



United Nations
Educational, Scientific and
Cultural Organization



"UNESCO Chair in Prevention of Radicalisation
and Violent Extremism", Université de Sherbrooke,
Concordia University, Université du Québec à Montréal

Evaluating Programs for Preventing Violent Extremism

A systematic methodological review

JULY 2022

**PABLO MADRIAZA, DAVID MORIN, GHAYDA HASSAN, VIVEK VENKATESH,
MAUDE PLAUDE, CAROLINE DELI, MÉLINA GIRARD, LOÏC DUROCHER-CORFA,
RAPHAËL GRIJALVA-LAVALLÉE, KAREN POULIN**



United Nations
Educational, Scientific and
Cultural Organization



"UNESCO Chair in Prevention of Radicalisation
and Violent Extremism", Université de Sherbrooke,
Concordia University, Université du Québec à Montréal

**THIS
REPORT
SHOULD BE
CITED AS:**

Madriaza, P., Morin, D., Hassan, G., Venkatesh, V., Plaudé, M., Deli, C., Girard, M., Durocher-Corfa, L., Grijalva-Lavallée, R., & Poulin, K. (2022). *Ce que nous savons de l'évaluation de programmes de prévention de l'extrémisme violent : Une revue systématique méthodologique des évaluations des programmes de prévention dans ce domaine*. Chaire UNESCO en prévention de la radicalisation et de l'extrémisme violents (Chaire UNESCO-PREV).

The authors sincerely thank François Champagne, Iris Boyer and Michael King for their careful reading of this report and the invaluable recommendations that have enabled us to improve it.

Funded by the
Government
of Canada

Canada

THE AUTHORS



Pablo Madriaza

Pablo Madriaza is a professor in the Department of Social Work at the Université du Québec en Outaouais. He has served as scientific coordinator for the UNESCO Chair in Prevention of Radicalization and Violent Extremism and as general coordinator for the Canadian Practitioners Network for the Prevention of Radicalization and Extremist Violence.



David Morin

David Morin is a full professor in the School of Applied Politics at the Université de Sherbrooke and a co-holder of the UNESCO Chair in Prevention of Radicalization and Violent Extremism.



Ghayda Hassan

Ghayda Hassan is a clinical psychologist, a full professor of clinical psychology at the Université du Québec à Montréal and a co-holder of the UNESCO Chair in Prevention of Radicalization and Violent Extremism.



Vivek Venkatesh

Vivek Venkatesh is a filmmaker, a musician, a curator and an applied-learning scientist. He conducts research and creation projects at the intersection of public pedagogy and critical digital literacy. He is a co-holder of the UNESCO Chair in Prevention of Radicalization and Violent Extremism and the director of the Centre for the Study of Learning and Performance and a full professor of inclusive practices in visual arts at Concordia University in Montreal.



Maude Plourde

Maude Plourde is a graduate student in the School of Criminology at the Université de Montréal and conducts studies specifically related to current issues of domestic security.



Caroline Deli

Caroline Deli is a doctoral student in the Department of Criminology at the Université de Montréal. She holds a master's degree in neuropsychology from the Université d'Aix-Marseille, France. Her doctoral research examines the process of radicalization of incels through the study of life trajectories.



Mélina Girard

Mélina Girard is a master's student in the Department of Criminology at the Université de Montréal and a research assistant with the UNESCO Chair in Prevention of Radicalization and Violent Extremism, the Canadian Practitioners Network for the Prevention of Radicalization and Extremist Violence, the College of Immigration and Citizenship Consultants and the Digital Arts Resource Centre. She is also a co-founder of Academic Journal of Criminology/Journal Universitaire de Criminologie and the Regroupement Étudiant en Criminologie profil Analyse.



Loïc Durocher-Corfa

Loïc Durocher-Corfa is a doctoral student in the Department of Psychology at the Université du Québec à Montréal and a research assistant for the UNESCO Chair in Prevention of Radicalization and Violent Extremism.



Raphaël Grijalva-Lavallée

Raphaël Grijalva-Lavallée is a master's student of social psychology under the supervision of Professor Roxane de la Sablonnière at the Université de Montréal. Raphaël studies phenomena of social polarization and conflicts between ethnic groups in periods of social crisis and, more specifically, the evolution of prejudices against people of Chinese origin in Canada.



Karen Poulin

Karen Poulin earned her master's degree in applied political studies at the Université de Sherbrooke. Her research interests include feminist issues and the status of women. To expand her intellectual horizons and keep on learning, she has worked as a research assistant for the UNESCO Chair in Prevention of Radicalization and Violent Extremism (UNESCO-PREV).

Table of contents

Executive summary	7
Introduction	10
1. Theoretical and methodological shortcomings of past reviews	13
2. Methodology of this systematic review	24
3. Findings about the studies reviewed	28
3.1 Statistics on the studies reviewed	28
3.1.1. Number of studies by year	28
3.1.2. Number of studies by continent and country	29
3.1.3. Number of publications in academic literature and grey literature	30
3.1.4. Number of evaluated programs, by type of extremism targetted	32
3.1.5. Number of studies by program prevention level	33
3.1.6. Number of studies by scope of interventions evaluated	36
3.1.7. Number of studies that reported their funding sources	37
3.2. Statistics on the studies' authors	38
3.2.1. Gender	39
3.2.2. Geographic origin	39
3.2.3. Disciplines	41
3.2.4. Professions	43
3.3. Methodologies of the studies reviewed	44
3.3.1. Evaluation types (objectives)	44
3.3.3. Methodological design (quantitative, qualitative or mixed)	48
3.3.4. Use of experimental and quasi-experimental methods	50
3.3.5. Use of repeated measurements	52
3.3.6. Description of participants and use of control groups	54
3.3.7. Data-collection tools	57
3.3.8. Use of direct and indirect indicators of violent extremism	58
3.4. Limitations and conflicts of interest in the studies reviewed	60
3.4.1. Studies that identified their own limitations	61
3.4.2. Studies with reported conflicts of interest and unreported potential conflicts of interest	61
3.4.3. Types of limitations described by authors	62

3.5. Quality of the studies reviewed	68
3.5.1. Quality of the qualitative studies	69
3.5.2. Quality of the quantitative descriptive studies	72
3.5.3. Quality of the experimental studies	74
3.5.4. Quality of the quasi-experimental studies	77
3.5.5. Quality of the mixed-methods (qualitative + quantitative) studies	80
3.5.6. Did the quality of PVE evaluations improve over the years covered by this review?	84
3.6. Case studies	85
3.6.1. Evaluations of PVE programs targetting right-wing violent extremism	85
3.6.2. Evaluations of online PVE programs	88
Recommandations	91
Conclusion	94
Références	97
Appendix A: List of evaluation studies included in this systematic review	112
Appendix B: Complete methodology of this systematic review	123

EXECUTIVE SUMMARY

This systematic review is part of Phase 1 of the PREV-IMPACT Canada project, supported by the Community Resiliency Fund of the Canada Centre for Community Engagement and Prevention of Violence, a part of Public Safety Canada. The goals of this project are to develop and implement Canadian models for assessing practices for prevention of violent extremism (PVE) and, ultimately, to build the capacity of key PVE stakeholders in Canada.²

¹ In this systematic review, we regard the concepts of prevention of radicalization to violence and prevention of violent extremism (PVE) as synonymous, but we use mainly the latter, for convenience.



This review presents the background, methodology and findings of a systematic review of past evaluation studies of PVE programs.

1.1 METHOD

In this systematic review, we included all studies that had used primary, evidence-based data to evaluate PVE programs and that had been published in English, French, or Spanish from 2001 through 2019. To select the studies for this review, we first conducted keyword searches in 21 scientific databases and on 228 websites of organizations that work in this field. Next, to eliminate any studies that did not meet our eligibility criteria, we screened the title and abstract of each of the studies identified in these searches. Each study was reviewed by two members of the research team. To ensure that there was sufficient agreement between them at this first screening stage, a Cohen's kappa coefficient was calculated, with a very positive result (kappa = 0.86). The studies that passed this first eligibility screening were then reviewed in depth in a second screening to confirm whether they met all of the eligibility criteria. Lastly, we used a specialized tool to assess the methodological quality of all of the studies that had passed this second screening, and been included in this review.

1.2 FINDINGS

Out of the total of 18,886 studies that we had found in the keyword searches, we ended up selecting 219 studies (as reported in a total of 211 publications) for this review. Most of these studies evaluated PVE programs in Europe (n = 98), Africa (n = 50) and Asia (n = 42). Only 18 evaluated programs in North America and 7 in Australia. Though these studies had thus evaluated programs on several different continents, they were concentrated in a small number of countries, including the United Kingdom (n = 50), Indonesia (n = 16), the United States (n = 15), the Netherlands (n = 12), Kenya (n = 9) and Pakistan (n = 9). The geographic origins of the authors of these studies were similarly concentrated: about half came from either the United States (24.2%) or the United Kingdom (23.5%), while only 5% came from countries in Africa. One major reason for this pattern is that the authors from Western countries had evaluated programs in both Western and non-Western countries. Indonesia was an exception: the authors of most of the studies conducted there were Indonesian.

The year 2016 marked a turning point in the average number of studies published on evaluations of PVE programs. Prior to that year, the average number of such studies published annually was 9; in 2016, it jumped to 30.

The majority (n = 127) of the programs evaluated in the studies in this review targetted all types of violent extremism rather than any particular type; programs specifically targeting violent extremism related to Islamism came next (n = 84), followed by programs specifically targetting right-wing extremism (n = 20). Classified by prevention level, primary and targetted primary prevention programs² were the most numerous (n = 136), followed by secondary prevention programs (n = 61) and tertiary prevention programs (n = 46). These findings show that in recent years, the PVE programs evaluated have tended to adopt a more universal, less specific approach.

² All efforts that seek to reduce or eliminate risk factors or encourage protective factors and that target a specific community that is not identified as being at risk. Example: universal prevention programs in Muslim communities.

When the studies reviewed are classified by the types of evaluations that they involved, the two most common are impact evaluations (n = 159) and process evaluations (n = 110), indicating that authors were more interested in learning about the effects of these programs rather than the factors involved in their implementation.

Out of the 219 studies, 55.3% used quantitative methods (either alone or in mixed-methods designs), 41.6% used mixed-methods designs, and 43.8% used purely qualitative designs. This review also identified 54 studies that used quasi-experimental designs and 6 that used experimental designs, which is more than were found in earlier reviews. However, such designs seem better suited for evaluating primary prevention programs, which are more universal, rather than tertiary prevention programs, which are more specific, and in which methodological and ethical issues arise that make experimental and quasi-experimental methods impractical.

Past reviews of the literature have identified some important limitations in PVE program evaluations. First, very few of them take repeated measurements—multiple observations of the same participants at two or more points in time. Out of all the studies in the present review, only 22.4% took repeated measurements. (Though this percentage is low, it still reflects an increase over the past few years.) A second frequently reported limitation of prevention-program evaluations is the failure to use control groups. (Our review found only 20 studies that had done so.) A third important limitation arises from how hard it is to measure violent extremism directly in evaluation studies. The majority of the studies that we reviewed (74%) used indirect indicators only,³ just 4.1% used direct indicators only, and about 20% used both kinds.

The quality of the methods used in evaluations of PVE programs is another important concern. To assess the quality of the methods used in the studies in this review, we used the Mixed Methods Appraisal Tool (MMAT criteria for their respective), which can be applied to studies employing a variety of methods—experimental, quasi-experimental, quantitative descriptive, qualitative and mixed. Very few of the studies included in this review met all of the criteria measured by the MMAT and, on average, the results for each of the methods evaluated were middling, with the quantitative descriptive studies receiving the lowest ratings. The most common problem with the quality of all of the evaluation studies is the limited transparency in their methods—in other words, the limited amount of information and details that they provide in their methodology sections. The quasi-experimental studies, for example, often lacked detailed information about the representativeness of their samples. The experimental studies received poorer ratings with regard to the randomization of their participants between their treatment and control groups and in the comparability of the groups at the outset. In contrast, the studies that used qualitative methods received higher ratings, even though the interpretation of the results was not always sufficiently substantiated by concrete data.

1.3 CONCLUSION

Evaluation of PVE programs is difficult, to be sure, but it is possible. The main reason that there have been so few robust evaluations to date is that the field is so new and that the need to act has been so urgent over the past 15 years. With very few exceptions, evaluating these programs is not actually that different from evaluating other complex programs for preventing violence. This may explain why the number of studies of this kind has increased considerably in recent years. This increase is encouraging, but there are still many challenges to be overcome, including the quality of the studies done and the number of studies done on certain specific topics. For example, very few evaluations of online PVE programs have been done to date; many of those that have been done were not independent of the programs in question, and the findings were not always very conclusive. There have also been very few evaluations of programs to prevent right-wing violent extremism.

³ Indicators that do not directly measure radicalization, violent radicalization or violent-extremist sympathies.



Introduction

The continuing rise of extremist groups in various parts of the world and the widening variety of forms, targets and perpetrators of extremist and terrorist attacks make it clear that the traditional security response to such phenomena is insufficient and may sometimes even be counterproductive. The reason is that the factors originally used to explain these phenomena—first perceived primarily in the form of “jihadist” terrorism in the specific context of Middle Eastern conflicts—have failed to explain the new waves of extremist and terrorist attacks in the West. Hence local factors and individuals’ varied trajectories have now become a growing concern and received growing attention in the scientific literature. This trend has been accompanied by a proliferation of new conceptual and applied approaches. More attention has begun to be paid to the idea of violent extremism, which has been conceptualized on the basis of the process of radicalization to violence. Interest has also grown in the broader concept of prevention of violent extremism.⁴

Gradually it is being recognized that violent extremism can be prevented by means other than the war on terror and other traditional security-based approaches. Although such approaches remain a pillar of strategies for countering terrorism and extremism, emerging new approaches (notably, psychosocial ones) to both prevention and intervention offer much promise. Their specific contribution is that they view prevention not as stopping individuals from committing terrorist attacks or other acts of violence, but rather as taking earlier steps to reduce or eliminate the risk conditions that may make individuals more vulnerable to violent radicalization, extremism or terrorism.

This new view of prevention has broadened the field of action for practitioners. They have begun not only to borrow tools and approaches from other disciplines, but also to expand the range of prevention programs and services that they offer. In Europe, for example, the Radicalisation Awareness Network (RAN) had identified over 200 promising prevention programs as of 2019. In Canada, the Canadian Practitioners Network for the Prevention of Radicalization and Extremist Violence (CPN-PREV) had identified 26 secondary and tertiary prevention programs as of 2020 (Hassan, Ousman et al., 2020).

However, this proliferation of programs and services has not been supported or accompanied by the development of a clearly defined, rigorously delimited conceptual and empirical foundation, in terms of the definitions of the concepts used, the factors explaining the emergence of radicalization and violent extremism and especially the solutions proposed for dealing with them. Heydemann argues that this “blurring of boundaries reinforces perceptions of CVE⁵ as a

⁴ In this systematic review, we treat the concepts of prevention of radicalization to violence and prevention of violent extremism (PVE) as synonymous, but use mainly the latter, for convenience.

⁵ Countering violent extremism

catch-all category” (2014, p. 10). This problem of definition in turn affects our understanding of the phenomenon of radicalization to violence: according to Neumann, radicalization is “what goes on before the bomb goes off” (2008, p. 4).

Against this background of conceptual and empirical flux, one observation has been made repeatedly in the literature: **there is very little evidence-based data or tangible proof regarding the effectiveness of the measures being taken to prevent violent radicalization and violent extremism**. Moreover, very few of the studies that have addressed this issue have applied a sound methodological framework to do so (Baruch et al., 2018; Bellasio et al., 2018; Feddes et Gallucci, 2015; Gielen, 2017; Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021; Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021; Hirschi et Widmer, 2012; Romaniuk, 2015). Two recent systematic reviews of studies evaluating prevention programs identified only 48 studies that met a minimum threshold for methodological quality, out of an initial database of more than 15,000 documents (Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021; Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021).

Among researchers, there is a fairly strong consensus about the reasons for these shortcomings. Because of the conceptual, methodological and practical difficulties involved in evaluating such prevention programs, past evaluation studies have had trouble in measuring these programs’ actual effects and have instead tended to focus more on their quantitative outputs (how many actions they have undertaken) (Bellasio et al., 2018; Feddes and Gallucci, 2015; Gielen, 2017; Lindekilde, 2012b; Mastroe and Szmania, 2016). The fuzziness of the conceptual framework makes it hard for many researchers and practitioners to determine what they should measure and what indicators of effectiveness they should use. Moreover, the experimental method, which is the gold standard for evaluations in many other disciplines, is hard to apply to prevention of violent extremism. The actors involved in prevention of violent extremism (PVE) and countering violent extremism (CVE) thus have little to guide them when it comes to evidence-based data or practices that offer promise or have proven effective for achieving these goals. Despite this deficit, there is a certain consensus among researchers, practitioners and policymakers about the need to develop suitable models for evaluating such programs. For practitioners, evaluations provide a means of improving their practices. For researchers, they provide a means of better understanding the mechanisms and processes that explain the success or failure of interventions. Lastly, for policymakers, evaluations can be used to guide public policy and to make more effective use of the limited sources of funding.

To fill these gaps, the UNESCO-PREV Chair has developed the PREV-IMPACT Canada project. Supported by the Community Resiliency Fund of the Canada Centre for Community Engagement and Prevention of Violence (a part of Public Safety Canada), the PREV-IMPACT Canada project aims to develop and implement Canadian models for assessing practices in primary, secondary and tertiary prevention of violent extremism (PVE) and, ultimately, to build the capacity of key PVE stakeholders in Canada. The first phase of this project fundamentally involves research. Its objectives are to:

- document and compare PVE evaluation strategies and tools in Canada and elsewhere based on existing evidence and practices;
- develop distinct, innovative evaluation models (logic models, strategies, tools, indicators, methodology) adapted to local primary, secondary and tertiary prevention programs to guide PVE policies and programs in Canada;
- test the evaluation models on three Canadian PVE programs.

The present systematic review is the first step in this project. It presents the background, methodology and findings of a systematic review of evaluation studies of programs for primary, secondary and tertiary prevention of violent radicalization and violent extremism that have been published in English, French and Spanish from 2001 through 2019.

THIS REVIEW IS DIVIDED INTO FOUR PARTS:

01

Part 1 summarizes the state of the art regarding evaluation of PVE programs and serves as the basis on which this review compares past evaluation studies of such programs and the advances that have been made in this field. Section 1.1 discusses the shortcomings that past literature reviews have found in evaluations of PVE programs. Section 1.2 discusses and explains the various types of difficulties that evaluators encounter in evaluating PVE programs. Section 1.3 briefly discusses the strengths and weaknesses of past literature reviews in this field.

02

Part 2 briefly describes the methodology that we used in the present systematic review, along with its limitations. (Appendix B presents this methodology in more detail.)

03

Part 3 presents the main findings of this systematic review. Section 3.1 provides various statistics on the evaluation studies that we included in this review and the programs that they evaluated. Section 3.2 presents the characteristics of the authors of these studies, while section 3.3 analyzes their methodologies. Section 3.4 looks at the limitations of these studies and section 3.5 at their methodological quality. Section 3.6 examines two case studies of evaluations of PVE programs—programs addressing right-wing violent extremism in one case, and online programs in the other.

04

This review ends with our recommendations and concluding remarks concerning evaluation of programs for preventing violent extremism.

01



Theoretical & methodological shortcomings of past reviews

In this section, we answer the following key questions:

- 1) What shortcomings has the scientific literature identified in past evaluations of programs for prevention of violent extremism (PVE)?**
- 2) What difficulties do evaluators encounter in conducting such evaluations, and what are the reasons for these difficulties?**
- 3) What have been the strengths and weaknesses of past literature reviews of such evaluations?**

We then conclude with a brief discussion about the state of evaluation in this field.

1.1. SHORTCOMINGS THAT PAST LITERATURE REVIEWS HAVE IDENTIFIED IN EVALUATIONS OF PVE PROGRAMS

Ever since terrorism and violent extremism became a formal field of study, the theoretical and methodological weaknesses of research in this field have been a persistent subject of debate and concern. In a review of terrorism studies from the early 1980s, Schmid and Jongman (1988, cited in Silke, 2001) reported some major methodological problems, in particular regarding data collection and analysis. For example, 92% of the studies reviewed used newspapers and public documents as primary sources. Silke (2001) analyzed articles published from 1995 to 2000 in the two most frequently cited scientific journals in this field: *Terrorism and Political Violence* and *Studies in Conflict and Terrorism*. His findings were similar to Schmid and Jongman's: an overrepresentation of studies that mainly used open secondary sources such as newspapers and public documents, along with anecdotal qualitative sources, few interviews (most of them unstructured) and very few quantitative analyses.

Following the terrorist attacks in the United States on September 11, 2001, Silke (cited in Neumann and Kleinmann, 2013) conducted another such analysis, of studies published from 2002 to 2004, and saw very little progress in these respects. In 2013, Neumann and Kleinmann conducted the most complete study yet in this field, this time examining studies on radicalization that had been published between 1980 and 2010. These authors found that although there had been clear improvement in this field, more than a third of the studies that they reviewed lacked rigour either methodologically (in their processes) or empirically (in the kind of data that they used), while a large fraction based their findings on secondary data.

More recently, Shuurman (2018) conducted another study, to evaluate the state of research on terrorism from 2007 to 2016. He reviewed all of the articles that had been published during that period in nine specialized journals and that had used primary data. He too observed improvement, especially regarding the use of primary data. But he confirmed that the field had yet to coalesce, in particular because most of the authors had contributed to it only once.

In reviewing evaluations of prevention programs, researchers have identified another set of methodological shortcomings. These can be summarized as a lack of systematization, consistency and harmonization in the methodologies, which are usually *ad hoc* and do not satisfy minimum scientific criteria (Davey et al., 2019;

Feddes and Gallucci, 2015; Lindekilde, 2012b; Marret et al., 2017). For example, no evaluation studies with experimental designs were identified in any past reviews, and very few studies with even quasi-experimental designs (Bellasio et al., 2018; Feddes and Gallucci, 2015; Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021; Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021; Mastroe and Szmania, 2016). Lack of empirical data is, in fact, a recurrent feature of these evaluations, which have focused more on describing the activities undertaken (output evaluation) than on assessing their effectiveness (outcome or impact evaluation) (Baruch et al., 2018; Marret et al., 2017; Romaniuk, 2015). In addition, though many of the programs that have been evaluated were designed to achieve results in the long term, the evaluations have been limited to the short term, using longitudinal models in only a few exceptional cases (Marret et al., 2017; Romaniuk, 2015). Studies in this field also rarely use control groups, which makes it hard to know with any certainty whether the positive or negative effects of the evaluated programs are due to the interventions or to other, concurrent factors (Marret et al., 2017). In some cases, this difficulty is due to limited access to information. For example, Davey et al. (2019) state that evaluations of online prevention programs largely try to measure the reach of the interventions and the engagement of the participants rather than changes in their attitudes or behaviours, in particular because of the limited information available on social-media platforms.

Cost-benefit analyses of PVE programs are also scarce (Marret et al., 2017). For example, only 19% of the samples analyzed by Feddes and Gallucci⁶ (2015) contained analyses of the economic costs of the programs. One last problem that researchers have identified relates both to lack of transparency in methodology and information sources and lack of independence among evaluators (Horgan and Braddock, 2010; Mastroe and Szmania, 2016; Williams and Kleinman, 2014). For example, the reported success rate for a set of recidivism-prevention programs was based on government sources, with no explanation of the methods used to measure the impact of these programs (Horgan and Braddock, 2010). In other cases, and particularly in non-governmental online initiatives, Davey et al. (2019) found that programs were rarely subjected to independent evaluation. This finding has not, however, been corroborated by other studies that focused on all types of prevention programs combined (Bellasio et al., 2018).

⁶ The authors do not, however, state whether this analysis included a cost-benefit analysis as such.

1.2 REASONS FOR DIFFICULTIES IN EVALUATING PVE PROGRAMS

The lack of empirical data on the effectiveness of past PVE programs can be attributed mainly to the challenges involved in conducting evaluations in a field that is so new and constantly changing. Holmer, Bauman and Aryaeinejad write that “Those challenges can be grouped into two categories: analytic challenges, such as establishing causality, addressing contextual variations, and developing valid indicators; and practical challenges, such as collecting relevant and reliable data” (2018, p. 4). Marret et al. (2017) offer a fairly similar assessment and identify two further shortcomings: there is no standardized methodology for evaluating such programs, and the practitioners who deliver them lack the knowledge to design and conduct appropriate evaluations of them.

The broadest and most complete analysis of these difficulties was performed by Bellasio et al. (2018), who conducted a systematic review of evaluations of programs for preventing and countering terrorism and violent extremism and analyzed the conceptual, methodological and practical difficulties identified by their authors. As shown in Box 1, Bellasio’s team classified these difficulties into five categories:

- 1) **inherent complexities of the field of counterterrorism and preventing and countering violent extremism (concepts and definitions, security concerns, etc.);**
- 2) **challenges associated with measuring real-world phenomena (rarity of events);**
- 3) **challenges associated with existing evaluation designs (difficulties in establishing causal links, lack of theories of change);**
- 4) **practical difficulties of conducting evaluations (difficulties in accessing information, financial constraints, etc.);**
- 5) **drawbacks and benefits of specific evaluation methods.**

Box 1. Difficulties involved in evaluation of programs for preventing violent extremism

1) **Inherent complexities of the field**

- Target groups
- Stakeholders
- Security concerns
- Interventions
- Concepts and definitions

2) **Challenges associated with measuring real-world phenomena**

- Rarity of events and lack of outcome measures.
- Lack of available outcome metrics
- Measuring long-term effects
- Accounting for social norms and expectations
- Tracking exposure to interventions

3) **Challenges associated with existing evaluation designs**

- Difficulties with claiming causality and conducting experiments and quasi-experiments.
- Challenges of adopting a longitudinal study approach
- Lack of theories of change

4) **Practical difficulties of conducting evaluations**

- Resource constraints
- Difficulty in accessing information about interventions and effects
- Access to data
- Difficulties with sample size

5) **Drawbacks and benefits of specific evaluation methods**

- Constraints of model-based investigations
- Constraints of survey instruments
- Importance of triangulation and strengths of qualitative methods

Source: Bellasio et al., 2018

Taking the above classification as our starting point and considering the findings of other authors who have researched PVE program evaluation, we decided that it is essential to describe these difficulties in more detail. We believe that the following seven categories provide an appropriate classification:

- 1. Conceptual difficulties and difficulties with understanding violent extremism**
- 2. Difficulties with design and implementation of prevention programs**
- 3. Difficulties with the actors involved**
- 4. Difficulties with funding**
- 5. Difficulties with limited number of cases and limited access to data**
- 6. Difficulties with methodology**
- 7. Difficulties with politicization of the phenomenon**

1.2.1. Conceptual difficulties and difficulties with understanding violent extremism

In the introduction to this review, we noted that the moment one considers evaluating a PVE program, the following difficulty arises: there is no consensus definition of violent extremism, so how can it be differentiated from other phenomena, and what kinds of empirical data can explain its emergence (Lindekilde, 2012b; Mastroe and Szmania, 2016; Ris and Ernstorfer, 2017)? Understanding of this phenomenon has improved in recent years, notably as the result of a number of published empirical studies, systematic reviews and meta-analyses.⁷ But the study of violent extremism is still marked by the empirical weaknesses mentioned earlier (Hirschi and Widmer, 2012; Lindekilde, 2012b; Ris and Ernstorfer, 2017). In the case of right-wing extremism, Hirschi and Widmer (2012) write that the nature of the problem has been defined with varying degrees of precision, that there are competing explanations and that the large number of differing explanations has been identified as an obstacle to evaluation. Other researchers state that because our understanding of the individual behaviours of violent extremists and the ties among them is limited, and because there are so many different pathways that can lead to violent extremism, it is impossible to design prevention programs with any precision, and that they are often designed with no sound empirical foundation (Baruch et al., 2018; Ris and Ernstorfer, 2017).

This conceptual problem may seem far removed from the realities of the field and of practice, but it has practical implications for designing programs (how to explain and induce changes in program participants), for implementing programs (inclusion criteria and reference criteria for at-risk individuals), and for developing indicators to measure programs' success. For example,

for tertiary prevention programs, recidivism is often cited as the only factual criterion to be considered (El-Said, 2015). But the conceptual and empirical boundaries of the phenomenon of recidivism are themselves the subject of debate, especially as regards "deradicalization" programs (Horgan and Braddock, 2010). What exactly does recidivism mean? Does it consist of committing another violent act, or rejoining an extremist group, or re-embracing radical ideas? What time scale should be considered? What factual indicators and information sources should be used to measure such changes in behaviour?

On the other hand, Ris and Ernstorfer (2017) argue that this insistence on a clear definition of the concept of preventing/countering violent extremism and on the exceptional nature of programs of this kind overlooks the fact that most approaches to preventing violent extremism are also based on experience acquired in other fields and can therefore be evaluated according to criteria that are often used elsewhere.

1.2.2. Difficulties with design and implementation of prevention programs

Another difficulty in evaluating prevention programs and in harmonizing methods of doing so arises from their wide variety and great specificity (Hirschi and Widmer, 2012; Lindekilde, 2012b; Marret et al., 2017; Mastroe and Szmania, 2016). These programs are highly heterogeneous; they pursue many different aims, and there is no common understanding about their approaches and objectives (Chowdhury Fink et al., 2013; Lindekilde, 2012b; Marret et al., 2017). Prevention programs are often deliberately designed to suit the local context and the specific traits of the individuals or groups that constitute their target populations. Hence these programs cannot be evaluated on a general basis, but only through a differentiated assessment of their specific local effects. The efficacy of these programs therefore depends greatly on the context in which they are implemented. According to Mastroe and Szmania (2016), this reality suggests that it might be hard for programs developed in one particular geographic area to be transferred to others. Some researchers also believe that there is very little in the way of well grounded evidence and findings as to what works, in what context and for what type of target group (Gielen, 2017; Ris and Ernstorfer, 2017).

Beyond these issues, the methods by which programs are to be evaluated are rarely defined when the programs themselves are being designed or implemented. Instead, evaluation methods often emerge late in the day, as an external process (not to say a foreign body), disconnected from the prevention programs themselves.

Lastly, in most prevention programs, the lack of theories

⁷ See, for example, Gill, Clemmow, Hetzel, Rottweiler, Salman, et al., 2020; Vergani, Iqbal, Ilbahar, and Barton, 2020; Wolfowicz, Litmanovitz, Weisburd, and Hasisi, 2019.

of change poses an additional difficulty for evaluation (Bellasio et al., 2018; Chowdhury Fink et al., 2013; Williams et Kleinman, 2014). Theories of change (Box 2) can be used to explain how the activities planned for a given prevention program are supposed to produce the desired effects and, if need be, what mechanisms underlie these actions. The application of a theory of change cannot compensate completely for the failure to incorporate an evaluation plan into a prevention program from the outset, but can be especially helpful for guiding the evaluation of that program and for choosing the most relevant indicators for that purpose.

Box 2. Theories of change

Connell and Kubisch define the theory-of-change approach as “a systematic and cumulative study of the links between activities, outcomes, and contexts of the initiative” (1998, p. 2). This approach was designed to evaluate and accommodate the multi-level, multi-dimensional impacts of comprehensive interventions in which the task of linking actions to outcomes is extremely complex, at a time when existing evaluation approaches were considered inadequate or inappropriate (Sullivan and Stewart, 2006). This evaluation model is part of the theoretical approaches to evaluation and is based on the idea that evaluators must help to identify the theory of action implicit in an intervention in order to define what should happen if the theory is correct (Sullivan and Stewart, 2006). Part of this task is to identify the indicators of short-, medium- and long-term change that will let the evaluators determine what elements they need to form an evaluative judgment. The theory-of-change approach is helpful for improving program planning, facilitating decisions about evaluation methods, and reducing the difficulties of causal attribution that are often the bane of evaluations of interventions of this kind (Mackenzie and Blamey, 2005).

1.2.3. Difficulties with the actors involved

Because prevention programs are so complex, they often involve a wide range of stakeholders. These programs have differing mandates and objectives and can also have differing needs with regard to evaluation (Bellasio et al., 2018; Chowdhury Fink et al., 2013). First of all, the wide variety of actors necessitates a complex evaluation to take their differing perspectives into account, which can increase its scope, time requirements and costs. The differing perspectives of these actors must be considered with regard not only to data collection but also to the

specific needs of the practice setting. For whom the evaluation is being done thus becomes a fundamental question.

Some funding agencies prefer that the impact of programs be evaluated according to a binary logic (whether they work or not) and expect an emphasis on communicating the project’s outcomes (Ris and Ernstorfer, 2017). The interest that some governments have in showing positive results not only may create conflicts of interest but also raises ethical questions about evaluation (Horgan and Braddock, 2010; Lindekilde, 2012b; Mastroe and Szmania, 2016). Prevention practitioners may perceive an evaluation as an opportunity as well as a constraint (Clement et al., 2021). They may regard it on the one hand as a means of improving their day-to-day practices, but on the other as a kind of audit in which their work is being monitored and its quality may be called into question. Lastly, program participants’ role in the program-evaluation process is often seen as nothing more than to supply information. (Participatory approaches are rare in this field.) But limiting program participants’ role in this way is problematic: they should actually be involved from the very start of the evaluation process, to reduce the risk of biasing its findings. However, as some past evaluations have shown, some prevention initiatives may potentially stigmatize participants (Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021), especially in the Muslim community, and so some program participants may be reluctant to participate in evaluation initiatives or to answer evaluation questions transparently. The situation becomes even more complex in the case of interventions that are mandatory (within the correctional system, for example) or connected with legal proceedings (Mastroe and Szmania, 2016).

For many researchers, it is important to know not only for whom an evaluation was done, but also who did it (Horgan and Braddock, 2010; Marret et al., 2017; Mastroe and Szmania, 2016). For one thing, the number of people with expertise in evaluation is limited, and they require additional training to produce evaluations that properly consider the specific characteristics of PVE programs. For another, the ethical issues raised above also show the need to instil a culture of transparency among both governments and researchers in order to ensure the reliability of the data communicated (Horgan and Braddock, 2010; Marret et al., 2017; Mastroe and Szmania, 2016).

1.2.4. Difficulties with funding

Funding for evaluations is a constant source of concern. In 2013, Chowdhury Fink et al. found that it was hard to obtain funding for program evaluations and that the amounts allocated for them were often modest. The high costs of evaluations may be the explanation. In an international study conducted in parallel with the present systematic review, 57 researchers and practitioners were asked about evaluation issues. The practitioners stated that program evaluations could sometimes cost more than the programs themselves (Madriaza et al., 2021). Bellasio et al. (2018) stated that lack of funding has a negative effect on the design, implementation and quality of the evaluations that are conducted. Funding has a notable impact on time constraints, which prevent data from being gathered at the most appropriate time and thus cause opportunities to collect relevant data to be missed. These constraints also reduce the time available for analyses and make it harder to design and carry out sound evaluations.

1.2.5. Difficulties with limited number of cases and limited access to data

One problem that researchers often mention is difficulty in accessing data and other relevant information. Although in certain Western countries, terrorism and violent extremism are matters of considerable concern, their actual incidence, as measured both by number of events (such as terrorist attacks) and number of individuals recruited by extremist groups, is low (Lindekilde, 2012b; Pistone et al., 2019). For instance, in one study, researchers had planned to evaluate three prevention programs in the French probation system, a known hotbed for recruitment and radicalization. But because the numbers simply were not large enough, the researchers had to give up on evaluating one of the programs, and had to evaluate the two others using only a limited number of cases, and without the control groups that had been planned (Madriaza et al., 2018b). In another study, Schuurman and Bakker (2016) evaluated a recidivism-prevention program in the Netherlands. (2016). Initially, these authors had planned to evaluate this program's impact, but they ended up having to evaluate its process instead, for reasons much like in the French study: at the time of the evaluation, only five individuals were enrolled in the program. As Baruch et al. (2018) point out, conducting scientific experiments with small numbers of cases is especially difficult in fields that are politically sensitive. In evaluations of on-line intervention programs, access problems are often due to the platforms' limitations regarding the type of information that can be collected (Davey et al., 2019).

Accessing information is also complex because of the very nature of the problem. When researchers ask to access the target group for a program and relevant

documents for evaluating it, the request may be denied on the grounds that this information is sensitive or even confidential—for example, because it relates to current criminal proceedings or the work of intelligence services or is considered important for national security (Hirschi and Widmer, 2012; Ris and Ernstorfer, 2017). For ethical reasons, social-service agencies will not disclose their clients' personal information, which makes such information inaccessible as well.

1.2.6. Difficulties with methodology

Methodological difficulties in conducting PVE evaluations are partly the cumulative result of the difficulties discussed in the preceding pages. But these evaluations by definition face the same challenge as in any other field where the focus is on prevention: how to demonstrate that a given behaviour or action has not taken place (Holmer et al., 2018; Lindekilde, 2012b; Madriaza and Ponsot, 2015; Mastroe and Szmania, 2016; Ris and Ernstorfer, 2017). In this regard, the small number of cases of violent extremism available for analysis poses an additional problem. In other fields, such as crime prevention, researchers can use the vast number of cases (crimes that are actually committed) to mathematically model the impact of a given prevention measure. This is not possible when it comes to prevention of violent extremism.

Beyond this fundamental problem, researchers agree that there are no clear, consistent, harmonized indicators for measuring the impact of PVE programs (Baruch et al., 2018; Davey et al., 2019; Feddes and Gallucci, 2015; Horgan and Braddock, 2010; Lindekilde, 2012b; Mastroe and Szmania, 2016; Romaniuk, 2015). For such programs, success is hard to define and observe, even, as noted earlier, in the case of rehabilitation of individuals who have committed terrorist acts. This is probably due to the difficulty of detecting and measuring the attributes of violent extremism (Baruch et al., 2018). Although there are not many cases, there is no single profile or pathway to violent extremism, so the number of indicators that might be used is potentially unlimited. In the case of right-wing extremism, Hirschi and Widmer (2012) believe that it is hard to clearly separate right-wing attitudes from other attitudes, especially more latent ones, even if the more obvious characteristics (tattoos, symbols, costumes, etc.) are relatively easy to study.

