because this creates a sense of detachment from the intended target and debases the persons thus designated.”	NO CITATION				
“When one has the feeling of not being personally responsible for violent behavior one can progress more quickly to such behavior:	ONLY CITED ONCE	Bandura (1990)	No – Literature Review	No – No mention of progression to violence more quickly	N/A – Not applicable to statement
	ONLY CITED ONCE	Bandura (1999)	No – Literature Review	No – Mention of escalation of level of punitiveness, but progression towards in terms of speed	N/A – Not applicable to statement
	In BA3, CM4	Bandura (2016)	No – Literature Review and Theoretical Development Book	No – no mention of progression to violence more quickly – does state that people act more cruelly due to diffusion of responsibility	N/A – Not applicable to statement

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

	ONLY CITED ONCE	Lieber, Efreom-Lieber, & Rate (2010)	Yes – Interviews (Conference Paper)	Fair – Theory underlying study – but not explicitly mentioned points within citation	Citing Bandura (1990) in Literature Review
	ONLY CITED ONCE	Bandura (1990)	No – Literature Review	No – No mention of progression to violence more quickly	N/A – Not applicable to statement
“Signs of dehumanization can be found in the way a subject describes his or her enemies or victims. The denial of human qualities of potential human targets can result in a total lack of identification with the victims. As a result, there is reduced moral inhibition to refrain from the use of violence”	ONLY CITED ONCE	Bandura (1999)	No – Literature Review	Yes	Citing Keen (1986) and Kelman (1973)
	In BA3, CM4	Bandura (2016)	No – Literature Review and Theoretical Development Book	Yes	Citing Bandura et al. (1975)
	ONLY CITED ONCE	Lieber, Efreom-Lieber, & Rate (2010)	Yes – Interviews (Conference Paper)	Fair – Theory underlying study – but not explicitly mentioned points within citation	Citing Bandura (1990) in Literature Review
“When terrorists dehumanize the target audience, the carnage is likely to be greater”	IN BA3, CM1	Klein (2016)	Yes – Data from Piazza	Yes	In Literature Review (Opinion of author - No citation given)
“When victims are denied human qualities or when it is emphasized that these opponents have no right to live, this leads to the psychological acceptance of the prospect of killing another human being”	ONLY CITED ONCE	Stanton (1996)	No – Lecture	Fair – Mentions dehumanization, but not psychological acceptance for killing (also note, work is on genocide and governments)	N/A – Not applicable to statement
“Dehumanization was applied in the Second World War to facilitate the Nazi policy of the extermination camps. It was used by the Hutu majority	NO CITATION				

Released by Department of Home Affairs
under the Freedom of Information Act 1992

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

against the Tutsi minority in Rwanda during the genocide of 1994.”					
“Dehumanization of opponents is described by Saucier (2009) and colleagues as the ‘universal’ theme of terrorists.”	IN BA1, BA2, BA3, BA4, BA5, SCI4, SCI5	Saucier et al. (2009)	Yes – Open source data	No – Not described as the ‘universal’ theme. Actually, the research suggests that there is no universal theme.	Dehumanization or demonizing of opponents is theme 14 of 16 in findings. Not ‘universal’.
“Recent experiments show that social connections and group dynamics make dehumanization possible”	IN BA3, SCI3	Waytz & Epley (2012)	Yes – Experiments on students	Yes (not terrorist related research)	In Findings of Item
“In such cases, individual responsibility becomes shared group responsibility.”	NO CITATION				
BA4 - Rejection of Democratic Society and Values					
“The rejection of a democratic society and its values is often cited as motivation for violent actions and declarations by militant extremists. It is a universal theme of militant-extremist or terrorist groups all over the world.”	NO CITATION				
“This is often expressed as “the current civil government is illegitimate”, or that “the modern world is disastrous”	IN BA1, BA2, BA3, BA4, BA5, SCI4, SCI5	Saucier et al. (2009)	Yes – Open source data	Fair However, quote “the current civil government is illegitimate” is incorrect – no mention of “the current” and “the modern world is disastrous” is also incorrect – actual	In Findings of Item

