CT MORSE COUNTER-TERRORISM
MONITORING, REPORTING AND SUPPORT MECHANISM

A PROJECT FUNDED BY **THE EUROPEAN UNION** | INSTRUMENT CONTRIBUTING TO STABILITY AND PEACE

Improving the Evaluation of Interventions to Counter and Prevent Terrorism and Violent Extremism

Author: Andrew Glazzard
Research Support: Michael Jones

Summary of Findings

- CT and P/CVE programming comes with a set of constraints that makes evaluation particularly challenging. This study identified 22 constraints which affect evaluation in this field. These can be categorised as conceptual, theoretical, technical, contextual, ethical and programmatic.

- The characteristics of high-quality evaluations in this field are little different from those in most others. However, there are features of evaluation design, data sources and collection, analysis/assessment, expertise, presentation, and the evaluation's effect which, according to evaluation experts, are particularly important to CT and P/CVE evaluations.

- There is a debate over whether evaluations in this field are weaker than those in any other. Some research shows that evaluations in this field are in aggregate less advanced than in other fields such as crime reduction, a consequence of this field's relatively recent emergence. However, evaluations in this field are subject to some specific challenges, such as evaluations remaining unpublished, a lack of support from donors for employing rigorous evaluation methods, and programmatic weaknesses (e.g. poor programme design making the task of evaluators more challenging).

- There are no overriding arguments in principle against using experimental and quasi-experimental methods in this field, although there are some practical challenges. These methods are not, though, always applicable. Evaluators need a wide range of methods and tools, including but not restricted to experimental/quasi-experimental methods. Qualitative/theory-based methods are effective and appropriate for many programmes, but when they are used it needs to be with greater rigour than is often the case at present.

- CT and P/CVE evaluation by its nature is multi-disciplinary. Much has been learned from fields such as peacebuilding, but evaluators should learn from – and adopt techniques pioneered by – a wide range of disciplines and fields, from criminology to internet studies.

Recommendations

The following recommendations have been developed to support donors and implementer in improving the evaluation and hence the impact of CT and P/CVE interventions.

1. Donors should commission CT and P/CVE interventions with a substantial research component that explicitly seeks to test mechanisms through experimental and quasi-experimental methods (whether embedded in a larger programme or standalone).

2. Where experimental and quasi-experimental methods are inappropriate or insufficient, donors and implementers should ensure that qualitative theory-based methods are applied rigorously and with advice/support from monitoring and evaluation specialists.

3. Donors and researchers should consider developing studies that, as a principal outcome, seek to apply to CT and P/CVE methods tried and tested in disciplines such as criminology.

4. Donors in the CT and P/CVE field should commit to publishing evaluations and where possible establish criteria and conditions for publishing evaluations with implementers and beneficiaries at the outset of a programme.

5. Donors and implementers should consider mixed evaluation teams that comprise different disciplines and specialisms. Where interventions are fully or partly designed to test the effectiveness of mechanisms, programme teams should include social scientists with expertise in designing and implementing research designs.

6. Donors and implementers in the CT and P/CVE field should commit to developing a monitoring and evaluation (M&E) design at the outset of a programme, and should ensure they have access to M&E expertise throughout a programme's lifecycle.

Introduction

This study was commissioned by CT MORSE, the monitoring and support mechanism for the European Commission's external actions for countering terrorism (CT) and preventing/countering violent extremism (P/CVE). The purpose of the study was to support the efforts of the EU, other donors, implementers and beneficiaries to improve the monitoring and evaluation of CT and P/CVE projects and programmes by examining best practice in this field, identifying shortcomings and opportunities for improvement, including scientific evaluation methods and approaches used in other related or analogous fields. At an early stage the research team chose to focus on evaluation as it was clear this was where shortcomings were more evident and more consequential: whereas monitoring is a more technical and programmatic challenge, evaluation goes to the heart of how (and whether) interventions work. Moreover, several toolkits and guidance documents have recently been published which are particularly focused on supporting programme design and implementation, including developing measures and indicators for monitoring frameworks.[1]

Methodology

The study sought to answer the following research question: How can evaluation of CT and P/CVE projects and programmes be improved? This was separated into five sub-questions: a) What challenges are specific or more acute in CT and P/CVE compared to other fields?; b) What are the characteristics of a high-quality evaluation in CT and P/CVE?; c) To what extent is evaluation in CT and P/CVE weaker than evaluation in other similar fields; d) To what extent can more advanced methods such as experimental and quasi-experimental methods be employed in evaluating CT and P/CVE projects and programmes; e) What can evaluators of CT and P/CVE projects and programmes learn from other fields?

The research team used qualitative methods given that the questions suggested an exploratory approach. Evidence for the study was developed from existing literature and from interviews with academic experts and evaluation practitioners with experience of evaluating a wide variety of CT and P/CVE projects and programmes commissioned by a range of donors, both national and multilateral. The literature review was based on a targeted search using standard academic search engines and databases which yielded a corpus of 25 papers (listed in the bibliography) comprising academic studies as well as 'grey' literature from thinktanks, NGOs, etc. Separately, semi-structured interviews were conducted with 15 experts and practitioners selected purposively and using a simple interview brief.