Changes in terms of radical ideas are harder to use as performance measures. Researchers therefore have a greater tendency to use indirect indicators (Marret et al., 2017) that are associated with extremist ideas and violent behaviour theoretically rather than empirically. The relationship between ideas and behaviours also remains hard to establish and prove (Holmer et al., 2018). It is often complicated to show that there is a cause-and-effect relationship between the intervention made

and the changes observed and that these changes are not in fact attributable to other factors (Holmer, 2013; Lindekilde, 2012b; Mastroe and Szmania, 2016). The upshot is a wide variety of evaluation approaches and methodologies that address many different aspects of the programs evaluated, often in a pragmatic fashion.

1.2.7. Difficulties with politicization of the phenomenon

Programs to prevent violent extremism do not operate in a neutral context. There is a sense of urgency about violent extremism, which is thought to be on the rise, and that feeling is fed by pressure from the media and concern from society at large. This fraught context has specific repercussions for evaluations. In particular, political pressures create an imperative not only to take action, but also to demonstrate that the action taken has been effective and that the approach used to evaluate it has been reliable. Some researchers have already expressed concerns about the reliability of the data communicated regarding certain “deradicalization” programs that are subject to tremendous political pressure (Horgan and Braddock, 2010; Mastroe and Szmania, 2016). Hirschi and Widmer provide a better explanation of the challenges that evaluators face in this context:

Incidents that have a right-wing extremist connection often have considerable public resonance, leading to an emotionalization of the phenomenon itself. The evaluation then stands faced with the challenge of so positioning itself that it can appear equally trustworthy to those who participate in a measure as well as to the wider circle of persons (for example in politics or in the media) who are interested in or affected by it. (2012, p. 172).

Consequently, many programs have been politically motivated rather than evidence-based. As a result, they have usually either had unclear, unrealistic objectives or been grounded in untested or overambitious theories of change that posed obvious difficulties for evaluators (Baruch et al., 2018; Ris and Ernstorfer, 2017). One good example of such a program was Pontourny, a “deradicalization centre” opened as an emergency measure as part of a French national strategy to turn marginalized youth away from jihad (Albert et al., 2020). Pontourny was one of the various failures of this strategy that received the most media attention. Practitioners and researchers agree that despite the pressures to which evaluations are subjected by politics and the media, the political will must be mustered to undertake evaluations that are independent and science-based and to draw the right lessons from them (Chowdhury Fink et al., 2013).

1.3. STRENGTHS AND WEAKNESSES OF PAST LITERATURE REVIEWS

To show why we decided to conduct our own systematic review of PVE program evaluations, we will now discuss and identify the strengths and weaknesses of past systematic reviews of evaluation studies in this and related fields. First we discuss these reviews in chronological order, and then we present our brief general conclusions about them and about the contributions that we felt that a new systematic review could make to the evaluation of PVE programs.

1.3.1 Chronological discussion of past reviews

Table 1 lists these past reviews and shows, for each of them, the number of studies that they included that we included in our own systematic review, the number that we excluded, and the reasons that we excluded them.

Lum, Kennedy and Sherley (2006) were one of the first teams to conduct a systematic review of counterterrorism programs. These authors reported that most of the programs discussed in the literature that they reviewed had never been evaluated, which indicated the lack of a factual basis for them. In fact, out of an original database of 20,000 titles, these authors found only seven counterterrorism programs that had been subjected to moderately rigorous evaluations.

Table 1. Systematic reviews and literature reviews of programs for prevention of violent extremism

Systematic reviews and literature reviews	Studies included	Studies excluded			
		CT*	NPD*	NE*	M*
Bellasio et al., 2018	28/48	7	3	2	8
Carthy, Doody, Cox, O'Hora, and Sarma, 2020	0/14	14			
Feddes and Gallucci, 2015	11/55	6	19	2	17
Gielen, 2017	25/73	4	38	3	3
Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021; Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021	47/51	2	1		1
Madriaza, Ponsot, Marion, Monnier, Ghanem et al., 2017; Madriaza and Ponsot, 2015	12/23	6	3		2
Mastroe and Szmania, 2016	16/43	7	14	1	5
Pistone et al., 2019	17/38	5	12	2	2
Pratchett, Thorp, Wingfield, Lowndes and Jabbar, 2010	1/18	4	6		7
Taylor and Soni, 2017	1/7	5		1	

CT: Studies classified as dealing with counterterrorism measures, not directly related to prevention or not dealing with any specific program.

SDP: Studies with no primary data or with anecdotal data

NE: Non-evaluation studies.

F: Publications inaccessible or merged with other publications that used the same sample and analysis

The first-ever review of PVE program evaluations was a “rapid evidence assessment” by Pratchett et al. (2010); it dealt mainly with programs in the United Kingdom. This review included 18 studies, the earliest of which dated from 1996,⁸ well before PVE programs had become a trend. Unlike in later reviews, Pratchett’s team enjoyed privileged access to evaluations by the United Kingdom’s Department for Communities and Local Government. According to the authors, none of these 18 studies made it possible to know which measures had worked best in changing communities’ attitudes toward violent extremism, because of a methodological limitation: all of these studies had a fundamentally qualitative design. Two years later, Christmann et al. (2012) found only two additional programs in the United Kingdom.

In 2014, the IMPACT Europe project published a report summarizing three studies, the last two of which addressed the subject of prevention (van Hemert et al., 2014). One of these two studies was based on a non-representative sample of 100 prevention programs from two databases covering the years 2000 to 2014. The goal of this study was not to review the evaluated programs, but rather to describe a set of variables present in the publications. The authors did, however, identify 52 programs that said that they had performed evaluations. The largest number were impact evaluations, followed by process evaluations and economic evaluations. But most

of these evaluations consisted of nothing more than simple feedback from the program participants.

The other prevention-related study discussed in the IMPACT Europe report did focus specifically on evaluation methods and served as the basis for the study by Feddes and Gallucci (2015). These authors conducted a more systematic literature review and identified 55 publications about program evaluations, involving 135 samples from 9 databases covering the years 1990 to 2014. Out of these 135 samples, only 16 came from programs that were subjected to empirical evaluations using primary data. The remainder involved anecdotal evaluations (49%)⁹ or theoretical ones (39%)¹⁰. Most of the evaluations that Feddes and Gallucci identified were a combination of impact and mechanism evaluations (46%),¹¹ followed by process evaluations (23%)¹² and economic evaluations (19%). Although most of the programs had been designed to produce long-term effects, the vast majority of the evaluations of these programs proved to be cross-sectional. In fact, only three of the samples came from studies that used quasi-experimental methods. Feddes and Gallucci also found a lack of theories of change in the programs evaluated (a finding that has recurred often in the literature since). These authors found a theory of change in only 12% of the samples, and in the vast majority of cases (60%), the evaluations were not based on any specific theory. Feddes and Gallucci (2015) was

⁸ This was a study by Knox and Hughes (1996) concerning community programs to strengthen the peace during the post-conflict period in Northern Ireland. It was not included in the present systematic review.

⁹ Publications that described only the program.

¹⁰ Publications that tested a theory by means of a literature review without using any qualitative or quantitative data.

¹¹ Evaluations that explain why a program worked (Feddes and Gallucci, 2015).

¹² Evaluations of the processes used to carry out programs (Feddes and Gallucci, 2015)

the first study of its kind and provided an important basis for the advancement of evaluation methodologies in this field.

In 2016, Mastroe and Szmania published a survey of the literature on evaluation metrics used in empirical studies of programs to counter violent extremism between 2005 and 2016. These researchers found 43 such studies. Although these studies were supposed to have an empirical focus, the authors found that only 22 of them (5 dealing with prevention and 17 with disengagement or deradicalization) used primary data. The 21 remaining studies dealt with the activities carried out to achieve the objectives of the evaluated programs (output evaluation). Most of these studies were descriptive in nature and, as in past reviews, none of them used experimental designs. The authors concluded that as of the time that they published their survey, because of the lack of empirical studies, there was very little consensus about the effectiveness of prevention programs. The authors questioned how much trust could be placed in published research that did not provide a transparent account of the methods used to make its findings.

Bellasio et al. (2018) is probably one of the most detailed studies ever done on evaluation of efforts to deal with violent extremism. It is similar to Feddes and Gallucci (2015) and Mastroe and Szmania (2016) in its specific focus on evaluation methodologies. For their systematic review, Bellasio's team produced an inventory of evaluations conducted between 2013 and 2018¹³ on strategies, policies and interventions in the fields of counterterrorism and preventing and countering violent extremism. These authors identified 48 such evaluations. The largest number were impact evaluations, followed by process evaluations. Twenty-four of the evaluations used qualitative methods. Compared with the other reviews mentioned above, this one found a larger number of evaluations that used mixed methods (14). Bellasio's researchers found the same lack of longitudinal studies as earlier reviewers: most of the studies in this review were cross-sectional and/or mid-term, and only six comprised ex post measurements. Bellasio's team devoted less attention to evaluation of methodological quality, but made a few observations worth noting. The vast majority of the evaluations that they reviewed were conducted by external evaluators, which is an indicator of the independence of the evaluations, and more than half were subjected to blind peer review or review by an independent panel. One interesting finding concerned the evaluation approach: Bellasio's team could not identify any clear theoretical approach in 33 of these studies. In 2018 and 2019, other reviews were conducted, but they did not focus specifically on the evaluation methodologies of the studies concerned. The scoping review by Pistone et al. (2019) identified 112 publications dating from 1989 to 2017.

Only 38 of them evaluated outcomes, and only 15 used primary data. The remainder only discussed or analyzed the interventions without measuring their effectiveness.

In 2020, Carthy et al. (2020) published a systematic review of 14 studies on the use of counter-narratives to prevent violent radicalization. This was the first systematic review to conduct a meta-analysis of the effects of measures of this kind. Surprisingly, although the title of this review might suggest otherwise, none of these studies specifically looked at violent radicalization, although the programs in question might be applied in this context.

In 2021, research teams led by Ghayda Hassan conducted two systematic reviews of programs to prevent violent radicalization, based on a public-health model. In Hassan, Brouillette-Alarie, Ousman, Kilinc et al. (2021), the team examined 33 studies of primary and secondary prevention programs, while in Hassan, Brouillette-Alarie, Ousman, Savard et al. (2021), they examined 18 studies of tertiary prevention programs. With the exception of Carthy et al. (2020), these were the first reviews to use a specific measure of methodological quality as an inclusion criterion, and the first to include only studies in which only primary data were analyzed, which can be regarded as an improvement in the quality of such studies. One major limitation of the two reviews by Hassan's teams, and of several other reviews described in the preceding pages, relates to publication bias: they looked only at studies published in specialized journals. In their review of studies of primary and secondary prevention programs, Hassan's researchers found more studies using mixed methodologies than in previous reviews (16), which was a recommendation in the specialized literature (Williams and Kleinman, 2014). Ten of the studies that they reviewed were exclusively qualitative, and only 7 used quantitative methods. As regards methodological quality, the studies included in this review received an average score of 6.7 on a scale of 0 to 10; the studies with a score of 3 or less were excluded.¹⁴

Lastly, Zeuthen (2021) recently published a new systematic review of the literature on tertiary prevention activities. The author found 15 studies that met her inclusion criteria, and 34 other publications consisting mainly of literature reviews in this field and related ones. Zeuthen states that this review included an evaluation of the quality of the studies included, but does not describe the method used for this purpose.

¹³ Eleven of these studies were recommended by experts and published before 2013.

¹⁴ This analysis was performed by the authors of the present systematic review on the basis of results presented in Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021 and Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021.

1.3.2 Conclusions about past systematic reviews and reasons for conducting this one

As the preceding discussion indicates, a sufficient number of relatively systematic reviews of PVE program evaluations have been done in the past for us to draw some conclusions about the state of evaluation in this field, the challenges faced in conducting reviews of this kind, and the gaps in current knowledge, limitations and apparent biases that the present review was designed to address.

Unquestionably, the first issue of concern in past reviews is that, partly because PVE evaluation is such a new field, reviewers have been forced to make some pragmatic compromises. According to Chowdhury Fink et al. (2013), the researchers and practitioners participating in a symposium on the difficulties of evaluating PVE programs, held in Ottawa in 2013, agreed on one point: the difficulties of evaluating such programs are obvious. Many actors have accordingly underscored the importance of taking a pragmatic approach to PVE evaluation. This same view has been expressed by other researchers as well (Marret et al., 2017; Romaniuk, 2015). Romaniuk put it this way: “Rather, a sense of pragmatism seems to prevail, with evaluators gathering the data they can with the resources available” (2015, p. 36).

This same observation applies to the literature reviews presented in the preceding section. The limited number of empirical evaluations and the pressing need to determine the effectiveness of prevention measures have led authors to include in their reviews some program-evaluation studies that did not use primary data or were based on anecdotal data. For instance, a significant share of past reviews have included evaluations that contained no data analysis. The concerns that Mastroe and Szmania (2016) expressed about the reliability of the data communicated regarding “deradicalization” programs also apply to studies that involve output evaluations or that use anecdotal data. The most recent studies, however, offer more promise in terms of empirical material and reflect improvement in the field. They have allowed more specific conclusions to be drawn about the effectiveness of the prevention measures that have been attempted.

A second issue of concern is the bias toward evaluating the effectiveness (impact) of PVE programs instead of conducting evaluations that are more comprehensive. Throughout this introduction, and especially in the preceding section, the primary focus has been on what we know about the effects of these programs, with very little focus on how they are applied, in what contexts they are the most effective, and what mechanisms underlying their operation can be mobilized to improve their performance. Among all the reviews of PVE evaluations that we have

been discussing, only Gielen (2017) attempted to go beyond this “what works?” approach and instead take a “realist” evaluation approach, examining what works, for whom, in what circumstances, and how.

This bias toward evaluating effectiveness or impact has led to an ongoing quest for rigorous quantitative methods, which are thought to be an indicator of evaluations that serve this purpose well. A good example is the use of the Maryland Scientific Methods Scale (MSMS) by Bellasio et al. (2018) to measure the quality of studies as a criterion for inclusion in their review—a tool that applies only to quantitative designs. This bias may be due to the newness of evaluation in this field, at a time when the effectiveness of these programs was probably the first question that needed to be answered. But as we shall see later in this review, this discipline is now mature enough to tackle other kinds of evaluation questions.

A third concern that still needs to be addressed is the publication bias in past literature reviews. The grey literature on evaluating PVE has proven invaluable and was considered in many of the reviews that we have just discussed. But as will be seen in the present systematic review, their searches were not exhaustive. Actors outside of academia have taken far more initiative in addressing these shortcomings.

A fourth important issue is assessing the methodological quality of PVE evaluations. Among the reviews that we examined, only Carthy et al. (2020), Hassan, Brouillette-Alarie, Ousman, Kilinc et al. (2021) and Hassan, Brouillette-Alarie, Ousman, Savard et al. (2021) applied specific tools to score the quality of the studies’ methodology and used the resulting scores to decide whether to include or exclude them in their reviews. As noted, Bellasio et al. (2018) used the Maryland Scientific Methods Scale (MSMS) for this purpose, but this scale measures the quality of studies solely according to what type of design they use. By default, it deems studies that use randomized trials to be of high quality, regardless of any other considerations.

A fifth and final concern regarding past reviews is their lack of specificity. Except for the review by Carthy et al. (2020) on counter-narrative measures, and the reviews concerning certain programs in the United Kingdom (Christmann et al., 2012; Pratchett et al., 2010; Taylor and Soni, 2017), most past reviews have tended to address PVE program evaluation generically, without considering the variety of evaluation measures that have been applied and the distinctive features of the intervention contexts. Gielen (2017) reviewed evaluation studies of no fewer than eight different types of CVE interventions and programs. As far back as 2011, Neumann asserted that the variety of these measures could potentially be

unlimited. In settings such as educational institutions and correctional systems, and in geographic areas such as North America, Africa, South Asia and Southeast Asia, prevention measures have been adapted to their local contexts and hence require specific types of evaluations.

The situation regarding evaluation of programs to fight right-wing violent extremism is similar: very few such programs have been evaluated, and they too present specific evaluation challenges (Bellasio et al., 2018; Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021; Hirschi and Widmer, 2012). Taylor and Soni (2017) reviewed the literature on experiences with the Prevent Strategy in the educational system in the United Kingdom, but most of the studies that these authors reviewed did not deal with specific programs, according to our criteria. The systematic reviews of primary, secondary and tertiary prevention programs by Hassan's research teams (Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021; Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021)

and the study by Zeuthen (2021) marked an advance in this regard, in that they focused on specific types of prevention. But the need to evaluate the effectiveness of these programs in specific settings remains a major issue in the field.

All of the preceding issues led us to conclude that it would be worthwhile to conduct a new systematic review of PVE evaluation studies, incorporating new knowledge and new ways of addressing this research challenge, in particular by searching more of the grey literature, assessing the quality of the evaluation methods that these studies used, and, of course, reviewing new empirical data and new studies that had not been considered in past reviews. Although at least three of those reviews (Bellasio et al., 2018; Feddes and Gallucci, 2015; and Mastroe and Szmania, 2016) focused on evaluation methods, they still had many of the shortcomings described above, which we have attempted to rectify in this new systematic review.

02



Methodology of this systematic review

The methodology that we used to conduct this systematic review is based on the review methods of the Campbell Collaboration.¹⁵ We adopted their definition of a systematic review as “a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyze data from the studies that are included in the review” (Moher et al., 2009, p. 1).

2.1. OBJECTIVES

The overall objective of this systematic review was to inventory all evaluations of programs for prevention of violent extremism (PVE) as reported in publications from 2001 through 2019. In addition to this overall objective, we had the following specific objectives:

1. Identify the methodologies used in evaluations of PVE programs
2. Identify the shortcomings in the literature on evaluation of such programs
3. Assess the methodological quality of the existing evaluation studies in this field
4. Make recommendations for the evaluation of PVE programs.

¹⁵ <https://www.campbellcollaboration.org>. Appendix B provides a more complete description of this methodology.

2.2. RESEARCH QUESTION

Our main research question was therefore, “On the basis of the literature, what are the main recommendations that can be made regarding evaluation of programs for prevention of violent extremism?”

2.3. INCLUSION AND EXCLUSION CRITERIA

This review targetted all studies published from 2001 through 2019 in which primary, evidence-based data were used to evaluate PVE programs.¹⁶ The purpose of such programs is to reduce or eliminate the risk conditions that may make an individual or group more vulnerable to radicalization and violent extremism or to recidivism.¹⁷ We included all studies whose purpose was to assess or judge a PVE program, project or strategy, even if they did not use the term “evaluation” explicitly. We did not use the target populations of these studies as an inclusion criterion. We thus targetted all evaluations of primary, secondary and tertiary PVE programs¹⁸ that attempted to change the attitudes, emotions or behaviours of individuals or groups; of their families, friends and acquaintances; and of practitioners who work in this field. We excluded evaluations of programs that work with direct or indirect victims,¹⁹ evaluations of counterterrorism measures, and studies that evaluated continent-wide strategies or provided overall assessments of a continent-wide approach.

Because one publication can discuss more than one study, the unit of analysis for this review was the individual published study rather than the publication. We regarded a publication as discussing more than one study if it a) discussed more than one sample that had been analyzed independently and b) presented independent results for each sample.

To be included in this review, the studies also had to have been written in English, French or Spanish (the languages read and spoken by the members of the research team).

2.4 VARIABLES CODÉES

Each study included in this review was coded according to a global coding frame composed of variables grouped under the following 20 main dimensions:²⁰

- General description of study
- Author(s) of study
- Prevention level
- Type of violent extremism targetted
- Evaluation type (impact, process, output, etc.)
- Evaluator type
- Methodological design according to
 - overall approach
 - manipulation of variables
 - program participants
 - number of observations
 - number of times observations taken
 - number of independent variables
 - number of dependent variables
- Data-collection tools
- Scope of intervention evaluated
- Sample
- Target population
- Target setting
- Type of indicators used or results obtained
- Types of effects
- Limitations of the study

To assess the methodological quality of the studies included in this review, we used the Mixed Methods Appraisal Tool (MMAT) (Hong, Pluye, Fàbregues, Bartlett, Boardman et al., 2018; Hong and Pluye, 2019). Unlike other evaluation tools, the MMAT can be used to evaluate all of the different kinds of studies that we included in this review: qualitative, quantitative descriptive, experimental, quasi-experimental and mixed designs. The MMAT consists of 25 variables divided into five groups representing the five kinds of studies just mentioned. This tool is used to assign each study a quality rating on a scale of 0 to 5.

¹⁶ Secondary data are data collected by someone other than the studies' authors or their teams. Examples of secondary-data sources in the social sciences include population censuses, data collected by government departments, organizational records, and other data that were originally collected for purposes other than the research in question.

¹⁷ See key definitions in Appendix B.

¹⁸ See key definitions in Appendix B.

¹⁹ The families of the individuals who engaged in this process may be regarded as indirect victims of extremist groups. But here we understand “victims” to mean individuals and their families who were the target of attacks, attempted attacks or other violent acts by extremist groups.

²⁰ A complete list of the definitions of the dimensions and variables included is presented in Appendix B.

2.5. SEARCH STRATEGY

Using the inclusion criteria, exclusion criteria and keywords that we had identified (see sections B2 and B5), we searched the following three bodies of material:

- scientific literature
- grey literature
- other sources.

2.5.1 Scientific literature

To search the scientific literature, we had a librarian with expertise in the social sciences and humanities apply our search criteria to 21 databases that contained not only published scientific articles and academic theses, but also a large volume of grey literature and conference papers. We also obtained access to the database used in a recent systematic review by the Canadian Practitioners Network for the Prevention of Radicalization and Extremist Violence (CPN-PREV) (Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021; Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021), and merged this database with the 21 others.

2.5.2 Grey literature

To reduce “publication bias” (Rothstein, Sutton and Borenstein, 2005) in our strategic review, we used Google to conduct an in-depth search of the grey literature. To identify additional documents, we also manually examined 228 websites of organizations involved in PVE, which we selected from the [UNESCO-PREV Chair’s map of centres of expertise in PVE](#). We also added other organizations in the course of this search. Table 32 provides a complete list of the selected organizations.

2.5.3. Other sources

In addition to identifying documents through the two searches just described, we compared our findings with other reviews that have been frequently cited in the literature (see Table 1). We also consulted 14 experts by e-mail to find out whether they knew of any other evaluation studies of PVE programs.

2.6 PROCEDURE

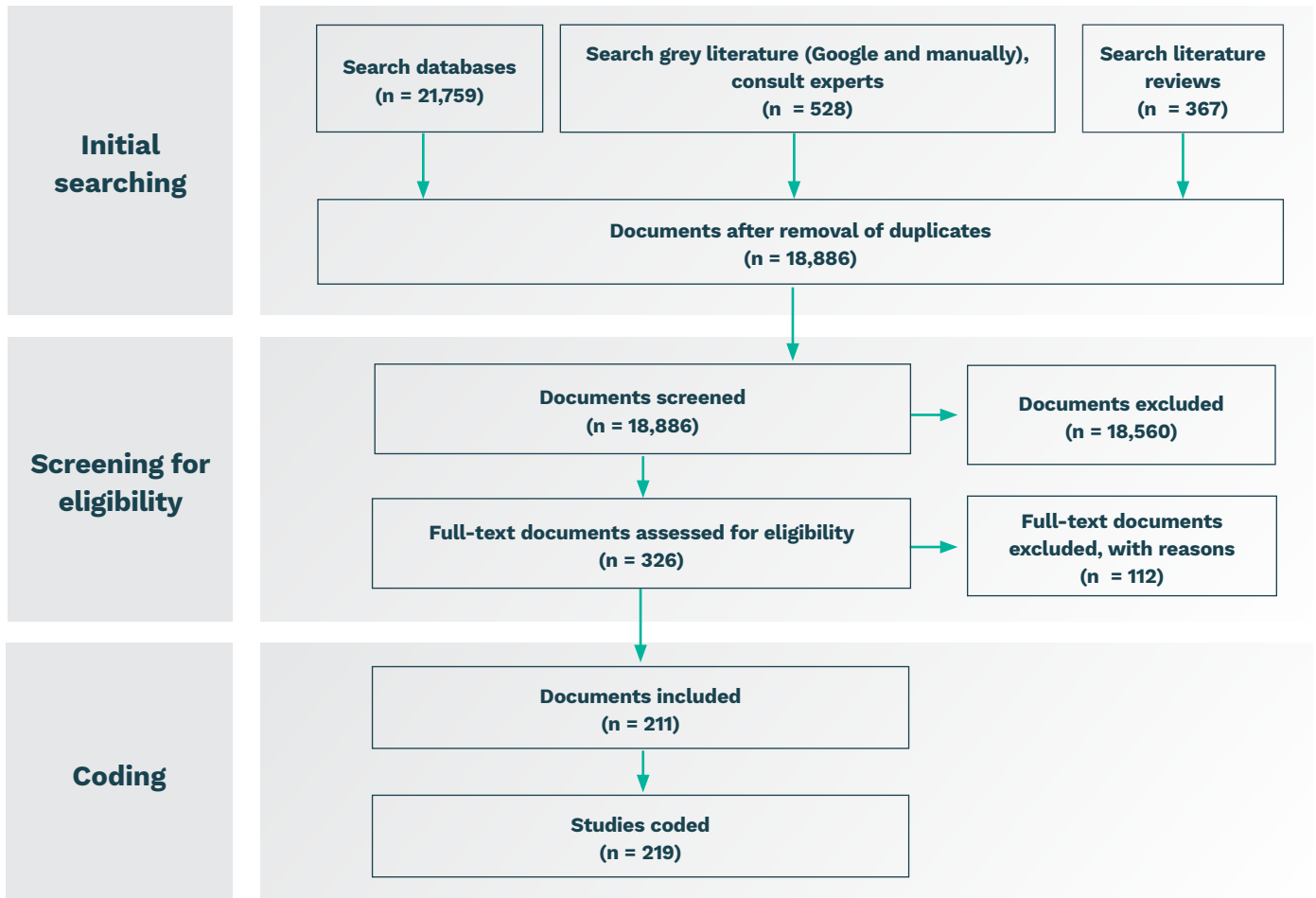
Before starting this systematic review, we trained the five research assistants who were working with us, to clarify the concepts and work methodology. To search the scientific literature, we then used two bibliographic databases. One of them came from the similar systematic review done recently by the CPN-PREV team (Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021; Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021), with which our review had certain keywords in common. This database covered all existing publications to January 2018. A new bibliographic review was conducted using the criteria and the 21 databases mentioned previously. While our librarian was collecting the scientific documents, the research assistants reviewed the grey literature on the websites of the organizations mentioned above. Once collection of data from the grey literature had been completed, the databases were merged and any duplicates were eliminated. Also, the 14 experts were contacted during this period.

To eliminate any ineligible studies, the principal investigator and the research assistants screened the titles and abstracts of all of the documents identified in the above searches. During this first phase, to ensure consistency, all team members coded the first 700 documents, analyzing and resolving any disagreements about how to code them. This phase also served as training for the team. Next, two coders reviewed each document. To ensure that there was sufficient agreement between the two coders, a Cohen’s kappa coefficient was calculated. During this initial coding, we worked iteratively: each pair of coders worked on a limited number of items. Then Cohen’s kappa was calculated. If its value fell below the minimum acceptable threshold of 0.6, the two coders reviewed their points of disagreement; if it was 0.6 or higher, they continued coding the next set of documents. The final kappa was 0.86.

The total number of publications selected was 211, but some publications discussed more than one study, so the total number of studies included in our systematic review was 219. (We regarded a publication as discussing more than one study if it discussed more than one sample that had been analyzed independently.)

We used the PRISMA model (<http://www.prisma-statement.org>) to record the results of our searches in the flow chart shown in Figure 1.

Figure 1. Prisma flow chart



2.7 LIMITATIONS

This systematic review involved certain limitations in its data collection and analysis that must be taken into account.

One of these limitations is a publication bias with regard to the languages in which the evaluation studies were written. A large share of the evaluation studies in this field—especially those dealing with programs to prevent right-wing violent extremism—are written in languages excluded from this review, such as German, Dutch, and the Scandinavian languages. We therefore did not have access to a significant number of these studies, which especially biases our results concerning evaluation of programs addressing this type of extremism.

Two other major limitations of this systematic review relate to our assessment of the methodological quality of the studies reviewed.

First, to make this assessment, we had to rely mainly on the information available in the publications themselves, many of which may have had to omit a great deal of

relevant information because of space constraints, especially in scientific journals and other academic publications. Hence our ability to analyze their quality was limited.

Second, the recommended procedure for the first step of assessing each study as a whole and its methodological quality in particular is to have two assessors perform these tasks blinded from each other. However, mainly because of the large number of publications included and the time and resource constraints that our team faced, we had to have just one person assess the methodological quality of each study. In fact, various persons assessed and coded various studies. (We did, however, have two people screen each study to determine its eligibility for inclusion in this review.) Note also that even though we used an excellent tool with detailed criteria to assess the studies' methodological quality, this task always involves an element of subjectivity that cannot be ignored.

03



Findings about the studies reviewed

3.1 STATISTICS ON THE STUDIES REVIEWED

Figure 1 in the preceding section is a Prisma flowchart showing the process by which we screened and selected the studies included in this review. Out of an original database of 18,886 scientific documents, we ended up including a total of **219 studies, discussed in a total of 211 publications**. The reasons for the discrepancy between these two numbers is that 13 of the publications discussed more than one study (for a total of 36), while 11 of the studies were discussed in more than one publication: for example, the report from the University of Amsterdam (2013) and the article by Feddes and Gallucci (2015) were based on the same study.²¹ The unit of analysis for this review was the individual study, so from here on we focus on the 219 studies that we included in it.

3.1.1. Number of studies by year

As both Gielen (2017) and Bellasio et al. (2018) observed, **2016 marked a turning point at which the number of evaluation studies published each year** began to rise. Prior to 2016, the number of such studies published annually averaged 9; in 2016, it jumped to 30 (Figure 2). This finding shows that despite the obstacles involved in evaluating interventions to prevent violent extremism, real efforts have been made to overcome them. An evaluation culture seems to have begun to develop among the actors in this field.

Figure 2. Number of evaluation studies, by year



²¹ We merged the information and considered only the more complete document in the final number of studies. When we found contradictions between the two documents, we gave precedence to the information in the more complete one.

3.1.2. Number of studies by continent and country

Most past literature reviews of PVE evaluation studies have found that a considerable number are conducted in European countries, and the present review found the same pattern (Figure 3). Of the 219 evaluation studies included in this review, nearly half dealt with European programs, and 22.8% were conducted in the United Kingdom, which was the pioneer in studies of this kind. Evaluations of various programs were also conducted in the Netherlands (n = 12), Germany (n = 5), Denmark (n = 5), France (n = 4) and Switzerland (n = 4).²² It should be remembered, however, that our review included only evaluation studies published in English, French or Spanish, and that the number of studies might actually be much higher if we counted studies published in other languages.

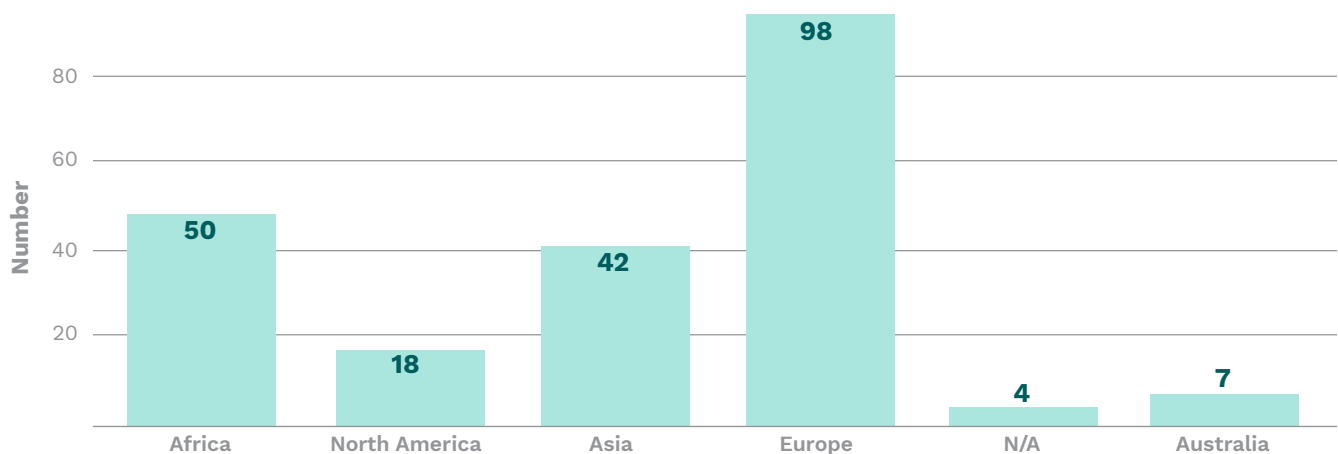
Besides finding the continued prominence of Europe in evaluation studies, this systematic review identified many evaluation studies that were conducted on other continents where violent extremism is a concern and that had not been identified in past reviews, including 50 from Africa and 42 from Asia. Within Africa, the countries where the greatest number of evaluation studies were conducted were Kenya (n = 9), Somalia (n = 6), Niger (n = 5) and Nigeria (n = 4). Within Asia, they were Indonesia (n = 16), Pakistan (n = 9) and the Philippines (n = 4). Notably, Indonesia was the country that accounted for the second largest number of evaluation studies in this systematic review, after the United Kingdom.

Another point to note is that 31 of the 50 studies done in Africa and 29 of the 42 done in Asia were published in

English, even though it was not the official language of the country in question. Almost half of the studies from Africa evaluated programs in countries where French is either an official language or a working language, but none of these studies was written in French. Having instead been published in English obviously gives these studies an advantage for the wider dissemination of their findings and for scientific exchanges. But it also poses an obstacle for many actors and practitioners in the field who cannot necessarily read English. The reason that these studies—especially those done in non-Western countries—were written in English may be that they were designed to meet the needs of the programs' funders, rather than to provide feedback to help the people who actually design these programs and deliver them in the field improve their practices. But that is only a hypothesis, because we have no way of determining whether any other methods of mobilizing knowledge were used to provide feedback to these people.

Surprisingly, although North America has a long tradition of evaluation in PVE and related fields, the United States and Canada accounted for very few of the studies in this review, although the United States did show a significant increase from 2016 on. A total of 15 studies from the United States were included in this review, which is a fair number. On the other hand, only three evaluations of PVE programs in Canada had been published as of 2019. But the CPN-PREV network's field study of secondary and tertiary prevention programs indicated that of the 26 tertiary programs identified in Canada, at least five had been or were going to be evaluated (Hassan, Ousman et al., 2020) (Table 2).

Figure 3. Nombre d'études par continent²³



²² This information is based on a single publication (Hirschi and Widmer, 2012) that was divided into seven different studies, four of which were included in this review.

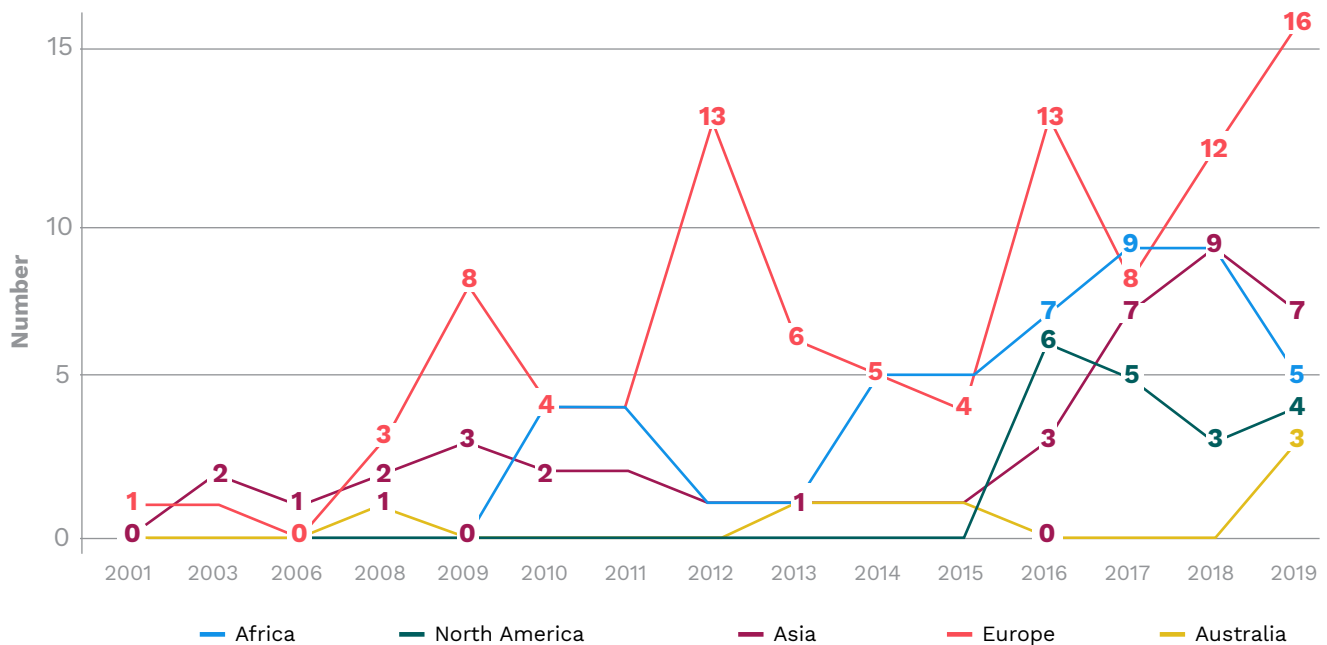
²³ N/A: studies of program (particularly online programs) that did not target a specific country or region. For a detailed description of the evaluations of online programs, see the section on case studies [page 85](#).

