Ministry
of Justice

Inter-rater reliability of the Extremism Risk Guidelines 22+ (ERG 22+)

Beverly Powis, Kiran Randhawa-Horne, Ian Elliott
Data and Analytical Services Directorate
Prison and Probation Analytical Services

Jessica Woodhams
University of Birmingham

Ministry of Justice Analytical Series
2019

Protecting and advancing the principles of justice

Analytical Services exists to improve policy making, decision taking and practice by the Ministry of Justice. It does this by providing robust, timely and relevant data and advice drawn from research and analysis undertaken by the department's analysts and by the wider research community.

Disclaimer

The views expressed are those of the authors and are not necessarily shared by the Ministry of Justice (nor do they represent Government policy).

First published 2019

Any enquiries regarding this publication should be sent to us at mojanalyticalservices@justice.gsi.gov.uk

This publication is available for download at http://www.justice.gov.uk/publications/research-and-analysis/moj

ISBN 978-1-84099-892-4

Acknowledgements

The authors would like to thank Carys Keane, Hannah Thompson and Christopher Dean for developing the case studies.

Thank you also to Alana Diamond and Helen Wakeling for comments on earlier drafts of the report.

The author

The views expressed are those of the authors and not necessarily shared by the Ministry of Justice (nor do they represent government policy).

Contents

List of tables

1. Summary

Given the increase in convictions for terrorist and terrorism related offences in the UK, there is an urgent need to develop reliable and valid assessment approaches to better understand the risk and needs of these individuals and inform their management through the criminal justice system.

Assessment tools used to make decisions about extremist offenders need to be reliable, consistently producing the same results from the same information, across different assessors. This is measured by examining the Inter Rater Reliability (IRR). While a number of assessment tools have been developed for extremists, none has been subject to an examination of IRR. This study examines the IRR of one measure, the ERG22+, which has been in use across Her Majesty's Prison and Probation Service of England and Wales since 2011. The ERG22+ is a structured professional judgement tool that assesses an individual along 22 factors that are grouped into three domains; Engagement, Intent and Capability. There are 13 Engagement factors, six Intent factors and three Capability factors.

The study examined two approaches to IRR; research reliability, which tested reliability in 'research' conditions between expert raters, and field reliability which tested reliability among practitioners using the measure in the field. For the research reliability, two experienced researchers rated 50 randomly selected convicted extremist cases using the ERG22+. For the field reliability, 33 trained practitioners rated two test cases specifically developed for the study against "gold standard" ratings. The study used a range of statistical approaches to measure IRR which were compared.

High levels of IRR were found between the researchers at the casewise, itemwise and domain levels, with ratings falling within the "excellent" range. Results were more mixed for the field reliability. Overall IRR across both case studies were found to be "moderate" to borderline "good", but varied considerably between raters. Differences were also observed in rating agreements at the domain level. Reliability for the domains of Engagement and Capability varied from "moderate" to "excellent", but were poor for the Intent domain. Poor agreement on a number of specific factors could explain the low overall field IRR for this domain. As the research reliability produced high inter-rater agreement for these factors, their reliability could be significantly improved with further assessor training and better definition of the items.

Analysis was conducted to compare ratings of “experienced” field raters from those who had limited experience of the ERG22+. Although the inexperienced raters appeared to use the ERG22+ effectively, their use was less consistent across the cases than that of experienced raters, especially for one case, where the experienced group performed statistically significantly better when compared to gold standard scores. This finding, along with the findings from this research reliability study, highlight the importance of raters having experience in the ERG22+ and the field of extremism.

The study had a number of limitations, including the use of only two research experts to rate cases in the research reliability study and only two case studies for the field reliability study. The research would have been improved with a greater number of raters and more cases respectively. In addition, the research and field raters had access to different information which may have influenced the ratings they produced. However, the study was strengthened by the use of two different approaches to measuring IRR which were subject to a range of statistical analyses.

This research is the first to examine the IRR of the ERG22+ tool for use with extremist offenders. The ERG22+ would appear to have IRR ranging from perfect to moderate, which can be improved when used by experienced practitioners and expert researchers. The report makes a number of suggestions that would further enhance the field IRR of the tool.

2. Introduction

Understanding and countering violent extremism is a high priority for the UK Governments and has become more pressing as terrorism-related arrests and subsequent convictions have increased in recent years (Politowski, 2016). Government statistics indicate that the number of prisoners detained for terrorism-related offences has continually increased since data recording began in 2009 (Home Office, 2018). This has meant that there is increasing pressure on the Criminal Justice System to effectively assess and manage these individuals.

The assessment of risk is considered to be an important first step in reducing risk (Campbell, French & Gendreau, 2007) and a range of risk assessment measures are widely used across judicial systems. Historically, risk assessment has followed two differing approaches each with their own limitations. The first-generation risk assessments used a clinical judgement approach where information on social, environmental, behavioural and personality factors related to harmful behaviour(s) were collected through detailed interviewing and observation (Campbell et al., 2007; Monahan, 1984; Quinsey et al., 1998). The second generation used an actuarial approach with the aim of standardising risk assessment. Statistical calculations, based on a set number of risk factors, are assigned numeric values to produce an overall score of risk (Campbell et al., 2007; Kemshall, 2001). However, both of these approaches have limitations, as they are thought to be either too unstructured or too rigid. A third approach that combines structured risk assessments and clinical interviews has been developed and is judged to be the most effective way to understand risk (Scott & Resnick, 2006). This approach is known as Structured Professional Judgement (SPJ) and uses empirically- derived risk factors that are considered using clinical judgement, giving assessors some flexibility and discretion in their determination of risk (Murray & Thomson, 2010). SPJ allows assessors to include additional information they feel is relevant to the individual.

