Ministry
of Justice

The Extremism Risk Guidance 22+

An exploratory psychometric analysis

Ian A. Elliott

Kiran Randhawa-Horne

Olivia Hambly

Data and Analysis Directorate

Ministry of Justice Analytical Series

2023

Data and Analysis 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 2023

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

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

ISBN 978-1-84099-985-3

Acknowledgements

The authors would like to thank our colleagues in HM Prison and Probation Service and the Counter Terrorism – Assessment and Rehabilitation Centre. We also thank our colleagues in MOJ Data and Analysis for their assistance in accessing relevant data. We would also like to thank two external academic peer reviewers for their constructive and informative comments as well as internal colleagues who also offered time and expertise in reviewing this study, particularly Philip Howard and Alana Diamond. We also thank Rory MacLeod in the Joint Extremism Unit Projects and Programme Team for some valuable and timely feedback. Finally, we would like to acknowledge the influence of our esteemed colleague Dr. Beverly Powis on the development and completion of this study.

Contents

List of tables

List of figures

1. Summary 1

2. Introduction 4
2.1 Study aims and objectives 4
2.2 Risk and need assessment 5
2.3 The Extremism Risk Guidance 22+ 6
2.4 Psychometric performance 8

3. Methods 10
3.1 Sample 10
3.2 Procedures 10
3.3 Limitations 13

4. Findings 15
4.1 Classification rates 15
4.2 Structural validity 16
4.3 Construct validity 19

5. Conclusions 24
5.1 Recommendations 27

References 28

Appendix A 34
Previous ERG22+ psychometric studies 34

Appendix B 35
Technical appendix 35
Descriptions of metrics and associated thresholds 37

List of tables

Table 1: A list of the ERG 22+ Items and the dimension to which they belong 7

Table 2: Reliability statistics for all items and for the four-dimension fit 20

Table 3: Construct validity metrics and subsequent classification for all items 21

Table B1: Frequency of offence type, where frequency is greater than 5 35

Table B2: Summary of overall responses to each ERG 22+ item in the sample 35

Table B3: The results of three linear regression analyses exploring the association between ERG 22+ dimension “point totals” and the overall SCJ classification 37

Table B4: Goodness-of-fit statistics for the one-, three-, and four-component models 38

Table B5: Item factor loadings for the one, three, and four-dimension models 39

Table B6: Factor loadings for a two-component Engagement dimension 40

Table B7: Descriptive data for classifications across the four dimensions in the adapted ERG 22+ structure 40

Table B8: Item-to-item correlations for the Ideological engagement domain 42

Table B9: Item-to-item correlations for the Non-ideological engagement domain 42

Table B10: Item-to-item correlations for the Intent domain 42

Table B11: Item-to-item correlations for the Capability domain 43

List of figures

Figure 1: Response rates across the whole ERG 22+ sample 15

Figure 2: Relationships between point total and overall classifications on each dimension 17

Figure 3: The number of components identified by the PA, MAP, and EMPKC tests 18

Figure 4: Distributions of point totals for each dimension of the four-dimension model 19

Figure 5: Difficulty parameters for each item per domain 23

1. Summary

Introduction and study aims

The U.K. Government considers countering extremism across the criminal justice system a prime focus and therefore improvements to assessments are key to our understanding of the risk posed by individuals of concern. Most assessments that exist across the criminal justice systems address individual's general risk and needs with the aim of preventing reoffending, protecting the public, and encouraging successful community reintegration.

Few assessments, however, exist specifically for extremism convictions. The Extremism Risk Guidance 22+ (ERG 22+) is a structured professional judgement assessment designed to inform operational decisions for individuals with extremism-related convictions, regardless of ideology. It consists of 22 items across three dimensions: Engagement, Intent, and Capability. Cases are assessed against each item and dimension and recorded as being “strongly present”/“significant”, “partly present”/”some”, or “not present”/“minimal”.

Previous Ministry of Justice studies have found that the ERG 22+ broadly measures what it is intended to measure and is used consistently across a variety of cases and different assessors. The aim of this study was to provide further analytical data that could be used to further judge the validity and reliability of the ERG 22+ and its constituent items.

Methodological approach and interpreting findings

A sample of 310 ERG 22+ reports was subjected to a variety of validity and reliability analyses designed to examine the structure and performance of the ERG 22+. Validity refers to the extent to which evidence provides a sound scientific basis for interpretations of scores derived from tests and assessments: does the information generated by the ERG 22+ aid clinicians in making decisions. Two forms of validity were examined in this study: structural validity (do the items group into the expected dimensions) and construct validity (do the items measure the intended theoretical concepts in the manner expected).

The main limitation of this study is that although it uses a large sample size relative to other studies of the extremist population, the sample size is smaller than those typically used to validate other assessments within and outside of criminal justice. The cases included represent a mix of ideologies and over a range of time during which their nature may have changed. Statistical validation techniques are typically designed to evaluate objective tests (e.g., educational exit examinations like GCSEs) rather than assessments based on professional judgement. They also often recommend direct alterations to tests whereas we have used them to explore performance more broadly to inform development. Finally, the implementation of the ERG 22+ has evolved over the timeframe of in this study (e.g., training, quality assurance), which may have affected our findings.

Key findings

Structural validity

- We explored whether the observed data grouped into the three expected dimensions of the ERG 22+ (Engagement, Intent, and Capability). Exploratory analysis suggested that while the items of the Intent and Capability dimensions do group together well statistically – some better than others – the Engagement dimension measured two concepts that we labelled Ideological engagement and Non-ideological engagement.

- When we separated the items of the engagement dimension into those two forms of engagement it statistically significantly improved the "goodness-of-fit" metrics (i.e., how well the observed data statistically "fits" into a predetermined theoretical structure). The adapted four-dimension structure was, relative to the original three-dimension structure, a *better* way to explain our data. Nevertheless, those metrics remained short of an acceptable statistical threshold – albeit closer – that would allow us to conclude that the four-dimension structure was, overall, a *good* way to explain our data.

- The subsequent four dimensions (Ideological engagement, Non-ideological engagement, Intent, and Capability) were, however, found to correctly measure one dominant construct each: a prerequisite for the statistical tests we planned to use to explore the items to judge their individual contribution to the overall assessment.

Dimension reliability

- Only the Intent domain exceeded acceptable thresholds for reliability – an estimate of the extent to which respondents consistently answer the same way for the same question – with Ideological engagement closely approaching that threshold. Reliability was below the acceptable threshold for Non-ideological engagement and Capability.

Item validity

- Individual dimension items varied in performance on metrics designed to measure whether the item generated useful information for decision-making (i.e., contributed positively to overall totals, provided unique information, and discriminates between people who have more of the attribute being measured and those who have less).

- Nine items were found to be statistically acceptable (very few concerns on any performance metrics). Eight items were statistically ambiguous (some issues of concern). Five items were statistically substandard (several metrics of concern). These were “Opportunistic involvement”, “Family and/or friends support extremism”, “Transitional Periods”, “Mental health issues”, and “Criminal history”. The items “Mental health issues” and “Criminal history” performed poorly on all metrics.

Conclusions

Overall, the current theoretical three-dimension structure of the ERG 22+ is not a good way to explain the data it generates. The findings, however, suggest this lack of support is not because the dimensions do not exist or that underlying theory is incorrect. Instead, it may be due to (a) the existence of sub-dimensions within the existing dimensions, (b) some items being misclassified as belonging to the wrong dimension, and/or (c) some items lacking consistent interpretation or not constructively contributing to decision making.

Five items were found to be statistically substandard. This does not mean that they are not relevant to risk in this population, it means that in our sample they did not generate information in a manner consistent with the other items in their respective dimension. If the ERG 22+ is designed to provide information on which inferences can be made about those being assessed these items are not contributing information to beneficially inform those decisions to the same extent as items that have performed well.

We recommended that these five items should be subjected to review to establish whether improvements can be made or if removal is more appropriate. It was also recommended that a further study be considered examining how assessors interpret ERG 22+ items to understand whether poor validity is a result of assessment design or practical application.

2. Introduction

Given the ongoing threat posed to national security by extremists, the U.K. Government considers work to tackle extremism across the criminal justice system, including prisons, as crucial in delivering the U.K.'s counter-terrorism strategy (Ministry of Justice, 2022). Both the 2018 CONTEST strategy and the U.K. Government's response to the Hall report (Ministry of Justice, 2022) highlight the challenges in detecting individuals who may be inspired to commit extremist attacks and pledge to improve assessments of the risk posed by individuals of concern. The Hall report (2022) also welcomed new bodies created within HM Prison and Probation Services to establish standards for risk assessment, risk reduction, and rehabilitation as well as seeking to improve the quality of relevant research.

Although the academic literature continues to grow, there is limited research exploring assessments and interventions for individuals with extremism convictions. Effective assessment and intervention are required to understand people's needs, prevent reoffending, and successfully reintegrate individuals into society (Powis et al. 2019a). The less robust literature on extremism compared to general or sexual violence also means there are comparatively fewer tools available (Logan & Sellers, 2020).

Assessments and interventions used within the wider offending population may be less effective with extremists due to their differing risks, needs, and motivations (Silke, 2014). Assessing risk in a valid and reliable manner is further complicated by the fact that violent extremists are not a homogenous group (Sarma, 2017; Silke, 2014). Although it is argued that violent extremists have similar criminal profiles to those with violent and non-violent convictions, the motivations for and the circumstances in which violent extremists commit offences are varied and complex (Powis et al. 2019a; Basra & Neumann, 2016; Dean, 2014). Understanding these complexities will likely assist in shaping the appropriate interventions and management actions needed to reduce risk of further extremist offences.

2.1 Study aims and objectives

The aim of this study was to use observed operational The Extremism Risk Guidance 22+ (ERG 22+) data to judge the validity and reliability of the ERG 22+ and its constituent items, in terms of its internal structure, construct validity, and criterion validity. Our objectives were to analyse an observed sample of ERG 22+ assessments to explore:

1. The levels of endorsement for the 22 items, via proportions of classifications.
2. The quality of the internal structure of the ERG 22+, via statistical factor analysis.
3. Reliability and validity statistics for the ERG 22+, via statistical techniques drawn from classical test theory (e.g., reliability estimates, item and total correlations) and item response theory (e.g., item difficulty, item discriminatory ability).