Table 2. 15 countries with the most studies in this review

Country	n	%	Country	n	%	Country	n	%
United Kingdom ²⁴	50	22.6	Pakistan	9	4.1	Niger	4	2.3
Indonesia	16	7.2	Australia	7	3.2	France	4	1.8
United States	15	6.8	Somalia	6	2.7	Nigeria	4	1.8
Netherlands	12	5.4	Germany	5	2.3	Philippines	4	1.8
Kenya	9	4.1	Denmark	5	2.3	Switzerland	4	1.8

A look at the year-to-year changes in the number of studies on each continent helps to explain the rise in the total number of studies since 2016. The number in Europe has risen and fallen but trended upward since 2001, while the number in Africa began trending upward in 2014. The numbers of studies in Asia and North America began rising considerably in 2016 (Figure 4).

Figure 4. Year-to-year changes in number of evaluation studies, by continent



3.1.3. Number of publications in academic literature and grey literature

A large majority of the studies included in this systematic review appeared in non-academic publications, which shows that interest in evaluation is not confined to academia and that many non-university institutions are involved in this activity (Table 3). But the academic literature has grown linearly since 2016 (Figure 5), which may be explained by the growing interest in quantitative approaches and experimental and quasi-experimental designs in recent years (see sections 3.3.3 and 3.3.4).

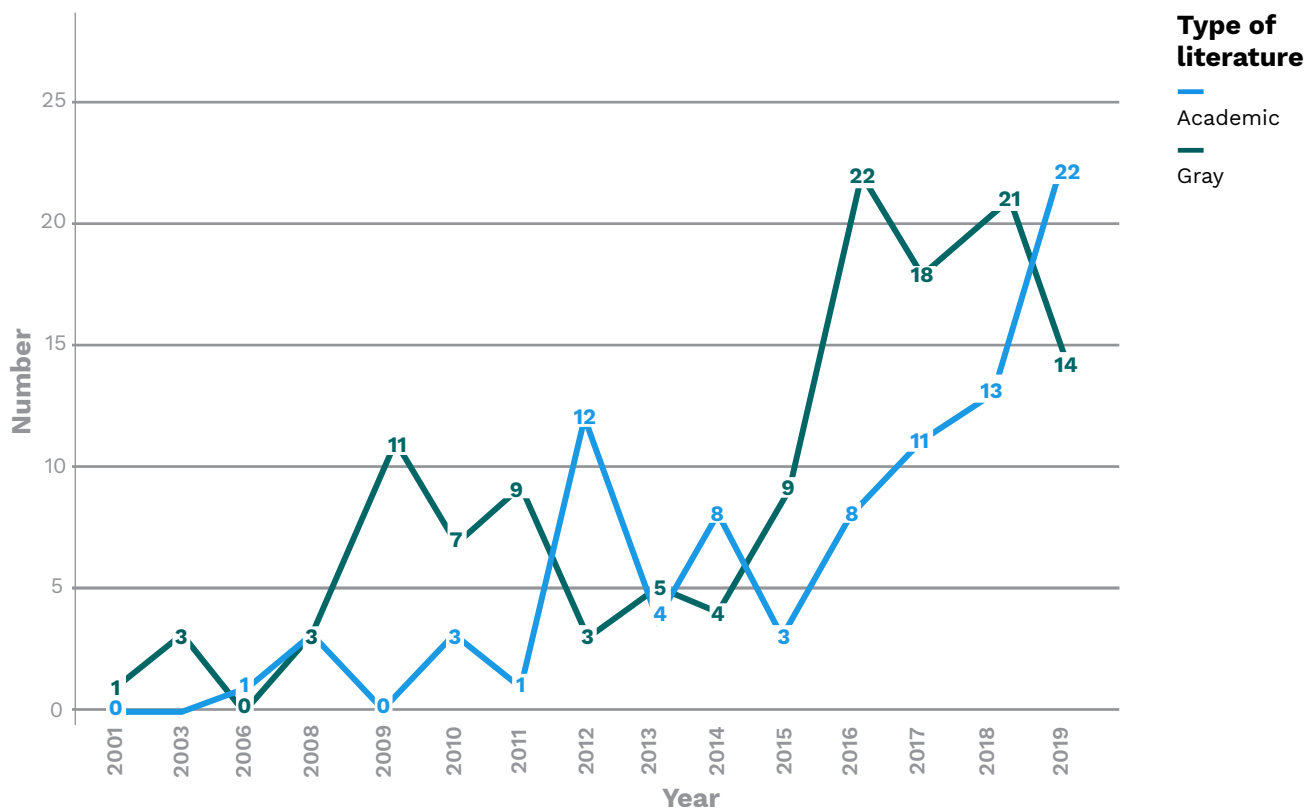
²⁴ The United Kingdom has four constituent countries: England, Scotland, Wales and Northern Ireland. Some of the studies in this review that were done in the United Kingdom indicated which of these four countries they were done in, while others did not. In this review, to avoid confusion, we classify all such studies as having been done in the United Kingdom.

Table 3. Publications in academic literature and grey literature, by continent²⁵

Continent	Type of literature		Total
	Academic	Gray	
TOTAL	89	130	219
Africa	22 % (11)	78 % (39)	100.0 % (50)
North America	27.8 % (5)	72.2 % (13)	100.0 % (18)
Asia	42.9 % (18)	57.1 % (24)	100.0 % (43)
Europe	47.5 % (47)	52 % (51)	100.0 % (99)
n/a	25 % (1)	75 % (3)	100.0 % (4)
Australia	100.0 % (7)	0.0 % (0)	100.0 % (7)

But the situation in this regard does vary from one continent to another. In Africa and North America, the overwhelming majority of the publications in this review (around three-quarters) came from the grey literature, including evaluations done for the United States Agency for International Development (USAID) or by non-governmental organizations such as Search for Common Ground. On the other hand, all the studies from Australia that were included in this review were published in traditional scientific journals, while those from Europe and Asia showed a more even balance between the academic and grey literatures.

Figure 5. Publications in academic and grey literature, by year



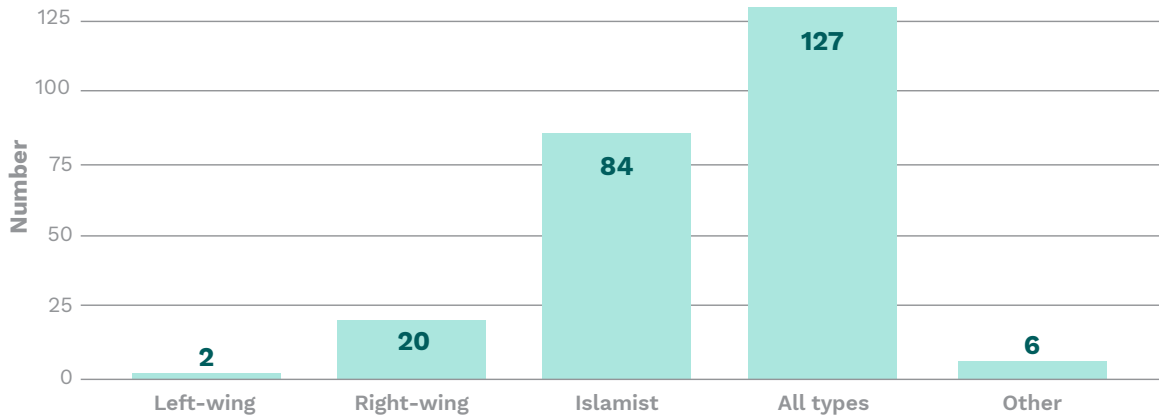
²⁵ Unless otherwise indicated, the percentages in the tables in this report are calculated across rows. Because some of the studies included in this review had none or more than one of the characteristics considered in any given table, the percentages in any given row may not always total 100%.

3.1.4. Number of evaluated programs, by type of extremism targetted

The majority of the programs evaluated in the studies in this review (n = 127) targetted all types of violent extremism rather than any one type in particular. Programs

that specifically targetted violent extremism associated with Islamism came second (n = 84), and those targetting right-wing violent extremism came third (n = 20).

Figure 6. Number of evaluated programs, by type of extremism targetted²⁶



The number of evaluated programs that targetted all types of extremism rose sharply starting in 2016 (Figure 7). Before then, the number of these programs averaged only 4 per year; since then, they have averaged 22 per year. Such programs accounted for 68% of all programs evaluated in Africa, 61.1% of those in North America and 60.2% of those in Europe.

The year-to-year trends for evaluations of such programs, as measured by number of publications, differed by continent. In Europe, this number began rising in 2014 and increased considerably in 2018 (n = 12) and 2019 (n = 13). In Africa, this number began increasing slightly in 2015 (n = 4) but has remained stable in recent years. In North America and Asia, a slight increase was seen starting in 2016 and 2017, respectively (n = 3).

One reason for this overall increase in programs that target all types of extremism, rather than any specific type, was a growing awareness of the effects of falsely associating Islam with terrorism and the potential for programs targetting Islamist extremism to stigmatize the Muslim community. In Europe, the negative evaluation of the United Kingdom's national Prevent strategy was one of the best known drivers behind this shift in approaches to preventing violent extremism. The first Prevent strategy (2007–2011) directly targetted the Muslim community and was regarded as a major contributor to the stigmatization of that community in the United Kingdom (Busher et al., 2019; Kundnani, 2012; Romaniuk, 2015). As a result of this negative evaluation, this national strategy was broadened in 2011 so as to address all forms of extremism (Busher et al., 2019). Many other countries, including Canada, took

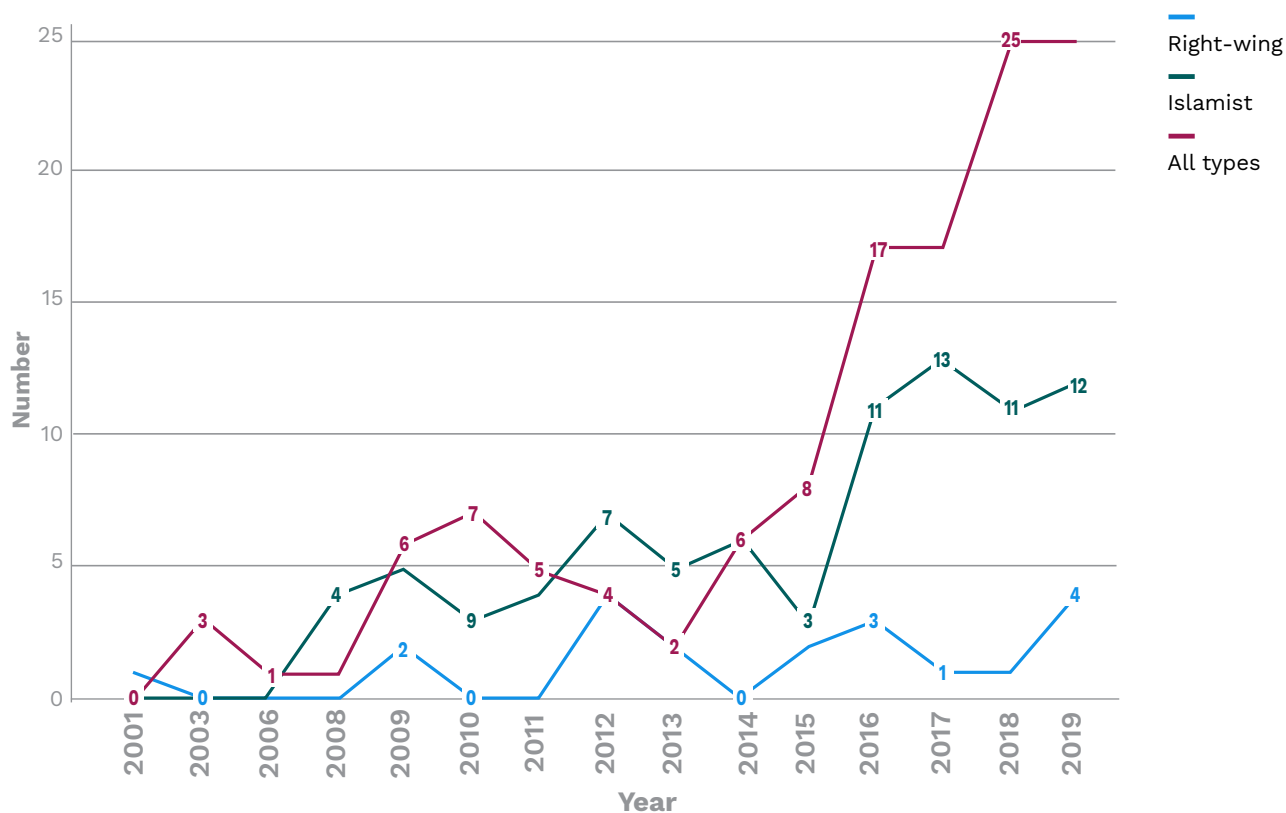
note of this development and decided to adopt more generalized intervention approaches themselves.

In Africa and Asia, many practitioners also are reluctant to use concepts such as radicalization or prevention of so-called Islamist violent extremism, for fairly similar reasons, as well as because of some special sensitivities due to the place of the Muslim religion in some of these societies (Madriaza et al., 2017). The use of such concepts can thus undermine trust between practitioners and the community and keep practitioners from carrying out their interventions (Madriaza et al., 2017). In such societies, prevention of violent extremism often takes the form of primary prevention programs, such as education and employability initiatives. Hence many practitioners regard the use of concepts such as radicalization or prevention of Islamist violent extremism as not very helpful or appropriate (Madriaza et al., 2017).

Notwithstanding the importance of this change in the approach to preventing violent extremism and evaluating PVE programs, the fact remains that they still usually target Islamist extremism. Apart from Western and Latin American countries where right-wing and left-wing extremism also exist, many of the countries whose programs fell into this “all-types” category in our review, particularly in Africa, were actually dealing with Islamist extremism only. In other words, if we consider the actual targets of the prevention programs and even more so of the evaluations, the vast majority of the evaluation studies in the literature today still deal mainly with “jihadist” or Islamist extremism.

²⁶ We coded this category on the basis of the information in the publications. When the publication specifically identified a particular type of extremism, we coded that type of extremism. When the publication did not identify any particular type of extremism or indicated that the program targetted radicalization or extremism in general, we coded it as “All types”. The numbers in the chart total more than 219 because several of the programs targetted more than one type of extremism.

Figure 7. Year-to-year changes in number of programs, by type of extremism targetted



These data can be confirmed by analyzing the year-to-year changes in the number of evaluation studies of programs that directly target Islamist violent extremism. In Europe, this number has been decreasing since 2016. In Africa and Asia, it has been relatively stable in recent years and continues to account for only a small percentage of the combined total for all types of PVE programs.

Second, there has been a moderate increase in the number of evaluation studies of programs targetting right-wing extremism, especially from 2015 to 2019, when 8 out of the 10 studies of this kind came from Europe (Figure 7). This is all the more important in that past literature reviews found only a small number of studies of programs targetting right-wing extremism and regarded this as a major shortcoming in the field (Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021; Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021; Hirschi and Widmer, 2012; Widmer et al., 2007).

Third, there have been almost no evaluation studies that specifically mention left-wing violent extremism; the two identified come from Europe.

Fourth and finally, although there is much public discussion about the potential role of social networks in preventing radicalization and violent extremism, very few online prevention programs—16 in total—have been the subject of evaluations (see section 3.6.2).

3.1.5. Number of studies by program prevention level

In this systematic review, we used a public-health model and the same classification system as Hassan, Brouillette-Alarie, Ousman, Kilinc et al. (2021) and Hassan, Brouillette-Alarie, Ousman, Savard et al. (2021) to classify the programs evaluated in the included studies as follows (Figure 8):

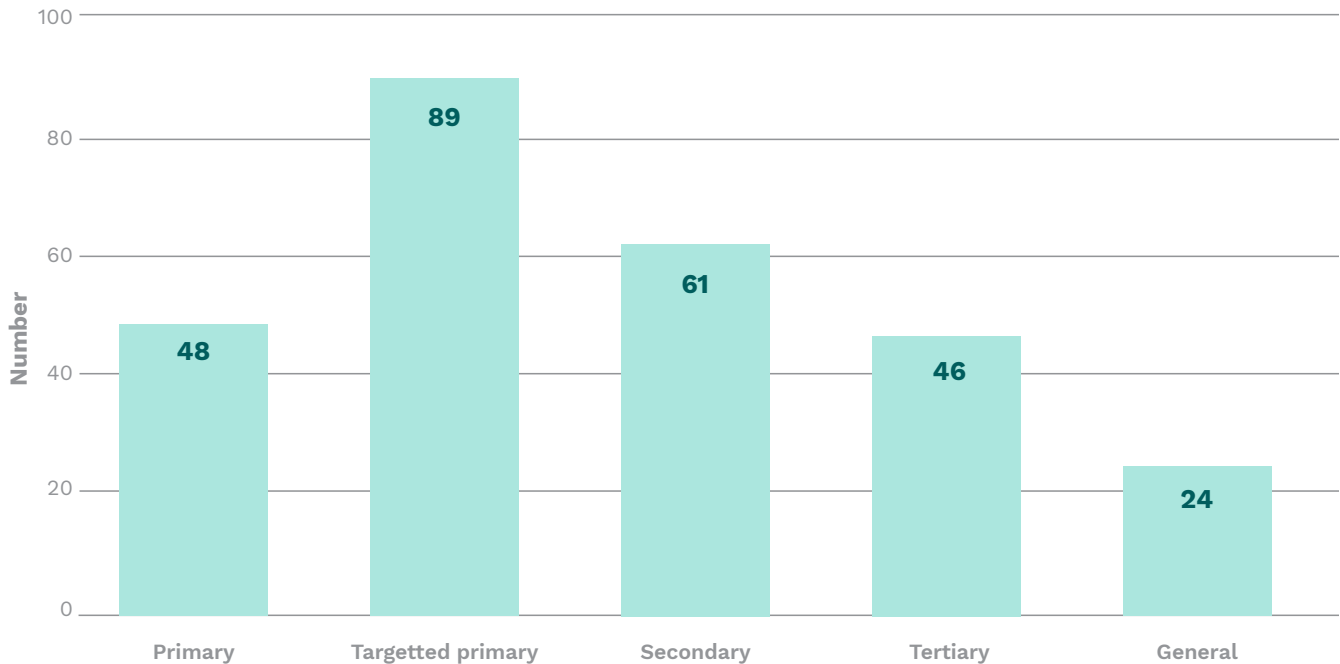
- Primary prevention programs (n = 48), which are universal and target the general population not identified as at risk;
- Targetted primary prevention programs (n = 89), which, though universal, target a specific population segment or community that is not considered at risk (for example, youth, or a Muslim community);

- Secondary prevention programs (n = 61), which target individuals or groups regarded as at risk or in the initial stages of the process of radicalization to violence;
- Tertiary prevention programs (n = 46), which target individuals or groups already engaged in the final phases of the process of radicalization to violence, or that belong to violent extremist groups or have committed acts associated with violent extremism;

- General prevention programs (n = 24), where the evaluation studies did not specify whether the programs operated at any of the preceding levels of prevention.

Thus the great majority of the programs evaluated were primary and targeted primary prevention programs.

Figure 8. Number of studies, by prevention levels of evaluated programs



As would be expected, the year-to-year changes in number of evaluated programs for each prevention level (Figure 9) generally follow the same pattern as for all evaluated programs combined: a continuous upward trend since 2016. But there are some differences from one prevention level to another. The upward trend is more obvious for primary prevention programs (except in 2019), targeted primary prevention programs and secondary prevention programs. As regards primary and targeted primary prevention programs, part of the reason for the upswing is the same as for the trend in types of extremism targeted—the shift toward a less specific, more generalized prevention approach. In contrast, the number of tertiary prevention programs evaluated has fallen sharply since 2016. Tertiary programs are usually delivered in correctional settings or through the probation system. As discussed in section 1.2, the difficulties in

evaluating such programs include the small number of cases involved and the difficulties in accessing data and participants in these settings, which might explain this decrease. But this is only a hypothesis, which may seem all the more surprising in that this decrease has occurred during a period marked by the gradual return of foreign fighters from the Syrian conflict zone. It should be remembered that this decline is in the number of programs evaluated, not in the number of programs implemented. Thus what we are seeing is a decreasing tendency to evaluate tertiary programs rather than a decrease in the number of tertiary programs, which could be evaluated at a later date.

Thus the great majority of the programs evaluated were primary and targeted primary prevention programs.

Figure 9. Year-to-year changes in number of programs by prevention level

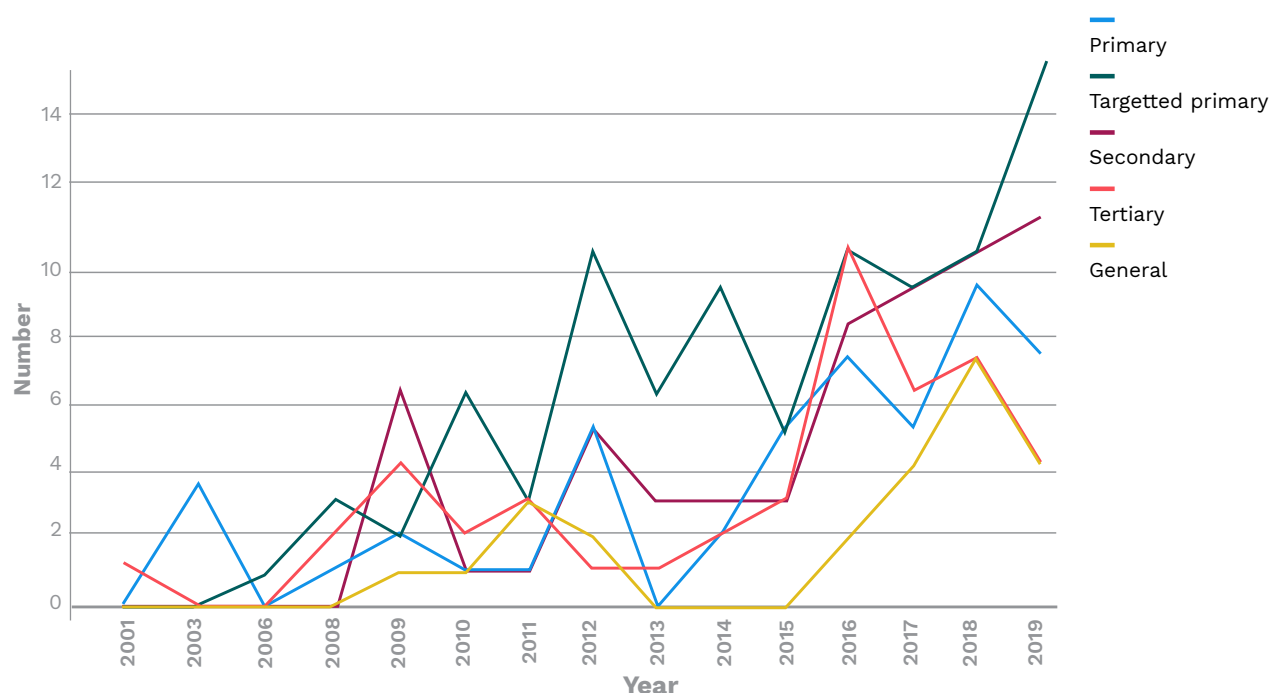


Table 4 shows the results for each prevention level by continent and by type of extremism targetted. The breakdown by continent thus shows each continent’s relative emphasis on the four levels of prevention. For example, targetted primary prevention programs account for a higher proportion of the total in Western cultures than in other parts of the world: the figures for North America, Europe, and Australia are 44.4%, 53.1%, and 66.7%, respectively. In Europe, secondary programs account for 32.7% and tertiary programs for 26.5%. In other words, programs in Europe tend to focus on the populations regarded as at risk, possibly because, compared with programs in other parts of the world, they predominantly attribute radicalization and violent extremism more to individual factors than to social ones.

Table 4. Percentage of studies by program prevention level, continent and type of extremism targetted²⁷

		Primary		Targetted primary		Secondary		Tertiary		General	
		n	%	n	%	n	%	n	%	n	%
Total		48	17.8%	89	33%	61	22.6%	48	17.8%	24	8.9%
Continent	Africa	20	40.0%	12	24.0%	13	26.0%	5	10.0%	8	16.0%
	North America	4	22.2%	8	44.4%	3	16.7%	1	5.6%	5	27.8%
	Asia	6	14.3%	13	31%	9	21.4%	11	26.2%	5	11.9%
	Europe	17	17.3%	52	53.1%	32	32.7%	27	26.5%	6	6.1%
	n/a	1	25.0%	0	0.0%	2	50.0%	2	50.0%	0	0.0%
	Australia	0	0.0%	4	66.7%	2	33.3%	1	16.7%	0	0.0%
Type of violent extremism	Right-wing	3	15.0%	5	25.0%	13	65.0%	8	40.0%	0	0.0%
	Islamist	11	13.1%	40	47.6%	25	29.8%	23	27.4%	6	7.1%
	All types	34	27%	48	38.1%	32	25.4%	18	14.3%	20	15.9%

²⁷ The sum of the numbers of studies may equal more than 219, because some studies were coded in more than one category. For the same reason, the sum of the percentages of the studies by continent and by type of extremism may exceed 100%.

In contrast, in Africa, primary prevention programs account for the highest number of evaluated programs. As mentioned earlier, other studies have shown that African prevention programs often focus on employability and education initiatives as a way of offering youth alternatives to recruitment and indoctrination by extremist groups that exploit socio-economic problems in their countries. (Madriaza et al., 2017). But the figures for Africa also show that programs on this continent tend to see violent extremism as a problem that affects all of society and that must be addressed in a more universal way. It should be remembered that a large share of these evaluated programs were funded and implemented by Western organizations and might reflect a more Western interpretation of the phenomenon of violent extremism in Africa.

The evaluated programs that target all types of extremism tend to operate at the most universal prevention levels: primary prevention (27%) and targeted primary prevention (38.1%). In contrast, the evaluated programs that specifically target right-wing violent extremism do not seem to apply a primary, universal approach but instead focus on secondary and tertiary prevention. This finding is

consistent with other systematic reviews that have found very few primary or secondary prevention programs that target right-wing violent extremism (Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021). Most of the programs found in the literature that target right-wing violent extremism are tertiary prevention programs, many of which use the EXIT approach, especially in Europe (Bjørge and Horgan, 2009). Programs addressing Islamist violent extremism tend to take a targeted primary prevention approach, focussed chiefly on Muslim communities.

3.1.6. Number of studies by scope of interventions evaluated

In this systematic review, we also classified each evaluation study into one of three categories according to the scope of the intervention that it evaluated: an entire national strategy, a part of a national strategy, or an individual program or project (Table 5). Unsurprisingly, this last category accounted for the greatest proportion of the evaluation studies overall (61.8%), while those for parts of national strategies accounted for 29.1% and those for entire national strategies for only 9.1%.

Table 5. Number of studies by scope of interventions evaluated, continent, prevention level and type of extremism targeted

		Entire national strategy		Part of national strategy		Program or project	
		n	%	n	%	n	%
Total		20	9.1%	64	29.1%	136	61.8%
Continent	Africa	2	4.0%	11	22.0%	37	74.0%
	North America	2	11.1%	4	22.2%	12	66.7%
	Asia	1	2.4%	11	26.2%	30	71.4%
	Europe	15	15.5%	34	35.1%	48	49.5%
	n/a	0	0.0%	0	0.0%	4	100.0%
	Australia	0	0.0%	2	28.6%	5	71.4%
Prevention level	Primary	4	8.3%	13	27.1%	31	64.6%
	Targetted primary	12	13.5%	25	28.1%	52	58.4%
	Secondary	6	10.0%	16	26.7%	38	63.3%
	Tertiary	5	11.1%	13	28.9%	27	57.4%
	General	5	20.8%	7	29.2%	12	50.0%
Type of violent extremism	Right-wing	1	5.0%	4	20.0%	15	75.0%
	Islamist	10	12.0%	30	36.1%	43	51.8%
	All types	13	10.3%	30	23.8%	83	65.9%

Evaluations of programs or projects also consistently accounted for the highest proportion of evaluations for every continent, every prevention level, and every type of extremism. But this proportion was not equally high in all cases. For example, it was less than 50% in Europe, where evaluations of parts of national strategies accounted for 35.1% of the total and evaluations of entire national strategies for 15.5%. The reason for this more even balance among the three categories on this continent was the number of evaluations that dealt with the British national Prevent strategy in whole or in part. On the other continents, evaluations of entire national strategies or parts of them accounted for about one-third of the total, which may reflect a growing response around the world to the need for evidence on efforts to coordinate PVE at the national level.

Evaluations of programs or projects also accounted for a much higher percentage of evaluations of interventions addressing right-wing violent extremism than of those addressing either Islamist violent extremism or all types of violent extremism. For evaluations of national strategies, the pattern was reversed: they accounted for a lower percentage of evaluations of interventions addressing right-wing violent extremism and a higher percentage of evaluations of interventions addressing all types of violent extremism at all levels of prevention.

The preceding findings do not provide a complete picture of all PVE programs, projects and strategies deployed around the world. But they do show how much more effort has been devoted to evaluating national strategies for preventing Islamist violent extremism than for preventing right-wing violent extremism. The reason may be that public policies on the former have been subjected to extensive public debate about their potential harmful effects on Muslim communities, whereas preventing right-wing violent extremism seems to have become a public-policy concern only more recently.

3.1.7. Number of studies that reported their funding sources

Table 6 provides more information on those studies in which the evaluators reported the sources of their funding. Whether an evaluation study reports its funding source is, in our view, an essential criterion for transparency, because this information lets readers identify potential conflicts of interest, as well as the ethical issues that the evaluators may have faced. In the present systematic review, **fewer than half of all the studies (43%) mentioned their funding source.** By continent, the highest percentages were for Africa (52%) and North America (50%), while the lowest was for Asia (35.7%).

Table 6. Number of studies that reported their funding sources, by continent, prevention level, and type of extremism targetted

		Studies that reported their funding sources	
		n	%
Total		95	43%
Continent	Africa	26	52.0%
	North America	9	50.0%
	Asia	15	35.7%
	Europe	40	41.2%
	n/a	2	50.0%
	Australia	3	42.9%
Prevention level	Primary	27	56.3%
	Targetted primary	40	45.5%
	Secondary	24	39.3%
	Tertiary	16	34.8%
	General	15	62.5%
Type of violent extremism	Right-wing	10	50.0%
	Islamist	27	32.5%
	All types	59	46.5%

By prevention level, the more specific the evaluated programs, the less often the funding source was reported, and the more universal the evaluated programs, the more often. Thus the general prevention level was the one for which the proportion of studies mentioning their funding source was highest (62.5%), while the tertiary prevention level was the one for which this proportion was lowest (34.8%).

The proportions according to type of extremism targeted did not show such a clear pattern. The proportion of evaluations reporting their funding sources was higher for programs targeting right-wing violent extremism than for those targeting Islamist violent extremism.

3.2. STATISTICS ON THE STUDIES' AUTHORS

As discussed in the introduction to this review, many researchers consider it especially important to know what person or organization was responsible for any particular PVE evaluation study, especially given the complexities of this field (Horgan and Braddock, 2010; Marret et al., 2017; Mastroe and Szmania, 2016). For this reason, in addition to describing the studies themselves in the preceding pages, we will now describe their authors (see Table 7). We have identified 389 authors who contributed to the 211 publications included in this systematic review.²⁸ Most of these publications (62.6%) identified several co-authors, while 29.6% showed only one author and 7.8% did not identify any author but simply gave the name of the institutions that had published them.²⁹ **The vast majority of the authors (89.5%) have only one publication each in our database on the evaluated programs**, while 10.5% of the authors have two publications and 3.3% have three publications on this specific subject. PVE program evaluation would thus not seem to be a highly specialized field. In an analysis of all of the articles published between 2007 and 2016 in 9 leading journals on terrorism, Schuurman (2018) noted the low degree of specialization in this field: he found that 72.2% of the authors had contributed only one article to these journals, while 13.4% had contributed only two. From this perspective, the authors of the evaluation studies in the current review can be considered more specialized than those in Schuurman's: 14.7% of his authors had only one publication in the field of security studies as sole author and 45.1% as co-author. But our database was larger than Schuurman's, which may explain this difference.³⁰ The percentage, however, is still very low, and the field of PVE program evaluation is still not highly specialized.

Table 7. Authors with two or more publications on evaluation in PVE and related fields

Author	n	Country of origin
Martin Manby	6	United Kingdom
Adrian Cherney	4	Australia
Allard Rienk Feddes	3	Netherlands
Anne Speckhard	3	United States
Beza Tesfaye	3	United States
James Khalil	3	United Kingdom
Lasse Lindekilde	3	Denmark
Paul Thomas	3	United Kingdom
Steven E. Finkel	3	United States
Chris A. Belasco	3	United States
Anne Aly	2	Australia
Anthony Sarota	2	United States
Bart Schuurman	2	Netherlands
Bertjan Doosje	2	Netherlands
Daniel P Aldrich	2	United States
David Schanzer	2	United States
Elisabeth (Lily) Taylor	2	Australia
Emma Belton	2	Australia
Jean-Camille Kollmorgen	2	United States
Jeffrey Swedberg	2	United States
Joel Busher	2	United Kingdom
Michele Grossman	2	Australia
Moli Dow	2	United Kingdom
Saul Karnovsky	2	Australia
Tinka Veldhuis	2	Netherlands
Tufyal Choudhury	2	United Kingdom
Cooper Gatewood	2	United Kingdom
Iris Boyer	2	United Kingdom
Alex Elwick	2	United Kingdom
Lee Jerome	2	United Kingdom
Jose Liht	2	United Kingdom
Sara Savage	2	United Kingdom
Oren Ipp	2	Sweden
Ardian Shajkovci	2	United States
Michael Neureiter	2	United States
John McCauley	2	United States
Louis Reynolds	2	United Kingdom
Therese O'Toole	2	United Kingdom
Daniel Nilsson DeHanas	2	United Kingdom
Tariq Modood	2	United Kingdom
Elena Savoia	2	United States
Marcia A. Testa	2	United States

²⁸ We included 211 publications, which discussed a total of 219 studies.

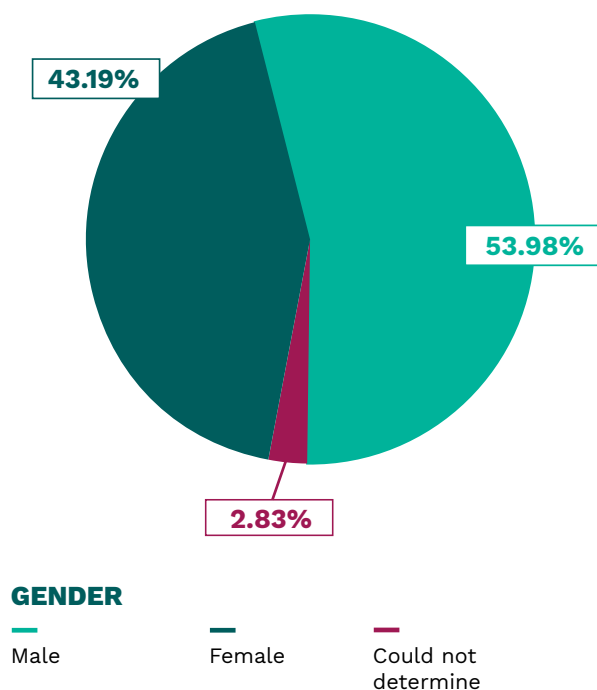
²⁹ One such publication was the 2008 report of the Audit Commission that evaluated the United Kingdom's national Prevent Strategy. The Audit Commission was an independent public corporation that was responsible for ensuring that public funds in the United Kingdom were spent economically, efficiently and effectively.

³⁰ We searched all of the authors' publications in fields such as radicalization, extremism, counterterrorism and security studies in the Quebec university system library database ("Sophia") and the first five pages of Google Scholar.

3.2.1. Gender

The majority of the studies' authors (54%) are male, but a fairly high proportion are female (43.2%); we were unable to determine the gender of the remaining authors. The gender balance varies fairly widely, however, from one continent to another. Among the authors from Africa, 78.9% are male and only 21.1% are female, whereas among authors from Asia, males account for 52.9% and females for 36.8%. Only Australia has a higher proportion of women than of men (59% versus 41%).

Figure 10. Percentage of evaluation studies' authors by gender



3.2.2. Geographic origin

As previously noted, many of the PVE programs evaluated in the studies in this systematic review took place in non-Western parts of the world, such as Africa and Asia, while only a small number took place in North America, and especially few in Canada. This pattern might be taken to mean that program evaluation has become democratized beyond the Western world, but when we analyze the number of the same authors' publications according to the authors' continents of origin (Table 8), the picture changes.

Table 8 shows, for the authors from each continent, the total and average number of publications in the field of security studies and the total and average number of publications included in this systematic review. About two-thirds of all the authors come from just two continents: Europe (39.1%) and North America (27.8%). The figure for Europe is unsurprising, given the large number of evaluations conducted in European countries. But the meaning of the figure for North America is less clear. Only 18 of the evaluation studies in this systematic review were conducted in North America, so we can deduce that the majority of the 108 North American authors (92 of whom were from the United States), were involved in evaluations on other continents. Our database includes only 3 evaluation studies that were done in Canada, but a total of 16 Canadian authors were involved in the 219 studies in this review. The situation is similar in Australia: only 7 of the evaluations in this review were conducted there, but 27 of the authors come from there.

Evaluation authors from North America are overrepresented, while evaluation authors from Africa are underrepresented.

In contrast, the number of African authors in this review—19—is far smaller than the number of evaluation studies conducted in Africa (50), which suggests that most of these evaluations were conducted by evaluators from elsewhere. The pattern in Asia is less dramatic and seems to reflect greater self-sufficiency in PVE program evaluation: our review includes 68 authors from Asia and 42 evaluations conducted in Asia.