It has been argued that it is very difficult for clinicians to be able to accurately predict risk of either recidivism or harmful behaviour (RTI International, 2018). Instead, their focus should be placed on understanding how to best manage risk (Murray & Thomson, 2010) through helping inform sentencing decisions, management and supervision after conviction (Kemshall, 2001). This is likely to be even more challenging for extremist offenders, given their low numbers. Because terrorist activities are relatively rare, there are insufficient numbers of offenders from whom predictive risk factors can be statistically tested and determined (Egan et al., 2016; Sarma, 2017; Scarcella et al., 2016). Little is also known about the personal and contextual circumstances of such individuals, as well as individual motivations for extremist offending, meaning that identifying a definitive set of risk factors to be developed into a tool is also likely to be difficult. In addition, extremist offenders are not homogenous (Sarma, 2017; Silke, 2014) and their motivations and circumstances in which they commit offences are varied and complex (Dean, 2014). Because of these difficulties, an SPJ approach has been recommended, because risk factors are identified systematically and with the flexibility to consider the individual context and additional factors that may be relevant (Monahan, 2012; Roberts & Horgan, 2008; Skeem & Monahan, 2011). It also allows the collation of information to be used for both assessment and management of the individual (Sarma, 2017).

Several measures have been developed to assess extremists, based on a combination of theories and detailed casework. In a review of these measures, Scarcella et al. (2016) found few had been assessed for their psychometric properties, with none having been tested for their IRR. One measure is the Extremism Risk Guidelines 22+ (ERG22+) (Lloyd & Dean, 2015; National Offender Management Service, 2011) which has been in mainstream use across Her Majesty's Prison and Probation Service of England and Wales since September 2011. The ERG22+ is intended for use with all offenders who have been convicted of an extremism or terrorist offence (regardless of their cause or ideology) and since its roll out, has been completed on all offenders convicted of a terrorism offence.

The ERG22+ adopts a structured professional judgement case formulation approach to analyse specific factors relating to an individual and the context around their circumstances that led them to offend. It is composed of 22 items, which focus on three domains, 'Engagement', 'Capability' and 'Intent'. There are 13 Engagement factors, six Intent factors and three Capability factors (NOMS 2011). These are listed in Table 1. The assessor considers and records whether each factor is strongly present, partly present or not present. The Intent and Engagement domains are each assigned an overall category of low, medium, or high. The Capability domain is categorised as being minimal, some and significant. The assessment is completed by psychologists or probation officers with experience of working in forensic settings who have received full training in its administration.

Table 1: 22 Factors of the ERG22+

Engagement	Intent	Capability
1. Need to redress injustice 2. Need to defend against threats 3. Identity, meaning & belonging 4. Need for status 5. Excitement, comradeship & adventure 6. Need to dominate others 7. Susceptibility to indoctrination 8. Political, moral motivation 9. Opportunistic involvement 10. Family and/or friends support extremism 11. Transitional periods 12. Group influence and control 13. Mental health Issues	14. Over-identification with group, cause 15. Us & them thinking 16. Dehumanisation of the enemy 17. Attitudes that justify offending 18. Harmful means to an end 19. Harmful end objectives	20. Personal knowledge, skills, competencies 21. Access to networks, funding, equipment 22. Criminal history

If assessments are to be used to make decisions about risk posed by extremists, they need to be reliable, consistently producing the same results from the same information, across different assessors. Risk assessments used in criminal justice settings are tools that will determine the levels and type of supervision and interventions an individual will be subject to, as well as monitor changes over time, so it is important that they produce the same results consistently. IRR measures the consensus between two or more assessors administering the same instrument to the same individual(s). Two approaches have been used to measure IRR of assessment tools used with forensic populations (Campbell, 2004; Wakeling et al., 2011). The first is *research reliability* that examines the IRR found between researchers. It has been suggested that IRR is likely to be high among researchers as they tend to administer the tool in large numbers (Campbell, 2004; Wakeling et al., 2011). The second is *field reliability* which examines the extent that practitioners using the tool operationally agree with each other when assessing an individual. Field reliability is considered to produce lower consensus among assessors than research reliability because of differing levels of experience with the tool and interpreting scores (Campbell, 2004; Doren, 2002; Webster et al., 2006).

This paper presents research undertaken to examine both research and field IRR of the ERG22+. There has been recent evidence regarding the structural properties of the ERG22+ with a study that examined the construct validity and internal consistency of the measure (Powis, Randhawa & Bishopp, 2019). The study concluded that the ERG22+ shows promise as a risk and need formulation tool, but that further research should be carried out on the measure, including a study of IRR. We are not aware of any other published studies that examine the IRR of any assessment tools for this offender group. This report therefore presents not only the first study of the IRR for the ERG22+ but also the first for any assessment of extremist offending.