It was not an aim of this study to explore alternative configurations of the ERG 22+. Given the limitations outlined in the following methods sections, our findings are intended to provide insight into how each item on the ERG 22+ is performing given real data. There are several interpretations for what these findings tell us about what is “good” versus “poor” performance of the ERG 22+ or its constituent items. Finally, the ERG 22+ is a structured clinical judgement tool and not a test. The intention of this study is to provide a data-driven approach using traditional metrics of validity and reliability from assessment evaluation to identify areas in which the ERG 22+ might be reviewed for performance and where the quality and relevance of that information for decision-making might be improved.

2.2 Risk and need assessment

There are a range of measures widely used across judicial systems to assess risk and accurate estimation is an important first step in reducing that risk (Campbell, French & Gendreau, 2007). Risk assessment has historically followed two approaches, each with strengths and limitations. The first generation of assessments centre around clinical judgement. Information on social, environmental, behavioural, and personality factors related to harmful behaviours were collected through detailed interviewing and observation (Campbell et al., 2007; Monahan, 1984; Quinsey et al., 1998). A second generation of assessments adopted an actuarial approach, with a greater focus on risk prediction rather than prevention and an aim of better standardising the assessment of risk. Actuarial risk assessment identifies risk factors that are statistically predictive of outcomes and calculates a numeric value for predicted risk (Campbell et al., 2007; Kemshall, 2001).

Terrorism and extremism, however, are relatively rare compared to other types of crime and there are insufficient numbers of individuals with those convictions from whom predictive risk factors can be statistically identified (Egan et al., 2016; Sarma, 2017; Scarcella et al., 2016). There has been an emphasis on understanding how best to *manage* risk rather than *predict* risk (Murray & Thomson, 2010). It has been argued that it is very difficult for clinicians to accurately predict risk of either recidivism or harmful behaviour (RTI International, 2018) and some suggest that focus should be placed on helping inform sentencing decisions, management, and supervision (Kemshall, 2001).

Logan and Sellers (2020) highlighted Structured Professional Judgement (SPJ) approach to risk assessment as good practice in risk assessment and management specifically for those with extremist convictions. The SPJ approach combines structured risk assessments and clinical interviews. It 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). It is recommended as a flexible approach, identifying risk factors systematically, considering individual contexts and any other additional factors that may be relevant (Monahan, 2012; Skeem & Monahan, 2011; Roberts & Horgan, 2008).

The SJP approach generates information for both assessment and management of the individual (Sarma, 2017). Additional benefits of the SPJ approach to practitioners in this field include clear audit trails to inform decisions and minimum information requirements that form the foundation of a coherent risk management strategy (Sarma, 2017). In assessing and managing risk for extremists, the Extremism Risk Guidance 22+ (ERG 22+) adopts the industry standard SPJ approach. As described, this case formulation approach analyses specific factors relating to an individual and the context around their circumstances that led them to offend. This study explores the psychometric performance of the ERG 22+, focusing on the validity and reliability of the 22 items.

2.3 The Extremism Risk Guidance 22+

The ERG 22+ has been available for use across Her Majesty's Prison and Probation Service of England and Wales since September 2011 (Lloyd & Dean, 2015; National Offender Management Service, 2011). It is intended for use with all individuals convicted of extremism related offences (regardless of their cause or ideology) and, since implemented, has been completed on all individuals convicted of an extremism related offence. It is important to emphasise that the ERG 22+ is a structured clinical judgement tool and is intended to provide decision-makers with information on which to make operational decisions about individual cases.

Table 1: A list of the ERG 22+ Items and the dimension to which they belong

Dimension	Item	Description	Assessor metrics
Engagement	1	Need to redress injustice and express grievance	**Item classifications:** Strongly present (2) Partly present (1) Not present (0) **Overall categories:** High Medium Low
	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	
Intent	14	Over-identification with group and/or 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	
Capability	20	Personal knowledge, skills, and competencies	**Item classifications:** Significant (2) Some (1) Minimal (0) **Overall categories:** Significant Some Minimal
	21	Access to networks, funding, and equipment	
	22	Criminal history	

The assessment is completed by registered psychologists or probation officers with experience working in forensic settings who have received full training in its administration. The ERG 22+ is designed to be collaborative, however, if an individual does not want to participate then other sources will be used and they are given the opportunity to review the final report upon completion. This tool provides a framework for facilitators in which to consider all the relevant information about an individual that could drive them to become involved in extremism and subsequent offending (Powis et al., 2019a).

Table 1 lists the 22 items of the ERG 22+ as well as the labels used to represent assessors' judgements for each item and domain. These items focus on three dimensions: "Engagement" with 13 items, "Intent" with 6 items and "Capability" with 3 items (National Offender Management Service, 2011). The Engagement dimension includes items related to the individual, their circumstances, and their beliefs about the group, cause and/or ideology (Powis et al., 2019a). Examples include "Need for status" and "Susceptibility to indoctrination". Although "Mental health issues" is also included as an item in this dimension, it noted that the link between mental health and risk remains unclear and complex. Additional specialist consideration of risk and need should also be taken if mental health is identified as a specific issue in an individual case (Powis et al., 2019a).

The Intent dimension includes items such as "Us & them thinking" and "Attitudes that justify offending". These items relate most heavily to an individual's potential to act on and readiness to support illegal activity and/or violence to further the goals of an extremist group, cause or ideology (Powis et al., 2019a). The Capability dimension refers to items that could facilitate illegal activity on behalf of a group, cause and/or ideology. These include an assessment subject's criminal history, access to criminal networks and their personal skills required to commit an offence. Lastly, the "+" suffix refers to anything else that may influence each of the dimensions and can be considered in the assessment.

2.4 Psychometric performance

Any assessments used to make decisions about the risk posed by extremists needs to fundamentally have good validity and reliability if it is to be considered useful and meaningful (Powis et al., 2019a). According to the American Educational Research Association and their collaborators (AERA, APA & NCME, 2014), validity 'refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests' and that the process of validation 'involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations' (p. 9). They argue that a sound argument for validity should integrate various metrics representing multiple strands of evidence that support a coherent account for the interpretation of the assessment. These strands include (but are not limited to):

1. The extent to which items and components reflect the true nature of the construct being measured (structural and construct validity).

2. The extent to which assessment scores predict performance on other variables or compare with other similar or divergent measures (criterion validity).

Early examinations of the structural and construct validity of the ERG 22+ have reported good overall internal consistency, but that improvements might be possible in some dimensions with further development (Powis et al., 2019a). That study also found that statistical techniques used to draw novel theoretical components from the data itself (rather than test the data for specific components) identified 5 components that broadly matched those in the existing ERG 22+ (see Appendix A). That study also found that the reliability of the ERG 22+ – its ability to produce the same result over multiple administrations of the test – was high. However, the reliability was found to be moderate for the Engagement and Intent domains specifically, and low for the Capability domain.

A second Powis et al. (2019b) study (see Appendix A) found that the inter-rater reliability (IRR) of the ERG 22+ – its ability to produce the same result across multiple administrators judging the data – was good both across different cases and across each of the ERG 22+ items, with both whole assessment and dimension level IRR in "excellent" range. Field IRR was lower than research IRR but still classified "moderate" to "good". Experience and training were found to be important to consistent use across field raters. Improvements in monitoring of the performance of assessors was recommended.

3. Methods

3.1 Sample

This study is a retrospective analysis of data held by HM Prison and Probation Service for 422 individuals subject to an initial ERG 22+ assessment between March 2011 and January 2020. These data were linked to Offender Assessment System (OASys) records. After data linking and after cases with missing data were removed, 350 ERG 22+ records remained. Finally, 40 cases involving assessments of females (11.4% of the sample) were removed due to concerns about the introduction of hidden and unknown effects due to the confounding influence of gender differences in ideology or behaviour. This left 310 records available for analysis (73.5% of data). Sample information can be found in Appendix B.

Of the sample, 80.6% were classified as having an Islamist-influenced extremist ideology, 11.0% classified as having an extreme right-wing ideology, and 8.4% classified as a mixture of animal rights activism, other religious ideology (e.g., Sikh-influenced), or other political ideology (e.g., extreme left-wing). The average age of the sample was 33.5 years and ages ranged from 18 to 67. Ethnicity was not recorded in 9.7% of cases. Where officially recorded ethnicity was provided, individuals in the sample were predominantly of aggregated Asian (49.7%) or white (22.6%) ethnicities, while smaller numbers were of aggregated black (9.4%), other (5.16%), or mixed (4.5%) ethnicities. For participants where the information was available, most (60.6%) had been charged with “Other offences against the State or public order”. More offence type information is provided in Appendix B.

3.2 Procedures

A series of quantitative psychometric analyses were conducted to examine the structure and validity of the ERG 22+. Since examination of the various forms of validity required different procedures and techniques, we have provided information on each of the various methodological approaches as a technical appendix (Appendix B). The two forms of validity examined in this study were structural and construct validity. All analyses were conducted using R (version 3.6.3). Descriptions of individual statistical tests and thresholds for acceptability can be found in Appendix B.

Classification rates

The first set of analyses examined the aggregate proportions of cases, per item, that were classified as being “strongly present,” “partly present,” or “not present” in the case of the Engagement and Intent dimensions or “minimal,” “some,” or “significant” in the case of the Capability dimension. These were presented as percentages.

Structural validity

The second set of analyses examined the extent to which our observed data support assumptions about the internal structure of the ERG 22+. Two structural assumptions were tested. The first was that the classifications are “ordinal”. Ordinal data (a) are categorical (i.e., assign individuals to a category rather than a number) and (b) form categories that have a natural rank order but where distance between categories is not strictly defined[1].

The use of numerical scores is not permitted in the routine administration or interpretation of the ERG 22+. However, for statistical purposes it was necessary for categorical classifications to be transformed into numerical format (low = “0”, medium = “1” or high = “2”) and for sums of those numbers to fairly mirror the classifications applied. In instances where classifications were combined, the case was removed as ambiguous and missing.

The second assumption was that the theoretical dimensions applied used to group the items (the Engagement, Intent, and Capability dimensions) are supported by the data. We used a combination of confirmatory factor analysis (CFA) and exploratory factor analysis (EFA) to examine if our data supported use of three dimensions. CFA is a form of factor analysis that is used to verify if a pre-determined structure of a set of variables exists in observed data. EFA is a form of factor analysis that reduces observed data down into a smaller set of dimensions depending on statistical relationships between variables. In that sense, CFA is “top-down” and theory-driven and EFA is “bottom-up” and data-driven.