Table 8. Numbers of publications by authors' continent of origin

Continent		Publications as sole author*	Publications as co-author *	Publications in the database**
Total	Mean	1.88	3.88	1.14
	Number of authors in our database	389	389	389
	Standard deviation	5.23	8.06	0.48
	Number of publications	732	1511	445
Africa	Mean	0.42	0.79	1.00
	Number of authors in our database	19	19	19
	Standard deviation	1.08	0.54	0.00
	Number of publications	8	15	19
North America	Mean	3.08	4.90	1.17
	Number of authors in our database	108	108	108
	Standard deviation	8.31	12.53	0.46
	Number of publications	333	529	126
Asia	Mean	1.09	2.51	1.00
	Number of authors in our database	68	68	68
	Standard deviation	3.77	5.00	0.00
	Number of publications	74	171	68
Europe	Mean	1.70	4.25	1.20
	Number of authors in our database	152	152	152
	Standard deviation	3.15	5.76	0.59
	Number of publications	258	646	182
Australia	Mean	2.19	5.07	1.30
	Number of authors in our database	27	27	27
	Standard deviation	4.14	6.42	0.67
	Number of publications	59	137	35

* Publications in the field of security studies.

** Publications included in this systematic review.

Out of the 10 countries with the greatest number of authors, 7 are Western countries, and nearly half of all the authors in this review come from just two countries: the United States (23.7%) and the United Kingdom (22.6%) (Table 9). This predominance of Western authors is consistent with the numbers of evaluations identified in this systematic review. The three non-Western countries that account for the greatest number of authors are

Indonesia, Pakistan and Kenya. Indonesia and Pakistan account for 42.6% and 25% of all the Asian authors, respectively. The Kenyan authors account for 52.6% of all of the African authors, followed by Moroccans (15.8%) and Nigerians (10.5%).

Table 9. Ten countries with the greatest number of authors in this systematic review

Country	n	%
United States	92	23.7%
United Kingdom	88	22.6%
Indonesia	29	7.5%
Australia	27	6.9%
Pakistan	17	4.4%
Canada	16	4.1%
Netherlands	16	4.1%
Kenya	10	2.6%
Belgium	7	1.8%
Italy	6	1.5%

The preceding analysis makes it clear that evaluation of programs to prevent violent extremism is a field that has been largely colonized by Western countries. The situation in Africa is especially striking. Many African PVE programs have been funded by the United States Agency for International Development (USAID), most likely implemented by U.S.-based organizations, and evaluated by U.S.-based evaluators and researchers who (as we shall see in section 3.4) often did not speak the language of the country concerned and wrote their evaluation reports in a different language as well.

3.2.3. Disciplines

As Table 10 shows, less than one-quarter (21.6%) of the authors of the evaluation publications in this review have political science as their primary discipline. This pattern holds for all continents and is fairly consistent with the trends observed in the broader field of terrorism studies (Schuurman, 2018) and, to a lesser extent, the study of violent extremism. Fairly far behind political science come, in descending order, the mental-health disciplines (11.1%), sociology and social work (9%), education (8%), security and peace studies (7.7%), criminology (6.7%) and public health and medicine (4.1%).

Some disciplines that might seem particularly relevant to PVE program evaluation have little or no representation in the above table. For example, only 2.6% of the authors specialize in communication (2.6%), a background that might be useful for assessing the impact of primary and secondary online and offline awareness campaigns. Another surprise is that only 0.8% of the authors have backgrounds in theology, a field often cited in secondary and tertiary prevention programs, especially those addressing Islamist violent extremism.

Table 10. Disciplines of the authors of the publications in this systematic review

Disciplines	n	%
Political Science	84	21.6%
Mental Health (Psychology/Psychiatry)	43	11.1%
Sociology/Social Work	35	9.0%
Education	31	8.0%
Security/Peace Studies	30	7.7%
Criminology/Police Science	26	6.7%
Public Health/Medicine	16	4.1%
Economics	13	3.3%
Communication/Literature	10	2.6%
Law	8	2.1%
Anthropology	6	1.5%
Administration / Management	6	1.5%
Demography/Geography	3	0.8%
History	3	0.8%
Theology	3	0.8%
Philosophy	3	0.8%

Although the above pattern generally applies throughout the world, some differences from continent to continent are worth noting (Table 11). Political scientists are well represented everywhere, but sociologists and social workers are not. They account for high percentages of the authors from Africa (25%) and Europe (16.8%), but much lower percentages of those from North America (6.6%), Asia (2%) and Australia (3.8%). Public health and medicine account for a fairly high percentage of the authors from North America only (11%), and for very low percentages of the authors from other continents (Africa, 0%; Asia, 4.1%; Europe, 2.9%; Australia, 0%). Similarly, in the mental-health disciplines, we find no authors from Africa and only one from Australia (3.8%); the percentages are much higher for North America (16.5%), Asia (14.3%), and Europe (14.6%). Even the combined figures for public health and medicine and mental health seem rather low overall, given the importance of these disciplines in secondary and tertiary PVE initiatives.

The field of education accounts for a high proportion of the authors from Australia (26.9%), lower percentages of those from Europe (10.2%) and Asia (12.2%), and much lower percentages for North America (4.4%) and Africa (0%). This seems surprising, in light of the many primary and secondary prevention programs delivered by the education sector, and suggests that the evaluations of these programs are being done by members of other disciplines.

Criminology accounts for 26.9% of the authors from Australia, 11% from North America and 6.6% from Europe; none of the authors from Africa or Asia are criminologists.

Table 11. Numbers and percentages of authors by discipline and continent

Discipline	Continent					
	Africa	North America	Asia	Europe	Australia	
Anthropology	n	0	1	0	4	1
	%	0.0%	1.1%	0.0%	2.9%	3.8%
Criminology/Police Science	n	0	10	0	9	7
	%	0.0%	11.0%	0.0%	6.6%	26.9%
Demography/Geography	n	0	1	0	1	1
	%	0.0%	1.1%	0.0%	0.7%	3.8%
Economics	n	2	0	9	2	0
	%	12.5%	0.0%	18.4%	1.5%	0.0%
Education	n	0	4	6	14	7
	%	0.0%	4.4%	12.2%	10.2%	26.9%
History	n	0	0	1	2	0
	%	0.0%	0.0%	2.0%	1.5%	0.0%
Mental Health (Psychology/Psychiatry)	n	0	15	7	20	1
	%	0.0%	16.5%	14.3%	14.6%	3.8%
Public Health/Medicine	n	0	10	2	4	0
	%	0.0%	11.0%	4.1%	2.9%	0.0%
Law	n	1	1	1	5	0
	%	6.3%	1.1%	2.0%	3.6%	0.0%
Political Science	n	4	25	12	39	4
	%	25.0%	27.5%	24.5%	28.5%	15.4%
Sociology/Social Work	n	4	6	1	23	1
	%	25.0%	6.6%	2.0%	16.8%	3.8%
Security/Peace Studies	n	3	11	5	7	3
	%	18.8%	12.1%	10.2%	5.1%	11.5%
Theology	n	0	1	0	2	0
	%	0.0%	1.1%	0.0%	1.5%	0.0%
Communication/Literature	n	1	1	5	3	0
	%	6.3%	1.1%	10.2%	2.2%	0.0%
Philosophy	n	0	1	0	1	1
	%	0.0%	1.1%	0.0%	0.7%	3.8%
Administration / Management	n	1	4	0	1	0
	%	6.3%	4.4%	0.0%	0.7%	0.0%

3.2.4. Professions

Well over half of all the authors of the studies in this review (63.3%) are academic researchers (professors or full-time researchers at universities; see Tables 12 and 13). By continent, this percentage is lower in Africa (30.8%) and higher in Europe (71.2%) and Australia (84%). Researchers at third-sector institutions such as foundations, NGOs and think tanks account for 16.1% of the worldwide total, while researchers from government account for 2.5%. Thus, overall, researchers account for over 80% of the authors of evaluation studies inventoried in this systematic review.

Independent consultants and consultants at private firms account for far smaller percentages of all the authors (5.7% and 5.4%, respectively; see tables 12 and 13). By continent, the proportions of authors who are consultants are highest in Asia (31.9%) and Africa (30.8%), lower in Australia (9.4%), and far lower in North America (5.6%) and Europe (4.3%). Directors, co-ordinators and managers of PVE program organizations account for 6.6% of all our authors, and a far higher percentage in North America (14.6%) than in Europe (4.3%), Africa (7.7%), Asia (0%) or Australia (0%). Only one of the authors (0.3%) is a PVE practitioner.

Table 12. Professions of publications' authors

Professions	n	%
Academic researchers (professors, doctoral students, postdoctoral fellows)	200	63.3%
Institutional researchers – third sector (foundations, NGOs, think tanks, etc.)	51	16.1%
Directors, co-ordinators, managers of PVE program organizations	21	6.6%
Consultants (independent)	18	5.7%
Consultants (private firms)	17	5.4%
Government researchers	8	2.5%
Practitioners	1	0.3%

Table 13. Professions of publications' authors by continent

		Continent				
		Africa	North America	Asia	Europe	Australia
Academic researchers	n	4	50	26	99	21
	%	30.8%	56.2%	55.3%	71.2%	84.0%
Institutional researchers – third sector	n	4	18	6	23	0
	%	30.8%	20.2%	12.8%	16.5%	0.0%
Directors, co-ordinators, managers	n	1	13	0	6	0
	%	7.7%	14.6%	0.0%	4.3%	0.0%
Consultants (independent)	n	3	3	4	6	1
	%	23.1%	3.4%	8.5%	4.3%	4.0%
Consultants (private firms)	n	1	2	11	0	17
	%	7.7%	2.2%	23.4%	0.0%	5.4%
Government researchers	n	0	3	0	4	1
	%	0.0%	3.4%	0.0%	2.9%	4.0%
Practitioners	n	0	0	0	1	0
	%	0.0%	0.0%	0.0%	0.7%	0.0%

Selecting the members of the evaluation team is unquestionably one of the most important steps in the evaluation process. This team's diversity and representativeness with respect to the program and the actors are important considerations. One of the most important findings in this review was that researchers account for such a high proportion of the authors of PVE evaluation studies. As discussed earlier, a large majority of the studies included in this review appeared in non-academic publications ("grey literature"), which might lead one to suppose that PVE evaluation has continued to develop largely outside the academic space. But as the figures just cited indicate, academics are very well represented among PVE evaluation authors, even though they may not always publish their findings in scholarly journals.

By and large, however, researchers are not the people who actually deliver the majority of PVE programs. This

raises several important questions. By whom and for whom are these evaluations being conducted? Are they intended primarily for the researchers' own purposes, and to some extent the funders'? Given that few if any PVE evaluation authors are PVE practitioners themselves, how relevant do practitioners consider evaluations that are conducted by researchers and hence in theory more remote from practitioners' concerns? To what extent are these evaluations suited, in terms of objectives and language, to the uses that practitioners might make of them? These questions are all the more important when one considers that, as indicated earlier, some of these evaluations are written in English but deal with programs delivered in non-English-speaking countries. Also, as will be seen in our discussion of the limitations of the studies in this review, one of the difficulties encountered in evaluating PVE programs has been the need for translation for researchers who did not speak the languages of the countries where the programs were delivered.

3.3. METHODOLOGIES OF THE STUDIES REVIEWED

This section discusses several characteristics of the methodologies used to conduct the evaluations included in this systematic review. These characteristics include: evaluation type (objective), type of evaluators (external versus internal), methodological design (quantitative, qualitative, or mixed), use of experimental and quasi-experimental methods, use of repeated measurements, participants and control groups, data-collection tools, and use of direct and indirect indicators of violent extremism.

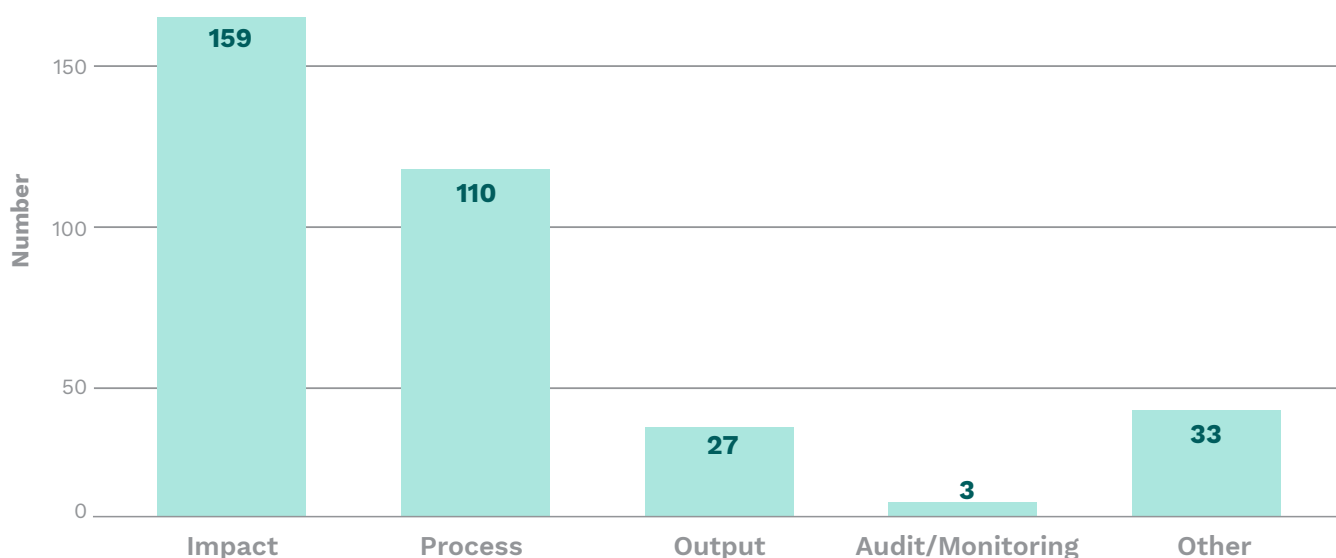
3.3.1. Evaluation types (objectives)

In this review, we defined five types of PVE program evaluations, according to their objectives:³¹ impact evaluations, process evaluations, and output, audit and monitoring evaluations. As Figure 11 shows, evaluations of the impact that PVE programs had on their target populations accounted for the largest number in this review (n = 159). Next came evaluations of the processes by which the programs were implemented (n = 110), which is consistent with the findings of past reviews (Bellasio et al., 2018; Feddes, 2015; Mastroe and Szmania, 2016). We classified a high proportion of all the evaluations (59.4%) as belonging to only one of these five types,

and another 32.4% as belonging to two. Among those evaluations that belonged to only one type, impact evaluations accounted for 52.2% and process evaluations for 35.5%. Among those evaluations that belonged to two types, 8 out of 10 were impact and process evaluations. Most of the evaluations that we classified as output, audit or monitoring evaluations were in fact also impact evaluations, or process evaluations, or both. Output evaluations, whose objective is to determine to what extent the planned program activities were carried out, seem to be an important but not fundamental component of PVE program evaluations.

³¹ A detailed definition of each evaluation type is given in Appendix B.

Figure 11. Number of evaluations by type



Turning to the year-to-year changes in number of evaluations by type (Figure 12), we see that most of the sharp increase from 2016 on can be attributed to impact evaluations. The number of process evaluations rose sharply from 2015 to 2016, remained relatively stable for two years, and then declined from 2018 to 2019. The number of output evaluations increased from 2016 to 2017, then fell steadily in the following two years. It should be noted that we coded evaluations as being of these types when they were so described directly by their authors. As we shall see later, authors' describing evaluations as being of a particular type does not necessarily mean that they followed a consistent methodology to achieve the stated objective of evaluations of that type. Thus the steady increase in the number of impact evaluations from 2016 on may have had more to do with the authors' having intended to evaluate programs' impacts rather than with their having actually done so.

Figure 12. Year-to-year changes in number of evaluations by type

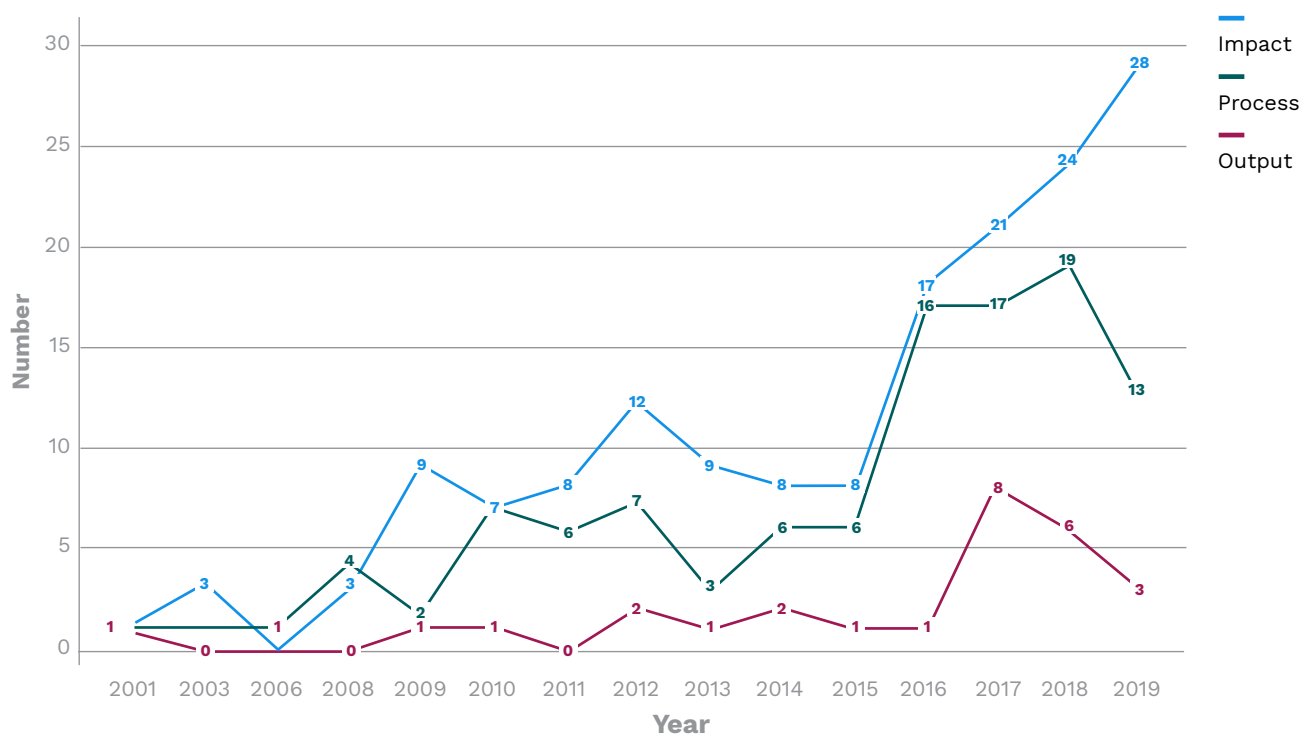


Table 14 provides more information on these three types of evaluations according to the continent where the programs were delivered, the prevention level that they attempted, and the type of extremism that they targeted. The patterns vary considerably from continent to continent. Though impact evaluations account for the highest proportion of evaluations on every continent, these percentages are highest in Australia (85.7%) and Africa (82%). The ratios of impact evaluations to process evaluations on these two continents are quite different, however. In Africa, 2.2 impact evaluations were conducted for every process evaluation. In Australia, the figure was 1.5, which is close to the ratio of 1.4 to 1 worldwide. For the other continents, the balance was closer to even (Asia = 1.3; Europe = 1.2; North America = 1.7). The heavy skew toward impact evaluations in Africa may be the

result of the types of evaluators and the funding sources. As we saw in the international qualitative study that we conducted in parallel with this systematic review (Madriaza et al., 2021), funders of PVE programs seem less interested in the processes by which they are implemented than in the impacts that they achieve. In contrast, practitioners are more interested in qualitative evaluations of such processes than in the results of quantitative methods of evaluating impact. As we have seen in the first two sections of the current systematic review, many programs in Africa receive their funding from outside of Africa, and most of the people evaluating these programs come from Western countries as well. As a result, evaluations are often planned from outside the programs, with a top-down perspective, in which the actors in the field have very little say.

Table 14. Evaluations by type, continent, prevention level and type of extremism targeted

		Impact		Process		Output	
		n	%	n	%	n	%
Total		159	72.6%	110	50.2%	27	12.3%
Continent	Africa	41	82.0%	19	38.0%	11	22.0%
	North America	10	55.6%	6	33.3%	3	16.7%
	Asia	32	76.2%	24	57.1%	3	7.1%
	Europe	67	68.4%	57	58.2%	9	9.2%
	n/a	3	75.0%	0	0.0%	1	25.0%
	Australia	6	85.7%	4	57.1%	0	0.0%
Prevention level	Primary	36	75.0%	16	33.3%	10	20.8%
	Targetted primary	71	79.8%	47	52.8%	8	9.0%
	Secondary	42	68.9%	33	54.1%	13	21.3%
	Tertiary	29	63.0%	33	71.7%	7	15.2%
	General	16	66.7%	10	41.7%	3	12.5%
Type of violent extremism	Right-wing	16	80.0%	6	30.0%	2	10.0%
	Islamist	61	72.6%	47	56.0%	10	11.9%
	All types	91	71.7%	62	48.8%	16	12.6%

The ratio of impact evaluations to process evaluations also varies with the prevention level of the programs evaluated and type of extremism that they target. As regards prevention levels, programs that target the broadest groups have the highest ratio of impact evaluations to process evaluations (primary prevention = 2.3; general = 1.6; targetted primary = 1.5), while for programs aimed at more specific groups, the ratio is more balanced (secondary = 1.3) or even reversed (tertiary = 0.9). In the introduction to this review, we alluded

to the difficulties of evaluating the impacts of tertiary prevention programs and the ethical issues involved in using experimental designs in such evaluations. As we shall see in section 3.3.4, the more sophisticated kinds of designs (experimental and quasi experimental) are more common in evaluations of the most broadly targetted programs, in which these methodological and ethical issues are less of a concern.

The ratios of impact evaluations to process evaluations

according to the type of extremism that the programs target follows a different pattern. This ratio is closer to even for programs targeting Islamist extremism (1.3) and all types of extremism (1.5), but far higher for programs targeting right-wing extremism (2.7). The relatively even ratio for programs targeting Islamist extremism is understandable, given the ethical issues involved in quantitative evaluations and the evaluators' possible interest in learning more from practitioners about the

realities that they face in the field, as often happens in process evaluations, especially in situations where the Muslim community has been highly stigmatized. The very high ratio of impact evaluations to process evaluations for programs targeting right-wing extremism is surprising. It might reflect the emphasis in the literature on the absence of evaluations of programs of this kind, but that is hard to know for certain.

3.3.2. External versus internal evaluators

In section 3.2 of this review, we presented the personal characteristics of the authors of the evaluation studies. Here we focus on how independent the persons or groups conducting the evaluations were from the programs that they were evaluating. For this purpose we have defined two main categories: external evaluations and internal evaluations (Table 15). External evaluations are conducted by individuals or groups that are external to the organization carrying out the program and the agency funding it. Internal evaluations, in contrast, are conducted by the people within the organization who

have been responsible for designing and implementing the program or who work for the agencies funding it or their partners. **By this definition, three-quarters of the evaluations were external** (n = 158) and only 51 were internal. For the rest of the studies, we did not have enough information about the authors to determine whether they should be classified as external or internal evaluations. We did not find any evaluations that we would have classified as participatory, meaning that many or all of the stakeholders (including program participants, practitioners and researchers) had contributed to them.

Table 15. Numbers of external and internal evaluations by continent, prevention level and type of extremism targetted

		External		Internal	
		n	%	n	%
Total		158	72.10%	51	23.30%
Continent	Africa	35	76.10%	11	23.90%
	North America	11	64.70%	6	35.30%
	Asia	31	77.50%	9	22.50%
	Europe	74	77.90%	21	22.10%
	n/a	1	25.00%	3	75.00%
	Australia	6	85.70%	1	14.30%
Prevention level	Primary	38	82.60%	8	17.40%
	Targetted primary	61	69.30%	27	30.70%
	Secondary	42	72.40%	16	27.60%
	Tertiary	31	73.80%	11	26.20%
	General	20	87.00%	3	13.00%
Type of violent extremism	Right-wing	11	57.90%	8	42.10%
	Islamist	58	71.60%	23	28.40%
	All types	97	80.20%	24	19.80%

The ratio of external to internal evaluations is an informative statistic. For all of the evaluations included in this review, this ratio is 3.1 external evaluations to every internal evaluation. For most of the continents considered individually, this ratio is slightly higher (Africa = 3.2; Asia = 3.4; Europe = 3.5), while for Australia it is far higher (6.0) and for North America it is much lower (1.8). The reason for this low ratio in North America may be that more of the programs delivered there have experimental designs and hence are evaluated by the same researchers who designed them.

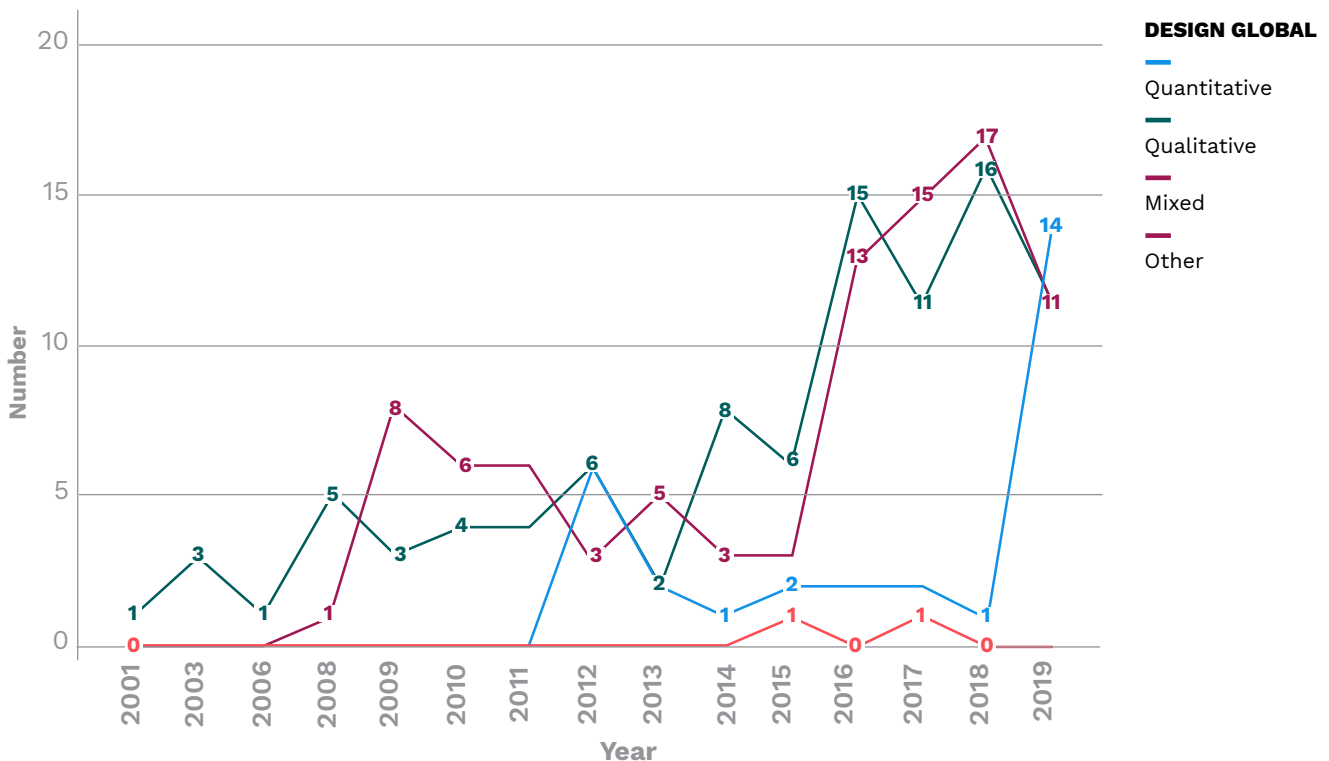
Comparing the ratio of external to internal evaluations according to the prevention levels of the programs evaluated, we see a clear trend. For the more broadly targetted programs, this ratio is higher than the ratio of 3.1 for all programs combined (4.8 for primary prevention programs and 6.7 for general prevention programs). For the more specifically targetted programs, it is lower (2.3 for targetted primary prevention programs, 2.6 for secondary prevention programs, and 2.8 for tertiary prevention programs). One explanation for this difference may be that evaluations of more specifically targetted programs are more sensitive and so require a greater diversity of viewpoints, both internal and external, in order to succeed. This same pattern is seen when we compare the ratios according to the types of extremism that the programs target. Once again, the ratios for

the more specific programs fall below the ratio for all programs combined (1.4 for programs targeting right-wing extremism and 2.5 for programs targeting Islamist extremism), while the ratio for the less specific programs (those targeting all types of extremism) is higher (4.2).

3.3.3. Methodological design (quantitative, qualitative or mixed)

In this section, we classify the evaluation studies according to their methodological design: qualitative, quantitative or mixed (quantitative and qualitative combined) (Figure 13 and Table 16). Studies with a purely qualitative approach and studies with a mixed approach were the most common (n = 96, 43.8% and n = 91, 41.6%, respectively). Only 30 of the studies (13.7%) were purely quantitative. The number of quantitative evaluations jumped exponentially from 2018 to 2019 (the last year covered in this review), when it exceeded the numbers of mixed and qualitative evaluations for the first time. This increase is attributable to the increase in academic publications, but may also be explained by the growing interest in recent years in impact evaluations (see preceding section), which usually take a quantitative approach. As Table 16 shows, Australia is the continent with the highest proportion of quantitative evaluations (28.6%), followed by North America (16.7%).

Figure 13. Year-to-year changes in number of evaluations, by methodological design



The number of purely qualitative evaluations began to rise steadily in 2014, slightly ahead of the other categories. But more recently, it has been surpassed by the number of mixed evaluations and especially by the number of quantitative evaluations in 2019. The reason may be that process evaluations, which mostly employ a qualitative approach, have become less popular in recent years. Qualitative approaches seem more characteristic of evaluations in Europe, where they account for 53.1% of all the studies included in this review.

Another interesting finding is the growing popularity of mixed designs, which, as discussed in the first section of this review, have been repeatedly recommended by

many experts. The number of evaluation studies with mixed designs began to increase considerably in 2016. As Table 16 shows, they account for a higher proportion of all evaluations in Australia (71.4%) and Africa (50%) than in Europe (34.7%). Bellasio et al. (2018) and Hassan et al. (2021) have already reported an increase in the number of mixed studies in recent years. But although the current systematic review covers approximately the same periods as these two, it shows an even higher proportion of mixed designs. No doubt part of the explanation is the publication bias in these two other reviews, which focused mainly on the academic literature, whereas 70 of the 91 mixed studies included in our systematic review appeared in the grey literature.

Table 16. Methodological designs of studies by continent and evaluation type

		Quantitative		Qualitative		Mixed	
		n	%	n	%	n	%
Total		30	13.7%	96	43.8%	91	41.6%
Continent	Africa	6	12.0%	18	36.0%	25	50.0%
	North America	3	16.7%	6	33.3%	8	44.4%
	Asia	5	11.9%	20	47.6%	17	40.5%
	Europe	12	12.2%	52	53.1%	34	34.7%
	n/a	2	50.0%	0	0.0%	2	50.0%
	Australia	2	28.6%	0	0.0%	5	71.4%
Evaluation type	Impact	29	18.5%	49	30.8%	81	50.9%
	Process	1	0.9%	66	60.0%	43	39.1%
	Output	2	7.4%	9	33.3%	14	51.9%

Table 16 also shows the interesting relationship between overall methodological design and evaluation type. As mentioned earlier, the stated objectives of the evaluations reviewed did not always match the results that they obtained or the tools that they used. Although qualitative methods can be used to evaluate the impact of programs, quantitative methods are more commonly used for this purpose. Conversely, quantitative methods can be used to evaluate the processes by which programs are implemented, but qualitative methods are more commonly used for this purpose, because it involves exploring elements that the evaluators do not always know in advance. In the studies included in this review, this pattern is maintained fairly intact. Almost all of the process evaluations (99.1%) were conducted using qualitative or mixed methodologies.³² Regarding impact

evaluations, the pattern was less clear-cut: 30.8% of the studies whose stated intent was to evaluate program impact used qualitative methods only. Although a large number of these studies reported results consistent with the objective of evaluating program impacts, in the remainder we were unable to confirm such consistency. In other words, as will be seen in our analysis of the quality of the methods used in the studies in this review, the results of studies do not always achieve the initial objectives stated in their introductions or their methodology sections. In such cases, the concept of impact evaluation seems to lose much of its original meaning and is thus applied to pretty much any type of program evaluation.

³² Just one study indicated the intent to evaluate the program's process but used only quantitative tools (Broadbent, 2013). In her methodology section, the author included some questions that pointed toward a process evaluation, but in her results section, the only data that she reported dealt with the effects of and satisfaction with the program.

3.3.4. Use of experimental and quasi-experimental methods

In this section, we discuss the extent to which the evaluation studies included in this systematic review were designed as scientific experiments, with experimental or quasi-experimental designs (Box 3). Unsurprisingly, the vast majority of these studies (n = 157) did not have such designs. Instead, as in past reviews, most of the studies had either qualitative or descriptive quantitative designs.

But in contrast with past reviews, the current systematic review did identify 6 experimental studies and 54 quasi-experimental studies.

Normally, such studies are conducted to assess programs' impact on the people who participate in them, and that was in fact the case for all 60 of these studies that we found. One of these experimental studies and 12 of these quasi-experimental studies also used

qualitative analyses to evaluate the processes by which the programs were implemented, so we have classified them as having mixed methodological designs as well. The number of experimental and quasi-experimental studies that we identified is encouraging for the field, particularly as regards evaluating programs' impacts. For example, Bellasio et al. (2018) identified only 6 quasi-experimental ex post evaluations in their inventory, while Hassan's teams, in their more recent inventory, identified only 4 (Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021; Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021).³³One reason that we found so many more experimental and quasi-experimental studies was our extensive search of the grey literature. Another was that the inclusion criteria for our inventory were more flexible than those in Hassan's two studies. As will be seen in our discussion of the methodological quality of the studies in our review, the use of experimental or quasi-experimental methods is unfortunately no guarantee of such quality. Bellasio et al. (2018) did not address this issue.

Box 3. Two examples of PVE program evaluations with experimental designs

The Effectiveness of an Educational Program for Developing Tolerance Values and Resistance to Intellectual Extremism at Secondary Level in Jordan (Al-Maqosi et al., 2019)

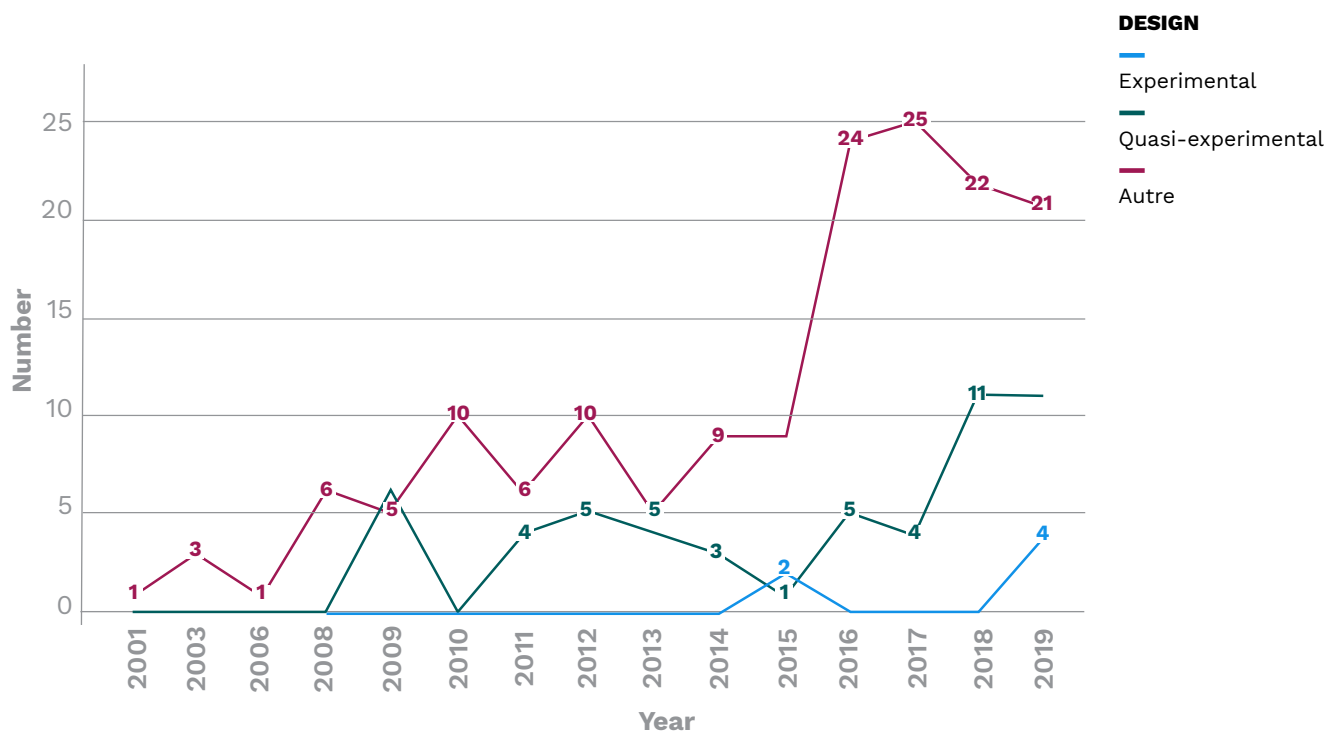
This study aimed to establish an effective educational program for high-school students in Jordan, focusing on the rights of non-Muslims under Islam in order to encourage the development of tolerance for various religions and intellectual resistance to extremist ideologies. In this program, the students attended 10 sessions lasting a total of 10 hours, in which they were educated about various fundamental human rights, such as dignity, security, justice, freedom of religion and freedom of expression. The study was conducted on a sample of 48 students in 11th grade at the Al-Arqam Islamic School, during the 2017-2018 academic year. An experimental method was used in which the sample was divided into two groups: 23 students in the experimental group and 25 in the control group. To measure the effectiveness of the educational program, the researchers developed and applied one scale to measure values of religious tolerance and another to measure resistance to intellectual extremism. The results showed statistically significant differences between the average scores of the experimental group and the control group on the measures for religious tolerance and resistance to intellectual extremism before and after the program. The average scores for the students in the control group were fairly similar before and after, but the average scores for the students who had participated in the educational program increased significantly.