3. Methods

3.1 Measuring Inter-Rater reliability (IRR)

IRR can be calculated using a number of statistical techniques, each with their respective strengths and weaknesses. The simplest method is simple percentage agreement, but this does not allow for either variance amongst raters nor does it correct for chance agreement (Bohrnstedt & Knoke, 1994; Cohen, 1960). A second technique is Cohen's kappa statistic (*k*) (Cohen, 1960), which reports the proportion of raters in agreement and corrects for chance agreement, but it does not allow for multiple raters and does not correct for other sources of variance. Where categorical or ordinal data is generated, k can be weighted in order to account for the magnitude of disagreement from "perfect agreement" to "worst disagreement" (e.g., Cohen, 1968; Fleiss, Cohen, & Everitt, 1969). A third technique is Fleiss' kappa statistic (referred to here by the notation k^F)(Fleiss, 1971), which reports the reliability of agreement between multiple raters on binary or nominal-scale ratings and corrects for chance agreement, but does not correct for other sources of variance. Fourth and finally, the intra-class coefficient (ICC) reports the ratio or degree of variance among raters (Shrout & Fleiss, 1979). There are multiple forms of ICC depending on the nature of the experimental design which can generate different results from the same data, meaning that analyses need to be chosen with due care and attention (McGraw & Wong, 1996). McGraw and Wong described 10 forms of ICC (adding to the 6 forms described by Shrout and Fleiss (1979)) based on the "model" (1-way random effects, 2-way random effects, vs. 2-way fixed effects), the "type" (single rater/measurement vs. the mean of *k* raters/measurements), and the "definition" of relationship considered to be important (consistency vs. absolute agreement) (Koo & Li, 2016). Scores can be classified in different ways, but the approach regularly advocated is that by Fleiss (1986) which classifies values greater than 0.75 as excellent, 0.61-0.75 as good, 0.4-0.6 as moderate, and less than 0.4 as poor (Altman, 1991; Fleiss, Levin & Paik, 2003; Vincent et al., 2011).

3.2 Research reliability

The aim of the research reliability analysis was to examine the use of the ERG22+ by two expert raters across a range of *cases*. Given both the varying rationales underpinning each of the approaches to measuring reliability and their respective strengths and weaknesses, a combination of statistics were utilised to compare and contrast the findings.

Participants

Two researchers with more than 5 years' experience in the use of the ERG22+ and specialist research experience with extremist offenders.

Procedure

The two researchers independently rated 50 cases using the ERG22+. First, a random sample of 50 individuals was computer generated from all those who have been convicted of an extremist offence in England and Wales and have been assessed using the ERG22+ between 2011 and 2016 (*n*=250). The scoring sheets were removed from each of the 50 completed ERG22+ assessments in the sample, with the background, interview and case file information that had been collated as part of the ERG22+ assessment process being retained. The cases included 42 individuals convicted of Islamist extremism offences, three Extreme Right Wing offenders, two Animal Rights extremists and three individuals convicted of extremist offences in support of other causes. The two researchers independently read the case information and scored each of the ERG22+ factors for every case, using the ERG22+ scoring sheet. They also gave a total overall rating for each of the three ERG22+ domains. Cases were scored individually, with no discussion between the researchers. Raters were provided with an ERG22+ manual to assist their judgements. Factors were scored as 2, 1 or 0 depending on whether the rater identified the factor as being strongly present, partly present or not, respectively. In addition, the three domains of Engagement, Capability and Intent were scored, with Intent and Engagement domains being scored as 0 for low, 1 for medium, and 2 for high. The Capability domain was scored as 0 for minimal, 1 for some and 2 for significant.

Analysis

Three forms of IRR statistics were generated to estimate the degree of consensus between the two raters, using the "irr" package in the statistical software R (R Core Team, 2012). IRR was calculated at both a "casewise" and an "itemwise" level. The casewise analyses compared each rater on each case (e.g. rater 1 on case 1 vs. rater 2 on case 1 and so forth) using simple percentage agreement and a weighted Kappa (Cohen, 1960, 1968) with a "squared" technique, whereby disagreements are weighted according to their squared distance from perfect agreement to account for the ordinal data. The itemwise analyses compared each rater on each item and scale (e.g. rater 1 on item/scale 1 vs. rater 2 on item/scale 1 and so forth) using simple percentage agreement, a weighted Kappa, and two forms of the intra-class coefficient were generated, with 95% confidence intervals: (1) a two-way random model of agreement based on a single rater (ICC[1]) to estimate the reliability of a single rater; and (2) a two-way random model of agreement based on the mean of each rater (ICC[2]) to estimate the reliability of the ratings. These analyses allowed us to judge the impact of both variability in cases and variability in items and scales on IRR.

3.3 Field reliability

The aim of the field reliability analysis was to examine the IRR of the ERG22+ across a range of *typical users*: clinicians working directly with the convicted violent extremist population. As with the research reliability analysis, a combination of statistics were used in order to compare and contrast the findings of each. Two analyses were conducted using two hypothetical test cases (Mr G and Mr H). The first analysis explored overall agreement between the raters on each case. The second explored the extent to which the raters agreed with an anticipated “gold standard” rating that the cases were intended to produce.

Participants

All those who had been trained in administration of the ERG22+ between 2014 and 2016 were invited to participate in the study (*n* = 45). Of those who were approached, 33 completed both cases representing a 73% participation rate. Analyses were conducted for both the full sample (*n* = 33) and for a sub-sample of raters who could be considered “experienced” raters (*n* = 25) having authored four or more ERG22+ assessments and worked directly and closely with the related population of extremist offenders for a minimum of three years.

Procedure

Two case studies, Mr G and Mr H were designed and created by three members of staff who were forensic psychologists from the Extremism National Clinical Team who had developed and designed the ERG22+ and were practiced in using the assessments. Mr G was an Extreme Right Wing case and Mr H was an Islamic extremist case. The assessments were based on real cases, modified and anonymised for the purpose of the study. Each rater was sent the two case studies and asked to score each of the 22 factors as “not present”, “partly present”, or “strongly present” using the ERG22+ scoring sheet. They were also asked to give a total overall rating for each of the domains; Engagement, Intent and Capability. Each rater scored each of the cases individually and were instructed not to discuss the cases with other raters. The scoring sheets were then returned to the researchers.