Tests of “unidimensionality” were also conducted. To be considered unidimensional, each dimension should measure a single attribute. The statistical tests used to examine construct validity required that ERG 22+ dimensions be “essentially” unidimensional,

[1] This distinguishes ERG 22+ classifications from “nominal” data with no natural order (e.g., “dog”, “cat”, “rabbit”), “interval” data with a fixed order that is equal and consistent (e.g., temperature in degrees Celsius), or “ratio” data: interval data with a fixed zero point (e.g., weight in grams; height in metres).

meaning that each should measure one dominant major factor and that any minor factors that are unintentionally[2] measured should have only a small influence on item scores.

Construct validity

The third set of analyses examined the extent to which assessments generate information that can aid inferences about assessed individuals or groups. For the ERG 22+, this means whether higher or lower classifications on items or dimensions help decide whether one individual has a more or less clinical need than others with different classifications. It also represents the extent to which the various items and domains representing different aspects of clinical need collectively contribute to classifications in the expected manner.

Construct validity was examined using metrics derived from two approaches to psychometric analysis: classical test theory (CTT) and item response theory (IRT). Classical test theory (CTT) is a body of psychometric theory that predicts outcomes of psychological assessment such as the difficulty of items based on the premise that a person's observed score on an assessment is the sum of a “true score” and an amount of error. As opposed to CTT, item response theory (IRT) assesses the design, analysis, and scoring of assessments that measure latent variables (abilities, attitudes, etc) based on the relationship between individuals' performances on each assessment item and their overall levels of performance. Unlike CTT, it does not assume that each item is equally “difficult”.

Three tests derived from classical test theory were conducted. Five estimates of reliability – the ability of an assessment test to produce consistent results across multiple administrations of an assessment – were generated. Item-to-total correlations examined nonrelevance, where items fail to contribute to the overall total. Item-to-item correlations examined redundancy*,* where multiple items duplicate and measure the same concept.

An IRT graded response model[3] was used to generate difficulty and discriminatory metrics. Difficulty values indicate how much of the latent variable (i.e., clinically judged need) is needed to increase the classification upwards by one category (e.g., from “none”

[2] For example, an assessment designed to test social problem solving might also unintentionally measure the test-taker’s abilities in other cognitive functions, like verbal comprehension.

[3] IRT metrics were produced using the “ltm” R package (version 1.1-1).

to “partly” and “partly” to “strong”). Second were “discriminatory parameter” values that indicates how well each item distinguishes between overall high and low scorers.

3.3 Limitations

The main limitation of this study is the small sample size, albeit large in comparison to other studies of a group that is highly unique and specialist in the wider population of individuals with related but non-extremist convictions. Sample sizes of 200 to 500 have been considered adequate for psychometric analysis (Goldman & Raju, 1986; Jiang et al., 2016), although other studies have indicated that sample sizes up to 500–1,500 are required to ensure optimum accuracy in findings (e.g., Kilmen & Demirtasli, 2012; Kutscher et al, 2019). Therefore, although this represents a large sample in the context of the wider related literature, the findings of this study should be interpreted with appropriate caution.

Because the assessments were conducted between 2011 and 2020 the nature of any underlying ideologies may have changed over time. The sample also contains a mix of ideological groups. Although the imbalance between those with Islamist-influenced ideologies and those with other ideologies should not affect how the assessment is administered or scored, it could result in (a) an underestimation of the validity of the ERG 22+ for Islamist-influenced extremists due to inclusion of other ideologies or (b) an underestimation of the validity of the ERG 22+ more broadly due to a lack of ideological diversity and difficulty generalising findings beyond Islamist-influenced extremism. The same rationale was the reason for excluded the even smaller number of female cases.

Psychometric evaluations often use test and validation subsets of the full dataset so that issues can be identified in a larger test dataset, modifications made, and the consequent new structure can be evaluated in a smaller validation dataset. This study sought only to identify issues for potential review, not to recommend a new structure for the ERG 22+. Therefore, some psychometric tests are used in an atypical way, albeit with precedent in other areas of criminal justice (e.g., Paquette & Cortoni, 2020), for example, for use of item response theory (IRT) to evaluate an assessment for individuals with sexual convictions.

Assessments of risk in criminal justice are typically focused on predicting reconvictions. This was not examined here (a) because this is not a non-discretionary (i.e., actuarial or statistical) risk assessment (for the difference between discretionary and non-discretionary assessments see Hart et al., 2016) and (b) because there are insufficient numbers to provide data on prison misconduct or post-release reconvictions for robust enough results.

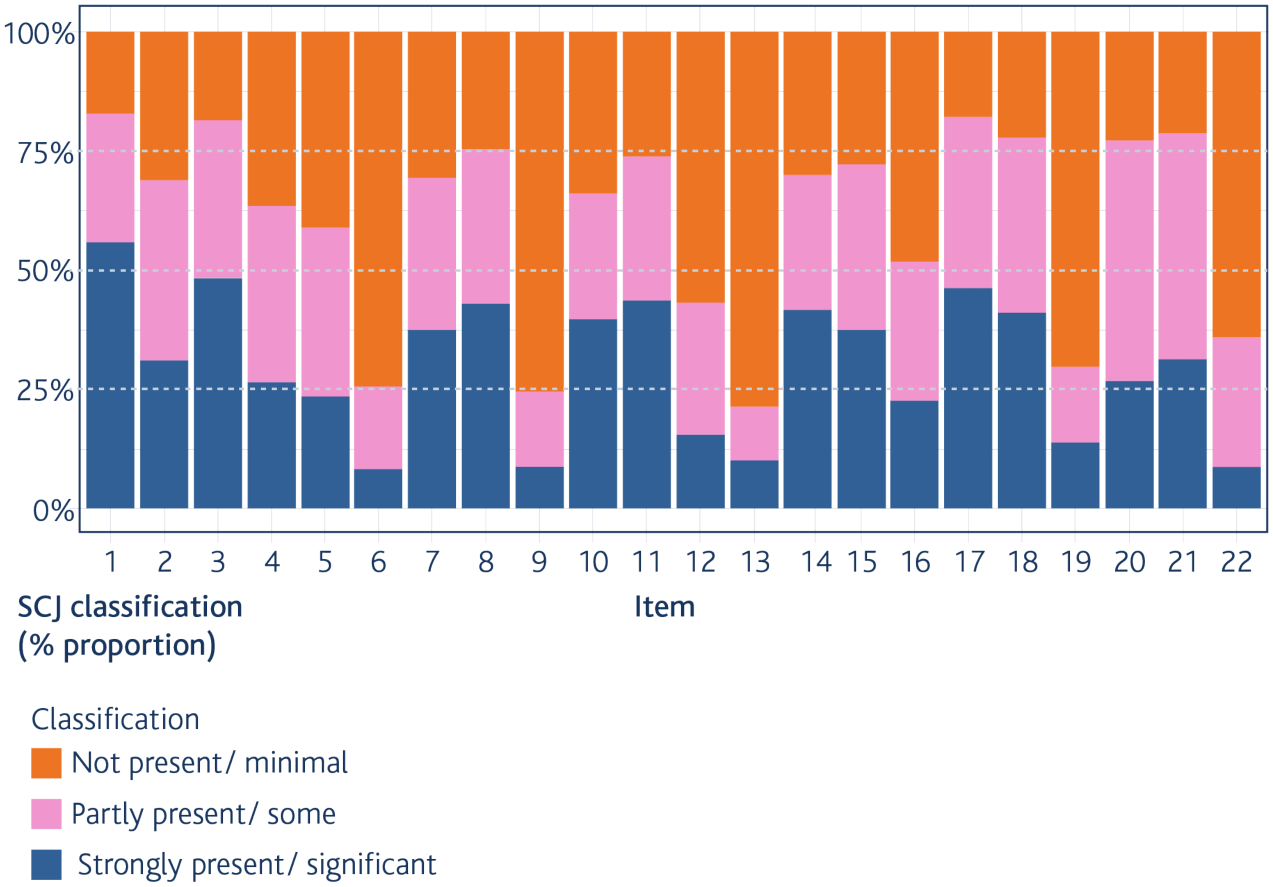
Finally, there are also operational issues in that various support structures around the ERG 22+ have evolved over the lifetime of the ERG 22+. There have been changes to the training curriculum during the time period covered in this study, which may have introduced some natural variability in the way in which the assessment is administered. Quality assurance processes too have evolved over that time meaning that confidence in the quality of the ERG 22+ assessments has increased as the ability to examine that quality has improved. Furthermore, the identification of and granting of access to sources of risk- and need-relevant security information with which to conduct ERG 22+ assessments may have also introduced some natural variability into the data over time.

4. Findings

4.1 Classification rates

Overall classification proportions were calculated (see Figure 1 and Table A1).

Figure 1: Response rates across the whole ERG 22+ sample



Seven items were classified “strongly present” or “significant” in more than 40% of cases:

- Need to redress injustice and express grievance (item 1)
- Identity, meaning & belonging (item 3)
- Political moral motivation (item 8)
- Transitional periods (item 11)
- Over-identification with group and/or cause (item 14)
- Attitudes that justify offending (item 17)
- Harmful means to an end (item 18).

Five items were classified “not present” or “minimal” in greater than 60% of cases:

- Need to dominate others (item 6)
- Opportunistic involvement (item 9)
- Mental health issues (item 13)
- Harmful end objectives (item 19)
- Criminal history (item 22)

Low endorsement is not necessarily a concern, as the item may be successfully identifying a small but important minority. It can, however, be indicative of psychometric nonrelevance (i.e., that the item does not add independent value to the assessment). Conversely, high endorsement is not necessarily always beneficial. If all cases receive high scores on an item, that item may not contribute to the identification of instances whereby clinicians judge issues of concern to be more or less present and consequently informing decisions about case management. These questions are examined in the following sections.