[1] See, in particular, the European Commission's Sector Indicator Guidance: Countering Violent Extremism (available at: https://kbb9z40cmb2apwafcho9v3j-wpengine.netdna-ssl.com/wp-content/uploads/2019/03/cve-ssi.pdf), and also Holdaway and Simpson (2018).

Findings

a) What challenges are specific or more acute in CT and P/CVE compared to other fields?

The study found broad agreement in the literature and among interviewees that CT and P/CVE programming comes with a set of constraints that makes evaluation challenging. Where some interviewees differ is whether these constraints add up to a uniquely challenging evaluation task. Some interviewees suggested that several constraints are either unique or particularly acute in the field, while others countered that many or most can be found in other fields concerned with preventing negative or risky outcomes in complex social contexts.

Analysis identified 22 constraints affecting CT and P/CVE evaluation, which were grouped into six categories: conceptual, theoretical, technical, contextual, ethical and programmatic.

Conceptual constraints include the lack of agreed definitions of key terms such as terrorism, violent extremism, radicalisation and resilience. Programmatic responses in the field are also beset by definitional problems – for example, some practitioners argue for a distinction between PVE and CVE, or between disengagement and deradicalisation, but there is no consensus on these debates. Furthermore, the scope and diversity of activities under these headings are wide. More broadly, two interviewees highlighted a gap of understanding between those practitioners with a development background and those from a security background. These two groups often seem to be have different perspectives, and can conceptualise the problem of terrorism/violent extremism very differently, with development practitioners focusing often on structural factors and communities, and security practitioners focusing on individuals and risk factors.

Theoretical constraints include the difficulty of measuring or estimating a non-event (i.e. the terrorism or radicalisation that CT and P/CVE interventions ultimately seek to prevent); the variety of pathways taken by those on a journey towards terrorism and the range of outcomes (from passive support across a wide range of ideologies/ movements, to participation in violence) so that evaluators are faced with a fuzzy or moving target; the variety of contexts in which CT and P/CVE programmes are implemented, meaning that comparisons between interventions may not be valid; the difficulty of translating abstract concepts (such as resilience or vulnerability) into measurable behavioural attributes; and the low base-rate problem, meaning that terrorist events in most contexts are rare and infrequent so CT and P/CVE programmes are ultimately addressing what one respondent called ‘needles in haystacks’. Several interviewees highlighted low base rates as a particularly important constraint for CT and P/CVE evaluation compared to other social or criminal problems which are usually present in greater volumes and where changes in data are easier to observe (see also Cherney and Belton 2019). The net result of these issues, one interviewee suggested, is that generating counter-factuals (i.e. what would have happened in the absence of the intervention) is unusually difficult in this field. Some evaluation designs propose proxy indicators as a solution, though one interviewee said that these tend to be more applicable at the community than the individual level.

Technical constraints are closely related to theoretical ones, but concern the difficulties of developing tools for measurement and analysis for evaluation in this field. These were: the lack of standardised methodology for data collection and analysis (Baruch et al 2018), the lack of existing validated measures for unobservable outcomes (Holmer et al. 2018), and the difficulty of evaluating interventions that are predicated on individualised interventions and treatment plans (Cherney and Belton 2019).

Contextual constraints relate to the contexts in which CT and P/CVE programmes are implemented. These were:

- the political sensitivity of the topic, which can lead to (a) lack of access to data, beneficiaries or stakeholders; (b) pressures on evaluators to comply with political imperatives; (c) donors and/or implementers relabeling or concealing a programme's CT or P/CVE purpose, potentially obscuring elements within the programme's causal chain; and (d) biased responses from respondents due, for example, to social desirability bias[2];
- security risks, especially when conducting research in fragile or conflict affected places, which can lead to access difficulties but also direct security risks (to staff and respondents);
- context complexity, where there are many factors (e.g. unrelated changes in the security environment, or the effects of other unrelated interventions) which could influence outcomes or impacts, leading to difficulties of attributing changes to the intervention rather than an external factor.

Ethical constraints concern the need to do no harm, directly or indirectly, intentionally or unintentionally. These were: exposing respondents to security threats; and the consequences of withholding treatment from beneficiaries (e.g. as part of an experimental evaluation). This latter constraint is somewhat controversial and is discussed further below.

Programmatic constraints derive from the ways in which CT and P/CVE interventions are designed and delivered. While there is obviously a great variety of projects and programmes in this field, interviewees identified problems which in their experience tended to recur: over-optimistic or unfocused programme designs which contain unachievable outcomes or contribution to impact (see also Harris-Hogan 2020); a lack of targeted or disaggregated data collection, so that evaluations struggle to identify how interventions have affected different groups – with obvious implications for gender issues (Gielen 2018); composite programmes containing a number of different interventions, making it difficult to identify the 'active ingredient', i.e. which intervention had what effect; a reluctance on the part of many implementers, donors and beneficiary governments to share data and learning publicly, creating problems for comparative analysis (Koehler and Fiebig 2019); and a lack of

[2] Social desirability bias is a well-known bias in population research. In simple terms it means respondents saying what they think the researcher either wishes to hear or what they believe will make themselves look better.

resources and/or capacity (knowledge and skills) for monitoring and evaluation within the programme.

b) What are the characteristics of a high-quality evaluation in CT and P/CVE?