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

				quote is “modernity is disastrous” or “the modern world is a disaster”	
“The leaders of the G8 summit in 2010 declared their opposition to this type of dismissive proposition. They emphasized the advantages of a democratic society, its values, freedoms and the legitimacy of the constitutional state”	ONLY CITED ONCE	G-8 Summit (2010)	No – Political Statement	Yes	Opinion in Political Statement, not Empirical
Terrorist groups promote the boycott of democratic elections and espouse right-wing or left-wing standpoints that reject democratically legitimized decisions.	NO CITATION				
They also refuse to recognize internationally accepted legal governments.	NO CITATION				
Some terrorist groups reject all ‘man-made’ laws and accept only those laws they regard as deriving from a higher religious authority.	NO CITATION				
BA5 - Expressed Emotions in Response to Perceived Injustice					
“Moral emotions are used to condemn others”	ONLY CITED ONCE	Haidt (2003)	No – Literature Review	Fair – Mention of contempt, anger, and disgust as “other condemning family” of emotions	Drawn from Rozin, Lowery, Imada, and Haidt (1999)
“Anger is one of the most common and powerful emotions associated with	NO CITATION				

Released by Department of Home Affairs under the Freedom of Information Act 1982

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

political violence and terrorism.”					
“This emotion occurs in response to a circumstance or event in order to regain control and/ or remove the reason for anger. Anger is directed at those considered to be causing it”	ONLY CITED ONCE	Cottam et al. (2016)	No – Textbook	Yes	Citations drawn from are Frijda (1986), Izard (1977)
“According to Huddy et al. (2007) anger is intensified when the responsible party is perceived to be unjust and illegitimate. Feelings of anger, hate, frustration and persecution can promote alienation from a society. These feelings can be exacerbated by anticipated or experienced disasters in the past or present.”	ONLY CITED ONCE	Huddy et al. (2007)	Yes – Telephone Survey	No – Study is regarding consequences to threat perception following terrorist attacks	N/A – Not applicable to statement
“The occupation of territories that are believed to belong to another group has led to terrorist activities in the Middle East, Africa and Europe. These terrorist actions are often the result of feelings of frustration and moral indignation. Strong emotions can facilitate the willingness to use catastrophic violence, even against ‘fellow citizens’”	NO CITATION				
“Vidino (2016) argued that many ISIS-supporters who have planned attacks in the United States and/or left to join ISIS in Syria, had	IN BA5, SCI1, SCI3, CMIntr	Vidino (2016)	No – Statement made to Senate	No – Only one case presented mentioned hate.	N/A – Not applicable to statement

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

developed feelings of hatred towards America. They found justification for the ISIS doctrine not only in the struggle in Syria but also as a result of injustices in their domestic environments.”

Document does state “They all represent, in the conspiratorial worldview they have adopted, proof of the evil nature of America and every other entity or idea IS opposes.”

“Hatred, anger and a sense of persecution about foreign-policy decisions have been identified as motivating factors in terrorist actions in the United States, Canada, Somalia, the United Kingdom and Australia. The murder of Corporal Cirillo by Michael Zehalf-Bibeau in October 2014 in Canada, was prompted by the deployment of Canadian troops in Afghanistan and other Islamic countries. Corporal Cirillo was a ceremonial soldier guarding the National War Memorial in Ottawa.”

NO CITATION

“The act known as ‘catastrophizing’, focusing on disasters in the past, present or future, is an established stimulus for violent acts”

IN BA1, BA2, BA3, BA4, BA5, SCI4, SCI5

Saucier et al. (2009)

Yes – Open source data

No – Although states that catastrophizing is one of the 16 facets examined, does not specifically mention that catastrophizing is “an established stimulus for violent acts”

N/A – Not applicable to statement

BA6 - Hostility to National Identity

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

<p>“Post (2005, 2007) states that a collective identity can be a motivating factor in terrorist acts.”</p>	<p>IN BA6, BA7, SCI5, SCI7, HAC1, CMIntr</p>	<p>Post (2005)</p>	<p>No – Statement</p>	<p>Fair – Mention of collective identity as “provides the greatest analytical power in understanding this complex phenomenon” but not as a specific motivating factor</p>	<p>Opinion of author – Not Empirical</p>
	<p>IN BA6, CMIntr</p>	<p>Post (2007)</p>	<p>Yes – Case Studies</p>	<p>Fair – Mention of collective identity “provides the most powerful lens through which to understand terrorist psychology and behavior”, not a specific motivating factor</p>	<p>Conclusions from Post (2005)</p>
<p>“It has been observed that terrorists subsume their individual identity to the collective identity of their organization or group.”</p>	<p>NO CITATION</p>				
<p>“They believe that whatever serves the group, organization or network is of primary importance.”</p>	<p>NO CITATION</p>				
<p>“This group loyalty can lead to hostility towards the national collective identity. Such hostility arises when personal and collective values are irreconcilable with national values. In such cases, feelings of loyalty are shifted to the</p>	<p>NO CITATION</p>				

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

group instead of to the country of residence.”