³³ Carthy et al. (2020) conducted a systematic review of studies with experimental and quasi-experimental designs. But as stated in the first section of the current review, we did not include any of the studies in that review, because none of them directly targetted prevention of violent extremism.

Evaluation and Analytical Services (EAS) Project for the Regional Peace and Governance Programs – Impact Evaluation of Peace through Development II (P-DEV II) Radio Programming in Chad and Niger – Final Report (Finkel et al., 2015)³⁴

The authors evaluated the radio component of a program called Peace Through Development II (P-DEV II), whose main goal is to counter violent extremism in Chad, Niger and Burkina Faso (this study focused on Chad and Niger). To achieve this goal, the program concentrates its efforts on empowering youth, increasing moderate voices and resilience to violent extremism, and strengthening local government. To these ends, the program strives to build the capacity of local radio stations to produce and broadcast content designed to counter violent extremism by providing them with equipment and technical assistance and training their staff. To evaluate the impact of the radio program, the authors conducted a longitudinal study of 750 individuals aged 15 to 30 in Chad. Each of these individuals was interviewed before the program was implemented, and another interview took place several months afterward. In addition, 525 of these 750 youth received messages encouraging and reminding them to listen to P-DEV II radio programming, while the remaining 225 did not. The results indicated that listening to this programming increased dissatisfaction with life in both countries. In Chad, listening to this programming also increased interest in local affairs and politics, had positive effects on diversity and inclusiveness and helped to reduce support for violence. In contrast, in Niger, no significant effects were observed with regard to diversity, inclusiveness or support for violence. Some different significant effects were obtained, notably increased trust in local government.

Figure 14. Year-to-year changes in number of studies that used experimental or quasi-experimental designs



As Figure 14 shows, studies with experimental designs began to emerge during the last year covered by this review, which may explain why none were found in the earlier reviews. In contrast, there have been at least a few quasi-experimental studies in most years since 2009, and their number began to rise starting in 2016. Interestingly, not all of the experimental studies have

been conducted in Western countries. In fact, as Table 17 shows, none were conducted in North America, while there were 3 in Africa, 2 in Europe and 1 in Asia. For the quasi-experimental studies, the pattern is similar: the largest numbers were conducted in Europe (25), Africa (16) and Asia (9).

³⁴ This study was divided in two, because the data collection, analysis and interpretation were different for each of the two countries concerned.

Table 17. Experimental or quasi-experimental designs, by continent

Continent	Experimental		Quasi- experimental		Other	
	n	%	n	%	n	%
Total	6	2.7%	54	24.7%	157	71.7%
Africa	3	6.0%	16	32.7%	30	61.2%
North America	0	0.0%	3	16.70%	15	83.3%
Asia	1	2.4%	9	21.4%	32	76.2%
Europe	2	2.1%	25	25.8%	70	72.2%
n/a	0	0.0%	0	0.00%	4	100.0%
Australia	0	0.0%	1	14.30%	6	85.7%

Table 18 shows the number of evaluations that used experimental or quasi-experimental designs, according to program prevention level and type of extremism targetted. Among evaluations of tertiary PVE programs, there were none with experimental designs, and only 13.6% had quasi-experimental designs; other design approaches accounted for the remaining 86.4%. As mentioned earlier, part of the explanation (particularly in the case of impact evaluations) may be the difficulties associated with collecting data and the ethical issues

involved in using experimental models in tertiary PVE programs. The greatest number of evaluation studies with experimental or quasi-experimental designs thus dealt with primary and targetted primary prevention programs and with programs that addressed all types of extremism. There were also a fair number of quasi-experimental evaluations of secondary prevention programs. At these prevention levels, it is easier to use program participants and control groups, because a greater number of individuals are involved.

Table 18. Experimental or quasi-experimental designs, by prevention level and type of extremism targetted

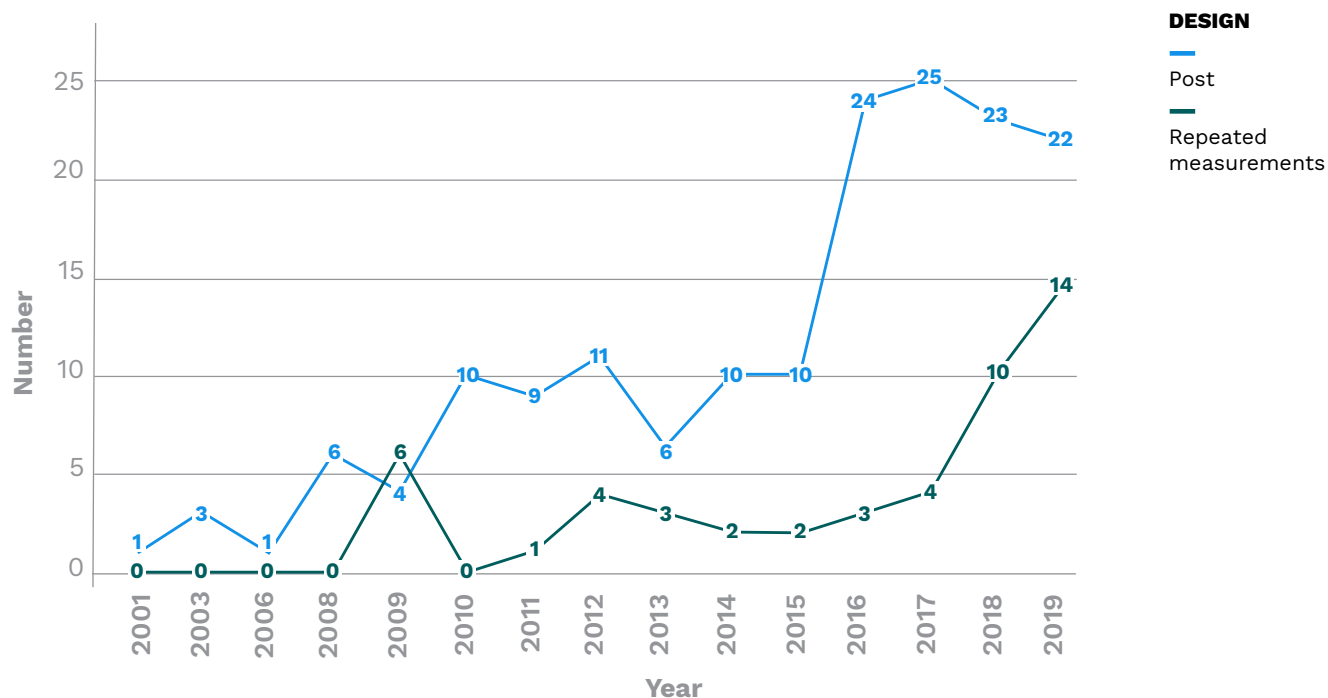
		Experimental		Quasi-experimental		Other	
		n	%	n	%	n	%
Prevention level	Primary	3	6.30%	11	22.9%	34	70.8%
	Targetted primary	5	5.60%	22	24.70%	62	69.70%
	Secondary	0	0.00%	16	26.7%	44	73.3%
	Tertiary	0	0.00%	6	13.6%	38	86.4%
	General	0	0.00%	8	33.30%	16	66.70%
Type of violent extremism	Right-wing	1	5.00%	5	25.00%	14	70.00%
	Islamist	1	1.20%	13	15.7%	69	83.1%
	All types	4	3.2%	37	29.6%	84	67.2%

3.3.5. Use of repeated measurements

Past literature reviews have often mentioned the lack of program evaluations in which repeated measurements are taken—in other words, where multiple observations are made of the same subjects at two or more different points in time. In the current systematic review, we classified as studies with repeated measurements all of the evaluation studies that took at least one measurement before and one measurement after the program intervention, regardless of whether the time between the first measurement and the last was measured in hours, days or months.

As Table 19 shows, **we found a total of 49 studies with repeated measurements** (43 of them with quasi-experimental designs and the 6 others with experimental designs), representing about one-quarter of all the studies included in this review. As is typical for such studies, all of them attempted to assess the effectiveness (impact) of the programs in question, and only 10.2% also used qualitative analyses to evaluate the processes by which the programs were implemented (in other words, employed mixed designs).

Figure 15. Year-to-year changes in number of studies with repeated measurements



As seen in Figure 15, according to our review, the first PVE evaluation studies with repeated measurements were published in 2009, and the number of such studies increased substantially in the last two years covered by this review. This finding shows that evaluations using repeated measurements are starting to become a standard in this field, which may be seen as a sign of improvement in the quality of the designs used, particularly for evaluating programs' impacts.

Table 19. Presence of repeated measurements, by continent, prevention level and type of extremism targetted

		Repeated measurements			
		No		Yes	
		n	%	n	%
Total		165	75,3 %	49	22,4 %
Continent	Africa	39	79.6%	10	20.4%
	North America	16	88.9%	2	11.1%
	Asia	32	78.0%	9	22.0%
	Europe	69	71.9%	27	28.1%
	n/a	3	100.0%	0	0.0%
	Australia	6	85.7%	1	14.3%
Prevention level	Primary	37	77.1%	11	22.9%
	Targetted primary	65	73.0%	24	27.0%
	Secondary	46	78.0%	13	22.0%
	Tertiary	35	85.4%	6	14.6%
	General	19	79.2%	5	20.8%
Type of violent extremism	Right-wing	11	64.7%	6	35.3%
	Islamist	72	88.9%	9	11.1%
	All types	91	72.2%	35	27.8%

The vast majority (n = 37) of the studies with repeated measurements used a simple pre-post design, meaning that they took measurements (collected data) on one occasion before the program intervention and another occasion afterward. Six other studies took measurements on at least one more occasion: 3 of them at some time between the pre- and post-measurements (between or midline measurements) and 3 of them at some time after the post-measurements (follow-up measurements). In one of these studies, the researchers attempted to examine the long-term effects of a training program intended to make vulnerable youth more resilient to radicalization (University of Amsterdam, 2013). In this evaluation, the researchers used a longitudinal model and took measurements at four different times (before the training, during the training, immediately after the training, and three months after the training). This is one of the rare examples of a longitudinal PVE program evaluation in which measurements were taken a significant time after the intervention had ended.

In another evaluation study, conducted in the United States using data from its federal Department of Homeland Security, the author used a difference-in-difference design³⁵ to examine whether community-engagement events held by this department were associated with a reduction in pro-ISIS content on Twitter (Mitts, 2017). She collected the content on Twitter before the events and 7, 14, 21 and 30 days afterward.

Lastly, we found four studies that took no measurements before or during the intervention but took one set of measurements immediately after and at least one follow-up set some time later. In one of these studies, Finkel et al (2015) evaluated the impact of one of the components of a program in Niger and Chad that worked to build the capacity of local radio stations to develop, produce and broadcast their own PVE content. The methodological design involved a longitudinal experimental panel in which a randomly selected group of individuals were encouraged to listen to the radio programs. Then their data were compared with those from a control group that had not received such encouragement. The data were collected in two “waves”: once right after the radio messages had been broadcast and again about 10 months later.

According to prevention level, evaluations with repeated measurements were most common for targetted primary prevention programs, accounting for 27% of all evaluations in this category, and least common for tertiary prevention programs, where the proportion was 14.6%. This lower percentage is explained by the ethical, methodological and practical difficulties discussed earlier. The reasons that targetted primary prevention programs have the highest percentage are not entirely clear, given that these

programs do not necessarily present fewer problems in this regard than primary prevention programs do.

The percentage of evaluations with repeated measurements was higher for programs targeting right-wing extremism than for those targeting Islamist extremism, possibly because there are so few programs of the former kind that even a slight change in number can create a large change in percentage.

3.3.6. Description of participants and use of control groups

A detailed description of the participants (subjects) is a standard part of most scientific studies and a quality requirement for scholarly journals. But not all of the evaluation studies included in this review provided such descriptions. In this section, we present information about those studies that did include such descriptions. Table 20 shows the number of studies that stated the number of participants and the number of studies that used control groups, by continent, prevention level and type of extremism targetted.

Out of the 219 evaluation studies in this review, only 128 (58.5%) stated the number of participants that they included, and only 22 (10%) used control groups. All of these 22 studies were impact evaluations; 6 of them used experimental designs, and the 16 others used quasi-experimental designs. Regardless of how an evaluation study is designed, the lack of a relatively detailed description of the participants is a clear indicator of a lack of methodological transparency. As will be discussed in the section on methodological quality, such a lack of transparency is one of the main defects of a considerable number of the studies included in this systematic review.

The percentage of the evaluation studies that failed to give any information on their samples varied by continent, prevention level, and type of extremism targetted. By continent, this percentage was 61.1% for studies conducted in North America and 54.8% for studies conducted in Asia, both of which are far higher than the figure of 41.6% for all continents combined. In contrast, the figures for Australia and Africa were lower (28.6% and 30%, respectively). By prevention level, no information on the sample was provided in the evaluations of 52.1% of the primary prevention programs, 47.5% of the secondary prevention programs, and 47.8% of the tertiary prevention programs. A description of the sample was also lacking in a high percentage of the programs targeting right-wing extremism (65%).

³⁵ A difference-in-difference design estimates the effect of an intervention by comparing the difference between the control group and the treatment group before the treatment with the difference after. The model used by Mitts (2017) is similar to a time-series model.

Table 20. Number of participants and number of studies using control groups

			n*	Total*	Mean	Std. dev.	Minimum	Maximum	
Total	Participants		128	219	233.54	415.7	3	2789	
	Control groups		22		343.7	409	1	1452	
Continent	Africa	Participants	35	50	387.9	552.3	14	2789	
		Control groups	11		468.9	486.5	27	1452	
	North America	Participants	7	18	199.3	269.7	26	767	
		Control groups	2		186.0	198.0	46	326	
	Asia	Participants	19	42	358.9	563.3	4	1657	
		Control groups	3		227.0	189.6	25	401	
	Europe	Participants	61	98	126.3	238.0	3	1113	
		Control groups	6		225.0	370.8	1	976	
	n/a	Participants	1	4	154.0		154.0	154.0	
		Control groups							
	Australia	Participants	5	7	49.2	39.2	16.0	117.0	
		Control groups							
	Prevention level	Primary	Participants	23	48	512.0	653.9	9.0	2789.0
			Control groups	6		710.5	544.3	102.0	1452.0
Targetted primary		Participants	57	89	226.1	361.2	5.0	1657.0	
		Control groups	10		451.6	425.0	25.0	1050.0	
Secondary		Participants	32	61	237.9	416.8	3.0	1644.0	
		Control groups	3		24.7	22.6	1.0	46.0	
Tertiary		Participants	24	46	171.9	311.4	4.0	1170.0	
		Control groups	2		128.0	179.6	1.0	255.0	
General		Participants	15	24	116.4	105.5	4.0	368.0	
		Control groups	5		119.8	28.6	85.0	152.0	
Type of violent extremism	Right-wing	Participants	20	151.9	269.1	3.0	747.0	20	
		Control groups		102.0		102.0	102.0		
	Islamist	Participants	84	216.5	384.7	3.0	1644.0	84	
		Control groups		183.1	178.9	1.0	484.0		
	All types	Participants	127	245.9	461.1	4.0	2789.0	127	

* n = Number of studies that stated the number of participants

Total = Total number of studies included in this review

In contrast, all of the studies that used control groups at least stated the number of participants in their sample. Whether a study uses a control group depends on its methodological design and the research questions that it investigates. The use of control groups is recommended for evaluating the effectiveness (impact) of programs, and in the current systematic review, all of the studies with control groups were impact studies, using experimental or quasi-experimental designs. The relative absence or presence of control groups therefore cannot be used as a quality indicator for all of the studies in this review. Keeping this consideration in mind, the use of control groups was more frequent in evaluations in Africa (22%) and North America (11.1%). It was also more common in evaluations of primary prevention programs (12.5%), targetted primary prevention programs (11.2%) and general prevention programs (20.8%)—in other words, the more

broadly targetted programs, which is consistent with the idea that such programs lend themselves more readily to the use of more sophisticated quantitative evaluation designs.

The size of the samples in the evaluation studies varies quite widely, from 3 to 2789 (standard deviation = 415.7) for all participants and from 1 to 1452 (standard deviation = 409) for members of control groups. This wide variability signifies that the mean is a less meaningful indicator. For example, the median number for all participants is 56, while the median number for members of control groups is 148.5. The median of 56 indicates that half of the evaluation studies that mentioned their sample used 56 or fewer participants to evaluate their programs (Table 21).

Table 21. Number of participants and number of studies using control groups, by methodological design

		n*	Total*	Mean	Std. dev.	Minimum	Maximum
Quantitatif	Participants	21	30	376.3	419.7	16	1657
	Control group	10		429.7	429.8	25	1050
Qualitatif	Participants	52	96	48.3	66.5	3	357
	Control group	1		1.0		1	1
Mixed	Participants	55	91	354.1	530.3	5	2789
	Control group	11		296.6	404.3	46	1452

* N = number of studies that stated the number of participants; Total = total number of studies included in this review

Obviously, the overall methodological design influences this variability. Qualitative studies require fewer participants than quantitative studies. As Table 21 shows, for the qualitative studies, the mean number of participants was 48.3 (median = 21.5), and variability in sample size remained a major issue, with a range from 3 to 357. Despite this variability, half of the qualitative studies dealt with 21 persons or more. Qualitative studies require a diversity of viewpoints, and these results may be evidence of the quality of the qualitative studies in this review. On the other hand, only 55% of the qualitative studies reported their samples.

The mean number of participants in the quantitative studies was 376.3 (median = 191), and 7 out of 10 quantitative studies mentioned their samples. One-third of the quantitative studies used control groups.

Lastly, 60.4% of the mixed studies mentioned or described their samples. When they did so, the variability was far more significant than in the case of the other types of studies, and the use of control groups was less extensive. As will be seen in the section on evaluation of methodological quality, this type of study was evaluated as being of lesser quality as regards the integration of the two methodologies.

3.3.7 Data-collection tools

As Table 22 shows, the vast majority of the studies used traditional data-collection tools, such as individual interviews (74%), surveys (49.8%), focus groups (32.4%) and, less often, direct observations (16.9%). The category

“Other” in this table includes secondary data,³⁶ taken, for example, from the records and documentation of the organizations involved in the evaluations or analyses of activity on online social networks.

Table 22. Data-collection tools used in studies, by prevention level, type of extremism targeted and type of evaluation³⁷

		Surveys		Interviews		Focus groups		Observations		Other	
		n	%	n	%	n	%	n	%	n	%
Total		109	49.8%	162	74.0%	71	32.4%	37	16.9%	73	33.3%
Prevention level	Primary	25	53.2%	32	68.1%	18	39.1%	8	17.0%	21	44.7%
	Targetted primary	53	59.6%	62	69.7%	30	33.7%	19	21.3%	23	25.8%
	Secondary	33	55.0%	48	80.0%	23	38.3%	6	10.0%	27	45.0%
	Tertiary	14	31.8%	39	86.7%	10	22.7%	5	11.1%	18	40.9%
	General	9	37.5%	20	83.3%	8	33.3%	6	25.0%	7	29.2%
Type of violent extremism	Right-wing	8	42.1%	11	57.9%	0	0.0%	4	21.1%	5	26.3%
	Islamist	35	42.2%	63	75.9%	30	36.6%	17	20.5%	29	34.9%
	All types	69	55.2%	97	77.0%	45	36.0%	19	15.1%	42	33.6%
Type of evaluation	Impact	98	62.0%	109	69.0%	59	37.6%	27	17.1%	49	31.0%
	Process	44	40.4%	103	93.6%	40	36.7%	22	20.0%	33	30.3%
	Output	13	50.0%	23	88.5%	17	65.4%	4	15.4%	19	73.1%

As Table 22 also shows, the relative proportions of the data-collection tools used in evaluations of PVE programs vary with their prevention level, the type of extremism that they target, and the aspect of the programs that they evaluate (impact, process, etc.).

In evaluations of programs that target Islamist extremism or all types of extremism, focus groups are an important tool. But that is not the case for evaluations of programs targeting right-wing extremism, where individual interviews are the main preferred tool. One likely reason is the limited number of participants in such evaluations (half of these evaluations questioned 15 people or fewer). Orban (2019), for example, explained that in his evaluation

³⁶ All of these secondary data are sources of information that supplement the primary sources and are therefore described here even if we included only the studies that used primary sources.

³⁷ Only the main categories were considered in this analysis.

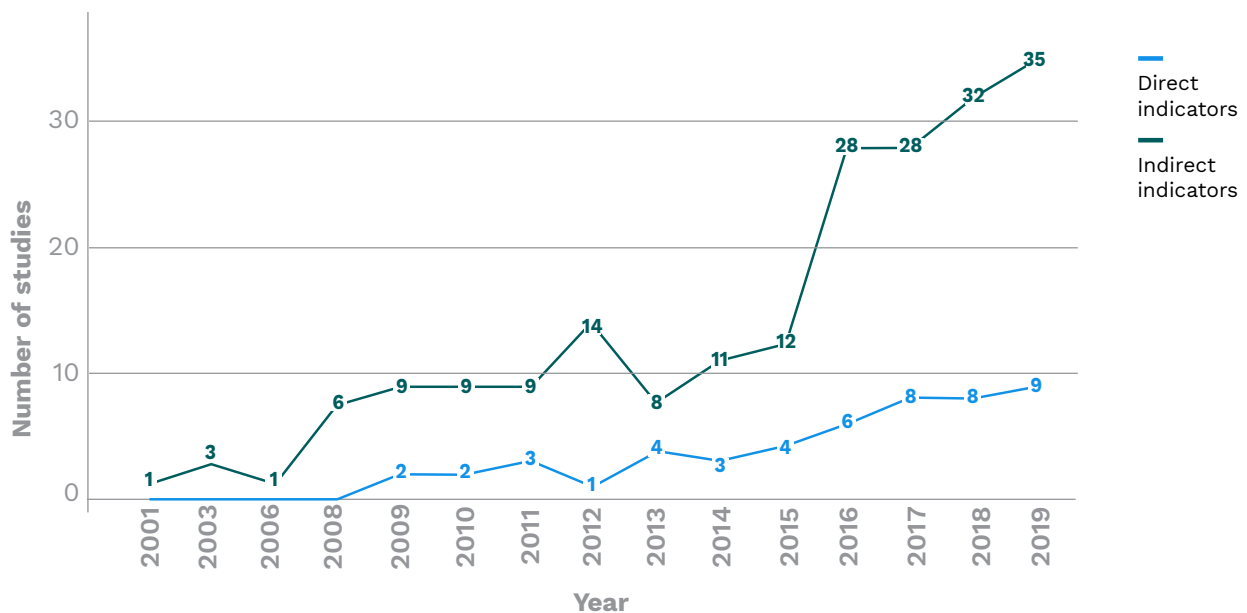
of a program targetting members of the extreme right, one of the main difficulties was in gaining the trust of the people being interviewed. A similar situation is seen in evaluations of tertiary prevention programs, in which more individual interviews tend to be conducted than in evaluations of primary prevention programs. Once again, the reason seems to be the small number of individuals concerned.

As mentioned earlier, impact evaluations are associated with a primarily quantitative approach. But the impact evaluations included in this systematic review employed a variety of data-collection tools. This tends to corroborate the importance given to mixed methods and to the viewpoint of the actors in the field. Interviews were conducted in 69% of the impact evaluations included in this review, but surveys were used in only 62%. This confirms that some of the evaluations used qualitative approaches to measure impact. The process evaluations in this review mainly used qualitative data-collection tools, while the output evaluations mostly used secondary data (Other = 73.1%).

3.3.8. Use of direct and indirect indicators of violent extremism

As mentioned in the introduction to this review, one of the difficulties in evaluating PVE programs lies in finding specific indicators for measuring their impact or performance (Figure 16). That is not always easy, because of the vagueness surrounding the definitions of radicalization and violent extremism. To overcome this problem, some evaluators use indirect indicators that do not attempt to measure radicalization, violent extremism or associated sympathies directly. Instead, they measure other factors—such as self-esteem, individual and community resilience and integrative complexity—that are theoretically associated with radicalization leading to violent extremism. In this section we explore the use of direct and indirect indicators in the studies included in this systematic review.

Figure 16. Year-to-year changes in use of direct and indirect indicators in PVE evaluations



The vast majority of the studies in this review (74%) used indirect indicators only, while only 4.1% used direct indicators only. Thus about 1 out of every 5 studies used both. As Figure 16 shows, the use of indirect indicators began rising sharply in 2016. The use of direct indicators also rose from 2016 on, but not nearly so much. In any given year, there was never more than one study that used direct indicators exclusively, and this pattern has remained stable over the years. This finding tends to

confirm the difficulty of measuring radicalization or violent extremism directly and the need to take a pragmatic approach in order to evaluate PVE programs with the resources available (see section 1.1).

As Table 23 shows, the percentage of studies using indirect indicators is quite high on every continent, in keeping with the worldwide pattern. But the percentage of studies using direct indicators shows significant

variation from one continent to the next. This percentage is higher in Africa (40%) and Asia (28.6%) than elsewhere, for two possible reasons. First, because these continents are more heavily subjected to acts of terrorism, they may have normalized the presence of terrorist groups, so that responses to direct questions about violent extremism may be less problematic ethically and methodologically. The second possible reason is less attractive: as mentioned earlier, a large portion of these evaluations, especially in Africa, are conducted for funders or by researchers from Western countries. We can speculate that such researchers may find it easier to ask direct questions about radicalization or violent extremism in Africa or Asia than they would back home.

Table 23. Number of studies using indirect and direct indicators, by continent

	Indirect		Direct	
	n	%	n	%
Africa	48	96.0%	20	40.0%
North America	18	100.0%	2	11.1%
Asia	38	90.5%	12	28.6%
Europe	91	93.8%	18	18.4%
n/a	4	100.0%	1	25.0%
Australia	7	100.0%	1	14.3%

Table 24 shows that there is no major difference in use of indicators by type of extremism or type of evaluation, except that in evaluations of the processes by which programs are implemented, direct indicators are used less often, which is consistent with the logic of such evaluations.

Table 24. Number of studies using indirect and direct indicators, by prevention level, type of extremism and type of evaluation

		Indirect		Direct	
		n	%	n	%
Prevention level	Targetted primary	46	95.8%	16	33.3%
	Secondary	86	96.6%	17	19.1%
	Tertiary	53	88.3%	21	34.4%
	General	40	88.9%	18	39.1%
	Générale	22	91.7%	4	16.7%
Type of violent extremism	Right-wing	17	85.0%	6	30.0%
	Islamist	77	92.8%	21	25.0%
	All types	123	97.6%	30	23.6%
Type of evaluation	Impact	149	93.7%	48	30.2%
	Process	106	97.2%	20	18.2%
	Output	26	96.3%	7	25.9%

The pattern for use of indicators according to the prevention level of the programs evaluated is slightly different. Almost all of the evaluations of the most broadly targetted programs (primary, targetted primary and general) tend to use indirect indicators. The evaluations of more specialized (secondary and tertiary) programs make slightly less use of indirect indicators and slightly more use of direct ones. This greater use of direct indicators to evaluate more specialized programs would seem to make sense, because in theory, tertiary programs are more likely to work with individuals who can be more readily classified as radicalized or at risk of becoming radicalized. The use of indirect indicators to evaluate more generalized programs also seems to make sense, because their target populations are not necessarily associated with radicalized groups. However, from a methodological standpoint, this approach raises some issues. The groups targetted by secondary and tertiary prevention programs are probably more politicized and more convinced of their ideas. These programs are often highly politicized and delivered to inmates of penal institutions. When direct indicators are used in such settings, the answers provided and results obtained are subject to serious bias. For example, Madriaza et al. (2018) reported having had these kinds of difficulties when evaluating a program in the French probation system.

3.4. LIMITATIONS AND CONFLICTS OF INTEREST IN THE STUDIES REVIEWED

Any study that uses scientific methods may have some limitations in its design or in the way that the data were collected or interpreted. It is considered best practice to mention such limitations in the published study, because they represent the boundaries within which the interpretation of the data is valid, reliable and possibly transferable. Conversely, if such limitations are not mentioned, the implication is that no problems at all were encountered in the research process or that the interpretation of the data is valid without regard to the research context. In PVE evaluation studies, pointing out any limitations is even more essential, because the objective is to transfer the lessons learned from the evaluation and improve practices in the field.

The nature of these limitations is indissociable from the question of whether a program that has been evaluated

positively in one prevention context can be reproduced successfully in another. Conflicts of interest relate to matters of research ethics and factors external or internal to the research process that may have affected a study’s findings. In a published study, identifying actual or potential conflicts of interest, just like identifying limitations, gives readers some perspective on the interpretation and reliability of the data.

In this section, we provide both a quantitative and a qualitative analysis of the limitations and conflicts of interest in the PVE evaluation studies included in this systematic review. First we provide statistics on these matters (Table 25), and then we describe the main limitations and conflicts of interest that the authors of these studies reported.

Table 25. Limitations and conflicts of interest in the PVE evaluation studies reviewed

		Identified limitations		Conflicts of interest reported		Potential conflicts of interest	
		n	%	n	%	n	%
Total		84	38.4%	14	6.4%	45	20.5%
Type of publication	Academic literature	28	31.5%	4	4.5%	23	25.8%
	Grey literature	56	43.1%	10	7.8%	22	16.9%
Type of evaluators	Internal	18	35.3%	3	5.9%	36	70.6%
	External	64	40.5%	10	6.4%	9	5.7%
Continent	Africa	25	50.0%	3	6.1%	10	20.0%
	North America	3	16.7%	0	0.0%	4	22.2%
	Asia	21	50.0%	4	9.5%	8	19.0%
	Europe	28	28.6%	7	7.1%	18	18.4%
	n/a	3	75.0%	0	0.0%	2	50.0%
	Australia	4	57.1%	0	0.0%	3	42.9%
Prevention level	Primary	16	33.3%	2	4.3%	9	18.8%
	Targetted primary	28	31.5%	5	5.6%	22	24.7%
	Secondary	34	55.7%	5	8.2%	15	24.6%
	Tertiary	23	50.0%	4	8.7%	9	19.6%
	General	5	20.8%	3	12.5%	4	16.7%
Type of violent extremism	Right-wing	8	40.0%	1	5.0%	6	30.0%
	Islamist	37	44.0%	4	4.8%	22	26.2%
	All types	42	33.1%	8	6.3%	20	15.7%
Type of evaluation	Impact	65	40.9%	10	6.3%	39	24.5%
	Process	43	39.1%	6	5.5%	17	15.5%
	Output	12	44.4%	2	7.4%	5	18.5%
Methodological design	Quantitative	14	46.7%	3	10.0%	8	26.7%
	Qualitative	30	31.3%	8	8.3%	6	6.3%
	Mixed	40	44.0%	3	3.3%	30	33.0%

3.4.1. Studies that identified their own limitations

As Table 25 shows, scarcely more than one-third of all the evaluation studies in this review (38.4%) identified their own limitations. Surprisingly, the proportion was lower (31.5%) in the academic literature, where the standards for publication are supposedly stricter. The proportion in the grey literature was higher (43.1%), but still less than half. Thus most of the evaluation studies in this review (nearly two-thirds) did not clearly identify any limitations, which represents a weakness in their methodology. The inherent difficulties of the evaluation process, discussed in the introduction to this review, suggest that some major limitations must have been observed and clearly recognized in the overwhelming majority of these evaluations. We can therefore speculate that the reason that so many of the authors do not report any such limitations is not so much that they did not encounter any difficulties as that they chose not to mention them. One possible reason for this choice is that the evaluators may be dependent to varying degrees on the people who commission and fund the prevention programs. The resulting pressure to make positive findings might lead the evaluators to hold back information about these problems so as not to diminish the impact of their findings. The proportion of internal evaluation studies that identified their own limitations was 35.3%, but the proportion for external evaluations, which are supposed to be less subject to such pressures, was not that much higher (40.5%), which suggests that this lack of information is fairly common across the entire field.

The proportion of evaluation studies that identified their own limitations varied considerably with the continent where they were conducted. About half of the studies did so in Australia (57.1%), Africa (50%) and Asia (50%). But the figures were far lower for North America (16.7%) and Europe (28.6%). In other words, in these parts of the world that have such great research and evaluation traditions, the limitations in evaluation studies were identified significantly less often. This finding may seem surprising, given that many of the difficulties involved in evaluating PVE programs have been identified by researchers based in Western countries. But it resonates with what many of them have said about the challenge of keeping program evaluations independent.

The proportion of evaluation studies that identified their own limitations also varied with the prevention levels of the programs evaluated. The more general or universal this level, the lower this proportion. Thus, limitations were identified in only 33.3% of the evaluation studies on primary prevention programs, 31.5% of those on targeted primary prevention programs, and 20.8% of those on programs whose prevention level we classified as “General”. In contrast, limitations were mentioned in 55.7% of the evaluation studies on secondary prevention programs and 50.0% of those on tertiary prevention

programs. It is understandable that authors of these last two types of studies say more about their limitations, because their field of action is more circumscribed, the programs’ desired effects are better identified, and the risks associated with the programs are greater. But in light of the negative effects of certain programs that specifically target certain communities, it is also possible that researchers who are evaluating more general programs simply encounter fewer difficulties in collecting and analyzing information than researchers who are evaluating more specific ones.

As regards methodological design, Table 25 shows a significant difference: limitations were mentioned in 46.7% of the quantitative studies but only 31.3% of the qualitative studies. This finding may seem surprising, given that qualitative studies face the obvious constraint that the information that they analyze depends on the subjectivity of the participants and the analysis.

Lastly, classified according to the type (objective) of the evaluation, information on limitations was provided in a slightly higher percentage of those studies that included a section dedicated to output evaluation than in those that evaluated programs’ impacts or processes. Output evaluations thus generally reflect a greater awareness of the limitations of this kind of evaluation.

3.4.2. Studies with reported conflicts of interest and unreported potential conflicts of interest

As Table 25 shows, only 14 (6.4%) of all the evaluation studies in this review report conflicts of interest. Most of these 14 studies come from the grey literature. Four of them allude to the existence of conflicts of interest but do not specify their exact nature. Khalil et al. (2019) state that they worked for an organization that had some influence over the program that they were evaluating. Sabir (2014) states that he had previously been wrongfully arrested and detained for acts of terrorism when he searched for information on the Internet, and that this was his motivation for undertaking the evaluation of the program in question. In short, in practice, with some very rare exceptions, most PVE program evaluation studies say little or nothing about the nevertheless essential issue of conflicts of interest.

In this review, we therefore also analyzed the evaluation studies for potential conflicts of interest that their authors did not mention. The results of this analysis should of course be regarded very cautiously, because it relied on the fragmentary information available in these publications. In any event, as Table 25 shows, we identified potential conflicts of interest in 20.5% of all the studies in this review. To take the analysis further, we classified these potential conflicts of interest into three distinct categories. The first category comprises conflicts that may exist when at least one of the people who

evaluated the program also participated directly in its development or implementation; this category accounts for 53.3% of the 45 cases that we identified. The second category comprises conflicts that may exist when at least one of the people who evaluated the program works for the organization that delivers it, without necessarily belonging to the team that designs and implements it (35.6%). The third category concerns the source of the funding for the evaluation: 6.7% of the evaluations were funded by the same organization that implemented the program.

Here, instead of providing percentages by category as we did for the studies that identified their own conflicts of interest, we consider it more helpful to look at the ratio between the number of studies where we identified potential conflicts of interest that the authors did not and the number where the authors reported such conflicts themselves. For all of the studies in this review combined, this ratio was 3.2 to 1. But the figures for various categories varied widely.

Among internal evaluations, for every study where the authors did report conflicts of interest, we found 12 other studies with potential conflicts of interest that they did not report. This is all the more disturbing in that internal evaluations are precisely the kind that should be most attentive to these issues. Among external evaluations, on the other hand, the number of studies with unreported potential conflicts of interest was smaller than the number that reported such conflicts themselves.

Among the three kinds of methodological designs, mixed studies (combining qualitative and quantitative methods) were the most extreme case: for every study where the authors did report conflicts of interest, we found 10 others with unreported potential conflicts of interest. For purely qualitative studies, on the other hand, the number with unreported potential conflicts was smaller than the number where conflicts were reported.

We found lower but still disturbingly high ratios for studies of programs targeting right-wing violent extremism and Islamist violent extremism and studies published in the academic literature rather than the grey literature: 6 to 1, 5.5 to 1, and 5.8 to 1, respectively. This last finding is surprising, given that so many academic journals require any conflicts of interest to be reported in the studies that they publish.

3.4.3. Types of limitations described by authors

In the preceding sections, we have emphasized how few of the studies in this review—only 84 out of 219—described their own limitations. But now we will examine the nature of the limitations that these 84 studies did describe. This information provides useful insights into the problems encountered in evaluating PVE programs in the field. As Box 4 shows, we have classified these limitations into five categories: methodology, analysis, evaluators, programs and practitioners, and externalities.

Box 4. Limitations identified by authors of evaluation studies

1. Methodology

a. Design

- i. Exploratory designs
- ii. Lack of pre- and post-measurements

b. Indicators

- i. Limited
- ii. Unsuitable

c. Data collection

- i. Data based on perceptions
- ii. Credibility of information obtained
- iii. Limited or unsuitable data-collection tools
- iv. Limited access to data
- v. Incomplete information

d. Participants

- i. Small samples
- ii. Unrepresentative samples
- iii. Lack of control groups
- iv. Homogeneous samples
- v. Social-desirability bias
- vi. Reluctance to share information

2. Analyses

- a. Analytical capacity
- b. Depth of analyses
- c. Generalization
- d. Types of analyses performed
- e. Types of results obtained
- f. Causal relationships
- g. Sensitive information

3. Evaluators

- a. Lack of independence
- b. Subjective judgments in analyses
- c. Problems of translation

4. Programs and practitioners

- a. Programs
 - i. Delivered over short periods
 - ii. Lack of clear definitions
 - iii. Funding
- b. Practitioners
 - i. Commitment
 - ii. Security
 - iii. Role of outside stakeholders

5. Externalities

- a. Time limitations
- b. Contextual and security limitations
- c. Budget limitations

1. Limitations regarding methodology

When authors did identify the limitations of their own evaluation studies, methodological limitations were the type they mentioned most often. These limitations had arisen in most phases of the evaluation process, such as designing the study, selecting the indicators and the sample participants, and collecting the data.