Most interviewees said (or suggested) that in principle there should be no difference between a high-quality evaluation in CT or P/CVE and one in any other field. However, all interviewees highlighted at least one attribute which they believed to be particularly important in demonstrating quality in this field. These responses were categorised into the following themes: evaluation design, data sources and collection, analysis/assessment, expertise, presentation, and effect. (In the summary below, the numbers in brackets represent how many interviewees raised each attribute.)

On **evaluation design**, interviewees specified:

- the need for a valid and reliable research design that at least shows the difference between baseline and end-line (1);
- clarity about limitations of an evaluation design (1), and the level of certainty about attribution of outcomes to the intervention (1);
- clearly stated evaluation questions (1);
- clearly articulated theories of change, whether these were derived from the programme or developed analytically by the evaluator (2), and that specify why and how an intervention will have an effect (1);
- compliance with accepted evaluation standards (e.g. with respect to ethics) (1) and the use of OECD DAC evaluation criteria (1);
- and a design that includes assessing risks and unintended outcomes (including the risk of doing harm) (1).

On **data sources and collection**, interviewees highlighted:

- the importance of triangulation of evidence (i.e. using multiple sources to establish a fact) (2) – an observation supported by Davies et al. (2017)
  who found that gang membership evaluations, which were in aggregate of higher quality than CT and P/CVE evaluations, relied on triangulation of evidence and the use of mixed methods;
- where possible, evaluations should use quantitative (1) or mixed (i.e. quantitative and qualitative) methods (1);
- evaluators should have reached relevant populations (e.g. beneficiaries) in sufficient number (2).

On **analysis and assessment**, interviewees raised:

- the need to assess how the context may have affected the intervention (2);
- the need to discriminate between results at different levels (e.g. differentiating between outputs and outcomes) (1). The literature also discusses the importance of disaggregating gender analysis (Gielen 2018).

On **expertise**, interviewees mentioned:

- the value of an evaluation using a mixed team in terms of security, development and evaluation specialism (1);
- the value of local researchers (1).

In terms of **presentation**, interviewees said:

- an evaluation should clearly separate content relating to process from that relating to impact, and not assume that effectiveness demonstrated by a process evaluation necessarily means the achievement of outcomes (2);
- evaluations should be clear about programmatic assumptions (1);
- they should report results comprehensively, not selectively (1);
- they should be specific about dependent variables (i.e. the intervention's intended outcomes) (1).

One interviewee said that the quality of an evaluation was ultimately a product of its **effect**, principally on the donor: a good evaluation will itself lead to outcomes (1).

c) To what extent is evaluation in CT and P/CVE weaker than evaluation in other similar fields?

Interviewees were divided on this question. Some asserted that CT and P/CVE evaluations were generally inferior to those in other fields, lacking the rigour seen in more established fields such as crime reduction. One interviewee (an experienced evaluator) added that every evaluation they had read and performed had been deficient in one respect or another. Another said that evaluations are improving, albeit from a low base, due to an influx of evaluators with experience of evaluation in international development. One interviewee, however, suggested that CT and P/CVE evaluations tend to be more rigorous than those in peacebuilding, for example, and have benefitted from substantial resource investment in recent years. Another commented that an assumption that evaluation needs to be more rigorous and robust is actually common to several fields and is not confined to CT and P/CVE: few fields are able to match the standards found in healthcare, where experimental methods are more applicable and which has benefited from decades of methodological development. Four interviewees also highlighted the lack of publicly available evaluations, the result presumably of sensitivities relating to security and confidentiality on the part of donors, implementers and beneficiaries – although some interviewees also suggested that donors and implementers were also often unwilling to expose the limited effectiveness of their project to public view. Refusing to publish evaluations weakens the field in two respects: it restricts the number of evaluations in the evidence base, and it deprives future evaluators, donors and implementers of the knowledge derived from the evaluation.

In the literature review there was some evidence that CT and P/CVE evaluation lacks some of the attributes usually associated with high-quality evaluation, namely the rigorous use of empirical data (Feddes and Galluci, Pistone 2015), and methodological clarity: in a sample of 48 CT and P/CVE evaluations, Bellasio et al (2018) found only 33 that set out the evaluation's method. One study concluded that evaluation in P/CVE is demonstrably weaker than it is in the analogous field of gang desistance. Davies et al. (2017) compared evaluations of P/CVE (126 studies) with those of gang desistance (67 studies). In evaluation quality, 64% of P/CVE were rated low, 37% medium, and 0% high. Gang desistance evaluations scored 15% low quality, 55% medium quality and 30% high quality; 63% of these evaluations were impact evaluations, while 21% evaluated process and impact. By contrast, only 49% of P/CVE evaluations addressed impact.