BA7 - Lack of Empathy and Understanding for those Outside One's Own Group

“Members of a group may only feel empathy or compassion for other members of this group and for a limited and selected group of other people.”

NO CITATION

“Terrorists, in contrast to psychopaths, are generally social beings and feel attachment to like-minded group members”

IN BA7, HAC2, HAC3, CM3 Sageman (2004)

Yes – 172 Cases (open source)

Yes (note findings are specific to Salafi Jihad)

In Theory and Findings

IN BA7, SCI3, SCI7, HAC2, HAC3, CMIntr Sageman (2008)

Yes – Case Studies

Fair – Focus of study is the social networks

In Theory and Findings

“Empathy for one's own group and attaching value to membership of this group is identified by experts as a motivating factor for terrorism”

IN BA7, HAC2, HAC3, CM3 Sageman (2004)

Yes – 172 Cases (open source)

No – No mention of empathy or attaching value to membership as a motivating factor

N/A – Not applicable to statement

IN BA7, SCI3, SCI7, HAC2, HAC3, CMIntr Sageman (2008)

Yes - Case Studies

No – No mention of empathy or attaching value to membership as a motivating factor

N/A – Not applicable to statement

IN BA6, BA7, SCI5, SCI7, HAC1, CMIntr Post (2005)

No – Statement

No – No mention of empathy for the group or attaching value to membership

N/A – Not applicable to statement

“They are able to show what can be referred to as ‘selective empathy’

IN BA7, PRMIntr Pressman (2009)

No – Literature Review

Yes

Within Comparative Analysis between Traits of Psychopathy, Dissocial PD, Antisocial PD, and Violent Extremism. Unclear Where

Released by Department of Home Affairs
Under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

				traits within VE cohort were drawn from (unclear empirical support).	
“They may feel sympathy and empathy for their ‘brothers’ and their own group, but not for those who fall outside this group”	NO CITATION				
“concluded that a commonly shared and increasingly stimulated ‘us-versus-them’ perspective was the single overlapping factor among their analysis of a diverse sample of 26 suspected Salafi-Jihadists in The Netherlands	IN BA7, CM2, CH1	van Leyenhorst & Andreas (2017)	Yes – Case Studies drawn from Police files.	Not specifically – Actually states underlying motivations differed, but does give overview of four narratives which infer the ‘us vs them’ perspective.	Narratives and inferences within Results.
SCI1 – Seeker, User or Developer of Violent Extremist Materials					
“The degree to which someone actively seeks or makes use of violent extremist materials or develops such materials influences the assessed risk for radicalization to violence and violent extremism, although research questions remain”	ONLY CITED ONCE	Aly (2017)	No – Literature Review	Yes – “Further questions need to be asked about how and why the terrorist message becomes meaningful to certain people, their media habits and the contexts within which they consume, engage with, and decode the terrorists’ message.” (element that shows ‘some questions remain’)	Statement by author – non-empirical and no citations backing statement. Opinion based.
	ONLY CITED ONCE	Conway (2017)	No – Literature Review	Yes – ‘No proven connection’ (element	Statement by author – non-empirical and no