Analysis

For the field reliability study, four forms of IRR statistics were generated using the “irr” package on R. First, simple percentage agreement was generated at the itemwise level (per item, per scale) and across both cases. Second, Cohen’s weighted kappa coefficient was calculated per item, per scale, per case, to measure each rater’s individual performance against a “gold-standard” score. Third, Fleiss’ kappa statistic was calculated for each item, scale, and overall to measure IRR across multiple raters. Fourth, two forms of the intra-class coefficient, with 95% confidence intervals, were calculated to measure variance across raters: (1) a two-way random model of agreement based on a single rater (ICC^{1}) to estimate the reliability of a single rater; and (2) a two-way random model of agreement based on the mean of each rater (ICC^{2}) to estimate the reliability of the ratings. ICC statistics were calculated for each item, as the mean for each scale, and as an average across all items.

4. Results

4.1 Research Reliability

Itemwise IRR was found to be high between the two research raters, with percentage agreements for overall ratings for the three ERG22+ domains of Capability, Engagement and Intent being 92%, 90% and 93% respectively (see Table A.1). A high level of agreement between assessors was also found for each of the ERG 22+ factors. Weighted kappa scores ranged between 0.81 and 1 and ICC^1 scores between 0.81 and 1, which are all considered to be "excellent" agreement (Fleiss, 1986). Table A.1 also provides a summary of weighted kappa and ICC scores, with 95% confidence intervals, by item and domain.

At the casewise level, a high level of agreement between assessors was also found. Mean agreement was 94.8% (SD = 8.0) with a pooled weighted kappa of 0.95 (SD = 0.1). Weighted kappa values were found to be under 0.90 for 6 cases, with only case 4 (0.36) and case 14 (0.78) falling below .80 (see Table A.2).

4.2 Field Reliability

Full sample

Using the classification system advocated by Fleiss (1986), overall IRR for Mr G was "moderate" (ICC^1 = 0.48; k^F = 0.47) across the full sample of raters. Levels of agreement were marginally improved for Mr H, with ICC^1 values suggesting a "moderate" level of reliability that was borderline "good" (ICC^1 = 0.6; k^F = 0.59). Analysis at the domain level, ranged from "moderate" to "excellent" for the Engagement and Capability domains across the two cases, with ICC^1 and k^F values falling between 0.57 and 0.79. However, the Intent domain appeared to show "poor" agreement, with ICC^1 and k^F values between 0.14 and 0.29 (see Table A.3 for full results).

Performance of the raters against the "gold standard" was judged by pooling the Cohen's kappa statistic of each individual rater versus the gold standard. On average, raters were found to have higher pooled kappa values versus the gold standard for Mr H (*M* = 0.65, *SD* = 0.21) than for Mr G (*M* = 0.73, *SD* = 0.17) (see Table A.4). A repeated-measures *t*-test, however, did not find the difference to be statistically significant (*t* (32) = -1.44, *p* = .16). The Mr G case generated a wide range of kappa values (*k* = 0.17 - 1.00) and an overall percentage agreement rate of 73.3%. Similarly, the Mr H case generate a wide range of values (*k* = 0.19 – 1.00) and a higher overall percentage agreement of 82% (see Table A.4).

These kappa statistics represent a good level of agreement (Fleiss, 1986); however, both the wide ranges of kappa values indicate that although field use of the ERG22+ against expected performance was reliable across the sample, some individuals clearly did not demonstrate expected performance against the gold standard. Analysis was therefore carried out to examine the percentage agreements for each item in each case (see Table A.5). This found that a number of factors in the Intent domain had particularly low levels of agreement. In particular, factor 19, "harmful end objectives" (51% and 50% agreement), Factor 14, "over-identification with group, cause or ideology" (56% and 52%), and Factor 18, "harmful means to an end" (66% for Mr G, but 83% for Mr H) had low levels of agreement.

Experienced vs. non-experienced comparisons

As a result of the wide range of expected values for performance against the gold standard, an unplanned secondary analysis was conducted that separated those raters considered to be "experienced" raters from those who had limited experience of the ERG22+ for the field reliability study. Consequently, two groups were generated from the sample: one comprised of "experienced" (*n* = 25) and one comprised of "inexperienced" (*n* = 8) (Please see methods for fuller description). The pooled Cohen's kappa statistics for each group along with minimum and maximum values are presented in Table A.6A. Subsequent multivariate ANOVA found a significant multivariate effect of expertise on pooled kappa ratings, (*F* (2,30) = 6.89, *p*< .01. Games-Howell post-hoc tests indicate that this appears to be a result of the experienced group performing significantly better versus the gold standard than the inexperienced group on the Mr G case (*p* <.01) but with no significant effect for the Mr H case (*p* = .932). This suggests that although the inexperienced raters appeared to use the ERG22+ effectively, their use was less consistent across the cases than that of experienced raters.

5. Discussion

Two highly experienced researchers rated 50 convicted extremist cases using the ERG22+ in study one to examine research reliability of the measure. Findings from this study showed high levels of inter-rater reliability between the researchers at both the casewise and itemwise levels. Kappa and ICC[1] scores all fell within the "excellent" range for classification. Analysis of the three ERG22+ domains of Engagement, Capability and Intent, also produced reliability statistics within the "excellent" range. IRR between the two raters across all 50 cases also fell within the "excellent" range with weighted kappas consistently above .90. This suggests that ERG22+ has high levels of research reliability and will produce reliable ratings when used by experienced assessors in research conditions.