Interviewees identified specific weaknesses in many CT and P/CVE evaluations, and some suggested explanations for why these weaknesses are evident. In particular, several interviewees suggested that evaluations, in this as in other fields, are only as good as the programme they evaluate: a poorly designed programme with an inadequate (or absent) theory of change (ToC), in which evaluation is a late-order activity, usually entails a process evaluation which is limited to assessing results at output level, or is based on an improvised ToC built retrospectively by the evaluators themselves. One interviewee added that common weaknesses in programme and evaluation design include ambiguous terminology, a lack of hierarchy in the results chain, indicators unconnected to the results chain, and vague statements of effect.

In terms of explanations for these weaknesses, several interviewees highlighted the field's relative immaturity. Two interviewees suggested that some weaknesses are donor-driven, one adding that donors in this field prefer simpler theories of change, bring their own assumptions to programming, and are unwilling to invest in rigorous evaluation methods, while the other added that donors are unlikely to be attracted to modestly framed objectives so tend to invest in overly ambitious programmes (see also Harris-Hogan 2020). A third interviewee commented that some donors are more interested in/supportive of evaluation than others. Two suggested that the conceptual and definitional constraints discussed above help to explain the field's relative deficiencies.

d) To what extent can more advanced methods such as experimental and quasi-experimental methods be employed in evaluating CT and P/CVE projects and programmes?

Experimental methods in evaluation are those conducted on a scientific basis usually involving a randomised control trial (RCT), in which a treatment group receives the intervention and a control group does not, with participants assigned randomly to each. Although they were developed in the natural sciences and are particularly associated with healthcare, RCTs are sometimes described as the 'gold standard' in evaluating social scientific interventions such as in P/CVE. However, interviews revealed a range of views on the applicability of experimental methods to CT and P/CVE evaluations. Some interviewees were sceptical and pointed out what they saw as significant drawbacks in applying RCTs to CT and P/CVE, while others were clear that RCTs could and should be used where possible. Some said that RCTs could be employed under some circumstances, and several of these added that often it would be more feasible to use quasi-experimental methods, where evaluators have access to a treatment group and a *comparison* group. (A comparison group differs from a control group in that its members have not been assigned randomly, and/or has received a different treatment rather than no treatment at all.)

Interviewees' arguments against using RCTs in this field were ethical, practical and theoretical. The ethical argument applies particularly to P/CVE: withholding a treatment from individuals identified as being at risk of radicalisation potentially exposes them and (if they are or become a security threat) the wider population to harm. The practical arguments are, firstly, that RCTs are often expensive to operate and hence require significant resource commitments from donors; secondly, that is difficult to run experiments and/or collect data to scientific standards in fragile and conflict affected environments ("the field is not a laboratory", said one interviewee); and, thirdly, that it is very challenging to run effective RCTs at the tail-end of a programme: RCTs generally need to begin at an intervention's planning stage. The theoretical arguments are: firstly, the low base-rate problem (see above) means that it difficult to design valid and reliable RCTs to measure outcomes that in the real world are actually very rare; secondly, RCTs are most effective at illuminating simple causal pathways (e.g. whether a vaccine prevents subjects contracting a disease) whereas terrorism and radicalisation are usually the product of complex, variable and unpredictable social and behavioural factors (see also Baruch et al. 2018), and P/CVE programmes in particular tend to be collections of several interventions; thirdly, it is difficult to isolate treatment/non-treatment in real-world environments so control groups may be exposed to spillover effects from the treatment or from another programme; and, fourthly, the breadth and variety of contexts for CT and P/CVE programming mean that an RCT's finding(s) may be specific to the context in which it was conducted and not comparable with or generalizable to others. Summing up these arguments, one interviewee said that RCTs could potentially show effectiveness, but are unlikely to demonstrate impact.

Those interviewees advocating RCTs in CT and P/CVE acknowledged that no method is right for every question but suggested that some of these arguments are overstated. For example, one said that the ethical issue does not arise as RCTs are usually not considered *at all* by donors and implementers, and if they are rejected it is more likely to be on practicalities or cost rather than ethical grounds. Another interviewee suggested that the ethical argument does not apply in the P/CVE field where the risk of radicalisation is unpredictable and interventions are generally untested or unvalidated: a single P/CVE intervention is unlikely to be the difference between someone conducting or not conducting a terrorist attack, and in any case P/CVE programmes can never treat everyone so there will always be at-risk individuals who do not receive the treatment. Advocates of RCTs also suggested that the theoretical arguments can be answered by effective and appropriate experimental design. One interviewee, for example, said that RCTs could and should be used to evaluate smaller-scale, shorter-term interventions. Two others added that RCTs are more feasible where the project or programme's objective is specifically to evaluate rather than implement interventions, while risks can be mitigated by appropriate consultation with stakeholders (such as beneficiary governments) and scaling of the intervention. One interviewee said that programme implementers often lack the research skills necessary to run or commission RCTs but that mixed teams including implementers and scientists are becoming more common.