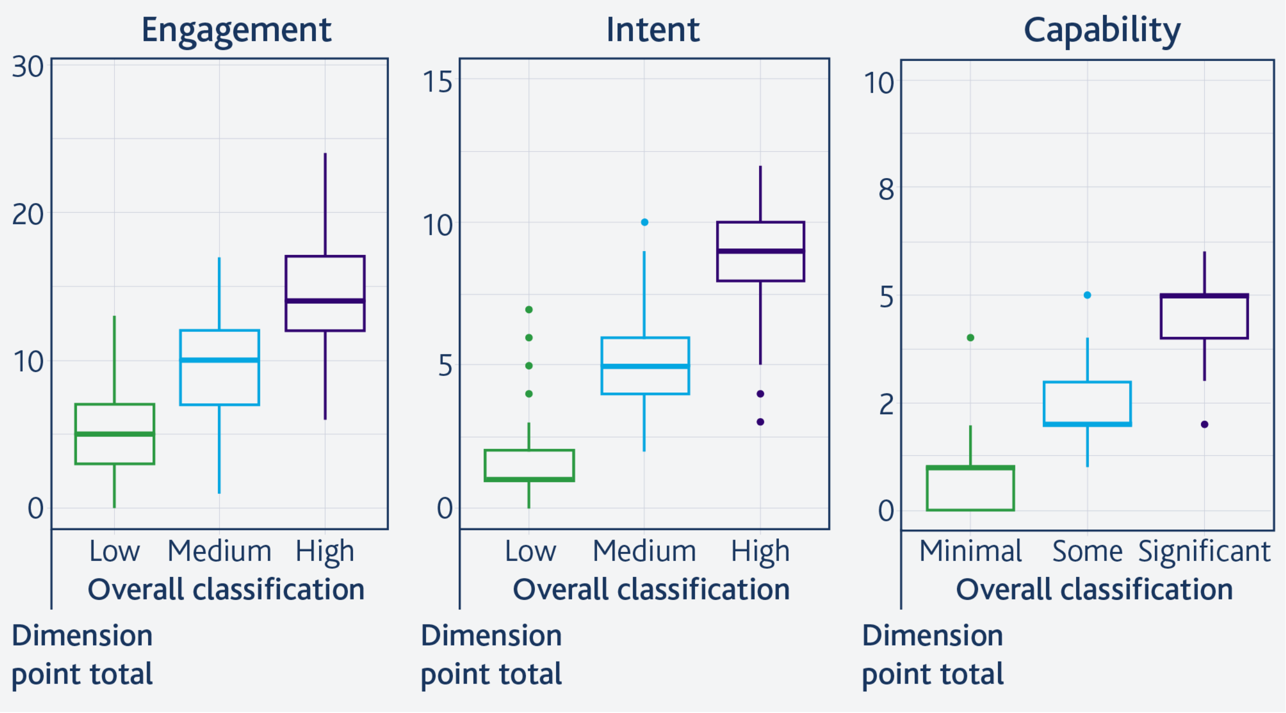
4.2 Structural validity

Ordinal nature of classifications

Linear regression analyses indicated that there were statistically significant predictive relationships between ERG 22+ point totals and overall classifications, on all three dimensions (see Table B1). This meant that an individual’s “point total” (i.e., sum of 0s, 1s, and 2s) predicted whether they were classified as “low”, “medium”, or “high” with higher classifications associated with larger point totals (see Figure 2). A change from “low” to “medium” was associated with around a 0.5 standard deviation increase in point total and a move from “low” to “high” around a 1 standard deviation increase. This supports the assumption that numerical point totals fairly represent categorical decisions of assessors.

To reiterate, our aim here was to establish whether sum totals of scores are of a nature that allows us to perform the statistical analyses planned. The use of “point totals” and overall numerical scores is not permitted in the routine administration of the ERG 22.

Figure 2: Relationships between point total and overall classifications on each dimension



Dimension structure

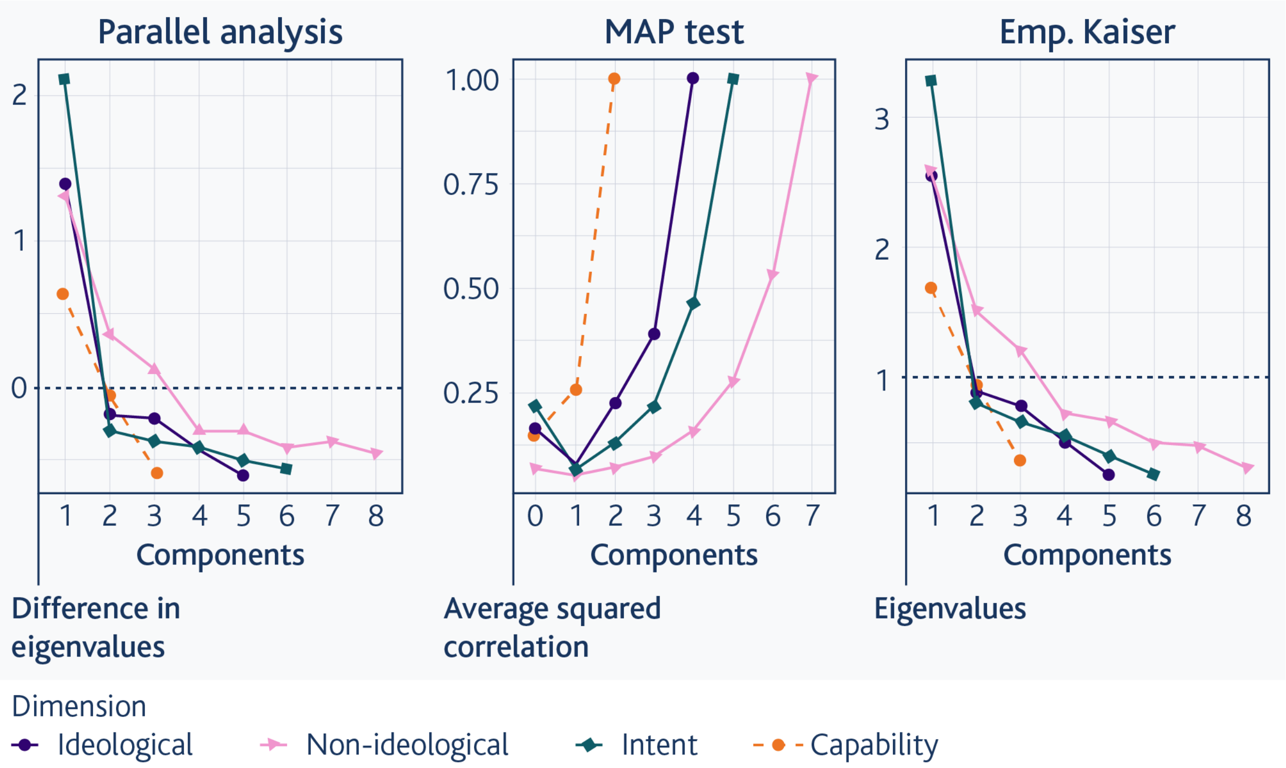
Our structural analysis found poor support for the existing three-domain structure of the ERG 22+. The CFA generated poor values for absolute fit when the observed data was separated into three dimensions, but the three-dimension fit was statistically better than when the observed data was coerced into one dimension (i.e., no separate dimensions).

This led us to conclude that (a) the three-dimension model did not meet the structural quality requirements necessary for us to apply tests from measurement theory and (b) the observed data may contain a theoretically plausible multi-dimensional model that would meet those requirements. We tested for multiple dimensions in the observed data using parallel analysis (PA: Horn, 1965), empirical Kaiser criterion (EMPKC: Braeken & Van Assem, 2017, Kaiser, 1960), and minimal average partial (MAP: Velicer, 1976) tests. The PA and the EMPKC recommended four dimensions and the MAP recommended three indicating that, at least, the observed data was multi-dimensional.

To remain as close to the theoretical foundations of the ERG 22+ as possible, we next explored whether or not the three original dimensions were themselves unidimensional, using PA, EMPKC, and MAP tests. All tests suggested that the Intent and Capability dimensions were essentially unidimensional. However, they suggested that the

Engagement domain might contain more than one dimension: PA suggested four, EMPKC four, and MAP two. To avoid "overfitting" (i.e., recommending a structure that is so specific to these data that we cannot generalise any further) we concentrated on the simplest solution: that the existing Engagement dimension is actually two dimensions.

Figure 3: The number of components identified by the PA, MAP, and EMPKC tests



An EFA limited to two dimensions, separated the items of the Engagement domain into components we labelled "Ideological" and "Non-ideological" engagement. Ideological engagement included redressing injustice and grievances (Item 1), defending against threats (Item 2), and, explicitly, having a "political and moral motivation" (Item 8). Non-ideological engagement included a need for identity, meaning, and belonging (Item 3), excitement, comradeship, and adventure (Item 5), and transitional periods (Item 11).

We then conducted further PA, EMPKC, and MAP tests on the Ideological engagement and Non-ideological engagement domains, to examine whether separating the engagement items in that manner was theoretically defensible. All tests suggested that Ideological engagement was essentially unidimensional. The MAP test suggested that Non-ideological engagement was also essentially unidimensional, but the PA and EMPKC tests suggested three dimensions. Since the Non-ideological engagement dimension contains 8 items it was considered (a) unlikely that three dimensions of 2 or 3 items would be either theoretically plausible or operationally useful and (b) predictable for a variety of constructs associated only by their not being ideological. The MAP findings also provided support that the dimension was unified enough to apply measurement theory (Figure 3).