In Study 2, a sample of 33 trained practitioners with differing levels of experience rated two test cases specifically developed for the study against "gold standard" ratings to examine filed reliability of the ERG22+. Using two different statistical methods to measure agreement, overall field inter-rater reliability across both case studies were found to be "moderate" to borderline "good".

One case study produced higher reliability than the other, suggesting one case was easier to assess, although both fell within the "moderate" classification range. When examining IRR at the ERG22+ domain level, differences were observed in rating agreements. Reliability for the domains of Engagement and Capability varied from "moderate" to "excellent" across the two cases depending upon the statistic applied, but was consistently poor for the Intent domain. The poor agreement on a number of factors is likely to explain the low overall field IRR for the Intent domain. Given that the research reliability produced high inter-rater agreement for these factors would suggest that their reliability could be significantly improved with further assessor training. It may also be beneficial to amend the definition of the items to improve clarity and differentiation, particularly between factors 18, "harmful means to an end" and 19, "harmful end objectives". "Harmful means to an end" is the factor that considers the extent to which an individual is prepared to commit harm to further their cause, whilst the "harmful end objectives" factors relate to the extent that the individual requires harm to be committed to fulfil their goals. For example, an ideological goal may be to destroy an entire race of people.

As a result of the wide range of performance against the gold-standard ratings, secondary analysis was conducted that separated those raters considered to be "experienced" from those who had limited experience of the ERG22+ and the ratings were compared. Although the inexperienced raters appeared to use the ERG22+ effectively, their use was less consistent across the cases than that of experienced raters, especially for one case, where the experienced group performed significantly better when compared to gold standard scores. This finding, along with the findings from study 1, highlight the importance of the expertise of the user being considered when judging the reliability of the tool. Knowledge and experience are particularly valuable when considering the field of extremism, where assessors are required to have a good level of political awareness of the group, cause or ideology in question to ensure they include political, cultural, and social context in the analysis (Lloyd & Dean, 2015). However, it has been acknowledged that, given the relatively low numbers of extremist offenders, assessors have more limited opportunities to gain this knowledge and maintain their expertise (Lloyd & Dean, 2015). This could especially apply to the assessment of offenders supporting ideologies where numbers in the criminal justice system are low. The study found the performance of the inexperienced raters was worse for the Extreme Right Wing case. This may be because, as there are lower numbers of convicted Extreme Right Wing offenders (Home Office, 2019), non-specialist practitioners may have less exposure to this group in their everyday work, so their understanding may be reduced. Less is known about Extreme Right Wing perpetrators in general, as research has tended to focus on Islamist extremists, (Briggs & Goodwin, 2018). However, as numbers of Extreme Right Wing offenders are increasing (Home Office, 2019), further research is needed to improve our understanding of this group. The findings may also suggest that the ERG22+ is easier to use consistently with some offender groups than others, but this would need to be tested further, before any conclusions could be made. Ongoing and top-up training for assessors would help to improve their knowledge and expertise in the field, as well as increasing their use of the measure. In addition, performance of assessors could be improved by regular monitoring, perhaps in the form of periodic assessments.

While the field ratings were found to be lower than the research ratings, the field reliability was still classified as moderate to good. This is encouraging given that attaining good levels of inter–rater agreement among practitioners working in the field has been found to be difficult. Studies of other forensic assessments have found significantly lower IRR when used in the field than reliability estimates based on agreement between trained and expert researchers (Boccaccini et al., 2008; Murrie et al., 2008; Vincent et al., 2012). This could be attributed to the environments in which practitioners are required to complete assessments, where they are juggling often busy workloads. However, attempts should be made to improve the field reliability of the measure, especially since these are the conditions in which actual cases will be assessed.

A strength of this study was the use of two methods to examine research and field reliability, including both case studies developed specifically for the study and 50 real cases. Field reliability was examined among raters themselves and by comparing with “gold standard” ratings, with ratings being compared against an ideal. This offers greater reliability to the findings, as while a pool of raters may agree amongst themselves, it may be that expert rating of the cases would provide different ratings. In addition, the study reports on a range of reliability statistics that further enhances the quality of the findings.

However, there were a number of limitations to the study which may have had an impact on the emerging findings. First, in study one, only two research experts rated the cases. The study would have been strengthened if more raters had been included, as this would have allowed greater comparison between different expert research raters. Secondly, for study two, field raters were asked to assess only two case studies. The study would have been improved with a larger number of cases being included, as it may have been that the cases developed for the study had some anomalies and ambiguities that made reliable rating particularly difficult. While the cases had been specifically designed for the study and efforts made to reduce any anomalies, they had been designed to be challenging for the assessors. In particular, one case was found to be more difficult to assess reliably, which may have been more ambiguous, especially to inexperienced raters. IRR was found to be considerably higher for study 1 research raters than study 2 field raters. It could be that the cases randomly selected to be included in study 1 were more straightforward than the two specifically developed case studies used in study 2.

A further limitation was that the research and field raters had access to different information which may have influenced the ratings they produced. Research assessors were given information that had been collected as part of the ERG22+ assessment process so there had been some level of organisation and selection of information that was deemed relevant to the assessment, especially in terms of the interviews conducted with offenders, which may have aided assessors in their rating. This was not so for the field case studies.