In discussing design limitations, some authors said that the exploratory design of their studies made their analyses overly general and prevented them from drawing any conclusions about the effectiveness of the programs that they were evaluating. For example, Glazzard and Reed (2018) drew their conclusions from a sample of local programs in Europe without conducting any fieldwork or a comprehensive analysis of each program. Study designs that did not include pre- and post-measurements also limited the authors' ability to evaluate the programs' impacts. Often, as in Johns, Grossman and McDonald (2014), the reason for this omission was that the evaluation began after the program was already under way.

Other methodological limitations identified by study authors included choosing indicators that measured only one aspect of the relevant reality. Garaigordobil (2012), for example, states that the approach to violence in her study was based on individual psychological characteristics and so did not explain this phenomenon fully, because there are social and psychosocial variables that should also be taken into account. In other studies, the indicators chosen were not appropriate for assessing whether the specified program objectives had been met. For instance, Van der Heide and Schuurman (2018) state that they had to use recidivism rates as an indicator of deradicalization. McDowell-Smith, Speckhard and Yayla (2017) report that they focused mainly on one aspect of a counter-narrative campaign (whether the counter-narratives were effective in supporting or increasing existing negative attitudes towards ISIS) instead of taking a broader perspective (such as whether they were effective in increasing participants' interest in ISIS).

Data collection is another essential phase of the evaluation process. Some authors of evaluation studies in this review mentioned challenges and difficulties that they had had with the following aspects of data collection and that might have limited their ability to interpret the data properly: a) the type of data collected, b) the credibility of the information obtained, c) the tools that they had used to collect these data, d) limited access to data and e) incompleteness of the data.

a) The type of data collected obviously has a considerable influence on the analyses. For example, some researchers believe that statements obtained from interviews, surveys or internal documents are more likely to reflect the perceptions of program designers, practitioners and participants than day-to-day operational reality (Gatewood and Boyer, 2019; Hirschi and Widmer, 2012;

Rodon, 2018; Wilson and Krentel, 2018). These perceptions are highly subjective, raising the question of how much these data can be relied upon.

b) The credibility of the information obtained in data collection for PVE evaluations is an obvious limitation. There are many possible sources of bias in such data: the sensitivity of the subject from a political and security standpoint (Sarota, 2017), inaccuracies in the memories of the people interviewed (Khalil et al., 2019) and sometimes of the researchers interviewing them (Tsuroyya, 2017), and contradictions in the information collected from different interviewees (Cockayne et al., 2015) or different documents (Vittum et al., 2016).

c) The kind of tools used to collect the data may also limit the usefulness of the results—for example, in surveys where only a small number of questions could be asked (Parker and Lindekilde, 2020) or where biases were introduced by the translation process, especially in non-English-speaking settings (Murtaza et al., 2018).

d) **Limited access to data is probably the problem that evaluators encounter most often in the field.** For example, many researchers report having been unable to gain access to key informants, such as program participants (Anindya, 2019; Jailobaeva and Asilbekova, 2017; Murtaza et al., 2018; Tines et al., 2017), members of a paramilitary faction (Dwyer and Maruna, 2011), members of the government (Jailobaeva and Asilbekova, 2017) and religious or community leaders (Muncy et al., 2015; Sarota, 2017; Wilson and Krentel, 2018). Harahap, Irmayani and Lubis (2019), for example, were unable to interview certain students whose parents were known terrorists, because it was considered taboo in the community. As a result, most of these evaluations were able to consider the views only of the program designers and practitioners. Some of these authors considered these sources sufficient, but others felt that the resulting data were incomplete, because they relied mainly on the perceptions of the people directly involved in the project. For example, Wilson and Krentel (2018, p. 4), identified inability to gather information from communities as one of the limitations of their evaluation, stating that “The evaluation team was unable to verify practitioners' self-reported data about their project activities through observation or follow-up interviews.”

Some evaluators provide the reasons for their problems in accessing key informants. These reasons include the short time provided to conduct evaluations and difficulties in coordinating the timing of evaluations with the availability of certain actors in the field. In conflict areas, insecure conditions and access restrictions imposed by government can pose an obstacle to conducting evaluations properly. Lastly, other researchers mention logistical problems, such as participants' having changed their mobile phone numbers or being away on vacation.

In other studies, documentation on programs was missing or incomplete, because the programs and organizations involved did not know how to organize the information that they produced (Ipp et al., 2014; Muncy et al., 2015; Rodon, 2018; Williams et al., 2016). Rodon (2018, p. 111) explains this situation as follows:

[Translated from French] The analysis identified limitations in the data collected that were due to the procedures by which they had been compiled. This illustrates just how much the work of evaluation depends on how rigorously the organizations conducting programs define what they can and cannot consider program outcomes. In other words, the degree of professionalism of the programs, particularly as regards their methods and procedures for compiling information, affects the scope of the findings that can be made and the conclusions that can be drawn from impact evaluations. (2018, p. 111).

In addition to this lack of systematization, lack of baseline data, especially at the start of programs, can prevent comparisons with the collected data (Ipp et al., 2014; Schumicky-Logan, 2017; Swedberg and Reisman, 2013). In some cases, the researchers had to rely on the memories of the actors in the field to reconstruct these data. This process introduces obvious biases into the evaluation inasmuch as these data cannot be verified independently (Ipp et al., 2014). Obviously, PVE programs are designed not to meet the needs of researchers and evaluators, but rather to take actions that may prevent violent extremism. But the lack of awareness of evaluation issues among some actors in the field can significantly affect the evaluation of these actions. This observation confirms the importance of including teams of evaluators from the earliest design stages of programs, and of making practitioners more aware of evaluation issues.

e) One last issue regarding data collection concerns incompleteness of the data gathered, due in particular to low response rates and loss of information. Low response rates seem to be a fairly common problem in PVE evaluations, particularly in connection with surveys (Beider and Briggs, 2010; Hiariej et al., 2017; Hirschi and Widmer, 2012; Schorn et al., 2010; Tines et al., 2017). For example, Hirschi and Widmer (2012) state that in one of the programs for preventing right-wing extremism that they evaluated in Switzerland, only four out of the 107 teachers whom they contacted completed their questionnaire. Schorn et al. (2010) explained the low response rate in their study as a case of “survey fatigue” in which the people in charge of the program terminated

the impact evaluation that these authors were attempting to conduct. Another possible reason for the reported low response rates is that most of the studies in question conducted their surveys online, by telephone or by email rather than in person. The researchers also mentioned lack of interest in the subject, lack of time and the problems of access to the field mentioned earlier.

The main cases in which data were incomplete because information had been lost occurred in evaluations of online PVE programs, when social media accounts were deleted. (This issue is discussed in more detail in Box 5.)

Lastly, many of these evaluation studies reported methodological limitations related to the participants. In this regard, the limitation most often cited was the sample size, which was often too small for the type of analysis that the researchers wanted to perform or the kind of results that they wanted to obtain.³⁸ As mentioned earlier, some studies did not state how many participants they had in their samples, while others evaluated programs with as few as three participants.³⁹ Independently of sample size, representativity is an important issue for many studies, especially quantitative ones. In fact, it is rare to find representative studies, and most of them are based on non-probability samples selected for their ease of access or determined by the participants' own desire to get involved. This leads to certain biases in the selection of participants, which can interfere with the interpretation of the data, and in particular with efforts to establish causal links between program actions and program outcomes. Lack of control groups is also frequently mentioned as an obstacle to establishing causal links. Only 22 of the studies in this review used control groups; as a result, many of the studies were exploratory, general and descriptive and hence could not readily be used to evaluate programs' impacts.

In qualitative studies, the issue is not so much one of representativity as one of homogeneity or lack of diversity among the actors questioned.⁴⁰ In some studies, certain key actors were missing, or the persons interviewed did not necessarily match the profile of the population that the practitioners had been trying to reach.

In addition to sampling problems, some authors identify a number of problems related to the participants' motivations. One such problem is social desirability bias: participants' desire to cast themselves in a favourable light and give the researchers the answers they want to hear (Azam and Bareeha, 2017; Hiariej et al., 2017;

³⁸ Awan, 2012b; Busher, et al., 2017; Cherney and Belton, 2019; Gatewood and Boyer, 2019; Hirschi and Widmer, 2012; Joyce, 2018; Kollmorgen et al., 2019; Kollmorgen and Barry, 2017; Madriaza et al., 2018; Mansour, 2017; Octavia and Wahyuni, 2014; Schorn, Moubayed and Auten, 2010; Sjøen and Mattsson, 2019; Tines et al., 2017; Warrington, 2018.

³⁹ Half of the studies reporting their sample size had 56 participants or fewer.

⁴⁰ Anindya, 2019; Awan, 2012b; Azam and Bareeha, 2017; Bala and Deman, 2017; Busher et al., 2017; Hirschi and Widmer, 2012; Kollmorgen et al., 2019; Mansour, 2017; Muncy et al., 2015; Sarota, 2017; Schorn et al., 2010; Sjøen and Mattsson, 2019; Tesfaye, McDougal, Maclin and Blum, 2018; Warrington, 2018.

Johns et al., 2014; Khalil and Ipp, 2016; Kollmorgen et al., 2019; Kollmorgen and Barry, 2017). For example, Azam and Bareeha (2017) state that in their evaluation of a rehabilitation program, participants' positive responses concerning the program might have reflected their motivation to be considered rehabilitated. Hiariej et al. (2017) felt that the information that they had obtained from the government officials who had participated in their evaluation was not entirely reliable, because they tended to provide responses that were socially desirable and probably politically neutral. In an evaluation of the impact of a vocational training program in Afghanistan (Kurtz, 2015, p 15), the evaluators noted that participant response biases might have included under-reporting of employment and income variables in hope of receiving additional training or resources, or over-reporting to validate the program organizations' perceived desires.

All of the preceding examples raise the question of how much the information supplied by program participants can be trusted. But on the other hand, it is essential for evaluators to gain participants' trust so that they can overcome their natural reticence and feel free to share their impressions about such a complex subject as violent extremism. In this regard, Kurtz (2015, p. 15) writes, "we acknowledge the possibility of bias in responses as respondents may be suspicious of why the data was being collected, especially if being gathered by a stranger" (2015, p. 15). Actors' reluctance to participate in PVE evaluations may also be the result of action or lack of action by the programs themselves. In their program evaluation, for example, Pipe et al. (2016) encountered hostility from community members, because at the time of the evaluation, the program had carried out only one of its five planned initiatives. Lastly, Kurtz (2015) writes that in the context of intensified violence and conflict in which his evaluation was conducted, fear of reprisals from government or opposition forces may have deterred some individuals from providing frank responses to certain questions.

2. Limitations regarding analyses

Once the data for an evaluation study have been collected, other kinds of limitations may arise when they are analyzed. One such limitation, when researchers have gathered large volumes of information (especially in qualitative studies) is that they may not have enough capacity to analyze them. For example, Dietrich (2018) assessed the impact of the "White Dove" radio project for countering violent extremism (CVE) in Nigeria, in which a CVE messaging hub was used to deliver three original radio series. The evaluators conducted direct interviews and focus groups with a total of 824 people and also documented listeners' reactions to the first 44 episodes of the radio series. Because the evaluators had thus gathered such a large volume of information, they were able to analyze only part of it, in particular the listeners' reactions.

In other cases, the analyses may not go into sufficient depth and hence may provide only a descriptive overview of the program and its capabilities. (Nicolls and Hassan (2014, p. 50), for example, regret that "it was not possible to uncover all disparities, factors, and influences" on the implementation and outcomes of the initiative that they were evaluating.) As a result, in such analyses, disaggregation of data into more specific categories is very limited.

Another limitation with some analyses is the tendency to over-generalize from data collected through a process that was subject to the limitations mentioned above (small samples, limited access to data, limited time to conduct the evaluation, etc.). In such cases, the findings and lessons learned actually apply only to the specific program, the specific people who participated in the evaluation, or the specific geographic areas where the program was delivered. This limitation in turn limits policymakers' ability to draw conclusions about the programs evaluated and their transferability to other parts of society and other parts of a country.

The type of analysis performed also influences the extent to which the findings from an evaluation can be generalized. For example, Williams et al. (2016) state that their use of inferential/probabilistic statistics introduces an element of uncertainty as to whether their findings can be generalized beyond the particular program that they were evaluating. In other cases, the problem may be with the results themselves, which may be statistically significant but have a low effect size (Parker and Lindekilde, 2020). In a causal impact evaluation, the effect size measures the strength of the effect of the intervention on the desired outcomes. The question here, and in many other evaluations, is thus whether the type of analysis that the researchers used was capable of establishing a causal relationship between the programs in question and their outcomes (Cherney and Belton, 2019; Demant, Wagenaar and van Donselaar, 2009; Education Development Center (EDC) and USAID, 2019; Wilson and Krentel, 2018).

Lastly, some analyses have limitations attributable to the inherent nature of the information that they are dealing with. For example, Cherney and Belton (2019) state that because of the sensitive information gathered in their evaluation, only one of the evaluators coded the data, which increases the subjectivity of the analyses performed.

3. Limitations regarding evaluators

In this systematic review of PVE evaluation studies, three limitations regarding the evaluators came up repeatedly: their limited independence from the programs that they were evaluating, the subjective judgments that entered into their analyses, and problems of translation. All three of these limitations introduced analytical biases that may have produced an inaccurate picture of the results actually obtained.

Many of the studies in this review raise the question of whether the evaluators had the necessary independence and distance from the program to evaluate it properly (Broadbent, 2013; Joyce, 2018; Rodon, 2018; Wilson and Krentel, 2018). It is not unusual for research teams to have been involved from the early development stage of the programs that they subsequently evaluate (Broadbent, 2013; Madriaza et al., 2018b; Wilson and Krentel, 2018). In fact, such involvement is actually recommended by practitioners and researchers, as a way of bridging the gaps between the teams and compensating for outside evaluators' lack of field knowledge (Madriaza et al., 2021). However, this involvement can lead the evaluators to form a positive opinion of a program while neglecting the aspects that should be improved.

Similarly, and especially in the qualitative studies in this review, the evaluators mention the limitations of analyses that are based on their *own judgments* and therefore reflect their own views. Azam and Bareeha (2017, p. 7) state this more clearly: "Since the analysis and tools utilized were qualitative in nature, the study is based on the author's own judgments and analysis, which is open to scrutiny."

Problems of translation were mentioned earlier, in the section on methodological limitations. Such problems typically arise when foreign evaluators conduct evaluations in countries where they do not speak the language and must therefore rely on translators or interpreters to conduct interviews in the field (Dhungana et al., 2016; Kollmorgen and Barry, 2017; Murtaza et al., 2018; Swedberg and Reisman, 2013). This limitation is more complicated because it involves researchers who did not have direct access to the field, who had to rely on the accounts of intermediaries, and who could not read documentation about the program if it was written in a language other than their own (which was usually English). This situation comes up repeatedly in programs funded by the United States Agency for International Development (USAID). Although USAID reports have been a very rich source of information for this systematic review, the translation issue does raise some questions about the usefulness of this kind of evaluation and the reliability of the data analyzed for the countries where USAID operates, particularly in Africa, the Middle East, and south and central Asia.

Lastly, in addition to the above three main types of limitations regarding evaluators, two other types were mentioned in the evaluation studies included in this review: gender bias and researchers' health problems. On the subject of gender bias, the studies' authors have surprisingly little to say, possibly because the gender balance among these authors is relatively good. Only Kollmorgen and Barry (2017) mention gender bias, noting that the only male member of their evaluation team had had to leave when the evaluation was just starting. Regarding evaluators' health problems, Bean et al. (2011)

mention that the time allotted for their evaluation was disrupted by health problems that the team experienced in the field.

4. Limitations regarding programs and practitioners

Another category of limitations involves the problems that evaluation teams encounter with the programs that they are evaluating and the practitioners who deliver them.

Many evaluators mention that the shortness of the period over which a program was delivered prevented them from drawing more definite conclusions about how effective it was (Demant et al., 2009; Dhungana et al., 2016; Jailobaeva and Asilbekova, 2017). One good example was a program for preventing right-wing extremism, evaluated by Demant et al. (2009). This program lasted less than a year, because it was a pilot project as part of the municipal PVE strategy of the city of Winschoten, in the Netherlands.

Another program-related limitation is that, as mentioned earlier, some programs do not document their activities systematically enough and do not make certain definitions clear enough for more detailed conclusions to be drawn. For example, Bastug and Evlek (2016) evaluated a pilot program for individual disengagement and deradicalization in Turkey. But this program failed to make any clear distinction between the measures that it took to achieve disengagement and the ones that it took to achieve deradicalization, so that the two types often overlapped.

A program's funding can also influence its evaluation. In evaluating a CVE training program in Australian schools, Harris-Hogan et al. (2019) were unable to analyze the results for certain jurisdictions in the country, because there was no central funding for this program from the federal government.

The three types of limitations that the studies' authors identified regarding practitioners involved their degree of commitment, their safety, and outsourcing of program activities to outside consultants. Regarding the practitioners' degree of commitment, Frenett and Dow (2015) describe a direct online intervention program (see Box 5) in which the practitioners had other, full-time jobs and were participating in other, similar projects, but had to spend a great deal of time on the interventions (which were exhausting) and received minimal remuneration for their efforts. Another issue that Frenett and Dow mention regarding this program is the practitioners' safety: they were exchanging messages directly, via Facebook, with persons who were considered extremists, which made the practitioners concerned for their own safety. This is a very important subject in the field of PVE but has been neglected in the literature (Madriaza et al., 2017). The last

point, regarding outsourcing of some program activities to private consultants, was raised by Schuurman and Bakker (2016) with regard to a reintegration program in the Netherlands. This outsourcing limited the researchers' ability to draw conclusions about the program, because

they did not know what influence these outside practitioners had on the people receiving the program's services.

Box 5. Difficulties in evaluating online PVE programs

Evaluating online PVE programs raises very different issues from evaluating conventional ones. A case in point was an evaluation by Amanullah and Harrasy (2017) of 18 counter-narrative video campaigns developed with the support of the Institute for Strategic Dialogue (ISD) to fight extremist recruitment and propaganda on social media platforms in Kenya. During the course of these video campaigns, an umbrella Twitter account used for anonymous campaign dissemination was unexpectedly blocked by Twitter, limiting the evaluation team's ability to collect data for certain videos.

In another ISD online project, practitioners interacted directly, via Facebook, with people who openly expressed extremist views online, but the evaluators encountered various problems with the new technologies involved (Frenett and Dow, 2015). First, during the project, the tool used to search Facebook profiles for candidates for interventions (Graph Search) became increasingly limited, which slowed the pace at which candidate accounts could be identified. The project was also affected by the removal of some profiles by Facebook. Out of the 154 profiles originally identified as candidates, 42 were removed by Facebook in the course of the project—most of them Islamist rather than Far Right profiles.

A similar situation arose with evaluation of the Redirect Method campaign implemented by Moonshot CVE (Helmus and Klein, 2018). The Redirect method uses Google AdWord to identify people who are conducting Google searches for violent extremist content, then exposes them to an advertisement in their search results that links to counternarrative videos. But Moonshot CVE's campaign was temporarily terminated several times when Google tried to limit advertising on racist search terms. Helmus and Klein also point out that any evaluation based on a social media program depends on the self-selected nature of users who choose the content that they consume, and that comparisons between users who were exposed to the campaign and users who were not may thus be biased.

These difficulties demonstrate the importance of coordinating with all key actors, including major technology firms, not only when implementing online PVE programs, but also when evaluating them.

5. Limitations regarding externalities

The last category of limitations identified by the studies' authors concerned factors that were external to the programs and to the evaluation process but nevertheless had a considerable effect on the interpretation of the data collected. These limitations fell into three categories: time limitations, political and security limitations, and budget limitations.

Time limitations

As stated previously, time is a key factor that affects the various dimensions and stages of an evaluation. It is one of the factors mentioned most often in the studies in this review (Azam and Bareeha, 2017; Bean et al., 2011; Boyle et al., 2016; Hiariej et al., 2017; Muncy et al., 2015; Schorn et al., 2010). Often, the evaluators state that they were given a very short time to collect enough information for the evaluation or to visit all the sites that they had to evaluate. Hiariej et al. (2017, p. 17) clearly illustrate the time pressures that their evaluators faced when using a

paper questionnaire to collect survey data at five different sites in Indonesia, in addition to reviewing the program documentation, conducting interviews and running focus groups:

The survey cannot be conducted simultaneously because the evaluation can only mobilize limited numbers of researchers and local assistants. The time allocated for the survey was around 5-6 days for data collection in each city. Added with 1-2 day interval between each survey, school and national holidays, and other technical delays on the field the survey needed almost 3 months just for collecting all distributed questionnaires.

Having to conduct program evaluations in a very short time has obvious implications for what they can achieve. In particular, they can evaluate only the immediate effects of programs and not their longer-term effects, because evaluating the latter would require longitudinal study designs.

But the time allotted for evaluations is only one form of time limitation. Another, described in the preceding section, is how hard it is to assess the impact of programs that are delivered over very short periods. Still another is timing—whether the actors who need to be interviewed in the field are available during the period scheduled for the evaluation. In some cases, ordinary situations such as school holidays or personal vacations can affect whether an evaluation proceeds smoothly (Hiariej et al., 2017; Nicolls and Hassan, 2014). In others, the issue is the limited availability of particular individuals whom the evaluators need to interview, such as government officials or the people who put the program in place (D. Parker and Lindekilde, 2020). It must be recognized that evaluation is often a low priority for the actors in the field, except for the evaluators themselves.

Lastly, time also becomes a factor when the evaluators have to rely on people’s memories to reconstruct the history of a program, with the attendant risk of memory bias (Hiariej et al., 2017; Ipp et al., 2014; Khalil et al., 2019; Khalil and Ipp, 2016; Kollmorgen and Barry, 2017; Tines et al., 2017; Vittum et al., 2016). This can happen, for example, when there is no actual baseline information against which to compare the results of a program and the evaluators have to rely on practitioners’ memories to attempt to reconstruct such information (Ipp et al., 2014). It can also happen when an evaluation is conducted several months after a program ends (Khalil and Ipp, 2016; Kollmorgen and Barry, 2017) or when the evaluators have had to ask questions about the participants’ life histories or their status before the program began (Khalil et al., 2019; Tines et al., 2017). For example, to evaluate a disengagement program, Khalil et al. (2019) had to gather information from people who had disengaged from Al-Shabaab several years earlier.

Political and security limitations

Although violent extremism is a political phenomenon, the political dimension is rarely addressed in PVE program evaluations. The only author in our review who mentions it is Sarota (2017), who believes that political biases were part of the reason that some respondents did not express their opinions in a fully open way. In contrast, many of the authors identified security as a key factor, particularly in countries where there is open conflict or that are in post-conflict situations. In such situations, the evaluators may be physically unable to visit some sites (Bean et al., 2011; Khalil et al., 2016; Pipe et al., 2016; Tesfaye et al., 2018) or unable to work directly in the field. For example, Swedberg and Reisman (2013, p. 14) write: “Due to the poor security situation and logistical challenges in parts of Somalia, the evaluation team may not be able to conduct the field work directly.” Lastly, insecure conditions may also make some respondents more leery of answering questions from people whom they do not know (Basse, 2018).

Budget limitations

Budget limitations were the third type of external limitation to which authors in this review alluded. But even though budgets are often regarded as a constraint on research, they were mentioned in only two of the studies in this review (Bean et al., 2011; Mansour, 2017). In both cases, the authors confined themselves to a brief reference to the impact that the amount allocated for their evaluations had on the extent of their data collection—that is, the number of interviews that they conducted or the number of sites that they visited. According to these researchers, because of this limitation, they were unable to fully evaluate the broader impact that the programs had had on the people whom they were intended to benefit.

3.5. QUALITY OF THE STUDIES REVIEWED

As noted at the start of this document, methodological quality has been an important concern both in security studies in general and in PVE program evaluation studies in particular. For example, in his initial analyses of counterterrorism studies, Silke (2001, 2006) showed that they generally tended to use secondary data and not to use quantitative designs. In the specific field of PVE program evaluation, a high proportion of past literature reviews either did not analyze methodological quality (Feddes and Gallucci, 2015; Gielen, 2017; Mastroe and Szmania, 2016), or did so using a general scale that did not necessarily account for the particularities of the various methodological designs used in this field (Bellasio et al., 2018; Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021) or included only quantitative studies (Carthy et al., 2020; Lum et al., 2006). These analyses thus incorporated a bias in favour of quantitative/experimental designs applied to a particular type of evaluation (impact evaluations), which in principle excluded any analysis of the quality of evaluation studies that used qualitative designs or involved process evaluations. For example, the review by Bellasio et al. (2018) applied the Maryland Scientific Methods Scale, which determines the quality of studies on the basis of the type of design that they employ and operates on the assumption that experimental designs provide better evaluations than any other kind. As we shall see in the following pages, this is far from true when it comes to evaluating PVE programs.

The 219 studies that we included in this review, all of which analyzed primary data, serve as clear evidence that the field of PVE program evaluation has evolved tremendously and its overall quality has improved substantially, at least as regards use of primary data. But that is not enough to draw any conclusions about the overall quality of these studies’ findings. To do that, we performed a detailed analysis of the methodological quality of these studies, and we present our findings in the following pages.

As stated in section 2.4, to assess the methodological quality of the studies in this review, we used the Mixed Methods Appraisal Tool (MMAT) (Hong et al., 2018; Hong and Pluye, 2019). The MMAT can be used to evaluate studies with the following five types of designs: quantitative descriptive, qualitative, experimental, quasi-experimental and mixed (in this last case, the MMAT also measures how effectively the qualitative and quantitative methods have been integrated).⁴¹ This tool comprises five scales—one for each design type—and each scale consists of five questions that represent methodological quality criteria. In our quality analysis of each study, we coded the answers to these questions as 1 for Yes, 0 for No, and “Can’t tell” when the study did not contain the information needed for us to make a determination.⁴² We then summed all of the coded values of 1,⁴³ thereby assigning the study a quality rating on a scale of 0 to 5. Given that each study could potentially use all five of the methodological designs just mentioned and that the number of designs used varied from one study to another, we did not analyze each study individually as a whole, but instead separately analyzed the sections dedicated to each type of design. We thus analyzed some studies more than once, depending on how many designs they used.

3.5.1. Quality of the qualitative studies

According to the MMAT manual, “(q)ualitative research is an approach for exploring and understanding the meaning individuals or groups ascribe to a social or human problem” (Creswell, 2013b, cited in Hong et al., 2018, p. 3). Out of the 219 studies in this review, we found 188 that had used a qualitative methodological design, at least in part.

Box 6. MMAT methodological quality criteria for qualitative studies

1. Is the qualitative approach appropriate to answer the research question?
2. Are the qualitative data collection methods adequate to address the research question?

3. Are the findings adequately derived from the data?
4. Is the interpretation of results sufficiently substantiated by data?
5. Is there coherence between qualitative data sources, collection, analysis and interpretation?

The mean quality score for these 188 qualitative studies was 2.86, just slightly above the midpoint on the scale (2.5). Most of these studies (59%) received a score of 2 or 3, but roughly one out of every four (28.2%) received a very high score (4 or 5 on the scale). As Figure 17 shows, the problem with the qualitative studies lies more in their interpretation of the data than in the appropriateness of their approach or the adequacy of their data-collection methods (the scores for these last two criteria were quite high). In particular, in many cases we found that their interpretation of their results was not sufficiently substantiated by their data. In contrast, for the questions of whether the findings were adequately derived from the data and whether there was coherence between data sources, collection, analysis and interpretation, about half of the studies did not provide enough information for us to tell whether the answer was Yes or No.

Whether these last three quality criteria have been met can be corroborated in whole or in part by direct quotes from the people interviewed. Hence the studies’ failure to provide such corroborating information may have been due to a lack of space. Qualitative studies generally need more space to develop ideas and provide direct quotes, which is not necessarily the case for quantitative studies. Scientific journals limit the length of the articles that they publish, which may directly affect how much detail the authors provide. This speculation is confirmed by the higher mean quality score received by the qualitative studies in this review that appeared in the grey literature as opposed to the academic literature. In the grey literature, space limitations are not a problem, and the results section of a study may include all of the details needed to determine whether the interpretation of the data was justified.

⁴¹ In this systematic review, we use the terms “experimental design” and “quasi-experimental design” as shorthand for the actual MMAT terms: quantitative randomized controlled trials and quantitative non-randomized trials, respectively. In sections 3.5.1 through 3.5.5, we briefly describe each design type before discussing the quality of the studies in this review that used it.

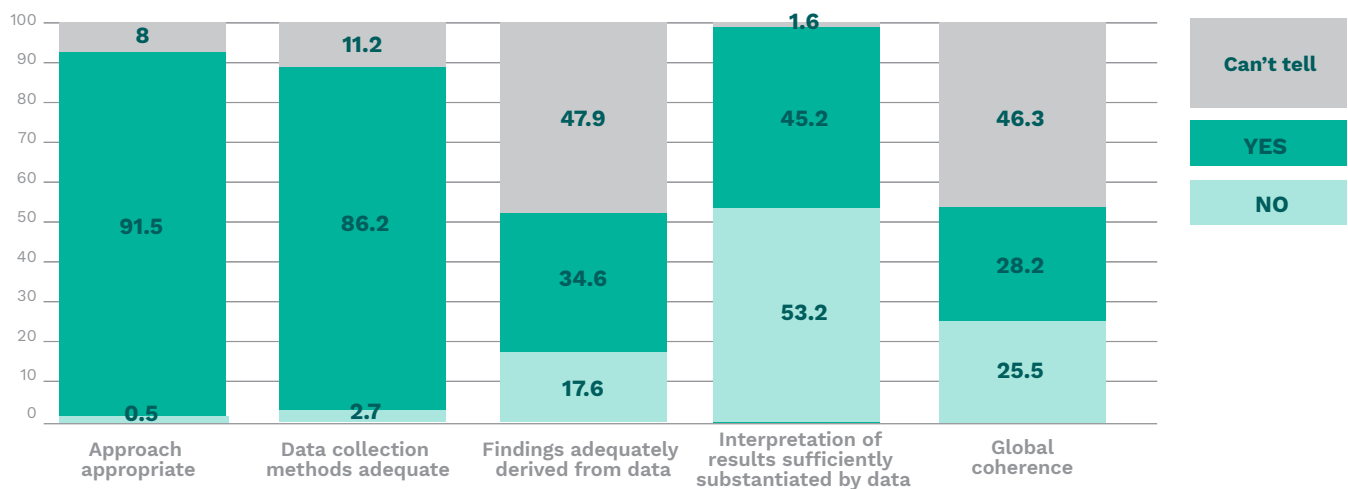
⁴² For example, when a study contained no methodology section, or only a very limited one, or when the nature of the study prevented us from determining whether the answer for a given criterion was Yes or No.

⁴³ We do not recommend using such scores as a criterion for inclusion in systematic reviews that examine the effects of the programs evaluated. We used these scores not as an inclusion criterion, but as a variable to describe the quality of the studies included in this review.

Table 26. Methodological quality scores for qualitative studies

Variable	Category	Number of studies	Mean	Std. dev.
Total		188	2.86	1.47
Type of publication	Academic literature	67	2.69	1.62
	Grey literature	121	2.95	1.38
Continent	Africa	43	3.00	1.29
	North America	15	3.13	1.68
	Asia	37	2.11	1.49
	Europe	86	3.03	1.44
	n/a	2	2.00	1.41
	Australia	5	3.60	1.34
Prevention level	Primary	38	2.63	1.63
	Targetted primary	75	3.12	1.50
	Secondary	54	3.00	1.29
	Tertiary	44	2.68	1.38
	General	21	2.76	1.58
Type of violent extremism	Right-wing	14	2.50	1.16
	Islamist	77	3.04	1.45
	All types	109	2.80	1.48
Type of evaluation	Impact	130	2.92	1.39
	Process	109	2.95	1.47
	Output	24	3.25	1.45

Figure 17. Percentages of qualitative studies meeting the applicable MMAT criteria



By continent, the qualitative studies that received the highest mean scores for methodological quality were done in North America (3.13), Europe (3.03) and Africa (3.0). The results for Asia were more disappointing. The high mean score for Africa is impressive, rivalling other parts of the world that have a longer evaluation tradition.

But the scores for Africa must also be viewed with caution, because they may be biased. The MMAT was developed in North America on the basis of criteria on which there is consensus among researchers there and in Europe, which may explain why the mean scores are even higher for these two continents than for Africa. As we

have seen in section 3.2.2, a large proportion of the authors who did the program evaluations in Africa came from Europe and, especially, from the United States. They were trained in this tradition and used these criteria to conduct their evaluations. But although the method was described adequately in many of these studies, other issues cast some doubt on the quality and interpretation of the data collected, particularly in studies in Africa. For example, the MMAT does not apply some other basic criteria that we consider equally fundamental for evaluating the quality of a study. Two such criteria, discussed in the preceding section, are whether the evaluators spoke the language of the country where they were conducting the evaluation and whether they had access to the information they needed to conduct it.

Among the mean methodological quality scores by program prevention level in Table 26, the two highest were for evaluations of targeted primary prevention programs and secondary prevention programs. By type of extremism targeted, the highest score was for evaluations of programs aimed at Islamist extremism. By type of evaluation, it was for output evaluations. This last finding is more puzzling, because output evaluations are often highly descriptive and quantitative. But the studies that attempted to evaluate program outputs (among other aspects) tended to be published in the grey literature, so that their higher quality scores may be explained as described earlier in this section.

Encadré 7. Deux exemples d'évaluations qualitatives

Evaluation of the Greater Boston Countering Violent Extremism (CVE) Pilot Program (Savoia et al., 2016)

Savoia et al. (2016) evaluated the overall framework of the Greater Boston Countering Violent Extremism (CVE) Pilot Program, which attempted to help various communities to build their resilience and capacity to prevent individuals, including young people, from being inspired and recruited by violent extremists. The purpose of this evaluation was to gather formative evaluation data, from a public-health perspective, regarding the goals of the program and recommendations on how the program should evolve. The evaluators had three specific objectives:

- Gather opinions both on program goals and on the overall initiative
- Identify recommendations for practice
- Develop a logic model for the evaluation of violence prevention activities aligned with a particular grant application.

Thus the evaluators were not assessing the activities that this program carried out, but rather its overall framework. To do so, they used a “snowball” technique to identify individuals with a variety of perspectives and experiences related to the program, and then they interviewed these individuals. Next, the evaluators used a coding system to analyze the information from the interviews so as to meet their three specific objectives. The published study provides good explanations of all the design, collection and analysis phases, except for the sub-dimensions addressed within each objective. The authors always illustrate their interpretations of the data with direct quotes from the people whom they interviewed.

Prevent in Southwark – 2009-2010 Evaluation Report (Rooke and Slater, 2010)

Rooke and Slater (2010) evaluated the delivery of a local project in the borough of Southwark in London, England as part of Prevent, the British national strategy for preventing violent extremism. This project prioritized engaging the Muslim communities in the borough, understanding their needs, and building a strong network of Muslim organizations working in partnership with local agencies through a dialogical community development approach to prevent violent extremism in the borough. The aims of this study included evaluating delivery of the project over the preceding year, evidencing any successes and exploring the reasons for them, and looking at value for money in relation to project delivery.

To evaluate this project, the evaluators conducted interviews with key staff, did two case studies (on a radio show for young Muslims in South London and a TV documentary on Islam in Southwark), and ran a focus group. According to the evaluation, this project was a success in 2009-2010, because it was able to address all of the priority areas and recommendations set out in the preceding year's evaluation. Its achievements included carrying out a significant number of youth-based projects, developing inter-generational work that brought young people together with older adults, and strengthening networks between Muslim groups in Southwark.

The evaluation report presented its entire methodological approach coherently and succeeded where some other qualitative studies have more problems: in the relationship between the data collected and their interpretation. The report did not, however, mention the limitations of this evaluation.

3.5.2. Quality of the quantitative descriptive studies

The MMAT defines quantitative descriptive studies as being “designed only to describe the existing distribution of variables without much regard to causal relationships or other hypotheses” (Porta et al., 2014, p. 72, cited in Hong et al., 2018). Out of the 219 studies in this review, we found 60 that had at least one quantitative descriptive section (see Table 27 and Figure 18).

Box 8. MMAT methodological quality criteria for quantitative descriptive studies

1. Is the sampling strategy relevant to address the research question?
2. Is the sample representative of the target population?
3. Are the measurements appropriate?
4. Is the risk of non-response bias low?
5. Is the statistical analysis appropriate to answer the research question?

Table 27. Methodological quality scores for quantitative descriptive studies

Variable	Category	Number of studies	Mean	Std. dev.
Total		60	1.87	1.44
Type of publication	Academic literature	22	1.27	1.12
	Grey literature	38	2.21	1.51
Continent	Africa	12	2.42	1.56
	North America	8	1.00	0.53
	Asia	11	1.45	1.29
	Europe	19	1.84	1.61
	n/a	4	3.25	0.96
	Australia	6	1.83	1.33
Prevention level	Primary	16	2.00	1.46
	Targetted primary	28	1.57	1.14
	Secondary	23	2.22	1.48
	Tertiary	12	2.33	1.78
	General	1	3.00	
Type of violent extremism	Right-wing	7	2.00	1.73
	Islamist	28	1.93	1.30
	All types	30	1.83	1.53
Type of evaluation	Impact	49	1.90	1.39
	Process	30	2.03	1.52
	Output	11	1.91	1.30

Figure 18. Percentages of quantitative descriptive studies meeting the applicable MMAT criteria



Compared with the qualitative studies, the 60 studies that had at least one quantitative descriptive section had a lower mean methodological quality score: 1.87, which is below the midpoint of 2.5 on the scale. The vast majority of these 60 studies (66.7%) had scores of 0 to 2. Scarcely 15% had very high quality scores (4 or 5). The main problem with these evaluations is their lack of transparency about their methods. For example, about 60% of these studies did not provide the information that would have let us tell whether the sample was representative, whether the risk of nonresponse bias was low or whether the statistical analysis was appropriate, while about 50% did not let us tell whether the sampling strategy was relevant. In many cases, these studies presented their results without making much effort to explain how they were obtained. This limitation is all the more disturbing in that 84% of these studies had assessing the programs' impacts as one of their overall objectives.⁴⁴

In the cases where we could in fact tell whether a given quality criterion was satisfied, the results were not encouraging either. The criterion met by the highest percentage of the quantitative descriptive studies (58.3%) was whether the measurements used were appropriate, meaning whether they were valid, reliable and well suited to answering the research question (this last aspect was the one we were best able to assess when scoring the studies).