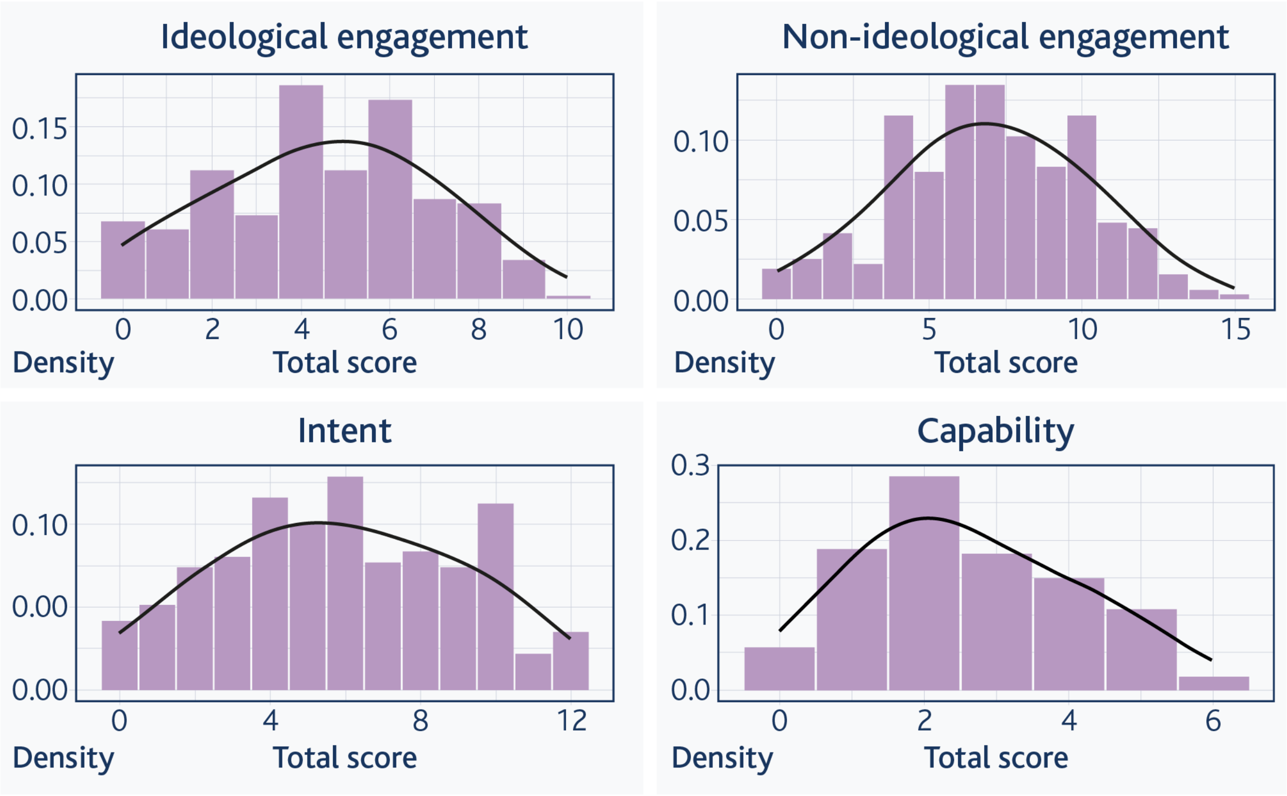
Finally, we used CFA to examine the absolute quality of fit for this adapted four-dimension structure and a chi-square difference test to examine the relative fit compared to the original three-dimension model. The four-dimension model had a statistically significantly superior fit to the three-dimension model with improved, but still insufficient, fit metrics (albeit where the 95% confidence intervals fell within the acceptable threshold). But the four dimensions were considered sufficiently unidimensional for measurement theory.

4.3 Construct validity

Normality

The point totals were broadly normally distributed for all four dimensions, with excess skew and kurtosis values within acceptable thresholds (see Figure 4 and Table B5).

Figure 4: Distributions of point totals for each dimension of the four-dimension model



Reliability

Thresholds for acceptable reliability were exceeded by the Intent scale, almost met by the Ideological engagement scale, but were not met by the Non-ideological engagement and Capability scales (see Table 2). Although alpha, omega total, and ten Berge and Zergers' mu, were lower, GLB^a and maximum split half metrics provide some potential reassurance for the internal consistency of the adapted Non-ideological scale. To aid interpretation of these metrics, a recent Monte Carlo simulation study of 30 reliability estimators found Guttman's lambda 2 and ten Berge-Zergers' mu to be consistently accurate (Cho, 2022).

Although these reliability statistics imply improvement in the Capability dimension from Powis et al.'s (2019) analysis, it is likely due to the reporting of Cronbach's alpha alone in that study. A common criticism of Cronbach's alpha is that it may be unduly affected by test length (e.g., Sijtsma, 2009, Taber, 2018) and the Capability scale has only three items. The alternative metrics presented here suggest those 2019 reliability findings were an underestimation, albeit reliability remains below the threshold for acceptability. Although the all-ERG 22+ reliability exceeded thresholds for acceptability, this should be interpreted cautiously given we know a one-dimension model (all items together) results in a poor statistical fit with our data.

Table 2: Reliability statistics for all items and for the four-dimension fit

Dimension	α	ω^t	GLB^a	λ^2	μ^2	Mean
Ideological	0.74	0.76	0.79	0.76	0.76	0.76
Non-ideological	0.65	0.67	0.81	0.68	0.68	0.70
Intent	0.83	0.84	0.88	0.84	0.84	0.85
Capability	0.57	0.67	0.67	0.62	0.64	0.63
All	0.85	0.86	0.97	0.86	0.86	0.88

Item correlations

Four items had item-to-total correlations below a Pearson's *r* value of 0.2, indicating potential non-relevance, where items fail to contribute to the overall total (see Table 3). These items were "Opportunistic involvement", "Family/friends support extremism", "Transitional periods", and "Mental health issues".

No two items had an item-to-item correlation greater than 0.8, indicating that no two items were obviously duplicating one another (see Tables B8 to B11). Two pairs of items had strong correlations (greater than 0.7). The first pair was “Need to redress injustice and express grievance” and “Political and/or moral motivation” (*r* = 0.73). The second pair was “Us and them thinking” and “Dehumanisation of the enemy” (*r* = 0.71). This indicates that they may, at least, be assessing a similar general underlying element of risk or need.

Three item pairs were found to have a weak negative correlation. The “Family/Friends support extremism” item negatively correlated with “Opportunistic involvement” (*r* = -.20) and “Mental health issues” (*r* = -.25). This indicated that when “Family/Friends support extremism” was endorsed to a greater extent, “Opportunistic involvement” and “Mental health” are endorsed to a slightly lesser extent. The third pairing indicated that “Susceptibility to indoctrination” negatively correlated with “Opportunistic involvement” (*r* = -04), indicating that when “Susceptibility to indoctrination” was endorsed to a greater extent, “Opportunistic involvement” was endorsed to very slightly lesser extent.

Discrimination and difficulty

The outcomes of the construct validity analyses are broadly positive (see Table 3), with 17 items found to have a “moderate” or above ability to discriminate between cases relatively higher and lower in the variable being measured (i.e., clinical need per dimension). Nine items had “high” or “very high” discriminatory ability. Only four items were found to have poor discriminatory ability: “Family/friends support extremism”, “Transitional periods”, “Mental health issues”, and “Criminal history”. Difficulty, the relative amount of clinical need required to obtain higher classifications, was well balanced across items (see Figure 5).

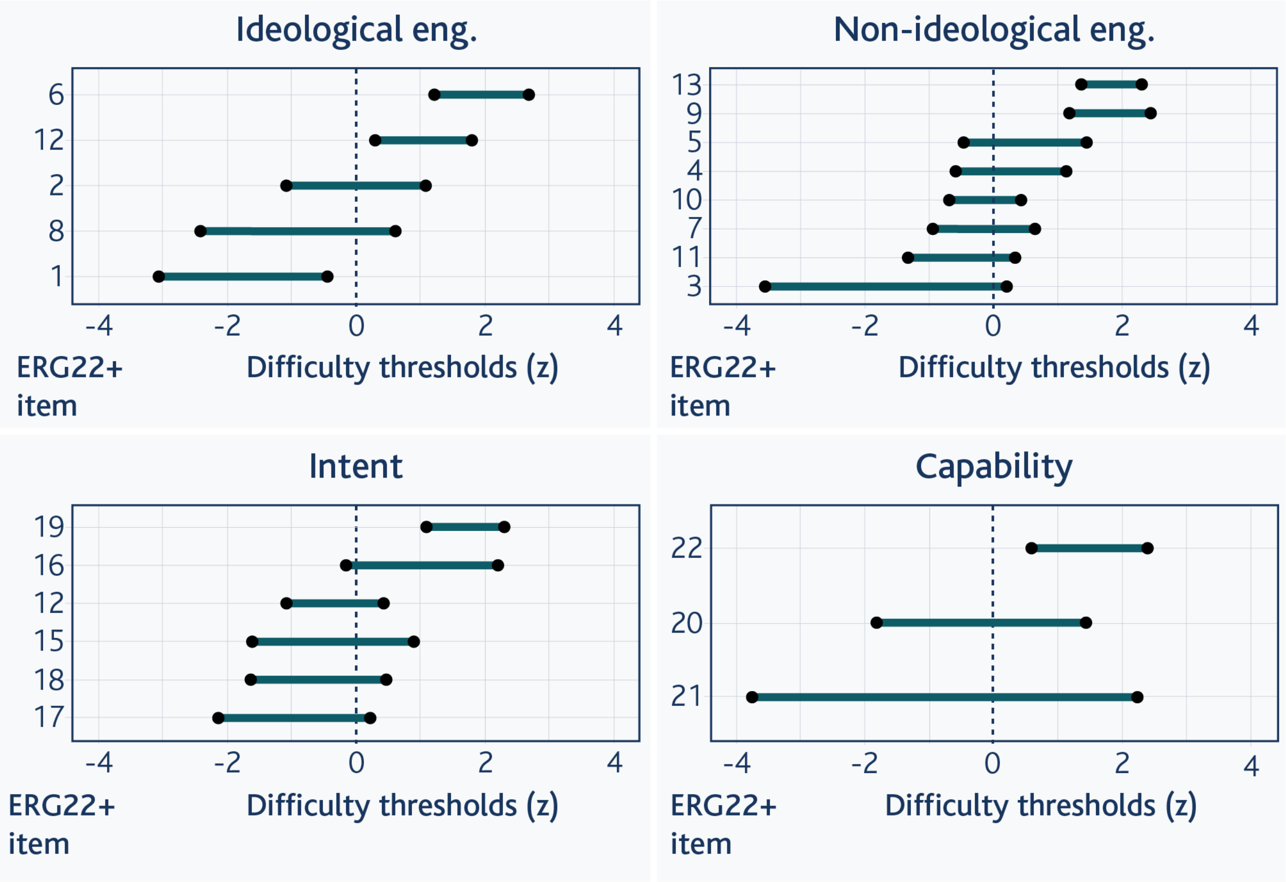
Table 3: Construct validity metrics and subsequent classification for all items

Item	Short name	CFA loading	Item-to-total	Difficulty parameters		Discrim.
				***b*1**	***b*2**	
1	Injustice and grievance	0.57	0.56	-3.08	-0.45	2.76
2	Defend against threats	0.45	0.45	-1.08	1.08	1.40
3	Identity, meaning & belonging	0.60	0.53	-3.56	0.20	3.62
4	Need for status	0.26	0.24	-0.59	1.13	0.67
5	Excitement, comrade. & adventure	0.38	0.44	-0.47	1.45	1.16

Item	Short name	CFA loading	Item-to-total	Difficulty parameters		Discrim.
				*b*1	*b*2	
6	Dominate others	0.21	0.25	1.21	2.67	0.85
7	Susceptibility to indoctrination	0.37	0.28	-0.94	0.64	0.95
8	Political/moral motivation	0.60	0.58	-2.43	0.61	3.08
9	Opportunistic involvement	0.07	0.11	1.18	2.43	0.43
10	Family/friends support extremism	0.15	0.07	-0.69	0.43	0.31
11	Transitional periods	0.40	0.37	-1.33	0.33	1.22
12	Group influence/control	0.22	0.21	0.29	1.79	0.52
13	Mental health issues	0.12	0.16	1.36	2.30	0.53
14	Over-identification with group	0.47	0.46	-1.09	0.43	1.22
15	Us and them thinking	0.55	0.57	-1.62	0.90	2.17
16	Dehumanisation of enemy	0.50	0.58	-0.17	2.20	2.39
17	Attitudes justify offending	0.48	0.54	-2.15	0.21	1.56
18	Harmful means to end	0.44	0.45	-1.63	0.47	1.30
19	Harmful end objectives	0.33	0.41	1.09	2.30	1.32
20	Knowledge, skills & competencies	0.47	0.42	-1.82	1.43	1.69
21	Networks, funds & equipment	0.56	0.45	-3.77	2.22	4.40
22	Criminal history	0.10	0.13	0.58	2.39	0.31

Item factor loading onto the dimensions was poor in absolute terms with only two loadings greater than 0.6 (“Political/moral motivation” to Ideological engagement and “Identify, meaning and belonging” to Non-ideological engagement). However, item loadings for only four items fell below a loading value of 0.2 (“Opportunistic involvement”, “Family/friends support extremism” and “Mental health issues” to Non-ideological engagement and “Criminal history” to Capability). This should be viewed in the context of the poor absolute fit of the four-dimension model, indicating that some items are “mis-specified” (i.e., allocated to an incorrect dimension or not a constituent of any dimension) and their loading values likely to be an incorrect estimation of their relationship with clinical need.