The studies presented in this paper are the first to examine the IRR of the ERG22+ tool for use with extremist offenders. The findings demonstrate that the tool has IRR ranging from perfect to moderate and is improved when used by experienced practitioners and expert researchers. A number of suggestions can be made from the study that would further enhance the field IRR of the tool. In particular, a number of factors in the Intent domain could be improved by refining their definitions. Training should be as specific as possible and more time dedicated to those factors that appear the most problematic. A period of supervision for those who are less experienced may also be beneficial in improving reliability. Clinicians using the ERG22+ to assess extremist offenders should also keep up to date with socio-political contextual factors that may be relevant to decision making. This could be addressed in regular, ongoing training. As with all newly developed assessment tools, the ERG22+ should undergo a continuous process of review and refinement. This is especially important in the field of extremism, which is a dynamic topic and our understanding of such offenders is continuously evolving and expanding. As the tool is developed, further validation should also be carried out to examine the psychometric properties of the assessment.

While the study highlights the importance of experience and knowledge of field of extremism along with practice in administering ERG22+, even non-experienced users performed satisfactorily which would suggest the ERG22+ is a reliable tool for use in the field.

6. References

Altman, D.G. (1991) *Practical statistics for medical research* (1st ed). London: Chapman and Hall.

Boccaccini, M., Turner, D., & Murrie, D. (2008). Do some evaluators report consistently higher or lower PCL-R scores than others? Findings from a statewide sample of sexually violent predator evaluations. *Psychology, Public Policy, And Law, 14*(4), 262-283. doi: 10.1037/a0014523

Bohrnstedt, G., & Knoke, D. (1994). *Statistics for Social Data Analysis* (1st ed.). Peacock Publishers.

Briggs, R. & Goodwin, M. (2018). We need a better understanding of what drives right wing extremist violence. *LSE Blogs*. http://blogs.lse.ac.uk/politicsandpolicy/archives/24325.

Campbell, T. W. (2004). *Assessing sex offenders: Problems and pitfalls*. Springfield, IL: Charles C Thomas

Campbell, M., French, S., & Gendreau, P. (2007). The Prediction of Violence in Adult Offenders. *Criminal Justice And Behavior, 36*(6), 567-590. doi: 10.1177/0093854809333610

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational And Psychological Measurement, 20*(1), 37-46. doi: 10.1177/001316446002000104

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*(6, Pt.1), 426-443. doi: 10.1037/h0026714

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220. http://dx.doi.org/10.1037/h0026256.

Dean, C. (2014). The healthy identity intervention: the UK's development of a psychologically informed intervention to address extremist offending. In A. Silke, *Prisons, Terrorism and Extremism* (1st ed., pp. 89-108). Oxon: Routledge.

Dernevik, M., Beck, A., Grann, M., Hogue, T. & McGuire, J. (2009). The use of psychiatric and psychological evidence in the assessment of terrorist offenders. *Journal Of Forensic Psychiatry & Psychology, 20*(4), 508-515. http://dx.doi.org/10.1080/13501760902771217

Doren, D. (2002). *Evaluating sex offenders*. Thousand Oaks: SAGE Publications.

Egan, V., Cole, J., Cole, B., Alison, L., Alison, E., Waring, S. & Elntib, S. (2016). Can you identify violent extremists using a screening checklist and open-source intelligence alone?. *Journal Of Threat Assessment And Management, 3*(1), 21-36. http://dx.doi.org/10.1037/tam0000058

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382. doi: 10.1037/h0031619

Fleiss, J. (1986). *Design and Analysis of Clinical Experiments*. Willey, New York.
Fleiss, J.L., Cohen, J., & Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 72, 323-327. http://dx.doi.org/10.1037/h0028106

Fleiss, J.L., Levin, B., & Paik, M.C. (2003) Statistical methods for rates and proportions, 3rd ed. Hoboken: John Wiley & Sons.

Hart, S.D., & Logan, C. (2011) Formulation of Violence Risk using Evidence-Based Assessments: The Structure Professional Judgement Approach. Forensic Case formulation, 83-106.

HM Government. (2018). CONTEST: *The United Kingdom's Strategy for Countering Terrorism*. London: Home Office.

Home Office (2018). *Operation of Police Powers Under the Terrorism Act and Subsequent Legislation. Statistical Bulletin 29/18*. London: Home Office.

Home Office (2019). *Fact Sheet: Right-Wing Terrorism*. https://homeofficemedia.blog.gov.uk/2019/03/19/factsheet-right-wing-terrorism

Kemshall, H. (2001). Risk assessment and management of known Sexual and Violent offenders: A review of current issues. Police research series: Paper 140. London: Home Office.

Koo, T. K., & Li, M. Y. (2016). A guideline for selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163. https://dx.doi.org/10.1016%2Fj.jcm.2016.02.012

Lloyd, M. & Dean, C. (2015). The development of structured guidelines for assessing risk in extremist offenders. *Journal Of Threat Assessment And Management, 2*(1), 40-52. http://dx.doi.org/10.1037/tam0000035

Liht, J., White, W., Savage, S., O'Neill, K. & Conway, L. (2011). Religious Fundamentalism: An Empirically Derived Construct and Measurement Scale. *Archive For The Psychology Of Religion, 33*(3), 299-323. http://dx.doi.org/10.1163/157361211x594159

Manganelli Rattazzi, A., Bobbio, A. & Canova, L. (2007). A short version of the Right-Wing Authoritarianism (RWA) Scale. *Personality And Individual Differences, 43*(5), 1223-1234. http://dx.doi.org/10.1016/j.paid.2007.03.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 46*, 1-330.

Murray, J., & Thomson, D. (2010). Clinical judgement in violence risk assessment. *Europe'S Journal Of Psychology, 6*(1). doi: 10.5964/ejop.v6i1.175

Monahan, J. (1984). The prediction of violent behavior. *American Journal Of Psychiatry, 141*(1), 10-15.