Released by Department of Home Affairs under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

				that shows 'some questions remain')	citations backing statement – Opinion based.
"It is known that those vulnerable to radicalization visit extremist websites both for general information and for instructions about bomb making"	ONLY CITED ONCE	Aly (2017)	No – Literature Review	Yes – Covers information, but not specifically bomb making.	Sentence starting "It is known" is citing work of Weimann and von Knop (2008)
	ONLY CITED ONCE	Conway (2017)	No – Literature Review	No – No reference to vulnerability or why individuals visit websites. Piece is a review of the state of research. Actually, is distinct in stating that the knowledge base is poor, not "well known"	N/A – Not applicable to statement
	IN SCI1, SCI3	von Behr et al. (2013)	Yes – 15 case studies	No – In fact vulnerable individuals in sample were not cited as searching for bomb making	N/A – Not applicable to statement
	ONLY CITED ONCE	Weimann (2004)	No – Opinion Piece	Yes	Not empirical findings
	ONLY CITED ONCE	Cornish, Hughes, & Livingstone (2009)	No – Literature Review	Fair – Mentions bomb making manuals, but no mention of those 'vulnerable to radicalisation' seeking out such information	Citing Stenersen (2008) and Emigh (2004)
"Violent extremists, leaders and ideologues are often engaged in activity related to	NO CITATION				

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

publicity, propaganda, fundraising, sharing information, recruitment, mobilization and the planning and coordination of attacks.”					
“The dynamic role of the media and Internet in propagating terrorist information and promoting extremist acts is widely recognized”	IN SCI1, HAC3, CMIntr	Bakker (2006)	Yes – 31 case studies, and 200 cases	Fair – Not explicit on “dynamic role”. Mention of one case where internet was critical for radicalisation and leakage, and mention of discussion of religious nature	One case from data identified
	ONLY CITED ONCE	Vennhaus (2010)	Yes – 2032 cases of foreign fighters from multiple sources	Yes	Not explicitly linked to data outcomes.
	IN BA5, SCI1, SCI3, CMIntr	Vidino (2016)	No – Statement made to Senate	Yes	No empirical support – Opinion based conclusion
“The ability of ISIS to directly and constantly reach Americans through social media has played a role in triggering mobilization towards extremist violence”	IN BA5, SCI1, SCI3, CMIntr	Vidino (2016)	No – Statement made to Senate	Yes	No empirical support – Opinion based conclusion
“Violent extremists, including ISIS, are experienced users of the Internet, and they have undertaken video productions and used social media to their advantage.”	NO CITATION				
“Jihadist websites have sections in Western languages which are	IN SCI1, MD6	Europol (2017)	Yes	Yes	Within Findings

Released by Department of Homeland Security under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

designed to win recruits from Europe and North America for jihad in conflict areas”					
“This is designed to have a radicalizing influence.”	NO CITATION				
“An important risk indicator is the degree to which individuals are engaged as users, as developers of radicalizing materials, as narrators of stories and as generators of messages and postings of terrorism-related material on the Internet.”	NO CITATION				
“The Internet creates opportunities to become radicalized and can contribute to self-radicalization or serve to radicalize or grooming others”	ONLY CITED ONCE	Lennings et al. (2010)	No – Literature Review	Fair – Although explicit mentions of radicalisation are not used. Piece is more about other functions, e.g. networking, recruitment, grooming.	Citing work of Conway (2006), Weimann (2004)
	IN SCI1, SCI3	von Behr et al. (2013)	Yes – 15 case studies	Yes	In Literature Review
“It can act as an ‘echo chamber’: a place where individuals find their ideas supported and echoed by other like-minded individuals. But the role of the internet must always be placed within the broader context of the individual’s personal history and social relationships and can’t be singled out as the only source of radicalization”	IN SCI1, SCI3	von Behr et al. (2013)	Yes – 15 case studies	Yes	Findings and conclusion

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

	ONLY CITED ONCE	ISD (2011)	Not Referenced	Unable to Source	
	ONLY CITED ONCE	Cornish (2008)	No – Report of Roundtable Discussion	Fair – Though mentions role as important, it does not mention strength of influence (much more holistic and open than citation suggests)	Not empirical findings
“identified five political ways in which the Internet is used by terrorists and terrorist groups. These uses are: (1) provision of information, (2) funding, (3) networks, (4) recruitment and (5) collecting information.”	ONLY CITED ONCE	Conway (2006)	No – Literature Review	Yes	Core framework of argument by author – Not Empirical Findings
	ONLY CITED ONCE	Conway (2007)	No – Description of policies and literature	No – Specifically about policies, not terrorist actions or internet use	N/A – Not applicable to statement
	ONLY CITED ONCE	Conway (2017)	No – Literature Review	No – No mention of five ways.	N/A – Not applicable to statement
SCI2 – Target for Attack Identified (Person, Group, Location)					
“A report to the US Congress in 2007 emphasized the importance of the specificity of a target. Identification of a target with precision is considered to increase the risk of action which would apply to violent extremism or terrorist attacks”	ONLY CITED ONCE	Perl (2007)	No – Review of policy	No – Piece is about US response and state sponsored terrorism. Only possible link is this quote: “Both timing and target selection by terrorists can affect U.S. interests in areas ranging from preservation of commerce to nuclear non- proliferation to	N/A – Not applicable to statement