Figure 5: Difficulty parameters for each item per domain



5. Conclusions

A positive association between the categorical clinical classifications on each dimension and their associated numerical “point total” (sum of item scores) was found. The ERG 22+ is a structured clinical judgement tool, however, individual cases can generate the same overall point total based on any combination of items: identical point totals do not equate to identical need. Nevertheless, evidence for transforming classifications into numbers supports their use for quantitative research and development.

The 310 cases of observed data did not appear to support the existing 3-dimension structure of the ERG 22+. Exploratory analyses indicated that the Intent and Capability dimensions appear to be broadly sound, but the Engagement dimension appears to be measuring more than one construct. Further exploration indicated that two forms of engagement are being measured, which we labelled “Ideological” and “Non-ideological”.

The ideological dimension appeared to consist of items related to political and moral motivation involving a sense of injustice, grievance, and threat. These concepts may also overlap with concepts measured in the Intent domain. The non-ideological dimension appeared to consist of items related to other individual needs (e.g., excitement, status, and control) and vulnerabilities to external factors (e.g., involvement by and indoctrination from family, friends, and associates, mental health issues). This dimension should be explored further, as analyses suggest it may be measuring a variety of non-ideological constructs.

Internal consistency also appears to be mixed. Only the Intent dimension was found to exceed acceptable thresholds for reliability (as defined as internal consistency, not test-retest reliability). The Ideological engagement dimension approached but did not meet the threshold. Consistency, however, was limited for the Non-ideological engagement and Capability dimensions. This, however, may simply be due to misspecification of items rather than a fault of theory. We have already noted that the variation in themes across the Non-ideological dimension and the limited number of items in the Capability domain means it would not take many mis-specified items to negatively affect internal consistency.

The item-level analyses indicate that items appear to be performing to different levels. Firstly, five items were classified “not present” or “minimal” in the majority of cases: “Mental health issues”, “Need to dominate others”, “Opportunistic involvement”, “Harmful end objectives”, and “Criminal history”. Although low endorsement is not necessarily a concern, it can, however, be indicative of psychometric nonrelevance.

Based on the various psychometrics tests, overall, the items can be categorised into three tiers:

Tier 1: Statistically acceptable items (none or very few metrics of concern).

- Need to redress injustice (Item 1: Ideological engagement domain)
- Need to defend against threats (Item 2: Ideological engagement)
- Identity, meaning & belonging (Item 3: Non-ideological engagement)
- Political, moral motivation (Item 8: Ideological engagement)
- Us & Them thinking (Item 15: Intent)
- Dehumanisation of the enemy (Item 16: Intent)
- Attitudes that justify offending (Item 17: Intent)
- Personal knowledge, skills, competencies (Item 20: Capability)
- Access to networks, funding, equipment (Item 21: Capability)

Tier 2: Statistically ambiguous items (some metrics of concern)

- Need for status (Item 4: Non-ideological engagement)
- Excitement, comradeship & adventure (Item 5: Non-ideological engagement)
- Need to dominate others (Item 6: Ideological engagement)
- Susceptibility to indoctrination (Item 7: Non-ideological engagement)
- Transitional periods (Item 11: Non-ideological engagement)
- Over-identification with group, cause (Item 14: Intent)
- Harmful means to an end (Item 18: Intent)
- Harmful end objectives (Item 19: Intent)

Tier 3: Statistically substandard items (several metrics of concern)

- Opportunistic involvement (Item 9: Non-ideological engagement)
- Family and/or friends support extremism (Item 10: Non-ideological engagement)

- Group influence and control (Item 12: Ideological engagement)
- Mental health issues (Item 13: Non-ideological engagement)
- Criminal history (Item 22: Capability)

“Mental health issues” (Item 13) and “Criminal history” (Item 22) performed poorly on all metrics. These are arguably the two most static variables in the ERG 22+. They are not endorsed frequently, do not appear to fit well in their given dimension relative to other items, do not correlate with overall point totals, do not discriminate between high and low overall scorers, and are not scored consistently between different assessors and/or cases. It is worth noting that information required to judge mental health and criminal history are recorded elsewhere (e.g., OASys, Police National Computer) and other assessments exist that might provide better data to inform case related decisions outside of the ERG 22+.

It is also worth exploring whether any items should be reverse scored. When converted into a numerical format, some items negatively correlated with others: if classifications on one increased, classifications on the other typically decreased. An example of this was the Mental health and Opportunistic involvement items. For a test to provide useful information for decision making, one item should not contradict another: increases in need identified by each item should contribute to increases in the amount of need identified overall, which, in turn, are then reflected in overall classifications. Since the ERG 22+ is a SPJ assessment, there will not be a “score” that can be reversed. Instead, if should be ensured that assessors are interpreting each item in a manner that results in it working parallel to and in the same direction as other items rather than contrary to other items.

Poor performance on psychometric metrics, however, does not mean that an item is not relevant to risk in this population. It means that in our sample they did not generate information in a manner consistent with the other items in their respective dimension. If the ERG 22+ is designed to provide information on which inferences can be made about the person or people being assessed these items are not contributing information to beneficially inform those decisions to the same extent as items that have performed well. It is possible that they might contribute positively to another dimension or in another form.

It is also worth noting that some of those items found to be substandard are, relative to the other items, more “static” in that they are historical, unchangeable, and to a greater extent binary (e.g., you either have a criminal history, or have received a mental health diagnosis, or not). It is therefore possible that some of them are simply randomly distributed across higher and lower need cases and therefore less statistically associated with overall ERG 22+ classifications. While a high-need case might indicate greater grievance, political motivation, pro-violence attitudes, need for excitement, and so forth, mental health, criminal history, or having family or friends involved in extremism may simply be incidental.

5.1 Recommendations

Items in Tiers 1 and 2 should be given a light-touch review to ensure that they continue to produce informative data. Tier 2 items could also be considered for more comprehensive review. Items in Tier 3 should be subject to comprehensive review to establish (a) if improvements can be made to them in their current state or (b) if they are candidates for removal. Lines of enquiry in that review may include (but are not limited to):

- Are expected interpretations being adequately communicated to assessors?
- Are assessors interpreting the underlying item concepts correctly and consistently?
- Is enough relevant information available to assessors to address the item purpose?
- Are items being interpreted as changeable (dynamic) or unchangeable (static)?
- Are any items related to concepts that are better addressed by other assessments?
- Are items being interpreted in terms of presence of risk or an absence of strengths?
- Are any items trying to measure multiple concepts that might be better separated?

Finally, an additional study should be considered focusing specifically on those items in Tiers 2 and 3. This study could utilise cognitive interview techniques to establish how assessment-facilitators and assessors are interpreting the items and formulating their responses (see Willis, 2004, for example). This type of study would provide more detailed insights into whether items are not performing as expected due to issues of theory and/or assessment design but are due to ambiguity in interpretation and/or assessor training. A study of this nature could also help ensure that new or amended items that are the consequence of any review processes are being interpreted correctly from their inception.

References

AERA, APA, & NCME. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Baker (2001). The basics of item response theory (2nd Ed.). Second Edition. College Park, MD. ERIC Clearinghouse on Assessment and Evaluation.

Barrett, P. (2007). Structural equation modeling: Adjusting model fit. Personality and Individual Differences, 42, 815–24.

Basra, R., & Neumann, P. R. (2016). Criminal pasts, terrorist futures: European jihadists and the new crime-terror nexus. Perspectives on Terrorism, 10(6), 25–40.

Bentler, P. M. & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletins, 88, 588–606.

Braeken, J., & Van Assen, M. A. (2017). An empirical Kaiser criterion. Psychological Methods, 22(3), 450–466.

Browne, M.W. & Cudeck, R. (199). Alternative ways of assessing model fit (pp. 136–162). In: K. Bollen & J. Long (Eds.). Testing structural equation models. Newbury Park, CA: Sage.

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

Campbell, M., French, S., & Gendreau, P. (2007). The Prediction of Violence in Adult Offenders. Criminal Justice and Behavior, 36(6), 567–590.

Cho, E. (2022). The accuracy of reliability coefficients: A reanalysis of existing simulations. Psychological Methods. Advance online publication. Retrieved June 8, 2022 from https://psycnet.apa.org/doiLanding?doi=10.1037%2Fmet0000475

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

Doren, D. (2002). Evaluating sex offenders. Thousand Oaks, CA: 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.

Goldman, S. H., & Raju, N. S. (1986). Recovery of one- and two-parameter logistic item parameters: An empirical study. Educational and Psychological Measurement, 46(1), 11–21.

Hall, J, QC. (2022). Terrorism in prisons (Presented to Parliament pursuant to Section 36(5) of the Terrorism Act 2006). London, U.K.: Crown Copyright.

Hart, S. D., Douglas, K. S., & Guy, L. S. (2016). The structured professional judgement approach to violence risk assessment: Origins, nature, and advances. In L. Craig & M. Rettenberger (Volume Eds.) and D. Boer (Series Ed.), The Wiley handbook on the theories, assessment, treatment of sexual offending (pp. 643–666). London, U.K.: Wiley.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). Applied statistics for the behavioral sciences (5th ed.). Boston, MA: Houghton Mifflin.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179–185.