Where RCTs are impossible or impractical, several interviewees advocated the use of quasi-experimental methods. Two argued for a greater use of longitudinal evaluation designs using comparison groups, one of whom added that techniques such as regression discontinuity design and difference in differences remain underused and under-explored.[3] One recommended the use of proxy indicators for attributes that would be hard to measure because they are unobservable or because of security concerns: a proxy for violent intentions, for example, might be expressed attitudes towards use of violence (e.g. Webber et al. 2017 used attitudinal surveys to study the effects of a disengagement programme for former LTTE members in Sri Lanka). Two interviewees recommended the use of mixed methods approaches, and highlighted the importance of triangulating findings using different techniques: according to one of these interviewees, experimental and qualitative/theory-based approaches are not mutually exclusive, and in evaluation data may generally lack meaning if it is not placed in a valid theoretical framework.

Although the focus of this question was experimental and quasi-experimental designs, several interviewees urged more effective use of qualitative methods. One interviewee said that outcome harvesting and process tracing were valuable methods in CT and P/CVE.[4] One recommended the use of multi-country studies examining the effect of similar interventions in dissimilar contexts, while another said that more evaluations should include multi-level (e.g. individual and community, or community and its wider geographical area) and/or multi-site studies. Two recommended case studies and more and better use of longitudinal designs, such as Webber et al.'s evaluation (2017) cited above; one of these also advocated returning to locations of programmes some years after they have been closed to investigate long-term outcomes and impact. The literature review also found support for more rigorous approaches to qualitative or mixed-methods evaluation. In particular, theory-based approaches to evaluation are widely recommended where experimental designs are difficult or where the treatment remains undeveloped and untested (Feddes and Gallcci 2015; Baruch et al. 2018). One such approach, realist evaluation, is being increasingly applied in CT and P/CVE. This seeks to understand not simply effectiveness in general but how interventions work, where and for whom. To this end, it assesses the relationship between the intervention's context, its mechanism and its

[3] Regression discontinuity design is a quasi-experimental method which uses a threshold above or below which a treatment is assigned: comparisons are made before and after the intervention between both groups. Difference in differences is a statistical technique where data from observation of a 'natural experiment' is used to imitate an RCT by comparing over time a group which was naturally (i.e. unintentionally) exposed to a treatment with one that was not exposed.

[4] Outcome harvesting is a participatory approach which, rather than evaluating the achievement of a predetermined outcome, collects information of what changes have occurred and works backwards to investigate whether they were caused by the intervention. Process tracing deduces what would be the observable results of a theory of change and then investigates whether those results can be or have been observed.

outcome(s) (Gøtzsche-Astrup 2018; Womg et al. 2013; Cherney and Belton, 2019; Gielen; 2018; Gielen 2019;).

e) What can evaluators of CT and P/CVE projects and programmes learn from other fields?

Several interviewees identified fields or disciplines which have established a degree of robustness in evaluation and which could be models for CT and P/CVE evaluations to imitate. In particular, two interviewees highlighted criminology, a discipline which has also been discussed in the research literature as one with valuable lessons for CT and P/CVE evaluators. Criminology is obviously applicable as terrorism usually entails criminality of various kinds, and even though much criminology focuses on more common crimes than terrorist ones, it has a wealth of validated instruments and methods for evaluating interventions in complex social environments (Feddes & Gallucci 2015). Public health approaches have a long history of being applied to criminal behaviour and are also gaining ground in relation to terrorism and violent extremism: one interviewee noted that public health approaches should also inform evaluation – a point echoed by Davies et al. (2017) in their study of the applicability of gang evaluation to CT and P/CVE. This study adds that the most robust gang evaluations have combined existing datasets from law enforcement with primary data collection (interviews, surveys etc.) Two interviewees noted that psychology, which has been a leading discipline in terrorism studies in the last two decades, has also provided substantial input to CT and P/CVE evaluation, e.g. in providing validated instruments for measuring attitudes and attitudinal change. Other disciplines mentioned by interviewees included cultural anthropology (to provide a greater awareness of context and cultural differences among beneficiaries) and systems theory (where methods for modelling complex ecosystems can be applied to P/CVE interventions particularly). Three interviewees specified collective impact studies – where a number of interventions in a single location are evaluated, and which can be informed by systems theory – as a valuable and underused approach.

Although not mentioned by interviewees, one study recommends that techniques used in evaluating marketing and advertising campaigns, especially in more insecure environments, could be applied in CT and P/CVE evaluation to improve data collection and conduct ‘in-field experiments’ (Holmer, Bauman and Aryaeinejad 2018). Williams and Kleinman (2014) and Horgan and Braddock (2010), meanwhile, advocate utilization-focused evaluation, which focuses on how an intervention is used by different stakeholders, and propose an approach called Multi Attribute Utility Technology (MAUT) drawn from decision theory. However, Gielen (2019) cautioned this latter approach may ‘lean too much towards the needs of policymakers’ and produces a somewhat limited ‘technical analytical discourse which revolves around the positive effects of the programme or intervention’.