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

				the Middle East peace process.”	
“Based on empirical experience with and research into violent actions, the CIA has also identified this indicator as essential”	IN SCI2, HAC6	Fein & Vossekuil (1998)	No – Report following from Results of Fein & Vossekuil (1997)	Not CIA, but USSS. Not relevant for generic terrorism. Very specific form of targeted violence. But does note that assassins will choose their target.	Based on Findings of ECSP study (Fein & Vossekuil, 1997)
“Intelligence and national security experts have indicated that the more specific the target (a person, group or location), the greater is the chance of an attack.”	NO CITATION				
SCI3 – Personal Contact with Violent Extremists (Informal or Social Context)					
“Individuals who have been involved in terrorist acts often have personal contact with other violent extremists”	NO CITATION				
“Personal contact influences ideas and facilitates radicalization and recruitment.”	NO CITATION				
“Extremist ideas are developed and adapted socially by means of direct contact”	ONLY CITED ONCE	Halloran (1967)	No – Theoretical/ literature review	No – About attitude formation and change. No mention of terrorism or violent extremism	N/A – Not applicable to statement
“Such contact can support the development of a sense of ‘brotherhood’”	NO CITATION				

Used by Department of Home Affairs
the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

"Personal contact with other violent extremists can push one to take part in violent extremist acts"	IN BA7, SCI3, SCI7, HAC2, HAC3, CMIntr	Sageman (2008)	Yes – Case Studies	Yes – However, only Jihadist	In Findings
"Personal contact with other violent extremists... can promote ideological commitment"	IN BA2, SCI3, SCI6, HAC2	Moskalenko & McCauley (2011)	Yes – Case studies	Yes	In Findings
		Moskalenko & McCauley (2017)	No – Review of Theoretical Models (Literature)	Yes	Work of Pyszczyński et al. (2010), Della Porta (2011), Hafez & Mullins (2015)
"Many foreign fighters in Syria and Iraq are known to have become increasingly isolated from their past social networks and they developed bonds with new networks involving those who shared and helped to reinforce their radical views"	IN SCI3, SCI7, SI	Weggemans, Bakker & Groen (2014)	Yes – Interview data of those close to 5 jihadists and a focus group (25 professionals) put into two 'composite' life stories	Fair – Is a stretch to say that 5 cases constitute 'many', and also evidence in study does not suggest many undertake active withdrawal from past networks, but does include new networks	In Findings
		NO CITATION			
"Young people develop their network of relationships not only amongst those with whom they play football or physically 'hang out', but also on social network sites and					

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

otherwise through the internet.”					
“The way people function psychologically in cyber-space is not the same as in direct face to face social relationships”	IN SCI3, CMIntr	Alderdice (2014)	No – Book Chapter	Yes – In fact the wording in Alderdice is almost exactly the same – “the way people function psychologically in cyberspace is not quite the same as in direct face-to-face social relationships.” (NOT IN QUOTATION MARKS IN VERA-2R)	Opinion based conclusion, based on Atran (2010) and own opinions of Breivik case.
“Just as the virtual community of IS-supporters acts as an echo chamber, so to do real-life connections with like-minded individuals reinforce and strengthen individual commitment”	IN BA5, SCI1, SCI3, CMIntr	Vidino (2016)	No – Statement made to Senate	Yes – In fact the statement “Just as the virtual community of IS-supporters acts as an echo chamber, so to do real-life connections with like-minded individuals reinforce and strengthen individual commitment” is exactly as written in Vidino (NOT IN QUOTATION MARKS IN VERA-2R)	No empirical support – Opinion based conclusion
	ONLY CITED ONCE	Horgan et al. (2016) Assuming is Horgan, Altier et al. due to addition of both	Yes – One case study (R/W)	Fair – Not IS and not explicit on the matter of connections of like-minded individuals reinforcing commitment though - Also note that exact wording is	Reference to virtual community in findings (Note paper is not IS related)