Hu, L. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1–55.

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. Frontiers in Psychology, 7, 109.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 1, 141–151.

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

Kilmen, S., & Demirtasli, N. (2012). Comparison of test equating methods based on item response theory according to the sample size and ability distribution. Procedia-Social & Behavioral Sciences, 46, 130–134.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2(3), 151–160.

Kutscher. T, Eid, M. & Crayen, C. (2019). Sample Size Requirements for Applying Mixed Polytomous Item Response Models: Results of a Monte Carlo Simulation Study. Frontiers in Psychology, 10, 2494.

Logan, C., & Sellers, R. (2020). Risk assessment and management in violent extremism: a primer for mental health practitioners. The Journal of Forensic Psychiatry & Psychology, 32(3), 355–377.

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

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. Psychological Methods, 23(3), 412–433.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes. Structural Equation Modelling, 11(3), 320–341.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum Associates.

Ministry of Justice (2022). Tackling Terrorism in Prisons: A Response to the Independent Reviewer of Terrorism Legislation's Review of Terrorism in Prisons. London, U.K.: Crown Copyright.

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, & Law, 18, 167–205.

Murray, J., & Thomson, D. (2010). Clinical judgement in violence risk assessment. Europe’s Journal of Psychology, 6(1), 128–149.

National Offender Management Service (2011). Extremism Risk Guidance. ERG 22+ structured Professional Guidance for Assessing Risk of Extremist offending. London, U.K.: Ministry of Justice.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York, NY: McGraw-Hill.

Paquette, S., & Cortoni, F. (2020). The development and validation of the Cognitions of Internet Sexual Offending (C-ISO) Scale. Sexual Abuse, 32(8), 907–930.

Powis, B., Randhawa, K., & Bishopp, D. (2019a). An examination of the structural properties of the Extremism Risk Guidance (ERG 22+): A structured formulation tool for extremist offenders. Terrorism and Political Violence, 33(6), 1141–1159.

Powis, B., Randhawa, K., Elliott, I., & Woodhams, J. (2019b). Inter-rater reliability of the Extremism Risk Guidance 22+ (ERG 22+). London, U.K.: Ministry of Justice.

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

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. Psychometrika, 74(1), 145–154.

RTI International (2018). Countering violent extremism: The application of risk assessment tools in the criminal justice and rehabilitation process. Durham, NC: RTI International. Retrieved from https://www.dhs.gov/sites/default/files/publications/OPSR_TP_CVE-Application-Risk-Assessment-Tools-Criminal-Rehab-Process_2018Feb-508.pdf.

Roberts, K., & Horgan, J. (2008). Risk assessment and the terrorist. Perspectives on Terrorism, 2(6), 3–9.

Sarma, K. M. (2017). Risk assessment and the prevention of radicalization from nonviolence into terrorism. American Psychologist, 72(3), 278–288.

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).

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74(1), 107.

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, U.K.: Routledge.

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. Current directions in psychological science, 20(1), 38–42.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. Research in science education, 48(6), 1273–1296.

ten Berge, J. M., & Socan, G. (2004). The greatest bound to reliability of a test and the hypothesis of unidimensionality. Psychometrika, 69, 613–625.

ten Berge, J. M., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. Psychometrika, 43(4), 575–579.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. Psychometrika, 41, 321–327.

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.

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.

Willis, G. B. (2004). Cognitive interviewing: A tool for improving questionnaire design. London, U.K.: Sage.

Appendix A
Previous ERG22+ psychometric studies

A 2019 study (Powis et al., 2019a) used principal components analysis (PCA) identified seven factors that accounted for 64% of the explained variance. The seven factors were “motivation and ideology” (23% of variance), “Identity and vulnerability” (10%), “Status” (8%), “Influence” (7%) and “Personal knowledge and influence” (6%). A subsequent multidimensional scaling model, with a relatively good fit (a coefficient of alienation of 0.23, where less than 0.20 is considered good), identified 5 components that broadly accord with the factors suggested by the PCA. These were “Identity & external Influence”, “Motivation & ideology”, “Criminality”, “Capability”, and “Status and personal Influence”.

The alpha coefficient for ERG22+ in the 2019 validation study was 0.80, indicating high internal consistency (Powis et al., 2019a). The Engagement scale generated an alpha coefficient of 0.65, the Intent scale an alpha coefficient of 0.79, indicating moderate reliability. The Capability scale generated an alpha coefficient of 0.46, considered low.

A study of IRR of the ERG 22+ found differences between “research” and “field” IRR (Powis et al., 2019b). Research IRR describes the reliability of the measure under laboratory conditions (i.e., a small number of trained assessors across multiple cases). It has been suggested that IRR is likely to be higher among researchers as they tend to administer the tool in large numbers (Campbell, 2004; Wakeling et al., 2011; Powis et al., 2019b). Research IRR was good both across different cases and across each of the ERG 22+ items, with both whole assessment and dimension level IRR in “excellent” range.

Field IRR describes the reliability of the tool under “field” conditions (i.e., multiple trained facilitators using the tool across a smaller number of cases). Field IRR for the ERG 22+ was lower than research IRR but still classified “moderate” to “good”. Experience and training were found to be important to consistent use across field raters. Improvements in monitoring performance of assessors was recommended (Powis et al., 2019b). However, differences in experience and interpretation mean field reliability is often observed to produce lower consensus (Campbell, 2004; Doren, 2002; Webster et al., 2006).

Appendix B
Technical appendix

Additional sample data

Table B1: Frequency of offence type, where frequency is greater than 5

Home Office offence code	Offence description	Frequency	Percentage
66	Other offences against the State or public order	188	60.6
3	Conspiracy to murder/Threats to kill	25	8.1
1	Murder	15	4.85
5	Assault with intent to cause serious harm	15	4.85
8	Racially or religiously aggravated harassment with/without injury	11	3.5
35	Blackmail	7	2.3
56	Arson endangering/not endangering life	6	1.9
Other*		41	13.2
NA	Unknown	2	0.7

* Including attempted murder (n = 1), manslaughter (n = 1), drug offences, weapon offences, burglary, robbery, false imprisonment, criminal damage, and obstruction of justice, where frequency was 5 cases or fewer.

Endorsement rates

Table B2: Summary of overall responses to each ERG 22+ item in the sample

Dimension & item	Response		
Engagement	Not present	Partly present	Strongly present
1	53 (17.1%)	84 (27.1%)	173 (55.8%)
2	96 (31.0%)	118 (38.0%)	96 (31.0%)
3	57 (18.4%)	104 (33.5%)	149 (48.1%)
4	113 (36.5%)	115 (37.1%)	82 (26.5%)
5	127 (41.0%)	110 (35.5%)	73 (23.5%)
6	230 (74.2%)	54 (17.4%)	26 (8.4%)
7	95 (30.6%)	99 (32.0%)	116 (37.4%)

Dimension & item	**Response**		
8	76 (24.5%)	101 (32.6%)	133 (42.9%)
9	234 (75.5%)	49 (15.8%)	27 (8.7%)
10	105 (33.9%)	82 (26.5%)	123 (39.6%)
11	81 (26.2%)	94 (30.3%)	135 (43.5%)
12	176 (56.8%)	86 (27.7%)	48 (15.5%)
13	243 (78.4%)	36 (11.6%)	31 (10.0%)
Intent	Not present	Partly present	Strongly present
14	93 (30.0%)	88 (28.4%)	129 (41.6%)
15	86 (27.7%)	108 (34.8%)	116 (37.5%)
16	149 (48.2%)	91 (29.4%)	70 (22.4%)
17	55 (17.7%)	112 (36.1%)	143 (46.2%)
18	69 (22.2%)	114 (36.8%)	127 (41.0%)
19	218 (70.3%)	49 (15.8%)	43 (13.9%)
Capability	Minimal	Some	Significant
20	71 (22.9%)	156 (50.3%)	83 (26.8%)
21	66 (21.3%)	147 (47.4%)	97 (31.3%)
22	198 (63.9%)	85 (27.4%)	27 (8.7%)

Ordinal nature of responses

Linear regression analyses indicated that there were statistically significant predictive relationships between ERG 22+ point totals and overall classifications, on all three dimensions (see Table B3). A one unit increase in overall classification from “low” to “medium” (or “minimal to “some”) was related to a 0.56 standard deviation increase in point total for engagement, a 0.76 standard deviation increase for intent, and a 0.69 standard deviation increase for capability. Furthermore, a one unit increase in overall classification from “medium” to “high” (or “some” to “significant”) was related to a 1.3 standard deviation increase in point total for engagement, a 1.67 standard deviation increase for intent, and a 1.64 standard deviation increase for capability.

Table B3: The results of three linear regression analyses exploring the association between ERG 22+ dimension “point totals” and the overall SCJ classification

Dimension	ΔR^2	*B*	SE *B*	*β*	*p*
Engagement	0.46				
Intercept		5.19	0.68		<.0001
Overall = Medium		4.33	0.73	0.51	<.0001
Overall = High		9.16	0.73	1.08	<.0001
Intent	0.69				
Intercept		1.68	0.24		<.0001
Overall = Medium		3.33	0.28	0.53	<.0001
Overall = High		7.34	0.30	1.11	<.0001
Capability	0.68				
Intercept		0.81	0.11		<.0001
Overall = Some		1.60	0.13	0.53	<.0001
Overall = Significant		3.74	0.15	1.06	<.0001

Descriptions of metrics and associated thresholds

Structural analysis metrics

Confirmatory factor analysis

CFA is a form of factor analysis that is used to test whether measures of an underlying construct are consistent with the assessment developers' interpretation of the nature of that construct: do the observed data support the developers' theoretical decisions about how to combine the necessary items into a coherent and interpretable whole. Each CFA produces “goodness-of-fit” metrics (e.g., Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and chi-square values) (see Table B4).

Table B4: Goodness-of-fit statistics for the one-, three-, and four-component models

	1-dimension	3-dimension	4-dimension
Chi-square	870.12	786.53	661.58
Degrees of freedom	209.00	206.00	203.00
Chi-square *p* value	0.00	0.00	0.00
CFI	0.58	0.63	0.71
TLI	0.54	0.59	0.67
2-Log likelihood	-7246.24	-7204.44	-7141.96
AIC	14580.47	14502.88	14383.93
BIC	14744.88	14678.50	14570.75
RMSEA	0.10	0.10	0.09
RMSEA lower 95% conf. interval	0.09	0.09	0.08
RMSEA lower 95% conf. interval	0.11	0.10	0.09
RMSEA *p* value	0.00	0.00	0.00

These estimate how well the model explains the observed data. Some are absolute and test the observed model against a statistically perfect model (e.g., RMSEA). Others are incremental and test the observed model against a baseline model (e.g., TLI). Table B5 presents the item factor loadings from CFA of the one-dimension (i.e., baseline), original three-dimension, and adapted four-dimension models.