The relationship between P/CVE and peacebuilding is much debated by practitioners, and interviewees also had a range of views on whether peacebuilding has more robust methodology. Two interviewees said that the two fields have strong similarities and P/CVE in effect already learns from peacebuilding, given that many implementers are involved in both. These interviewees added that peacebuilding evaluations have made valuable progress in developing frameworks for measuring positive attributes (i.e. protective factors or resilience to conflict) and some P/CVE evaluators have already shown an interest in applying this. Another interviewee endorsed this point, highlighting the theoretical appeal of measuring a positive achievement rather than attempting to construct a counter-factual case.

Two interviewees said that CT and P/CVE evaluations could draw techniques from internet research and use data-driven approaches to investigate impact on a larger scale (e.g. the aggregate impact of several project/programmes in one location) and to provide a richer picture of contexts to specific interventions. Techniques such as sentiment analysis of social media communications, for example, can potentially reveal how community attitudes change over time. Another suggested that, counter-intuitively, CT and P/CVE evaluators have not learned as much from terrorism studies (e.g. in mapping terrorist social networks) as might be expected.

More generally, some interviewees discussed the implications for this question of the multi-disciplinary nature of P/CVE interventions, and the fact that programmes often comprise several interventions of different types. Given that P/CVE programmes include interventions with a more established record of evaluation – such as employment generation and microfinance – it should be possible to use validated evaluation techniques on those specific interventions.

Discussion and Recommendations

The interviewed experts had different perspectives on whether CT and P/CVE is a uniquely challenging field for evaluation, but this study's identification of 22 different constraints suggests that the challenges are, in combination, substantial. Of course, every field faces challenges and it is part of the skill of the programme designer and evaluator to design a method that will overcome or mitigate them. This study's interviews and literature review suggest that no challenge is insuperable in and of itself. But the CT and P/CVE field is still at a relatively early stage in development of valid and reliable evaluation methods. This study, then, supports the observation of Davies et al. (2017) that the P/CVE field is still at a stage of theory generation, whereas other more mature fields (such as gang desistance) have moved on to theory testing. Given that P/CVE in particular is of fairly recent provenance, this conclusion is perhaps unsurprising.

If evaluation in this field is to make progress, however, it is important that its methods become more robust. On the important question of the utility and appropriateness of experimental methods, this study concludes that, although ethical and security risks are often cited as an obstacle to experimental methods, there is no reason in principle to rule out RCTs and, given the status that they have in evaluation methodology more generally, there are obvious advantages to the field in adopting them more widely: this would help move CT and P/CVE evaluation from theory generation to theory testing. This study therefore supports Bellasio et al.'s conclusion (2018) that they should be used where possible and practicable. However, there are many instances where they are not the right method, and, like any method, they have limitations: experimental and quasi-experimental methods are ways of testing mechanisms, but the mechanism is only one aspect of an intervention.

As one interviewee said, experimental/quasi-experimental and qualitative methods are not mutually exclusive: accepting that CT and P/CVE evaluation needs more of both, and using them in combination where appropriate, seems to us to be the obvious way forward. This study agrees with those who advocate a more methodological approach to qualitative evaluation in this field, and in particular the use of theory-based approaches. Among the range of theory-based approaches, the authors were particularly persuaded by the merits of realist evaluation, which investigates the relationship between context, mechanism and outcome. The authors would also endorse statistical analysis of existing, relevant data sets to illuminate contexts and to triangulate findings. Taken together, increased use of experimental and quasi-experimental methods, more rigorous use of theory-based methods, and more statistical analysis will enable this field to make rapid progress.

Recommendation (1): donors should commission CT and P/CVE interventions with a substantial research component that explicitly seek to test mechanisms through experimental and quasi-experimental methods (whether embedded in a larger programme or standalone).

Recommendation (2): where experimental and quasi-experimental methods are inappropriate or insufficient, donors and implementers should ensure that qualitative theory-based methods are applied rigorously and with advice/support from monitoring and evaluation specialists.

The CT and P/CVE field has much to learn from related fields and disciplines (notably criminology, public health approaches to social problems, and peacebuilding). But while studies such as this have advocated applying learning from these to CT and P/CVE fields, it does not seem to be happening in practice to any extent, with research often making broad references to methods without detailing how they can be applied. There is also room for a wider analysis of evaluation techniques in potentially analogous fields to identify further applicable methods.

Recommendation (3): donors and researchers should consider developing studies that, as a principal outcome, seek to apply to CT and P/CVE methods tried and tested in disciplines such as criminology.

In addition, there are a number of practical steps that would improve evaluation in this field. At the most basic level, more evaluations and their supporting data need to be published. An unpublished evaluation benefits only a small circle of donors and implementers who have access to it. The scarcity of published evaluations relative to other fields is a major reason why CT and P/CVE evaluation is often perceived as weak. The onus should be on donors to find ways to overcome the real but not insurmountable sensitivities in his field.