Released by Department of Home Affairs under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

		wording and reference in version.		drawn from Vidino (2016)	
"Intragroup differences exist, also in time people may have different roles"				Yes – Mention of movement through roles in the one case under scrutiny	In Findings
"The internet is not replacing the need for individuals to meet in person during their radicalisation, the offline world plays an important role in the radicalisation process"	IN SCI1, SCI3	von Behr et al. (2013)	Yes – 15 case studies	Yes	In Findings
"Although lone actor terrorists may not seem to be members of a group, contact through the Internet can provide networks and a source of information and inspiration for terrorist acts"	IN BA3, SCI3	Waytz & Epley (2012)	Yes – Experiments on students	No – Not relevant to statement "Although lone actor terrorists may not seem to be members of a group, contact through the Internet can provide networks and a source of information and inspiration for terrorist acts"	N/A – Not applicable to statement
	IN SCI3, CMIntr, MDIntr, MD3	Gill & Corner (2017)	No – Literature Review	No – Not relevant to statement "Although lone actor terrorists may not seem to be members of a group, contact through the Internet can provide networks and a source of information and inspiration for terrorist acts"	N/A – Not applicable to statement

SCI4 – Expressed Intention to Commit Acts of Violent Extremism

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

“Writings by most terrorist groups indicate a relationship between emotion and the definition and realization of goals.”	NO CITATION				
“Extreme anger can serve as additional motivation to realize goals and to choose violence”	IN BA1, BA2, BA3, BA4, BA5, SCI4, SCI5	Saucier et al. (2009)	Yes – Open source data	No	N/A – Not applicable to statement
“Anger is often mixed with moral emotions, through which the intention to use violent acts in service of the higher cause are justified and supported.”	NO CITATION				
SCI5 – Expressed Willingness and/or Preparation to die for a Cause or Belief					
“The willingness to die for a cause or belief involves the concurrence of organizational, societal and individual causes and facilitators”	IN SCI5, HAC3, CM6	Merari (2010)	Yes – Interviews	Yes	Citing Hafez (2004), and inferred in findings
	ONLY CITED ONCE	Corte Ibanez (2014)	Not Cited (presumed correct name is de la Corte Ibáñez) Found piece on “The Social Psychology of Suicide Terrorism” - Literature Review	Yes	Evidence drawn from Elster (2006), Fields (1979), Silke (2006), Sageman (2004; 2008)
	IN SCI5, CM3	Atran, Sheikh, & Gomez (2014)	Yes – Although the piece itself is a note and not the main item where the	Yes	Evidence from Whitehouse et al. (2014), and inferred in findings

Released by Department of Justice Affairs under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

	IN SCI5, MDIntr		interviews and surveys were conducted		
		Lankford (2014a)	No – Overview of Book	Yes	Evidence from Burger (1999), Freedman and Fraser (1966), Grossman (1995), Johnson (1986), Waller (2002)
		Lankford (2014b)	No – Review of Book	Yes	Opinion of author
“Martyrdom and a culture linking martyrdom attacks to heavenly rewards as the atoning for previous sins or to attain paradise stimulate suicide terrorism”	ONLY CITED ONCE	Corte Ibanez (2014)	Not Cited (presumed correct name is de la Corte Ibáñez) Found piece on “The Social Psychology of Suicide Terrorism” This piece is a Literature Review	Yes – In fact the text states “atoning for previous sins or to attaining paradise” (p.24) which matches the text on p.66 of VERA-2R (NOTE, NO QUOTATION MARKS USED) “as the atoning for previous sins or to attain paradise”	No citations used in statement – Direct opinion of authors.
	IN BA2, SCI5	Atran (2003)	No – Literature Review	No	N/A – Not applicable to statement
	ONLY CITED ONCE	Hafez (2007)	Incorrect Citation – Assumed correct is Hafez (2006) Yes – analysis of last will and testaments of	Fair – Discusses martyrdom and the rewards, however does not link how martyrdom stimulates or causes suicide terrorism	No citation, opinion of author (regarding martyrdom). Not applicable for stimulation argument