Thresholds for "goodness-of-fit" are subjective and are often based more on intuition than on statistical justifications[4] (Marsh et al., 2004). For example, Browne and Cudeck (1993) suggested that an RMSEA value less than 0.05 indicates a “close fit” and that a value less than 0.08 suggests a “reasonable fit” between the model and the observed data. Bentler and Bonett (1980) suggested that a TLI of greater than 0.90 indicates an "acceptable" fit. Based on a relatively more rigorous simulation study, Hu and Bentler (1999) suggested that an RMSEA less than 0.06 and a TLI greater than 0.95 indicate a relatively good fit between the model and the observed data. Analysis of variance (ANOVA) can then be used to explore whether one model generated from factor analysis is better able than another to model and describe the observed data.

[4] The appropriateness of thresholds for fit are still extensively debated (e.g., Barett (2017) and responses).

Table B5: Item factor loadings for the one, three, and four-dimension models

Item	Short item name	1 dimension	3 dimension	4 dimension
1	Injustice and grievance	0.50	0.53	0.57
2	Defend against threats	0.41	0.43	0.45
3	Identity, meaning & belonging	0.22	0.22	0.60
4	Need for status	0.30	0.28	0.26
5	Excitement, comradeship & adventure	0.21	0.20	0.38
6	Dominate others	0.24	0.24	0.21
7	Susceptibility to indoctrination	0.18	0.19	0.37
8	Political Moral motivation	0.52	0.56	0.60
9	Opportunistic involvement	-0.11	-0.12	0.07
10	Family/friends support extremism	0.22	0.21	0.15
11	Transitional periods	-0.01	0.01	0.40
12	Group influence/control	0.25	0.25	0.22
13	Mental health issues	-0.06	-0.07	0.12
14	Over-identification with group	0.46	0.46	0.47
15	Us and them thinking	0.53	0.55	0.55
16	Dehumanisation of enemy	0.48	0.50	0.50
17	Attitudes justify offending	0.48	0.48	0.48
18	Harmful means to end	0.45	0.46	0.44
19	Harmful end objectives	0.31	0.32	0.33
20	Knowledge, skills & competencies	0.24	0.47	0.48
21	Networks, funds & equipment	0.28	0.56	0.56
22	Criminal history	0.08	0.11	0.10

A CFA generated poor values for absolute fit when we coerced the observed data into three dimensions (CFI = 0.63, TLI = 0.59, RMSEA = 0.101 [95% CI: 0.094, 0.108]). However, the three-dimension fit was better than a CFI in which we the observed data was coerced into one dimension (i.e., no separate dimensions) (CFI = 0.58, TLI = 0.54, RMSEA = 0.095 [95% CI: 0.088, 0.102]). Although still a poor fit in absolute terms, a chi-square difference test indicated that, relatively speaking, the three-dimension model was superior to the one-dimension model (X^2 difference (df = 3) = 83.6, p < .001).

Table B6: Factor loadings for a two-component Engagement dimension

Item	Short item name	Component 1	Component 2
1	Injustice and grievance	0.78	-0.01
2	Defend against threats	0.59	0.03
3	Identity, meaning & belonging	0.22	0.71
4	Need for status	0.36	0.39
5	Excitement, comradeship & adventure	0.21	0.65
6	Dominate others	0.44	0.27
7	Susceptibility to indoctrination	0.20	0.36
8	Political Moral motivation	0.86	-0.12
9	Opportunistic involvement	-0.33	0.39
10	Family/friends support extremism	0.27	0.07
11	Transitional periods	-0.04	0.57
12	Group influence/control	0.39	0.11
13	Mental health issues	-0.31	0.48

An EFA using principal axis factoring, a varimax rotation, and constrained to two dimensions, separated the items of the existing Engagement domain into those related to what we labelled “ideological” and “non-ideological” motivations for engagement (see Table B6). These two dimensions accounted for a cumulative 35.4% of the variance in Engagement classifications. The Ideological engagement dimension accounted for 19.9% of the variance in Engagement classifications and the Non-ideological engagement dimension accounted for 15.5% of the variance in Engagement classifications. Normality of statistical distribution metrics for the four-dimension structure are provided in Table B7.

Table B7: Descriptive data for classifications across the four dimensions in the adapted ERG 22+ structure

Dimension	Mean	SD	SE	Min	Max	Skew	Kurtosis
Ideological	4.5	2.42	0.14	0	10	-0.11	-0.77
Non-ideological	6.97	3.05	0.17	0	15	-0.07	-0.40
Intent	5.86	3.17	0.18	0	12	0.03	-0.90
Capability	2.59	1.47	0.08	0	6	0.29	-0.70

An additional CFA was used to examine the absolute quality of fit for this adapted four-dimension structure and a chi-square difference test to examine the relative fit compared to the original three-dimension model. The adapted four-dimension model generated improved but still poor overall statistics (CFI = 0.71, TLI = 0.67, RMSEA = 0.085 [95% CI: 0.078, 0.093]), albeit where the lower confidence interval fell below the threshold for acceptable fit (0.078). The four-dimension model also had a statistically significantly superior fit, relatively speaking, to the three-dimension model (X^2 difference (df = 3) = 125.0, p < .001). Nevertheless, we had reason to believe that the four dimensions in the adapted model were essentially unidimensional and so we could justly apply tests from measurement theory to explore construct validity and why model fit was statistically poor.

Classical test theory metrics

Construct validity was examined using metrics derived from two approaches to psychometric analysis: classical test theory (CTT) and item response theory (IRT). Classical test theory (CTT) is a body of psychometric theory that predicts outcomes of psychological assessment such as the difficulty of items based on the premise that a person's observed or obtained score on an assessment is the sum of a “true score” and some amount of error. Reliability, derived from classical test theory, is defined as the extent to which an assessment returns the same results consistently when used in the same context and on repeated occasions.

Reliability

Typically, general reliability – the ability of an assessment test to produce the same results across multiple administrations of an assessment – is examined by assessing individuals at different time intervals (known as “test-rest” reliability). Nevertheless, it is possible to estimate that reliability using data from one administration of an assessment using relevant statistical tests. Recent reviews suggest that a range of these statistical tests should be used together to provide sufficient evidence of reliability (Cho, 2022; McNeish, 2018; Revelle & Zinbarg, 2009; Sijtsma, 2009) and we calculated five reliability analyses based on their cumulative recommendations: Cronbach’s alpha (α: Cronbach, 1951; Kuder and Richardson, 1937), omega total (ω^t: McDonald, 1999), the algebraic greatest lower bound (GLB^a: ten Berge & Socan, 2004), lambda 2 (λ^2: Guttman, 1945), and ten Berge and Zergers’ mu-2 (μ^2: ten Berge & Zergers, 1978).

Item correlations

Item-to-item correlations examine whether items correlate strongly with one another. High item-to-item correlations (a Pearson's *r* value greater than 0.80: Hinkle et al., 2003) can be an indicator of *redundancy*: instances where separate items measure the same concept. Item-to-total correlations examine how well item scores correlate with total scores. Low item-to-total correlations (a Pearson's *r* value of less than 0.20: Nunnally & Bernstein, 1994) can be an indicator of *nonrelevance*: instances of items failing to contribute positively to the overall score.

Table B8: Item-to-item correlations for the Ideological engagement domain

	Item 1	Item 2	Item 6	Item 8	Item 12
Item 1	1	-	-	-	-
Item 2	0.523	1	-	-	-
Item 6	0.294	0.281	1	-	-
Item 8	0.731	0.549	0.386	1	-
Item 12	0.303	0.182	0.146	0.234	1

Table B9: Item-to-item correlations for the Non-ideological engagement domain

	Item 3	Item 4	Item 5	Item 7	Item 9	Item 10	Item 11	Item 13
Item 3	1	-	-	-	-	-	-	-
Item 4	0.327	1	-	-	-	-	-	-
Item 5	0.462	0.397	1	-	-	-	-	-
Item 7	0.467	0.034	0.199	1	-	-	-	-
Item 9	0.162	0.096	0.301	-0.039	1	-	-	-
Item 10	0.168	0.183	0.140	0.138	-0.200	1	-	-
Item 11	0.529	0.036	0.297	0.366	0.081	0.049	1	-
Item 13	0.201	0.074	0.234	0.060	0.325	-0.247	0.407	1

Table B10: Item-to-item correlations for the Intent domain

	Item 14	Item 15	Item 16	Item 17	Item 18	Item 19
Item 14	1	-	-	-	-	-
Item 15	0.417	1	-	-	-	-

	Item 14	Item 15	Item 16	Item 17	Item 18	Item 19
Item 16	0.413	0.709	1	-	-	-
Item 17	0.483	0.470	0.487	1	-	-
Item 18	0.421	0.474	0.402	0.497	1	-
Item 19	0.344	0.416	0.554	0.494	0.257	1

Table B11: Item-to-item correlations for the Capability domain

	Item20	Item21	Item22
Item20	1	-	-
Item21	0.632	1	-
Item22	0.118	0.1751	1

Item response theory metrics

Item response theory (IRT) assesses the design, analysis, and scoring of assessments that measure latent variables (abilities, attitudes, etc) based on the relationship between individuals' performances on each assessment item and their overall levels of performance. Unlike CTT, it does not assume that each item is equally “difficult”.

Difficulty

First were “difficulty” values indicating how much of the latent variable is needed to increase the classification upwards by one category. These constituted estimated threshold values for the standardised “amount” of latent trait (i.e., clinical need) required to move up one category (e.g., from “none” to “partly” and “partly” to “strong”). Items where both thresholds are below zero might be considered very low difficulty, whereas those where both thresholds are above zero might be considered very high difficulty.

Discriminatory ability

Second were “discriminatory” values that indicates how well each item distinguishes between overall high and low scorers. According to Baker (2001) the acceptable thresholds for discriminatory parameter values can be interpreted as: 0.0 (none); 0.01–0.34 (very low); 0.35–0.64 (low); 0.65–1.34 (moderate); 1.35–1.69 (high); >1.70 (very high); and infinity (perfect).