Recommendation (4): donors in the CT and P/CVE field should commit to publishing evaluations and where possible establish criteria and conditions for publishing evaluations with implementers and beneficiaries at the outset of a programme.

Several respondents pointed to the value of combining different types of expertise in evaluations, such as security and development specialists, and/or social scientists and experts in programmatic evaluation. As CT and P/CVE are by their nature multi-disciplinary, it makes sense to reflect this in the composition of evaluation teams. But the more fundamental point here is that evaluation potentially serves different purposes for donors and implementers (accountability and learning) and researchers (evidence).

Recommendation (5): donors and implementers should consider mixed evaluation teams that comprise different disciplines and specialisms. Where interventions are fully or partly designed to test the effectiveness of mechanisms, programme teams should include social scientists with expertise in designing and implementing research designs.

Finally, several interviewees told us that weak monitoring and evaluation are often a product of poor programme design. Evaluators will struggle to do more than assess outputs if they are brought in at a programme's conclusion; evaluation is fundamental to programme design and should be front and centre of donors' and implementers' considerations at the outset. Where donors and implementers lack expertise, this should be embedded in programmes with research strands and/or specialist evaluators, drawn from research institutes or universities where necessary.

Recommendation (6): donors and implementers in the CT and P/CVE field should commit to developing an evaluation design at the outset of a programme, and should ensure they have access to the evaluation expertise throughout a programme's lifecycle.

Disclaimer - This publication was produced with the financial support of the European Union. Its contents are the sole responsibility of the author and do not necessarily reflect the views of the European Union.

Annex: Evaluation Reports Mentioned by Interviewees

Experts interviewed for this study were asked if they could identify evaluation studies which they considered to be high quality. Those which are in the public domain are listed below. Several were mentioned which were not in the public domain, which have not been included.

Aly, A., Taylor, E. and Karnovsky , S. (2014) 'Moral Disengagement and Building Resilience to Violent Extremism: An Education Intervention'. Studies in Conflict & Terrorism, 37(4), 369-385.

Berman, E., Felter, J. H., Shapiro, J. N., & Troland, E. (2013). Modest, secure, and informed: Successful development in conflict zones. *American Economic Review*, *103*(3), 512-17. Available at: https://www.nber.org/papers/w18674.pdf.

Brett, J., & Kahlmeyer, A. (2017). Strengthening Resilience to Violent Extremism–Strive Horn of Africa: Evaluation Report. Available at: https://ct-morse.eu/wp-content/uploads/2017/04/170124-STRIVE-evaluation-Report-Final.pdf

Fisher, T., Range, D. & Cuddihy, J. (2020) 'Evaluation of 'Strengthening Resilience to Violent Extremism (STRIVE II) in Kenya': Final Report'. Available at: https://kbb9z40cmb2apwafcho9v3j-wpengine.netdna-ssl.com/wp-content/uploads/2018/01/evaluation-of-strive-ii-final-report-september-2020-version-for-publication_lm-2.pdf

Khalil, J., & Zeuthen, M. (2014). A case study of counter violent extremism (CVE) programming: Lessons from OTI's Kenya transition initiative. *Stability: International Journal of Security and Development*, *3*(1). Available at: https://www.stabilityjournal.org/articles/10.5334/sta.ee/

Mercy Corps (2016). Critical Choices: Assessing the Effects of Education and Civic Engagement on Somali Youth's Propensity towards Violence. *Mercy Corps: Portland, OR, USA*. Available at: https://www.mercycorps.org/research-resources/effect-education-civic-engagement-somali-youth

Bibliography

a) Studies Included in the Targeted Literature Review

Baruch, Ben, Tom Ling, Rich Warnes and Joanna Hofman (2018) 'Evaluation in an Emerging Field: Developing a Measurement Framework for the Field of Counter Violent Extremism', IMPACT Europe.

Bellasio, Jacopo, Joanna Hofman, Antonia Ward, Fook Nederveen, Anna Knack, Arya Sofia Meranto, Stijn Hoorens (2018) 'Counterterrorism Evaluation: Taking Stock and Looking Ahead', RAND Corporation.

Chauhan, Leah (2017) 'Deradicalisation Scientific Insights for Policy', Flemish Peace Institute.

Cherney, Adrian and Emma Belton (2019) 'Assessing Intervention Outcomes Targeting Radicalised Offenders: Testing the Pro Integration Model of Extremist Disengagement as an Evaluation Tool', Dynamics of Asymmetric Conflict.

Cherney, Adrian, Emma Belton, Ellen Leslie, Dr Jennifer Bell, Lorraine Cherney & Lorraine Mazerolle (2018) 'Countering Violent Extremism: Data Collection and Analysis Manual', Countering Violent Extremism Centre, University of Queensland.

Davies, Matthew, Richard Warnes and Joanna Hofman (2017) 'Exploring the Transferability and Applicability of Gang Evaluation Methodologies to Counter-Violent Radicalisation' RAND Corporation Europe.