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

			suicide bombers		
“identified elements of what he calls the “suicide production line”, the last phase of which involves making a video.”	IN BA6, BA7, SCI5, SCI7, HAC1, CMIntr	Post (2005)	No – Statement	Yes – Actual wording is “suicide terrorist production line” not “suicide production line”	Statement made by author
“discusses the point at which an individual may find it difficult to turn away from martyrdom.”				Yes	Statement made by author
“Some actions can be restricted through situational limitations (such as detention), but the willingness to die for the higher cause may remain resolute.”	NO CITATION				
	IN BA1, BA2, BA3, BA4, BA5, SCI4, SCI5	Saucier et al. (2009)	Yes – Open source data	Yes – But actual wording is “Glorification of dying for the cause” Not for the “higher” cause	In Findings
“It corresponds to the militant-extremist and terrorist theme of the ‘glorification of dying for the higher cause’”	ONLY CITED ONCE	Francis (2016)	No – Theoretical	No – Piece is on religion for explaining terrorism – no mention of suicide terrorism	N/A – Not applicable to statement
	IN BA2, SCI5, CM1, CM3	Dawson & Amarasingam (2017)	Yes – 25 Interviews with Foreign Fighters	No – Piece concerns foreign fighters and does not cover ‘glorification of dying’	N/A – Not applicable to statement
This indicator proved to be a consistent and important indicator in a study of terrorists motivated by different causes	ONLY CITED ONCE	Beardsley & Beech (2013)	Yes – Evaluation of VERA	No – This indicator is not within VERA, and not mentioned in piece	N/A – Not applicable to statement

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

“Some violent extremists only intend to cause material damage to infrastructure or property and are not interested in killing others or in martyrdom or dying for the higher cause. The planning of attacks with martyrdom as a goal is likely to represent a higher risk and threat level.”

NO CITATION

SCI6 – Planning, Preparation of Acts of Violent Extremism

“point to differences between individuals who are dedicated to radical or extreme beliefs and do not progress to committing unlawful violent acts and those who in fact do so.”

IN BA2, SCI3, SCI6, HAC2

McCauley & Moskaleiko (2011)

Yes – Case studies

Yes

Within Findings

IN SCI6, SCI7, HAC2

Atran (2010)

Yes – Interviews

Yes

Within Findings

“The difference can be characterized as ‘talking the talk’ rather than ‘walking the walk’ or taking action”

NO CITATION

“found that about 83 percent of lone actor terrorists attracted the attention of their social environment before their deed. In 58 percent of the cases, relatives were conscious of the fact that the perpetrator was not only speaking about the proposed action but also that he demonstrated behavior that indicated research, planning, and preparation for a violent attack. It was found that 59 percent of the perpetrators

IN BA1, SCI6, MDInt, MD3

Gill (2015)

Yes – 111 cases

No – Statistics are actually drawn from Gill, Horgan, & Deckert (2014) (119 as opposed to 111) as reads:

“In 58% of cases, other individuals possessed specific information about the lone actor’s research, planning, and/or preparation prior to the event itself. Finally, in a

N/A – Not applicable to statement

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

also produced statements within letters, videos, or manifestos which identified their preparation and planning.”				majority (59%) of cases, the offender produced letters or made public statements prior to the event to outline his/her beliefs (but not his/her violent intentions). These statements include both letters sent to newspapers, self-printed/disseminated leaflets, and statements in virtual forums.”	
“In the delineation of the progressive stages of risk, the statement of intent precedes an action”	ONLY CITED ONCE	Horgan (2003)	No – Literature Review on Disengagement	Unable to Source	
	ONLY CITED ONCE	Horgan (2005)	No	Unclear – No specific mention of statement of intent in the preparation phase	N/A – Not applicable to statement
	IN SCI6, HAC1, HAC2, PRMIntr	Horgan (2008)	No – Literature Review	Unclear – No specific mention of planning/preparation or statement of intent	N/A – Not applicable to statement
	ONLY CITED ONCE	Horgan (2012)	No – Review of methodology (interviews)	No	N/A – Not applicable to statement
	ONLY CITED ONCE	Taylor & Horgan (2006)	No – Theoretical Framework	No	N/A – Not applicable to statement
SCI7 – Susceptibility to Influence, Control or Indoctrination					
“Some people are more open to influence and indoctrination	NO CITATION				