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

Murrie, D., Boccaccini, M., Johnson, J., & Janke, C. (2008). Does interrater (dis)agreement on Psychopathy Checklist scores in sexually violent predator trials suggest partisan allegiance in forensic evaluations?. *Law And Human Behavior, 32*(4), 352-362. doi: 10.1007/s10979-007-9097-5

National Offender Management Service. (2011). *Extremism Risk Guidance. ERG22+ structured Professional Guidelines for Assessing Risk of Extremist offending*. London: Ministry of Justice.

National Offender Management Service. (2011). *Extremism Risk Screening (ERS)*. London: Ministry of Justice.

Politowski, B. (2016). *Terrorism in Great Britain: The Statistics. Briefing Paper,7613,9th June*. London: House of Commons Library.

Powis, B., Randhawa, K. and Bishopp,D. (2019). An examination of the structural properties of the Extremism Risk Guidelines (ERG22+); a structured formulation tool for extremist offenders. *Terrorism and Political Violence*. https://doi.org/10.1080/09546553.2019.1598392.

Pressman, 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. http://dx.doi.org/10.1108/14636641211283057

Pressman, E. and Flockton, J. (2014). Violent extremist risk assessment: issues and applications of the VERA-2 in a high-security correctional setting. In A. Silke, *Prisons, Terrorism and Extremism* (1st ed., pp. 122-144). Oxon: Routledge

Quinsey, V.L., Harris, G.T., Rice, G.T. & Cormier, C.A. (1998). *Violent offenders; Appraising and managing risk*. Washington DC: American Psychological Association.

R Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Roberts, K. & Horgan, J. (2008). Risk assessment and the terrorist. *Perspectives on Terrorism, 2*(6), 9.

RTI International. (2018). *Countering violent extremism: The application of risk assessment tools in the criminal justice and rehabilitation process*. Literature Review. Prepared for First Responders Group Department of Homeland Security Science and Technology Directorate. Retrieved from https://www.dhs.gov/sites/default/files/publications/OPSR_TP_CVE-Application-Risk-Assessment-Tools-Criminal-Rehab-Process_2018Feb-508.pdf.

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

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. http://dx.doi.org/10.1371/journal.pone.0166947

Scott, C., & Resnick, P. (2006). Violence risk assessment in persons with mental illness. *Aggression And Violent Behavior, 11*(6), 598-611. doi: 10.1016/j.avb.2005.12.003

Silke, A. (2014). Risk assessment of terrorist and extremist prisoners. In A. Silke, *Prisons, Terrorism and Extremism: Critical Issues in Management, Radicalisation and Reform* (1st ed., pp. 108-121). London: Routledge.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. doi: 10.1037//0033-2909.86.2.420

Skeem, J. & Monahan, J. (2011). Current Directions in Violence Risk Assessment. *Current Directions In Psychological Science, 20*(1), 38-42. http://dx.doi.org/10.1177/0963721410397271

Vincent, G., Guy, L., Fusco, S., & Gershenson, B. (2011). Field reliability of the SAVRY with juvenile probation officers: Implications for training. *Law And Human Behavior, 36*(3), 225-236. doi: 10.1007/s10979-011-9284-2

Wakeling, H., Mann, R., & Milner, R. (2011). Interrater Reliability of Risk Matrix 2000/s. *International Journal Of Offender Therapy And Comparative Criminology, 55*(8), 1324-1337. doi: 10.1177/0306624x10386933

Webster, S., Mann, R., Carter, A., Long, J., Milner, R., & O'Brien, M. et al. (2006). Inter-rater reliability of dynamic risk assessment with sexual offenders. *Psychology, Crime & Law, 12*(4), 439-452. doi: 10.1080/10683160500036889

Appendix A: Additional tables

Table A.1: Research reliability percentage agreement, weighted kappa, and ICC^2 for ERG22+ items and domains

Domain	Factor	Agreement (%)	Weighted *k*	ICC^1	95% CI	ICC^2	95% CI
Engagement	1.Need to redress injustice	100%	1.00	1.00	-	1.00	-
	2.Need to defend against threats	96%	0.97	0.97	0.95-0.99	0.99	0.98-0.99
	3.Identity meaning and belonging	94%	0.95	0.95	0.91-0.97	0.97	0.95-0.99
	4. Need for status	98%	0.93	0.93	0.88-0.96	0.96	0.94-0.98
	5. Excitement, comradeship & adventure	90%	0.91	0.91	0.84-0.95	0.95	0.92-0.97
	6. Need to Dominate others	96%	0.81	0.81	0.7-0.89	0.90	0.82-0.94
	7. Susceptibility to indoctrination	94%	0.95	0.95	0.91-0.97	0.97	0.95-0.98
	8. Political, moral motivation	98%	0.95	0.95	0.91-0.97	0.97	0.95-0.99
	9. Opportunistic involvement	96%	0.84	0.85	0.74-0.91	0.92	0.85-0.95
	10. Family and/or friends support extremism	86%	0.86	0.86	0.77-0.92	0.93	0.87-0.96
	11.Transitional periods	88%	0.90	0.91	0.84-0.95	0.95	0.91-0.97
	12. Group Influence and Control	94%	0.88	0.88	0.8-0.93	0.94	0.89-0.96
	13. Mental Health Issues	96%	0.88	0.89	0.81-0.93	0.94	0.89-0.97
	Total for Engagement	*92%*	*0.91*	*0.91*	*0.85-0.95*	*0.95*	*0.92-0.97*
Intent	14. Over-identification with group, cause	94%	0.96	0.96	0.93-0.98	0.98	0.97-0.99
	15. Us & Them thinking	94%	0.95	0.95	0.92-0.97	0.98	0.96-0.99
	16. Dehumanisation of the enemy	98%	0.98	0.98	0.97-0.99	0.99	0.99-1
	17. Attitudes that justify offending	100%	1.00	1.00	-	1.00	-
	18. Harmful means to an end	96%	0.97	0.97	0.95-0.98	0.98	0.97-0.99
	19. Harmful end objectives	96%	0.97	0.97	0.95-0.98	0.99	0.98-0.99
	Total for Intent	*90%*	*0.97*	*0.97*	*0.94-0.98*	*0.99*	*0.97-0.99*
Capa bility	20. Personal knowledge, skills, competencies	94%	0.94	0.94	0.89-0.96	0.97	0.94-0.98
	21. Access to networks, funding, equipment	94%	0.94	0.94	0.89-0.96	0.97	0.94-0.98
	22. Criminal history	98%	0.85	0.85	0.74-0.91	0.92	0.85-0.95
	Total for Capability	*98%*	*0.91*	*0.91*	*0.84-0.95*	*0.95*	*0.91.0.97*
	Overall	*93%*	*0.93*	*0.93*	*0.87-0.96*	*0.96*	*0.93-0.98*