Dawson, Laura, Charlie Edwards, and Calum Jeffray (2014). Learning and Adapting: The Use of Monitoring and Evaluation in Countering Violent Extremism: a Handbook for Practitioners, RUSI.

Feddes, Allard and Marcello Gallucci (2015) 'A Literature Review on Methodology Used in Evaluating Effects of Preventive and De-Radicalisation Intervention', Journal for Deradcialization, Number 5.

Gielen, Amy-Jane (2018) 'Exit Programmes for Female Jihadists: A Proposal for Conducting Realistic Evaluation of the Dutch Approach', International Sociology, Volume 33, Issue 4.

Gielen, Amy-Jane (2019) 'Countering Violent Countering Violent Extremism: A Realist Review for Assessing What Works, for Whom, in What Circumstances, and How?', Terrorism and Political Violence, Volume 31.

Gotzsche-Astrup, Oluf (2018) 'The Time for Causal Designs: Review and Evaluation of Empirical Support for Mechanisms of Political Radicalisation', Volume 39.

Harris-Hogan, Shandon (2020) 'How to Evaluate a Program Working with Terrorists? Understanding Australia's Countering Violent Extremism Early Intervention Program', Journal of Policing, Intelligence and Counter Terrorism, Volume 15, Issue 2.

Helmus, Todd C., Miriam Matthews, Rajeev Ramchand, Sina Beaghley, David Stebbins, Amanda Kadlec, Michael A. Brown, Aaron Kofner, Joie D. Acosta (2017) RAND Program Evaluation Toolkit for Countering Violent Extremism, RAND Corporation.

Hofman, Joanna and Alex Sutherland (eds) (2018) 'Evaluating Interventions that Prevent or Counter Violent Extremism: A Practical Guide', RAND Corporation.

Holdaway, Lucy and Ruth Simpson (2018) 'Improving the impact of preventing violent extremism programming: A toolkit for design, monitoring and evaluation', UNDP.

Holmer, Georgia, Peter Bauman and Kateira Aryaeinejad (2018) 'Measuring Up: Monitoring and Evaluating P/CVE Programs', United States Institute of Peace.

Horgan, John and Kurt Braddock (2010) Rehabilitating the Terrorists?: Challenges in Assessing the Effectiveness of De-radicalization Programs, Terrorism and Political Violence, 22:2, 267-291.

Khalil, James and Martine Zeuthen (2016) 'Countering Violent Extremism and Risk Reduction: A Guide to Programme Design and Evaluation', Whitehall Report, Royal United Services Institute.

Koehler, Daniel and Verena Fiebig (2019) 'Knowing What to Do: Academic and Practitioner Understanding of How to Counter Violent Radicalization', Perspectives on Terrorism, Volume 13, issue 3.

Mastroe, Caitlin and Susan Szmania (2016) 'Surveying CVE Metrics in Prevention, Disengagement and De-Radicalization Programs', Report to the Office of University Programs, Science and Technology Directorate, Department of Homeland Security. College Park, MD.

Mazerolle, Lorraine, Adrian Cherney, Elizabeth Eggins and Angela Higginson (2020) 'Police Programs That Seek to Increase Community Connectedness for Reducing Violent Extremism Behaviour, Attitudes and Beliefs', PROTOCOL, Campbell Collaboration.

Murphy, L (2016) Critical Choices: Assessing the Effects on Education and Civic Engagement on Somali Youths' Propensity Towards Violence', Mercy Corps.

O'Halloran, Patrick, 'The Challenges of Evaluating Attitudinal Change: A Case Study of the Effectiveness of International Countering Violent Extremism (CVE) Programs' Canadian Political Science Association Conference, April 2017.

Pistone, I, E Eriksson, U Beckman, C Mattson and M Sager (2019) 'A scoping Review of Interventions for Preventing and Countering Violent Extremism: Current Status and Implications for Future Research', Journal for Deradicalization, Volume 19.

Williams, Michael and Steven Kleinman (2014) 'A Utilization-Focused Guide for Conducting Terrorism Risk Reduction Program Evaluations', Behavioral Sciences of Terrorism and Political Aggression, Volume 6, Issue 2.

b) Other Works Consulted

Bowers, Kate, Paul Gill, Ruth Morgan and Sarah Meiklejohn and Shane D. Johnson (2018) 'Challenges for EMMIE as a Realist Evaluation Framework ()Graham Farrell and Aiden Sidebottom'. In *Realist Evaluation for Crime Science: Essays in Honour of Nick Tilley*, Routledge.

Webber, David, Marina Chernikova, Arie W. Kruglanski, Michele J. Gelfand, Malkanthi Hettiarachchi, Rohan Gunaratna, Marc-Andre Lafreniere, and Jocelyn J. Belanger (2018) 'Deradicalizing detained terrorists', *Political Psychology* 39, no. 3.

Westhorp, Gill (2014) 'Realist Impact Evaluation: An Introduction', Methods Lab, Overseas Development Institute.