Released by Department of Home Affairs under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

than others and such receptivity can increase risk.”					
“Charismatic leaders can persuade ‘true believers’ to carry out suicide attacks”	IN BA6, BA7, SCI5, SCI7, HAC1, CMIntr	Post (2005)	No – Statement	Yes	Statement made by author
“This is the case with secular terrorists such as the Tamil Tigers (LTTE) and Kurdish separatists (PKK) as well as with Al Qaida and the Islamic State (ISIS).	NO CITATION				
“Personal, social, and cultural identities may play a role in religiously and ethnically motivated terrorism”	ONLY CITED ONCE	Schwartz, Dunkel, & Waterman (2009)	No – Theoretical Review	Yes	Theories put forward by authors in Review
“found that suicide terrorists had characteristics that converge on an avoidant and dependent personality style and ego structure that is susceptible to social influence, especially by persons perceived as authoritative.”	IN SCI7, MD1	Merari et al. (2009)	Yes – 40 interviews	Yes	In Findings
“Charismatic leaders claim to provide security and dependency among its members, but also a sense of control, identity and meaning”	ONLY CITED ONCE	Hofmann (2016)	Yes – Open Source Data	No – No mention of provision of security, dependency, sense of control, identity, or meaning	N/A – Empirical element regarding success of attacks in groups led by charismatic leaders
“The influence can depend on the communication skills and the recognition by followers”	ONLY CITED ONCE	Hofmann (2016)	Yes – Open Source Data	Yes	Info drawn from Literature Review – specifically Bendix & Weber (1977)
	IN SCI7, SI	Beevor (2017)	No – Develops a theoretical model	Fair – About charismatic authority,	Citing various works, including Rinehart (2006)

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

				but no mention of communication skills	
	IN SCI7, SI	Rinehart (2006)	Yes – Three Case Studies	Yes	Conclusion of findings
				Yes	Citing Keppel (2004), Precht (2007). Some inference in description of Al-Awaki (author's own work)
"Overcoming moral doubts of potential recruits about using violence to achieve a noble goal is a key challenge"	In SCI7, HAC5, SI	Gendron (2017)	No – Literature review	Yes – Exact wording in text matches VERA – 2R (BUT NOT IN QUOTATION MARKS) "Overcoming the moral doubts of potential recruits about using violence to achieve a noble goal is a key challenge"	Opinion of author – No citations attributed
"Ultimately, such leadership aims to motivate followers to perform acts of violence and self-sacrifice to further the organization's goals."	NO CITATION				
"A Dutch study on foreign fighters showed that all those studied came under the influence of 'charismatic persons' or 'inspiring figures' to some degree"	IN SCI3, SCI7, SI	Weggemans, Bakker, & Grol (2014)	Yes – Interview data of those close to 5 jihadists and a focus group put into two 'composite' life stories	No – Not all. Concluding remarks are "In the researched cases individuals were confronted with, and shared their radical ideologies... Sometimes these	N/A – Not applicable to statement

Released by Department of Home Affairs
under the Freedom of Information Act 1982

APPENDIX 3 – Assessment of Citations utilised in the VERA-2R

Testing the Reliability, Validity and Equity of Terrorism Risk Assessment Tools.

The content of this deliverable may be shared within agencies but is not to be reproduced or distributed further.

<p>“This information is important for the assessment of overall risk as well as for criminal justice interventions”</p>					networks consisted of old friends... new friends... or they encountered charismatic persons”.	
	In SCI7, HAC2	Weine et al. (2009)	No — Literature review	Yes – However, study notes that, at the time of writing, the risk factors discussed are “working hypotheses” as “To date no rigorous research has been conducted to investigate either those adolescent boys and young men who mobilized or their families” (p. 185).	No citations – Opinion of authors	
	IN BA7, SCI3, SCI7, HAC2, HAC3, CMIntr	Sageman (2008)	Yes – Case Studies	Fair – Only touches upon RE Charisma or Inspiring	No citations – opinion of author outside of case studies	
	IN SCI7, HAC2	Atran (2008)	No — Testimony	No — Testimony is focused on the need for more field-based research. Only brief mention of leadership. Findings suggest that notions of charismatic leaders are “exaggerated”	N/A – Not applicable to statement	

Released by Department of Home Affairs
under the Freedom of Information Act 1982