Table A.2: Research reliability percentage agreement and Cohen's kappa for ERG22+ cases

Case	% agreement	Weighted *k*	Case	% agreement	Weighted *k*
1	100	1.00	26	100	1.00
2	91	.91	27	100	1.00
3	95	.97	28	100	1.00
4	64	.36	29	100	1.00
5	95	.98	30	100	1.00
6	100	1.00	31	100	1.00
7	95	.96	32	100	1.00
8	100	1.00	33	100	1.00
9	91	.93	34	95	.97
10	95	.95	35	100	1.00
11	100	1.00	36	100	1.00
12	100	1.00	37	82	.85
13	100	1.00	38	100	1.00
14	86	.78	39	95	.97
15	73	.80	40	100	1.00
16	100	1.00	41	100	1.00
17	100	1.00	42	91	.94
18	95	.96	43	91	.92
19	86	.86	44	95	.97
20	73	.81	45	95	.96
21	95	.97	46	100	1.00
22	86	.90	47	95	.96
23	95	.93	48	91	.93
24	100	1.00	49	100	1.00
25	91Tab	.94	50	100	1.00

Table A.3: ICC and Fleiss' kappa values for ERG22+ domains from both cases in the full sample

Case	% agree	Domain	Fleiss *k*	ICC[1]	95% CI	ICC[2]	95% CI
Mr G	78.3	Engagement	0.57	0.59	0.41-0.80	0.98	0.96-0.99
	59.2	Intent	0.14	0.18	0.06-0.59	0.88	0.69-0.98
	80.1	Capability	0.52	0.62	0.29-0.98	0.98	0.93-1.00
	73.3	*Overall*	*0.47*	*0.48*	*0.34-0.66*	*0.97*	*0.95-0.98*
Mr H	82.2	Engagement	0.63	0.65	0.48-0.84	0.98	0.97-0.99
	78.6	Intent	0.25	0.29	0.12-0.72	0.93	0.81-0.99
	86.5	Capability	0.71	0.79	0.48-0.99	0.99	0.97-1.00
	81.8	*Overall*	*0.59*	*0.60*	*0.47-0.76*	*0.98*	*0.97-0.99*
Total	77.6						

Table A.4: Field rater performance versus the "gold standard" for both cases

	Mr G	Mr H
% agree	73.3%	82.0%
k mean	0.65	0.73
Median	0.65	0.79
Max	1.00	1.00
Min	0.17	0.19

Table A.5: Item level agreement ratings for both cases in the full sample

Domain	Factor	Mr G	Mr H
Engagement	1.Need to redress injustice	83%	88%
	2.Need to defend against threats	62%	83%
	3.Identity meaning and belonging	78%	100%
	4. Need for status	66%	62%
	5. Excitement, comradeship & adventure	66%	83%
	6. Need to Dominate others	88%	94%
	7. Susceptibility to indoctrination	73%	83%
	8. Political, moral motivation	69%	94%
	9. Opportunistic involvement	51%	94%
	10. Family and/or friends support extremism	100%	51%
	11.Transitional periods	94%	100%
	12. Group Influence and Control	88%	54%
	13. Mental Health Issues	100%	83%
Intent	14. Over-identification with group, cause	56%	62%
	15.Us & Them thinking	69%	100%
	16. Dehumanisation of the enemy	66%	83%
	17. Attitudes that justify offending	62%	94%
	18. Harmful means to an end	51%	83%
	19. Harmful end objectives	51%	50%
Capability	20. Personal knowledge, skills, competencies	100%	83%
	21. Access to networks, funding, equipment	78%	88%
	22. Criminal history	62%	88%

Table A.6: IRR statistics versus the gold standard (k) and versus each other (*k*^{F}, ICC^1, ICC^2) for the experienced and inexperienced groups

	Mr G		Mr H	
	Experienced	**Inexperienced**	**Experienced**	**Inexperienced**
% agree	78%	67%	82%	82%
k mean	0.73	0.45	0.72	0.73
Median	0.77	0.45	0.79	0.78
Max	1.00	0.64	1.00	0.90
Min	0.20	0.17	0.19	0.53
Fleiss kappa	0,55	0.29	0.59	0.57
ICC^1	0.57 [0.43-0.73]	0.32 [0.17-0.52]	0.60 [0.46-0.76]	0.58 [0.42-0.75]
ICC^2	0.97 [0.95-0.98]	0.79 [0.62-0.90]	0.97 [0.95-0.99]	0.92 [0.85-0.96]