For the two criteria concerning the sample, we found that the sampling strategy was relevant to the research question in 43.3% of the studies, and the sample was representative of the target population in 28.3%. This last figure should be interpreted cautiously, however, because scoring a study for this MMAT criterion does not necessarily involve calculating the representativeness of the sample in relation to the target population. It is more a matter of assessing things such as whether the respondents matched the target population or whether the study clearly described the sample and the target population. Thus, even though this criterion was very broadly defined, only 28.3% of the studies met it.

For the 41.7% of the studies for which we could tell whether the risk of non-response bias was low, the response rates themselves were low, which obviously limited the quality of the statistical analyses concerned, even though in about one-third of these studies, these analyses were appropriate for answering the research question.

Among the 60 quantitative descriptive studies, the mean methodological quality score was clearly better for those published in the grey literature (just as was the case for the qualitative studies), those evaluating programs in Africa, and those evaluating secondary and tertiary prevention programs. The results for Africa are also consistent with those for the qualitative studies, and the

⁴⁴ As explained in section 3.3.1, we classified an evaluation study as being an impact evaluation when its authors had made an explicit statement to that effect somewhere in the study itself (objectives, research questions, statement of intent, etc.). We did not base this determination on our own judgment.

same caveats apply to analyzing these results. The African studies did better regarding the appropriateness of their measurements for answering their research questions and much worse regarding the representativeness of their samples. For 70% of the African studies, we could not tell whether the sample was representative.

In contrast, among the 60 quantitative descriptive studies, those of programs in North America received the lowest quality scores. For example, in 7 out of these 8

studies, we could not determine the response rate at all, and in the one remaining study, the response rates were low. Compared with evaluations of programs operating at other prevention levels, evaluations of secondary and tertiary prevention programs received higher scores for the appropriateness of their measurements (which follows the same general trend) and the relevance of their sampling strategies for addressing the research question.

Box 9. A PVE program evaluation with a quantitative descriptive design

An evaluation of the Prevent program In English schools (Joyce, 2018)

Joyce (2018) explored teachers' perceptions and attitudes regarding the implementation of the Prevent program in schools in West Yorkshire, England. He used a sequential explanatory mixed methods design that included a descriptive quantitative phase and a qualitative phase. The target population for the quantitative phase consisted of teachers at 10 primary schools and 2 secondary schools. The final sample consisted of 38 teachers who completed a 14-item questionnaire. Descriptive statistics were used to analyze the quantitative data. From the standpoint of the MMAT, this study had all of the characteristics of a good quantitative descriptive study. More precisely, the sampling strategy was relevant to the research questions, the measurements were appropriate, the analyses were appropriate for the proposed approach, and, in general, the methodology was transparent and well explained. It would have been better if the study had provided more information about the characteristics of the target population (for example, how many teachers in total were working at these schools) so that we could tell more accurately whether the sample was representative.

Two implications of the study's findings, according to Joyce, are that teachers need the core components of the Prevent program to be far clearer and need better training and ongoing support in their efforts to implement anti-radicalization strategies.

3.5.3. Quality of the experimental studies

Experimental studies are studies in which the participants are assigned randomly to a control group or an intervention group (also known as a treatment group or experimental group)—in other words, where the researchers determine which participants will receive the intervention (Hong et al., 2018). As stated previously, in this systematic review, we identified 6 PVE evaluation studies that can be classified as experimental studies (see Table 28).

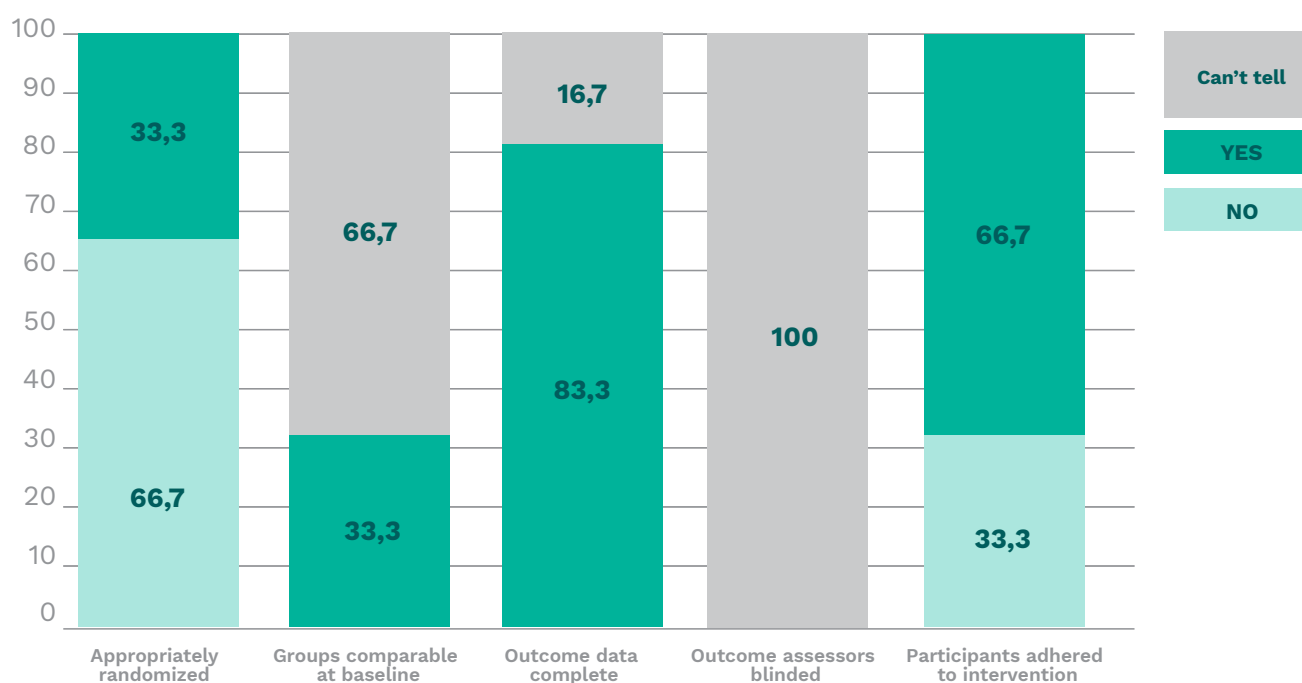
Box 10. MMAT methodological quality criteria for experimental studies

1. Is randomization appropriately performed?
2. Are the groups comparable at baseline?
3. Are there complete outcome data?
4. Are outcome assessors blinded to the intervention provided?
5. Did the participants adhere to the assigned intervention?

Table 28. Methodological quality scores for experimental studies

Variable	Category	Number of studies	Mean	Std. dev.
Total		6	2.17	1.47
Type of publication	Academic literature	3	2.33	1.53
	Grey literature	3	2.00	1.73
Continent	Africa	2	1.00	0.00
	North America	0		
	Asia	1	1.00	
	Europe	2	3.00	1.41
	n/a	1	4.00	
	Australia	0		
Prevention level	Primary	3	2.00	1.73
	Targetted primary	5	1.80	1.30
	Secondary	0		
	Tertiary	0		
	General	0		
Type of violent extremism	Right-wing	1	2.00	
	Islamist	1	1.00	
	All types	4	2.50	1.73
Type of evaluation	Impact	6	2.17	1.47
	Process	1	2.00	
	Output			

Figure 19. Percentages of experimental studies meeting the applicable MMAT criteria



The growing number of experimental studies is probably one of the greatest advances in PVE program evaluation in recent years. This trend shows that, under certain conditions, experimental designs can be used to evaluate PVE programs, and especially those that attempt primary prevention or that target all types of violent extremism rather than one type in particular. But our methodological quality scores for the six experimental studies (Figure 19) should be interpreted cautiously. The MMAT is a comprehensive tool, but it was designed to evaluate any studies in any field, so long as they meet certain conditions. Some of the five MMAT quality criteria for experimental studies are less suited than others for evaluations of social interventions such as PVE programs. One such criterion is whether the people who assess the outcome of the intervention are blinded. In randomized clinical trials for a new medication, this blinding means that the people assessing the effects of the medication do not know which participants received the medication and which ones received the placebo. But in evaluations of social interventions, the people who are assessing the outcome (by completing a questionnaire) are often the subjects who received the intervention. These people of course know whether or not they participated in the intervention, so blinding is often impossible. In the current review, we could not tell whether such blinding had been done in any of the six experimental studies.⁴⁵ We therefore modified the scale and scored these studies according to only 4 of the 5 relevant MMAT criteria. Thus, whereas for the other study designs, the quality scores ranged from 0 to 5, and the midpoint on the scale was 2.5, for the experimental studies, the range was 0 to 4 and the midpoint was 2.

The mean quality score for the six studies was 2.17, slightly above this midpoint. Four of the six studies had scores of 1 or 2, and the two others had scores of 4. Thus the overall quality of these studies was middling. Their scores were low for two of the criteria: whether the random assignments to the intervention group and the control group were made appropriately, and whether these groups

were comparable at baseline. These are two fundamental aspects of the quality of experimental studies. The MMAT judges randomization very strictly: “A simple statement such as ‘we randomly allocated’ or ‘using a randomized design’ is insufficient to judge if randomization was appropriately performed” (Hong et al., 2018, p. 4). In other words, the researchers must have had a predetermined randomization scheme and explained it clearly in their published study. In four of the six experimental studies in this review, this explanation was inadequate. Regarding the comparability of the groups at baseline, the situation was similar, but for four of the six studies, we could not even tell whether this criterion had been met. Yet this knowledge is vital for determining whether the observed changes were due to the intervention or to the characteristics of the groups studied.

The 6 studies scored better for the two other criteria: whether the participants adhered to the assigned intervention and whether the outcome data were complete (that is, the extent to which all of the participants contributed to almost all of the measurements). For four of the six studies, adherence was not a problem for the evaluation. For five of the six studies, we determined that the outcome data were complete (for the sixth study, we could not tell). The MMAT does not provide a single standard or threshold for judging whether a study’s outcome data are complete, but does suggest using the same standard for all of the studies considered. In this review, we used the lowest threshold mentioned in these studies (80%) as our standard.

The small number of experimental studies included in this review prevents us from comparing them further on any variables. But we can observe that among these studies, the three published in academic journals seem to be of higher quality than those in the grey literature. In this regard, peer review seems to be a key factor for ensuring sound methodological quality.

Box 11. Three PVE program evaluations with experimental designs

Voices for Peace’ Impact Evaluation of a Radio Drama to Counteract Violent Extremism in the Sahel Region in Burkina Faso (Bilali, 2019)

Voices for Peace was a 5-year intervention to reduce vulnerability to violent extremism in the Sahel region of West Africa (Burkina Faso, Niger, Cameroon, Mali and Chad). Its goals were to: 1) denounce violent extremism and reduce support for it; 2) raise awareness about the factors that contribute to violent extremism and to youth’s recruitment into violent extremist groups; 3) increase people’s engagement in behaviours that counter support for violent extremism; and 4) encourage participatory governance. In this study, the authors evaluated a part of the *Voices for Peace* project that consisted of educational entertainment in the form of a radio drama presented to listeners in Burkina Faso. The drama, entitled *Wuro Potal*, focused on violent extremism (specifically, violence inflicted on a fictional community by an armed group), collaboration between the population and the security forces/military, governance and corruption, and migration. In order to evaluate the effects of this drama, a sample

⁴⁵ One exception is the use of awareness or training activities as the “placebo” for the control group.

composed of 132 villages was randomly and evenly divided between an intervention group (66 villages) and a control group (66 villages). In each village in the intervention group, 22 participants listened to 52 episodes of the drama over 12 weeks. The results showed that the intervention reduced justification of extremist violence to only a small extent, but increased willingness to collaborate with the police and security forces, awareness of governance and violent extremism, and people's beliefs in their personal and collective ability to bring about positive change and improve conditions in their communities.

A former right-wing extremist in school-based prevention work: Research findings from Germany (Walsh and Gansewig, 2019)

This article summarized the results of an evaluation of the impact of a PVE program delivered in German schools by a former right-wing extremist. The program was presented to pupils in grade 8 or higher and consisted of four lessons (totalling three hours) on the topics of violent extremism and crime. The former extremist first addressed theoretical aspects of these topics, then told the pupils about his own experience. An open discussion followed. The goal of this primary prevention program was, among other things, to reduce extreme right-wing attitudes and delinquent behaviour among youth. In order to evaluate the program, a sample composed of 564 pupils from 50 school classes was randomly and evenly divided between a treatment group and a control group. The data gathered through questionnaires and observations in class do not suggest that this prevention program influenced right-wing extremist attitudes and delinquency. But the authors did not consider these results surprising, among other reasons because the participants' opinions and behaviour could not be expected to change following the delivery of a single, three-hour prevention measure.

Preventing Extremism with Extremists: A Double-Edged Sword? An Analysis of the Impact of Using Former Extremists in Danish Schools (D. Parker and L. Lindekilde, 2020)

In this study, at the request of Danish authorities, the authors evaluated the effectiveness of an initiative funded by the Danish government. In this initiative, former extremists visited schools, local theatres and youth centres across Denmark to talk about their experiences to large groups of young people aged 13 to 20 and thus raise their awareness regarding violent extremism. More specifically, the former extremists emphasized the negative impacts that violent extremism had had on their lives and described how they were first exposed to extremist ideologies. In this way, the intervention tried to address and shed light on the process of radicalization and its harmful effects, with the ultimate goals of countering extremist narratives and increasing young people's critical thinking. In order to evaluate this program, a questionnaire was completed by 1931 Danish youth who had been randomly assigned to an experimental group (976 individuals) and a control group (955). The results indicated that the project was effective in increasing the participants' ability to recognize extremist ideas and recruitment methods. But the authors found a small decrease in political tolerance among the youth who had participated in the program. In short, the evaluation provided some support for the primarily theoretical assertions in the literature to the effect that former extremists can be credible, effective partners for implementing counterterrorism strategies. But the program's negative effects point to the risk that such initiatives may have unintended consequences, especially when they seek to influence attitudes.

3.5.4. Quality of the quasi-experimental studies

Quasi-experimental studies are defined as any quantitative studies that estimate the effectiveness of an intervention but do not use randomization to allocate the participants to groups that will be compared (Hong et al., 2018). In the current systematic review, we identified 54 studies that met this definition.

Box 12. MMAT methodological quality criteria for quasi-experimental studies

1. Are the participants representative of the target population?
2. Are measurements appropriate regarding both the outcome and intervention (or exposure)?
3. Are there complete outcome data?
4. Are the confounders accounted for in the design and analysis?
5. During the study period, is the intervention administered (or exposure occurred) as intended?

Table 29. Methodological quality scores for quasi-experimental studies

Variable	Category	Number of studies	Mean	Std. dev.
Total		54	2.48	1.69
Type de publication	Academic literature	17	2.76	1.82
	Grey literature	37	2.35	1.64
Continent	Africa	16	2.88	1.54
	North America	3	4.00	0.00
	Asia	9	2.78	2.11
	Europe	25	1.96	1.62
	n/a			-
	Australia	1	2.00	
Prevention level	Primary	11	2.36	1.80
	Targetted primary	22	2.86	1.81
	Secondary	16	1.94	1.44
	Tertiary	6	1.50	2.35
	General	8	2.63	1.19
Type of violent extremism	Right-wing	5	0.60	0.89
	Islamist	13	3.46	1.61
	All types	37	2.16	1.55
Type of evaluation	Impact	54	2.48	1.69
	Process	12	2.25	1.60
	Output	5	2.00	0.00

Figure 20. Percentages of quasi-experimental studies meeting the applicable MMAT criteria



The mean quality score for these 54 quasi-experimental studies was 2.48 out of 5, just slightly below the midpoint on the scale. The distribution of scores was fairly even: about one-third (27.8%) of the studies received scores of 0 or 1, another third (38.9%) received scores of 2 or 3, and the remaining third (33.4%) received scores of 4 or 5 (see Table 29 and Figure 20). The two most problematic criteria were whether the participants were representative of the target population and whether confounders were accounted for: in both cases, the answer was Yes for only about one-third of the studies.

The MMAT's representativeness criterion for quasi-experimental studies is slightly stricter than the one for quantitative descriptive studies but is still fairly broad. It does not require the sample to be calculated. If the study provides a clear description of the target population and the sample (with inclusion and exclusion criteria) and at least attempts to achieve a sample of participants that represents the target population, that is enough for this criterion to be met. But for 42.6% of these studies, we still could not tell whether this criterion was satisfied, which represents a serious deficiency in their methodological transparency. In this case, the explanation was not a lack of space, as it was for the qualitative studies, because over two-thirds of the quasi-experimental studies were published in the grey literature, yet received a lower mean score than those published in the academic literature. The explanation is that in over one-third of these studies, the sample was not described at all.

As regards accounting for confounders, the problem is somewhat less serious: almost 40% of the studies did not account for confounders, while for 27.8%, we could not tell whether they did or not. Confounders influence both the dependent variable (the effects of the intervention)

and the independent variable (the intervention itself), so failure to account for them may lead to incorrect interpretation of the causal link between an intervention and its effects.

Though the percentages of the quasi-experimental studies that met the three other quality criteria were higher, they still were not excellent. For about one-third (27.8%) of these studies, we could not tell whether they had met these criteria or not. For only one of these three criteria—whether the intervention was administered as intended—was the percentage of studies fairly high (70.4%). As regards the two other criteria, we found that 59.3% of the studies had complete outcome data (that is, most of the participants had contributed to most of the measurements), and 50% had used appropriate, validated measurements to answer the research questions.

The quality scores for the quasi-experimental evaluation studies differed according to other variables that we considered. All three studies from North America received a score of 4, which is very high, and the 13 evaluations of programs targeting Islamist violent extremism had a high mean score (3.46). In contrast, the five evaluations of programs targeting right-wing violent extremism had a very low mean score (0.6), notably because they were missing information needed to determine whether the quality criteria had been met. As we have described, this was the case for a high proportion of the studies that received low scores. The six evaluations of tertiary prevention programs also received low MMAT scores, which may be explained by the problems already mentioned in section 1 of this review, such as the small number of cases, the lack of control groups, and the type of indicator used.

Box 13. Two PVE program evaluations with quasi-experimental designs

Preventing Violent Extremism through Value Complexity: Being Muslim Being British (Liht et Savage, 2013)

Liht and Savage (2013) developed and evaluated a program called Being Muslim Being British, designed to prevent violent extremism in young British Muslims by developing their ability to understand other people's views and values in more complex ways (exercise greater integrative complexity). This intervention was pre- and post-tested with 81 young Muslim males and females across seven pilot groups around the United Kingdom. The evaluation tested two hypotheses: that as a result of the intervention, the participants would a) think in more complex ways and b) care about a greater quantity of values (show greater value pluralism) when working on social issues underpinned by conflicting values. The main indicators for testing these two hypotheses were integrative complexity and conflict-resolution style. Both hypotheses were tested on two sets of verbal data that were quantified in two ways: first, through coding of the participants' written responses to six moral dilemmas to which they were exposed before and after the intervention, and second, through coding of group discussions that took place during group activities at the beginning and end of the intervention. These data were coded by two trained coders who were blind to the pre-intervention and post-intervention conditions. This coding was validated by calculating inter-coder reliability (Cohen's Kappa) between the two coders. The results of this evaluation showed that the participants' integrative complexity had increased significantly post-intervention and that their conflict-resolution styles had become more collaborative and compromising.

Mindanao Youth for Development (MYDev) Program (Education Development Center and USAID, 2019)

The MYDev program, based in the Philippines, was originally an employability program that provided experiential training and post-training support to improve life skills and increase civic engagement and employability among vulnerable, out-of-school youth in conflict-affected areas in that country. The program was granted a one-year extension to include a fourth objective related to changes in young people's perceptions concerning violence and violent extremism. USAID engaged the Education Development Center (EDC) to conduct a quasi-experimental impact evaluation to better understand the MYDev program's contribution to these objectives. The EDC conducted this evaluation with two cohorts of young people: the first concerned the evaluation of the first three objectives, and the second included the young people from the extension year. In both cases, measurements were taken before and after the intervention, but the evaluation team used a comparison group for the first cohort only. For the second cohort, the evaluators followed a proportional stratified random sampling approach to select 789 youth for the intervention group. The data were gathered at the start of the training programs and 4 to 6 months after they ended. To collect the data, the evaluators used two tools that had been validated and adapted for the Philippine context: the Youth Employment Survey and the Youth Perceptions Survey, which measured a total of eight indicators combined. The Youth Employment Survey measured the youth's life skills, work-readiness skills and leadership skills and their perceptions of gender roles in the workplace. The Youth Perceptions Survey measured the youths' perceptions of their governments and their communities, as well as their perceptions of violence and their resilience skills. Among this study's findings were that the youth who participated in this program showed desirable changes in their perceptions of violence, including violent extremism, along with an improvement in their resilience skills.

3.5.5. Quality of the mixed-methods (qualitative + quantitative) studies

As stated in section 3.3.3, when studies use a combination of quantitative and qualitative methods, they are said to employ mixed methods or mixed designs. In this systematic review, we identified 90 studies that used mixed methods. The MMAT quality criteria for mixed-methods studies emphasize how well they integrate their qualitative and quantitative methods, as discussed below. The separate quality scores for the quantitative (descriptive, experimental or quasi-experimental) components of these studies and their qualitative components have been discussed in sections 3.5.2, 3.5.3 and 3.5.4.

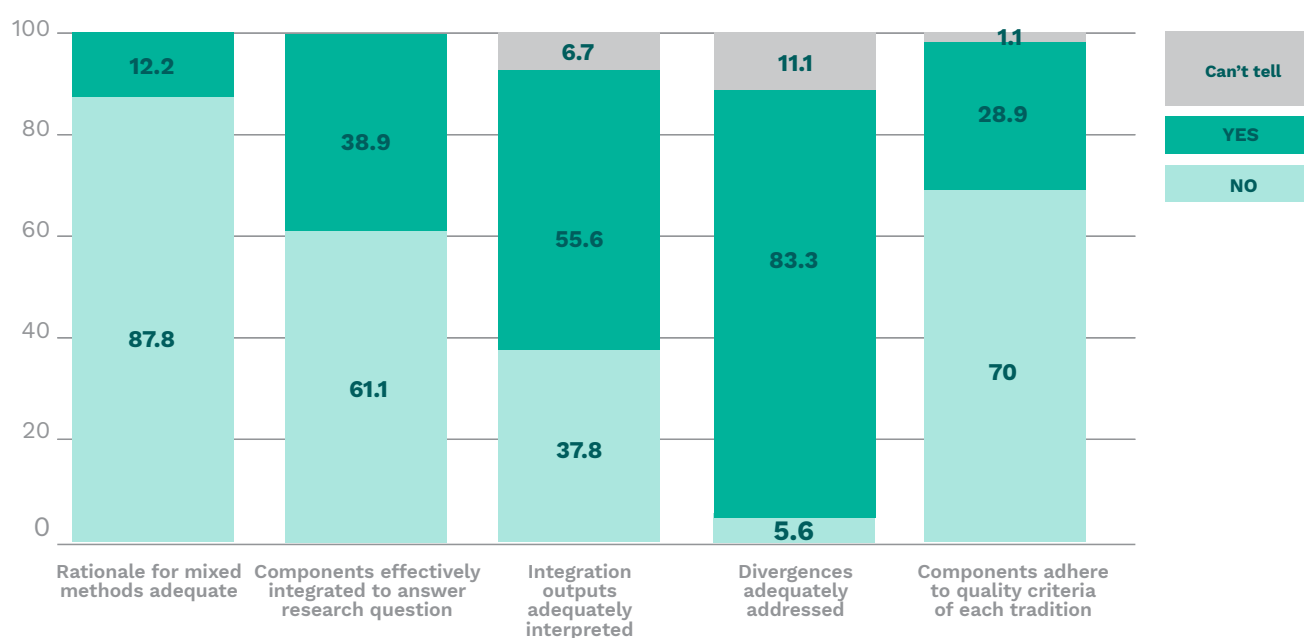
Box 14. MMAT methodological quality criteria for studies using mixed methods

1. Is there an adequate rationale for using a mixed methods design to address the research question?
2. Are the different components of the study effectively integrated to answer the research question?
3. Are the outputs of the integration of qualitative and quantitative components adequately interpreted?
4. Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?
5. Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?

Table 30. Methodological quality scores for studies using mixed (quantitative + qualitative) methods

Variable	Category	Number of studies	Mean	Std. dev.
Total		90	2.19	1.47
Type of publication	Academic literature	20	1.90	1.48
	Grey literature	70	2.27	1.46
Continent	Africa	25	2.52	1.23
	North America	8	1.63	1.06
	Asia	16	1.63	1.50
	Europe	34	2.38	1.61
	n/a	2	2.50	2.12
	Australia	5	1.80	1.64
Prevention level	Primary	21	2.14	1.62
	Targetted primary	41	2.27	1.41
	Secondary	32	2.28	1.46
	Tertiary	16	2.25	1.88
	General	6	2.50	0.84
Type of violent extremism	Right-wing	7	1.86	1.21
	Islamist	35	2.34	1.63
	All types	54	2.06	1.31
Type of evaluation	Impact	80	2.23	1.41
	Process	42	2.43	1.53
	Output	14	2.36	1.34

Figure 21. Qualité méthodologique des études mixtes par indicateur (%)



The mean quality score for all of the studies that used mixed methods is 2.19 out of 5, below the mid-point on the scale. Almost half of these studies (42.2%) received quality scores of only 0 or 1, while only one-fifth received very high scores (4 or 5). The MMAT criteria for mixed-methods studies can be seen as falling into three categories: rationale for using mixed methods, adherence to the quality criteria of each method, and integration of the qualitative and quantitative components. In all cases, the problem of methodological transparency that we had with all of the other study designs was practically non-existent here: for almost every one of the mixed-methods studies, we were able to tell whether each of the five criteria had been satisfied.

The main problem that we found with the methodological quality of these studies was probably the lack of an adequate rationale for using mixed methods. Most of these studies simply stated that they had used mixed methods, with very few giving their rationale for choosing this approach or citing its advantages over other approaches. Mixed-methods designs are being used more and more in the social sciences and have begun to be regarded as a necessity for answering complex research questions. This normalization of the mixed-methods approach may explain why the authors of these studies did not provide an explicit rationale for using it.

The fifth MMAT quality criterion for mixed-methods studies is whether their components adhered to the quality criteria for the respective methods involved—quantitative (descriptive, experimental or quasi-experimental) and qualitative. The scores for this criterion thus depended directly on the scores that these components received on the other applicable MMAT scales. In the MMAT, according to Hong et al. (2018, p. 8): “The premise is that the overall quality of a mixed methods study cannot exceed the quality of its weakest component”. Hence, for a mixed-methods study to receive a high methodological quality score, both its quantitative component and its qualitative component must receive high scores. In the current systematic review, only about 30% of the mixed-methods studies satisfied this criterion.

As regards integration of the qualitative and quantitative components, the mixed-methods studies vary. They do best on the criterion of whether they adequately address divergences and inconsistencies between quantitative and qualitative results. (For a study to meet this criterion, either there must be no divergences or, if there are any, the study must not only report them but also explain them.) More than 80% of the mixed-methods studies in this review met this criterion.

On the criterion of whether the outputs of the integration of qualitative and quantitative components are adequately interpreted, the studies did not score so well: only about half of them met this criterion. This criterion is also relevant to the rationale for using mixed methods,

because it “shows the added value of conducting a mixed methods study rather than having two separate studies” (Hong et al., 2018, p. 7).

The last criterion related to integration of the two components is whether they have been effectively integrated to answer the research question. The studies did less well on this criterion: only 38.9% of them satisfied it. This is the broadest of the integration criteria. It measures the studies’ ability to produce an overall picture by combining the results of the two methods. Seen in this light, mixed-methods evaluations of PVE programs still need a bit of improvement before they can unlock all of the potential of combining methods.

The scores for the mixed-methods studies also vary according to the variables that we analyzed. As with qualitative and quantitative-descriptive evaluation studies, mixed-methods evaluation studies published in the grey literature receive higher quality scores than those published in the academic literature. With complex quantitative designs, such as experimental and quasi-experimental designs, the pattern is the opposite.

By continent, quality scores are higher for the mixed-methods studies in Africa and Europe and lower for those in Asia and North America.

The prevention level that the evaluated programs target does not seem to be a key factor in the quality of their evaluations. The evaluations of tertiary and general prevention studies received slightly lower scores. The effect of prevention level on evaluation quality is more pronounced for the other study designs, especially in the case of tertiary prevention programs, for reasons explained in the preceding section.

The pattern according to type of violent extremism targeted is also repeated here: evaluations of programs targeting Islamist violent extremism systematically scored higher than other evaluations, especially of programs targeting right-wing violent extremism. The explanation may be accumulated experience: there have been many more programs targeting the former than the latter. The fact remains that evaluations of programs addressing right-wing violent extremism still have a long way to go before they can be useful to public policymakers.

Lastly, the quality of mixed-method evaluation studies does not seem to vary tremendously according to type of evaluation (impact, process, or output). But for other study designs, the pattern is different. In studies with more complex quantitative designs (experimental and quasi-experimental), impact evaluations consistently receive higher quality scores, which is consistent with the research questions that they attempt to answer.

Box 15. Two PVE program evaluations with mixed-method designs

“If Youth Are Given the Chance” Effects of Education and Civic Engagement on Somali Youth Support of Political Violence (Tesfaye et al., 2018)

The Somali Youth Learners Initiative (SYLI) was a Mercy Corps PVE program that focused on increasing access to secondary education and opportunities for civic engagement among Somali youth as a way of reducing the likelihood that they would support or join armed groups. Tesfaye et al. (2018) conducted an impact evaluation of this program using a quasi-experimental mixed-methods design. In addition to testing this causal relationship, the evaluators tested six variables that they hypothesized might mediate between the program intervention and the desired effects: potential to be disappointed by livelihood prospects, social isolation, belief in one's capacity to effect community change, confidence in nonviolent means of change, confidence in the federal government, and confidence in the state government. The evaluators also controlled for certain variables: gender, wealth and hunger indices, level of violence in the community, level of urbanization, and duration of intervention implementation.

The mixed methods consisted of a survey and of interviews with key informants. The survey participants were 1220 in-school and out-of-school youth ages 15 to 24. The evaluators divided them into three groups: a control group of 283 respondents who had not participated in the SYLI program, a treatment group of 215 respondents who had participated only in its education component, and another treatment group, of 722 respondents who had participated in both the education and the civic engagement components. The evaluators then conducted interviews with key informants: 40 youth ages 18 to 30. For this phase of data collection, the interviewers used a guide that incorporated semi-structured questions and storytelling components. The survey and the interviews had been designed from the outset to work together. The themes identified in the interviews were compared with the quantitative results from the surveys so as to pinpoint areas of agreement and disagreement between the data sources and identify alternative explanations.

The researchers found that overall, receiving improved access to secondary education supported by the SYLI program, with or without participating in the civic engagement component of this program, helped to reduce support for political violence among Somali youth. But when the evaluators tested their hypotheses about what might have led to the observed reduction in support for political violence, they were able to confirm only some of them. The findings from this study raised other important questions for future research, such as whether opening a school is more important as a short-term signal or a mechanism of longer-term change, and the extent to which education is important because it increases longer-term employment opportunities, as opposed to creating immediate resistance to extremist propaganda.

The role of self-help efforts in the reintegration of ‘politically motivated’ former prisoners: implications from the Northern Irish experience (Dwyer and Maruna, 2011)

“Sometimes I wish I was an ‘ex’ ex-prisoner”: release and reintegration: the experience of ‘politically motivated’ former prisoners in Northern Ireland (Dwyer, 2010)

The 2011 study by Dwyer and Maruna evaluated the practices that community mutual-aid organizations applied to the process of reintegrating “politically motivated” former prisoners from the conflict in Northern Ireland. To conduct this evaluation, the authors used a mixed-methods design that combined 35 interviews with former prisoners, members of mutual-aid organizations and non-governmental organizations and government representatives with a survey of 69 former Republican and Loyalist prisoners. The authors collected additional data through numerous meetings and informal correspondence with civil servants, academics and members of the community/volunteer sector, in particular human-rights organizations.

In her 2010 study, Dwyer's mixed approach was consecutive and exploratory: in other words, she used qualitative data to explore some themes that she then verified by means of quantitative data. The survey was thus used to triangulate the data collected in the interviews and was analyzed descriptively. This consecutive approach, planned from the outset, thus facilitated the integration of the data collected.

Dwyer found that the mutual-aid organizations had facilitated the reintegration of the former prisoners and promote a feeling of solidarity based on the fight against stigmatization.

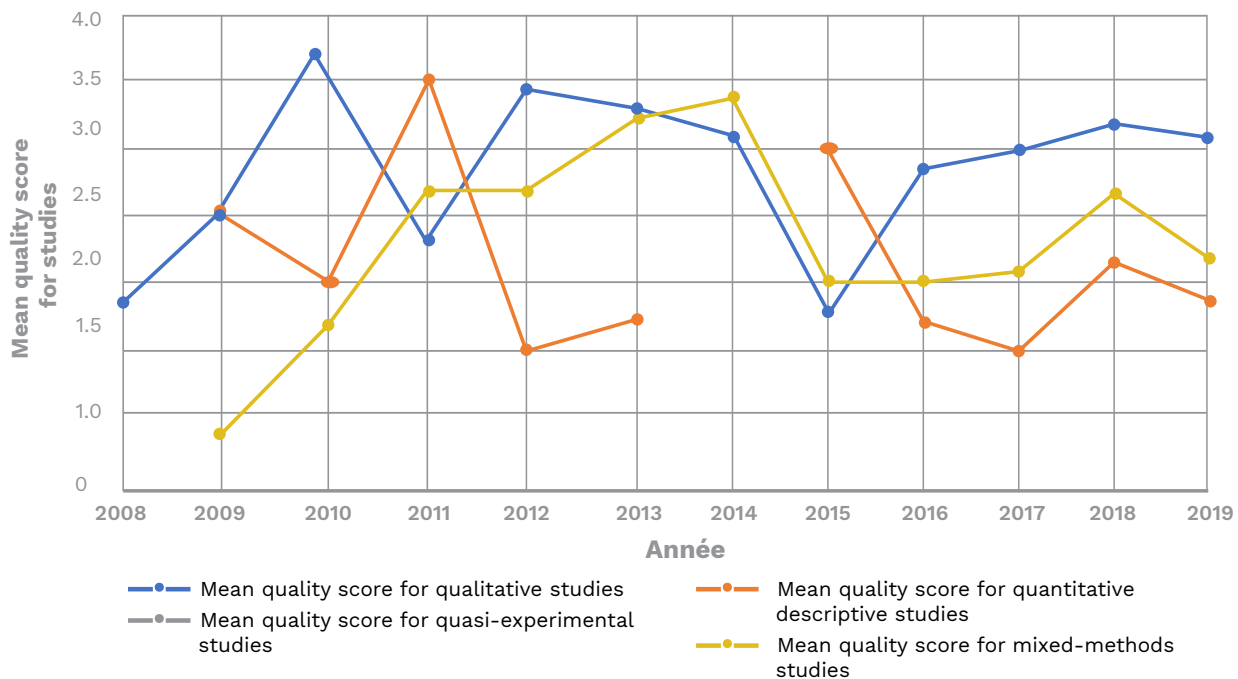
3.5.6. Did the quality of PVE evaluations improve over the years covered by this review?

To conclude our discussion of the quality of the PVE evaluation studies included in this review, we will now examine whether this quality showed an improving trend over the years in question.

This is not an easy question to answer, and the answer may be biased from the outset, because the number of studies for each year and each type of design was relatively small. To reduce the risk of bias, we excluded all studies from years prior to 2008 and all experimental studies (because there were so few of them), and we retained the data for all years from 2008 on in which there were at least two studies for each of the four remaining types of methodological designs. Figure 22 shows the year-to-year changes in the mean quality scores for each design type. For all four, there were two distinct periods: 2008

to 2014 and 2015 to 2019. For the quantitative descriptive studies, quality fluctuated substantially during the first period and showed a relative decline during the second. (Quantitative descriptive studies are not recommended for evaluations of program impact, because they provide a snapshot of reality at one point in time and do not make any comparisons between measurement times or groups. The lessons that can be learned from such studies are relatively limited.) For the three other design types, the quality scores tended to be better during the first period. During the second period, the scores for quasi-experimental studies tended to improve, while those for qualitative and mixed-methods studies tended to stabilize. These last results are more encouraging, because this is also the period when the overall number of evaluation studies started to increase.

Figure 22. Year-to-year changes in mean quality scores for each study design type





3.6. CASE STUDIES

Before presenting our recommendations and conclusion, we will discuss two kinds of PVE program evaluations that are often cited in the literature as posing particular challenges: evaluations of PVE programs that target right-wing violent extremism and evaluations of online PVE programs. As we shall show, these two kinds of evaluations require even closer attention from researchers, evaluators and policymakers.

3.6.1. EVALUATIONS OF PVE PROGRAMS TARGETTING RIGHT-WING VIOLENT EXTREMISM

Evaluating PVE programs that target right-wing violent extremism remains one of the biggest challenges in this field, especially in North America and Europe, where there is genuine concern about the rising tide of right-wing extremist groups. As stated in the introduction to this review, to date there have been few evaluations of PVE programs aimed at right-wing violent extremism, even though one of the best known experiments in reintegrating right-wing extremists (the EXIT program) began well before the latest wave of right-wing extremism began. Indeed, the most recent systematic reviews found only one study of a secondary prevention program (Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021) and five studies of tertiary prevention programs (Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021) targeting right-wing violent extremism. In the current systematic review, we found a total of 20 such studies⁴⁶ discussed in 17 publications; 80% of these studies were of programs in Europe. The actual figure may be even higher, because some studies of programs targetting right-wing extremism may have been excluded from this review because they were not written in English, French or Spanish. This is especially true for evaluations of programs for preventing right-wing violent extremism in the Scandinavian countries, Germany and the Netherlands, where such programs have been in place since the 1990s.

⁴⁶ This figure includes only studies that clearly indicated that they had evaluated a program targetting right-wing violent extremism. It does not include studies that evaluated programs that could be regarded as targetting all types of extremism.

a. Characteristics of the evaluated programs

Compared with the other kinds of PVE programs whose evaluations are discussed in this review, the programs targeting right-wing violent extremism were more focused: 60% of them targetted only this type of extremism, 75% of them were specific projects, 65% of them involved secondary prevention, and 40% of them involved tertiary prevention. In other words, most of these programs targetted at-risk populations. This observation is confirmed by the fact that a high proportion (65%) of these programs targetted specific individuals, while only 25% targetted a societal group (such as youth) or society as a whole.

Most of the evaluated programs (55%) concentrated on disengaging individuals from extremist groups or settings and then reintegrating them into society. These programs were evaluated primarily by means of interviews conducted after the program intervention. Other methods used less extensively included quantitative descriptive analysis and qualitative analysis of interactions on the Internet (two studies, described in detail in the following section), and there was one study with a quasi-experimental design. In this study, the evaluators assessed the impact of a program consisting of a series of workshops delivered over six to eight weeks to eight groups of young persons who had disengaged from extremist groups. These workshops explored topics such as identity and culture, diversity, migration and asylum, and understanding of extremism (i-works research ltd., 2013). The participants completed questionnaires before and after the workshops.

About 25% of the evaluated programs focused on actions in schools or in partnership with schools. For example, Walsh and Gansewig (2019) evaluated a program in which former extremists presented lessons in schools to prevent right-wing violent extremism. These authors used an experimental method in which they compared the intervention group with a control group. In other cases, the evaluation methods have been limited to interviews conducted or questionnaires administered after the program was completed, and the analyses have been primarily descriptive.

b. Objectives (types) of program evaluations

The objectives of these program evaluations differed in some respects from those of the evaluations of programs targeting other types of extremism (see section 3.3.1). Evaluation of impact was the most commonly stated objective for all of the studies in this review, but was a stated objective for a higher proportion of the programs targeting right-wing violent extremism: 80% of these studies stated that they had evaluated the programs' impact, whereas for the other studies, the figure was about 70% on average. Similarly, only 30% of the studies of programs aimed at right-wing violent extremism stated that they had evaluated the programs' processes, whereas for the other studies, the percentages averaged 47%. Thus evaluations of programs targeting right-wing extremism seem to be less concerned with the processes that these programs follow than in the outcomes that they achieve.

The majority of these program evaluations (55%) were conducted by external evaluators, which is lower than the corresponding percentages for evaluations of programs targeting Islamist violent extremism (71.1%) and those that target all types of extremism (80%).

c. Methodological designs of program evaluations

In terms of methodological design, these evaluation studies were fairly evenly distributed among those using qualitative methods (30%), those using quantitative methods (35%) and those using mixed methods (35%).

Among those studies that used quantitative methods, 7 had a quantitative descriptive section, 5 had a quasi-experimental section, and 1 used an exclusively experimental design. Although the percentage of studies using qualitative methods was similar across all of the evaluation studies, the percentage of studies using quantitative methods was higher for evaluations of programs targeting right-wing violent extremism compared with other PVE programs. This finding is consistent with the efforts that have been made to assess the impact of these programs. But these efforts have nevertheless been limited: only two studies used a control

group, and 55% of them took measurements only after the program was over. Moreover, most of these studies (70%) did not use any indicators that directly measured radicalization, violent extremism, or sympathies for either, while 80% used indirect indicators that did not measure these things directly.

d. Quality of program evaluations

Just as in section 3.5, we assessed the quality of these program evaluations according to the MMAT criteria for their respective methodological designs.

The majority of the 14 studies that used qualitative methods scored well on appropriateness of approach (92.9%) and adequacy of data-collection methods (78.6%). But only two of the studies showed sufficient coherence between data sources, collection, analysis and interpretation and had derived their findings from the data adequately.

The 7 studies that had quantitative descriptive designs showed weaknesses similar to those discussed in section 3.5.2, particularly as regards the risk of non-response bias (only 14.2% of these studies had an acceptable response rate), the sampling strategy (mentioned in only 42.9% of the studies) and the appropriateness of the statistical analysis for answering the research question (42.9%).

The 5 studies that used quasi-experimental designs to evaluate programs targeting right-wing violent extremism were comparatively less robust than a) the studies that used other methodological designs to do so, and b) the studies that used quasi-experimental designs to evaluate other kinds of programs. For example, in none of these 5 studies did we find that the measurements were appropriate or that confounders had been accounted for, and in only one of them did we find that the other MMAT quality criteria had been met.

The MMAT quality criteria for studies using mixed (quantitative and qualitative) designs emphasize how well the two approaches have been integrated, rather than how well one or the other has performed. In this review, the 7 mixed-methods studies of programs targeting right-wing violent extremism showed varying degrees of integration, depending on the criterion used. In 5 of these 7 studies, the outputs of the integration of the qualitative and quantitative components had been adequately interpreted and divergences and inconsistencies between quantitative and qualitative results had been adequately addressed. But none of the 7 studies presented an adequate rationale for using a mixed-methods design, only 2 of them effectively integrated the components to answer the research question, and in only 1 of them did the different components of the study adhere to the quality criteria for the corresponding methodological design.

Compared with past reviews of PVE program evaluations, the current systematic review included more evaluations of programs targeting right-wing violent extremism, which may be a promising sign. But much work still needs to be done to determine how useful these programs actually are. This observation applies to programs that focus on disengaging individuals from extremist groups, which are the most common type of programs targeting right-wing violent extremism. But it applies even more to programs that focus on other approaches, such as online programs and programs in educational settings, about which we still know very little. In addition, we have two methodological concerns. The first consists of the additional difficulties involved in evaluating programs of this type, as discussed throughout this report (difficulties in accessing information, ethical issues, etc.). The second is the poor quality of the evaluations that have been done to date, as discussed in the preceding section. Future reviews that incorporate evaluations of such programs in Scandinavian countries, the Netherlands and Germany may shed more light on the questions that remain unanswered in this regard.

3.6.2. EVALUATIONS OF ONLINE PVE PROGRAMS

Online programs are assuming growing importance in PVE and related fields. We are therefore devoting this separate section to a closer examination of the evaluations of online PVE programs that were included in this review and to the challenges involved in conducting such evaluations.

Like evaluations of other kinds of PVE programs, evaluations of online PVE programs are relatively scarce. Of the 219 studies in this review, 14 (6.4%) actually evaluated purely online PVE programs, while 2 (0.9%) evaluated PVE programs that combined an online component and an offline component, 7 (3.2%) described the evaluated programs as having an online component but did not evaluate it, and 4 (1.8%) were offline programs but targetted the digital space (specifically, digital literacy). In the following pages, we focus mainly on the 16 studies evaluating programs that were purely online or that combined online and offline components. But we sometimes also mention the offline programs that targetted the digital space.

a. Characteristics of the evaluated programs

Most of these studies (13 of them) evaluated online counternarrative programs, which consisted mainly of awareness campaigns for their target populations. For example, the Extreme Dialogue program in Canada operated for 16 months and produced a series of awareness-raising films that were posted on the Internet with the goals of reducing the appeal of extremism to young people and offering a positive alternative to the increasing amounts of extremist material on the Internet (SecDev Foundation, 2016). Almost all of the evaluations of such online awareness campaigns used the same approach: they assessed the campaigns' reach by measuring how many people watched the videos or were exposed to the content, and their impact by analyzing the various interactions generated by the content (number of shares, number of "likes", content of comments, and so on). In other words, these evaluations used data collected over the Internet.

Two other studies (McDowell-Smith et al., 2017; Speckhard et al., 2019) evaluated the same campaign by collecting data directly from two different samples: a group of American college students and a group of Somali-American students. The authors had the students watch two videos that were going to be posted online (entitled "A Sex Slave for You—A gift from Abu Bakr al-Baghdadi" and "Rewards for Joining the Islamic State"), then held focus groups with the students and analyzed the focus-group discussions quantitatively.

Only one of the evaluations of online programs in this review conducted both interviews and focus groups with youth program participants and program staff to understand the program's impacts in more detail (Jailobaeva and Asilbekova, 2017).

Two other studies in this review (Davey et al., 2018; Frenett and Dow, 2015) evaluated two programs that intervened directly, through social media, with persons who were considered at risk or who had extremist profiles. Both of these programs were designed, implemented and evaluated by the Institute for Strategic Dialogue (ISD). Their primary intervention consisted of a direct, personalized, private conversation between an intervention provider (generally a former extremist) and an intervention candidate, with the goal of deradicalizing them, or disengaging them from an extremist movement, or dissuading them from consuming or sharing extremist content. The evaluation procedure for both programs was also the same: the evaluators measured the initial response rates, number of sustained engagements (conversations that included five or more messages between the candidate and intervention provider), and indications of potential positive impact in the content of the conversations. In both cases, the number of interactions was too small to allow an effective evaluation of the impact of these programs.

The evaluations of online PVE programs also tended to regard any kind of interaction as positive. Frenett and Dow (2015), for example, included shifts in behaviour such as candidates' changing their privacy settings or blocking the intervention provider. Only five or more messages had to

be exchanged for an engagement to be coded as sustained, and they could be the first five exchanges to start the conversation.

Lastly, another study evaluated a national government policy through an analysis of documents and interviews with government officials (Warrington, 2018).⁴⁷

The main goal of the four offline programs that targetted the digital space (Colibaba et al., 2017; Gatewood and Boyer, 2019; Parker et al., 2018; Reynolds and Parker, 2018) was to develop critical thinking and digital literacy. For example, the Digital Resilience educational program evaluated by Reynolds and Parker (2018) aimed to provide young people with the knowledge, skills, attitudes and behaviours that they need to be positive digital citizens in the 21st century. This program focused on the challenges of online radicalization and exposure to extremism, from effectively dealing with hate speech to identifying active disinformation. Compared with evaluations of the kinds of online programs discussed above, evaluations of such offline programs tend to be more complex, mainly involving focus groups and surveys and repeated measurements before and after the interventions.

Of the 16 evaluations of programs that were entirely online or that had an online component that was in fact evaluated, 6 were of programs in North America, 4 in Asia, 2 in Africa and 1 in Europe. (The 3 other programs did not target a specific country or continent.) The distribution by country was quite diverse, with the United States and Canada accounting for the highest numbers of evaluated studies (4 and 2, respectively). In Asia, online programs were evaluated in Indonesia, Iraq, Kyrgyzstan and Pakistan. The distribution of these 16 online programs by prevention level was: targetted primary, 5; secondary, 6; tertiary, 5. The majority of these programs (11) targetted Islamist violent extremism, followed by programs targetting right-wing violent extremism (4) and those targetting all types of violent extremism (4).

b. Objectives (types) of program evaluations

All of the studies evaluating online programs are fairly recent: the earliest was published in 2015, and most of the others in 2016 and 2017. The objective of most of these evaluations (81.3%) was to assess the impact of the programs in question, while two assessed the processes by which the programs were implemented. Most of these evaluations (11) were done by internal evaluators—that is, by the same teams that designed or implemented the programs.

c. Methodological designs of program evaluations

According to strict methodological definitions, most of the evaluations of online programs in this review (93.8%) used quantitative descriptive designs to analyze metrics from the social networks used to deliver the programs, while incorporating a few elements of qualitative analysis, such as analyses of comments made on the Internet by people exposed to counternarrative campaigns, or analyses of messages exchanged in direct interventions. As indicated above, few of the evaluations of online programs used traditional data-collection tools such as survey questionnaires (4 studies), focus groups (4 studies) or interviews (5 studies). Instead, they often quantified the qualitative data that they have gathered, in order to obtain a quantitative overview of the evaluation. All of the studies used indirect indicators and only four used indicators that measured violent extremism or related variables directly. The study designs were not highly sophisticated either: data were collected and analyzed after the interventions were completed, with no control groups, and the focus was on describing the information collected. Transparency was not a strong point either: only five of these evaluations clearly stated how many people participated in them. Although almost all of these studies attempted to measure the programs' impact or effectiveness, and despite their use of quantitative analyses, they did not use any measures of the association between the interventions and their effects.

d. Quality of program evaluations

The MMAT is more appropriate for assessing the quality of traditional studies in which the number of participants is stated and the tools used to collect the data can be identified. Very few of the 16 evaluations of online PVE programs in this review followed this traditional

⁴⁷ A single program can target more than one type of extremism.

model. But some observations can be made from the data collected. For example, 14 of these evaluations analyzed the data from a qualitative standpoint, 15 used quantitative methods, and 13 used mixed methods. In this case, the qualitative approach and the data-collection methods were, for the most part, adequate to answer the research questions. On the other hand, only about half of these studies used data to substantiate their interpretations of the results. The coherence of their overall procedures was poor, as was the consistency between their stated conclusions and the data that they had gathered. None of the 15 studies that used quantitative methods employed experimental or quasi-experimental designs; instead, as mentioned above, they used quantitative descriptive designs.

The main risks in evaluations of online PVE programs are that the numbers of participants are low and the analysis is based not on individuals, but on actions on social media, such as views and content shares. As a result, none of the quantitative descriptive studies met the MMAT criterion of having a low risk of non-response bias, and very few met the other criteria, such as having a relevant sampling strategy or an appropriate statistical analysis. Having a sample that was representative of the target population was the only MMAT criterion that most of these studies satisfied, even though, as mentioned before, these criteria are fairly flexible.

The 13 online PVE program evaluations that used a mixture of quantitative and qualitative methods did a poor job of integrating these two components. None of these 13 studies provided an adequate rationale for using a mixed-methods design, and very few of these studies integrated their quantitative and qualitative components effectively. Only 66% of them adequately interpreted the outputs of the integration, and only 54% adequately addressed the divergences between the quantitative and qualitative results.

In conclusion, despite the importance of the Internet in radicalization and violent extremism, and despite the considerable efforts that have been made to deploy online programs for countering and preventing these phenomena, methods of evaluating such programs still seem to be in their infancy, and very few conclusions can be drawn from the evaluations of such programs that have been done to date.

Most of these evaluations accorded great importance to information readily available on the Internet, such as social media metrics, and very little importance to a coherent assessment of the impact that online programs have on the attitudes, emotions and behaviours of the individuals whom they target. Interpretations of social media metrics are also very readily biased. For example, a “like” in reaction to a post or a video can be interpreted in many ways. On Facebook, for example, it is not unusual to see a post about someone’s recent death where the reactions are “Like” icons rather than “Sad” emojis. In this context, the Likes may be interpreted not as a positive reaction to the news, but rather as a token of support. So how is one supposed to evaluate a program on the basis of indicators that can be interpreted in many different ways? This brings us to the question of the methodological validity of these evaluations. Do they really measure what they are supposed to measure?

These evaluations did not pay very much attention, or at least not very much systematic attention, to the implementation factors that might facilitate or impede an intervention. When the studies did cite such factors, they did so more in the form of lessons learned from the implementation process and the experience of the designers/evaluators, rather than results from a systematic analysis of information from various sources.

One final consideration regarding evaluations of online PVE programs is that they can raise ethical issues, especially when the programs involve direct interventions online. In this review, the evaluations of the two online PVE programs that involved such interventions simply stated that the information resulting from the participants’ interactions with the intervention providers was analyzed. But the evaluators did not state whether they had obtained the participants’ consent to analyze this information that they had shared privately.

Recommandations

Chaque public cible sera illustré par les icônes suivantes :



Évaluatrices et évaluateurs



Chargées et chargés de programmes



Baillleurs de fonds



Actrices et acteurs gouvernementaux



Conceptrices et concepteurs de programmes



Intervenantes et intervenants



Usagères et usagers des programmes



1.1. GENERAL

- Develop and encourage an evaluation culture within organizations working in the field of PVE.
- Encourage independence and better quality in evaluations of programs for preventing right-wing violent extremism and of on-line PVE programs.

1.2. FUNDING

- When the budget for any prevention program is first being planned, a specific line item equal to at least 10% of the total budget should be set aside for evaluation.
- This budget item must be sufficient to ensure a high-quality evaluation: it must be large enough to hire prevention and evaluation experts, to acquire the materials needed to collect and analyze the data, to cover travel costs and to provide enough time to conduct the evaluation properly. The amount of time needed for the evaluation should be estimated according to how large the program is, how many components it has, how complex it is, and how large a geographic area it covers.

1.3. EVALUATION TEAM

- Assemble an evaluation team that is diverse and representative with regard to the program and the actors involved. It should include external members to ensure the evaluation's independence and internal members to ensure an in-depth knowledge of the program (encourage the inclusion of practitioners among the internal members of the team).
- Encourage the use of local teams to conduct evaluations. If an evaluation is to be done by an international team, encourage the addition of local evaluators to the main team.



- c) Involve the evaluators as early as possible, ideally while the program is being designed, but no later than the start of program implementation.
- d) Make sure that all members of the evaluation team speak the language of the country in which the evaluation is to be conducted and have an in-depth knowledge of the local context.

1.4. DESIGNING THE EVALUATION PROTOCOL



- a) Use a clear theory of change and translate it into concrete, specific objectives from which the evaluators can determine what change indicators to measure.



- b) Encourage consistency throughout the evaluation process, so that results are measured by means of indicators that are appropriate for the evaluation's methodology and objectives.



- c) Work with a long enough time horizon to collect and analyze enough data. The evaluation schedule must allow time for reconstructing the theory of change, taking pre- and post-measurements as appropriate, analyzing the data, writing the report, and mobilizing the knowledge generated by the evaluation.



- d) Make sure to document program activities from the very outset so that the evaluators can rely on a rich source of official information.



- e) When using indirect indicators (indicators that do not measure radicalization, violent extremism or sympathies for these phenomena directly), always describe the relationship between these indicators and violent extremism.



- f) Always make sure to prepare, at the very outset, a diversified knowledge-mobilization plan that includes the report, an executive summary and, at a minimum, a workshop or a presentation to the program team.

1.5. METHODOLOGY



- a) For evaluating complex programs, encourage the use of mixed (quantitative and qualitative) designs, because they make it easier to evaluate both impact and process.



- b) Encourage process evaluations for learning more about the factors that influence the implementation of programs.



- c) Encourage the use of repeated measurements in program evaluations (take several observations of the same subjects at two or more different points in time).



- d) Avoid using quantitative descriptive designs as your only evaluation method.



- e) Do not ask respondents to fill out too many different evaluation tools, because that could decrease your response rates from key actors or generate such a large volume of data that you cannot analyze them all in sufficient depth.



- f) When applying quantitative methods, use samples with a large enough number of participants to perform the statistical analyses that are necessary and appropriate to answer the evaluation question.



- g) When using experimental designs, make sure from the outset that your randomization methods are well described and your treatment and control groups are truly comparable.



h) When using quasi-experimental designs, make sure that your samples represent the target population, and consider any confounding factors that may influence your results.



i) When using qualitative designs, make sure to question a diversified sample of stakeholders and participants from inside and outside the program.



j) When using qualitative designs, make sure that your interpretation of the data is supported by concrete evidence.



k) When using mixed designs, make sure that the quantitative and qualitative components of your evaluations are well integrated.



l) When evaluating the impact of tertiary prevention programs, do not use experimental designs.



m) When conducting impact evaluations, encourage the use of experimental and quasi-experimental designs with a control group, when the ethical conditions and context of the program allow.

1.6. DURING THE EVALUATION



a) Make sure that the evaluation is conducted in a setting where the participants and the evaluators can feel safe.



b) Always try to build trust between yourselves and the participants so that they will share information that they would not share otherwise.



c) Always make wise use of any time that practitioners spend on the evaluation, because having to spend too much time on it may undermine their commitment to it.



d) Facilitate the evaluators' direct access to the evaluation participants and to the information that is relevant for the evaluation.

1.7. AFTER THE EVALUATION



a) Write the evaluation report in the normal language of work in the country where the evaluation was conducted.



b) Be transparent about the methodology used and describe it in detail.



c) Always describe any limitations or conflicts of interest involved in the evaluation, and when there are none, always say so.



d) Always describe the participants in the evaluation.



e) Always make sure to implement a diversified knowledge-mobilization plan that includes the report, an executive summary and a workshop or a presentation to the program team.



Conclusion

Evaluating PVE programs is difficult, but it can be done.

Much of the past literature on this subject has focused so much on the methodological and practical difficulties of PVE program evaluations—the various pitfalls and dead ends—as to make such efforts seem impossible. The idea that violent extremism was somehow such an exceptional phenomenon that it required exceptional approaches in all respects, including program evaluation, probably contributed to such perceptions. In a sense, the pragmatic approach to PVE program evaluation discussed in section 1 of this report reflects these perceptions—the idea that we must simply evaluate what we can with the data that we have, inventing our methodologies as we go along.

But more recently, such perceptions of the exceptionality of violent extremism have begun to dissipate. The factors explaining violent extremism are not all that different from the factors explaining analogous phenomena, such as criminality (Wolfowicz et al., 2019), and programs for preventing violent extremism are not all that different from those for preventing other kinds of violence (Madriaza and Ponsot, 2015). **As we have seen throughout this report, the evaluation of PVE programs (including evaluation methods and objectives) is, with some exceptions, not very different from the evaluation of other complex programs for preventing violence.** The actual reason for the limited number of PVE program evaluations in past years and past literature reviews is rather that the process of radicalization to violence, the phenomenon of violent extremism, and ways of preventing them have only become the focus of research over the past 15 years or so, as the need to address these issues has become more urgent and overwhelming and government initiatives to do so have multiplied.

Another reason that so few PVE program evaluations have appeared in past literature reviews may be that they have had a publication bias. A large proportion of past reviews and analyses focused on the academic literature while ignoring the grey literature, which we found to be an extremely rich source of information in the current systematic review. What we observed in this review was that the year 2016 marked a kind of watershed for PVE program evaluation. Starting around then, the focus of prevention programs began to shift from trying anything that might work to designing interventions that could

be evaluated and that could generate more evidence and knowledge that could be applied to practice in future. The 219 studies included in this review vividly demonstrate this trend.

Still another reason that we found more evaluation studies in this review may have been our search strategy. We searched for all studies that evaluated an individual program, regardless of the quality of the evaluation methods used. Our objective was to provide a snapshot of the realities of PVE program evaluation today. If we had confined our systematic review to evaluations of PVE program impacts, we would probably have excluded a high proportion of these 219 studies.

Despite the trends just described, PVE evaluation does continue to involve certain distinct challenges that should be recognized. For example, it is still extremely difficult to measure violent extremism directly, because of the problems involved in defining and operationalizing this phenomenon. Evaluators have instead tended to use indirect indicators that have unquestionably been better suited to the specific contexts of the programs concerned. This practice paradoxically raises interesting questions about the distinctiveness of such programs that try to influence a phenomenon that is so hard to measure. The direct relationship between these indicators and violent extremism can therefore be theoretical only.

Interestingly, in this review we found a number of studies that had used experimental designs to evaluate the impact of PVE programs and that had not been inventoried in past reviews. Experimental designs appear to be easier to apply to evaluate primary and targeted primary PVE programs, which address the general public or entire communities, as opposed to secondary or tertiary PVE programs, which address individuals or groups at risk of becoming involved or already involved in violent extremism. For secondary and tertiary programs, various ethical and practical issues arise that reduce the possibility that experimental designs will become the standard for evaluations. One such practical issue is that as programs become more specific and focus on narrower populations, access to information sources and individuals falls off sharply.

Quasi-experimental evaluation designs, particularly those in which measurements are taken before and after the program intervention, therefore constitute a viable, more sophisticated alternative for measuring PVE program impacts. As we found, such designs have been used to evaluate PVE programs at every prevention level (primary, targeted primary, secondary and tertiary).

Mixed-methods designs (combining quantitative and qualitative methods) are also starting to become a standard for PVE evaluations. Such designs let evaluators measure and quantify the effects of PVE programs in a coherent way while also capturing information that can be used to evaluate the processes by which programs are implemented.

The value of qualitative designs should not be overlooked either. In an exploratory international study on improving PVE program evaluations, in which four of the current authors were involved (Madriaza et al., 2021), the PVE program practitioners and evaluators interviewed reported that qualitative studies provide information and context that are far more meaningful for PVE practice and PVE practitioners.

Mixed-methods studies that incorporate quasi-experimental designs may represent a practical solution for improving the quality of PVE programs in the field.

More complex quantitative designs now offer some promise for evaluating the impact of PVE programs. But it is critically important to recognize that in a number of fields, and especially in PVE, there is some confusion between evaluation in general and impact evaluation in particular. In other words, there is a tendency to assume that all program evaluations are program impact evaluations. As discussed, in the majority of the evaluation studies included in this review, evaluating the program's impact was the stated objective, but the methods used were not always consistent with meeting that objective. This was especially the case for the studies that used quantitative descriptive designs. These studies focused mainly on quantitatively describing the programs' activities, their participants' satisfaction and other related variables that cannot readily measure the impact that these programs had on their users.

This same bias can be seen in earlier literature reviews that limited their understanding of program evaluations to efforts to determine whether programs had achieved their desired results. Such information is essential for the study of PVE and for government policy on PVE. But few reviews have explored the process factors explaining why a program does or does not succeed. Process evaluations are the key to understanding these mechanisms, but to date no specific reviews of process evaluations have been conducted to understand what factors may contribute to or stand in the way of a program's success. The studies by Gielen (2017) and Veldhuis (2015) have shed some

light on this subject, but still do not suffice to provide an understanding of the mechanisms that influence this process.

Our analysis of the methodological quality of the evaluation studies also gave us some useful information about the challenges that PVE evaluations must meet. Very few of the studies met all of the MMAT quality criteria for their respective methodological design categories. On average, the quality scores for the studies in each category were middling, with quantitative descriptive studies scoring the lowest. This last finding confirms that quantitative descriptive designs should be avoided in program evaluations; such designs are the least suited to evaluating the changes that a program achieves.

Each of these design categories has its own strengths and weaknesses, however. The studies with qualitative designs were among the categories scoring the highest, but in many of them, the interpretation of the results was not sufficiently substantiated by the data. This weakness can exacerbate the subjective biases already present in qualitative designs.

The main problem with the quantitative descriptive studies in this review was their lack of transparency about the methods that they used. Most of the studies in this category either did not provide enough information for us to assess their overall quality or provided information that was very incomplete.

The studies with experimental designs performed better in general, but displayed some problems with regard to whether they had made the random assignments to the intervention group and the control group appropriately, and whether these groups were comparable at baseline, both of which are fundamental criteria for a good experimental design. More of the experimental studies scored positively, however, for two other criteria: whether the participants adhered to the assigned intervention and whether the outcome data were complete.

Quasi-experimental studies, which have almost all of the same characteristics as experimental studies except for randomized assignment of participants, have shown growing promise in recent years. Two areas for improvement are the extent to which the participants in the sample represent the target population and the extent to which confounders are accounted for in the design and analysis. Confounders can alter the association between an intervention and its effects and so must be accounted for in order to tell whether a program intervention actually had an impact on the program participants.

Lastly, in studies using mixed (quantitative and qualitative) methods, a common problem is that these two components are insufficiently integrated. Nevertheless, our findings suggest that mixed methods may become a standard for PVE evaluations.

The most common methodological-quality problem across all of the design categories may be the studies' limited transparency about their methods—that is, the amount and level of detail of the information provided in their methodology sections. Such transparency is essential for determining whether the results presented are reliable.

As discussed at various points in this report, the level of prevention at which PVE programs operate, and hence their degree of universality or specificity, seems to strongly influence the type of evaluations that tend to be done on them. For example, for more universal (primary and targeted primary) PVE programs, the proportion of impact evaluations is higher than the proportion of process evaluations. For tertiary PVE programs, the reverse is true. A higher proportion of evaluations of more universal PVE programs have external evaluators and include both pre- and post-intervention measurements; among studies with qualitative or quasi-experimental designs, evaluations of more universal PVE programs receive higher quality scores. Evaluations of more narrowly targeted prevention programs use more direct indicators but give less information about their own limitations. The challenges posed by PVE program evaluations thus vary with the degree of universality or specificity of the program; hence each program may require an evaluation model tailored to its specific realities and context.

This review also produced some interesting findings about the numbers of evaluations of programs targeting the various types of extremism. In the most recent years covered by this review, many evaluations continued to be done of programs targeting “jihadist” or Islamist violent extremism, but it was the number of evaluations of non-specific programs targeting all types of extremism that grew most significantly. The main reason for this trend was the negative evaluations that programs targeting radical Islamism exclusively had been receiving; very likely, the same numerical trend also occurred in programs that were not evaluated. The number of evaluations of programs targeting right-wing violent extremism is still relatively small, but has been trending upward slightly. The number of evaluations of programs targeting left-wing violent extremism is even smaller. As noted several times earlier, these findings may be biased, because we were able to access evaluation studies in English, French and Spanish only, whereas such studies may be more common in other languages—in particular, in evaluations of programs targeting right-wing violent extremism in northern Europe.

Evaluation of programs to prevent right-wing violent extremism thus remains one of the main issues in the field of PVE, especially in light of the growing concerns over right-wing extremist groups, particularly in North America and Europe, and the hateful discourse that they propagate. From a methodological standpoint, the evaluations of such programs have paradoxically

used qualitative methods to measure their impact on their participants, whereas quantitative methods would probably be more appropriate. Also, very few of these evaluations have used control groups, most of them have evaluated programs only after they had ended, and we consistently rated their methodological quality lower than that of evaluations of other types of programs. A great deal more certainly needs to be done to conduct more evaluations of programs of this type and to improve the quality of such evaluations.

Our analysis by continent found that though Europe accounted for by far the highest number of PVE evaluation studies in this review, considerable numbers had also been done in Africa and Asia. The number done in North America was quite low, even though many program evaluations have been done there in related fields, such as crime prevention. Within each continent, the studies tended to be concentrated in certain countries. In North America, the situation in Canada is especially concerning, because very few PVE evaluation studies have been published there—almost all of the North American studies were done in the United States. In Europe, fully half of the evaluation studies came from the United Kingdom. In Asia, one-third of the studies came from Indonesia, and in Africa, one-fifth came from Kenya.

Despite this relative balance in the number of studies done on various continents, the Western approach to evaluation is overwhelmingly predominant on most of them. This pattern is perhaps most obvious in Africa. Scarcely any of the authors of the evaluations of PVE programs in Africa were African themselves. Instead, most of these researchers came from the United States. (Indeed, though our literature search found only 15 PVE program evaluations done in the United States, a total of 95 of the authors of the studies in this review came from that country.) As noted in our discussion of the limitations of the studies that we reviewed, many of these foreign researchers did not speak the languages of the African countries in which they conducted their evaluations. All of the evaluation reports for African PVE programs were written in English, even though French was the predominant European language in many of the countries where these programs were delivered.

In several sections of this report, we have attempted to answer the question, “For whom were these evaluations carried out?” The purpose of an evaluation is not to write a report, but to provide a body of knowledge that will enable better decisions to be made and better programs to be delivered. We do not know whether any other forms of knowledge mobilization were used to feed information back to the teams in the field, but at least in the specific case of the African studies, the evaluations seem to have been done more out of administrative necessity, for the benefit of the funding agencies, than out of any genuine interest in applying evidence-based findings to develop better interventions.



Références

- Abu-Nimer, M., & Nasser, I.** (2017). Building peace education in the Islamic educational context. *International Review of Education*, 63(2), 153-167. <https://doi.org/10.1007/s11159-017-9632-7>
- Admo, N., Wood, A., & Ducol, B.** (2018). *Une place de choix pour dire et se dire, Évaluation d'implantation et d'impacts d'un projet de prévention de la radicalisation menant à la violence* (p. 54). Collège de Maisonneuve, Centre de prévention de la radicalisation menant à la violence et Institut Pacifique. https://info-radical.org/wp-content/uploads/2018/10/UNE_PLACE_DE_CHOIX_POUR_DIRE_ET_SE_DIRE.pdf
- Albert, A., Cabalion, J., & Cohen, V.** (2020). *Un impossible travail de déradicalisation*. Erès Editions.
- Aldrich, D. P.** (2012). Radio as the Voice of God: Peace and Tolerance Radio Programming's Impact on Norms. *Perspectives on Terrorism*, 6(6), 34-60.
- Aldrich, D. P.** (2014). First steps towards hearts and minds ? USAID's countering violent extremism policies in Africa. *Terrorism and Political Violence*, 26(3), 523-546. International Bibliography of the Social Sciences (IBSS). <https://doi.org/10.1080/09546553.2012.738263>
- Algristian, H., Choiriya, D. D., Abdillah, D. S., Ulya, A., Sodali, H. A., Muhammad, A. R., & Handayani, H.** (2019). Why does de-radicalization seem a utopia? Evaluation on "Children of the Country" program. *Journal of Public Health in Africa*, 10, 148-151. <https://doi.org/10.4081/jphia.2019.1211>
- Ali, Y., & Saragih, H. J. R.** (2018). *Implementation of Contra-Radicalization in Alkhairaat Educational Institutions*. 810-817.
- Al-Maqosi, Y. A., Al-Bataineh, M. T., & Al-Kilani, A. M.** (2019). The Effectiveness of an Educational Program for Developing Tolerance Values and Resistance to Intellectual Extremism at Secondary Level in Jordan. *Journal of Educational and Psychological Studies [JEPS]*, 13(4), 628-642. <https://doi.org/10.24200/jeps.vol13iss4pp628-642>
- Aly, A., Taylor, E., & Karnovsky, S.** (2014). Moral disengagement and building resilience to violent extremism: An education intervention. *Studies in Conflict and Terrorism*, 37(4), 369-385. International Bibliography of the Social Sciences (IBSS). <https://doi.org/10.1080/1057610X.2014.879379>
- Amanullah, Z., & Harrasy, A.** (2017). *Between Two Extremes, Responding to Islamist and tribalist messaging online in Kenya during the 2017 elections* (p. 24). Institute for Strategic Dialogue (ISD). <https://www.isdglobal.org/wp-content/uploads/2018/02/Between-Two-Extremes-Feb-2018-ISD.pdf>
- Anindya, C. R.** (2019). The Deradicalisation Programme for Indonesian Deportees: A Vacuum in Coordination. *Journal for deradicalization*, 18, 217-243.

- Arce, D., & Sandler, T.** (2005). Counterterrorism: A Game-Theoretic Analysis. *The Journal of Conflict Resolution*, 49(2), 183-200. <https://doi.org/10.2307/30045107>
- Audit Commission.** (2008). *Preventing Violent Extremism: Learning and Development Exercise. Report to the Home Office and Communities and Local Government.* Audit Commission.
- Awan, I.** (2012a). Muslim communities, conflict and terrorism: A study of Alum Rock. *Safer Communities*, 11(4), 195-204. <https://doi.org/10.1108/17578041211271463>
- Awan, I.** (2012b). 'I am a Muslim not an extremist': How the prevent strategy has constructed a "suspect" community. *Politics and Policy*, 40(6), 1158-1185. International Bibliography of the Social Sciences (IBSS). <https://doi.org/10.1111/j.1747-1346.2012.00397.x>
- Azam, Z., & Bareeha, F.** (2017). Mishal: A Case Study of a Deradicalization and Emancipation Program in SWAT Valley, Pakistan. *Journal for Deradicalization*, 11, 1-29.
- Badurdeen, F. A., & Goldsmith, D. P.** (2018). Initiatives and Perceptions to Counter Violent Extremism in the Coastal Region of Kenya. *Journal for Deradicalization*, 16, 70-102.
- Bala, A., & Deman, H.** (2017). *Bottom-Up Approach to Countering Violent Extremism in Tunisia Final Evaluation Report* (p. 28). Search for Common Ground. https://www.sfcg.org/wp-content/uploads/2018/07/Final_External_Evaluation_Report_on_Bottom-Up_Approaches_to_CVE_Project_-_SFCG_Tunisia.pdf
- Barkindo, A., & Bryans, S.** (2016). De-Radicalising Prisoners in Nigéria: Developing a basic prison based de-radicalisation programme. *Journal for Deradicalization*, 7, 1-25.
- Baruch, B., Ling, T., Warnes, R., & Hofman, J.** (2018). Evaluation in an emerging field: Developing a measurement framework for the field of counter-violent-extremism. *Evaluation*, 24(4), 475-495.
- Basse, Y. O.** (2018). *Final Evaluation Kallewa Manio: An Integrated Approach to Counter Violent Extremism in Diffa* (p. 51). Search for Common Ground. https://www.sfcg.org/wp-content/uploads/2019/05/Final_Evaluation-Kallewa_Manio-June_2018.pdf
- Bastug, M. F., & Evlek, U. K.** (2016). Individual Disengagement and Deradicalization Pilot Program in Turkey: Methods and Outcomes. *Journal for Deradicalization*, 8, 25-45.
- Bean, S., Hill, P., Sany, J., & Riveles, S.** (2011). *USAID/West Africa Peace through Development (PDEV): Program Assessment Report* (p. 87). USAID and EnCompass. https://pdf.usaid.gov/pdf_docs/PDACR829.pdf
- Beider, H., & Briggs, R.** (2010). Promoting community cohesion and preventing violent extremism in higher and further education. *Institute of community cohesion.*—URL: http://www.cohesioninstitute.org.uk/live/images/cme_resources/public/documents/publications/promoting-community-cohesion.pdf http://safecampuscommunities.ac.uk/uploads/files/2013/05/promoting_community_cohesion.pdf
- Bellasio, J., Hofman, J., Ward, A., Nederveen, F., Knack, A., Meranto, A. S., & Hoorens, S.** (2018). *Counterterrorism evaluation: Taking stock and looking ahead.* RAND. www.rand.org/t/RR2628
- Bilali, R.** (2019). 'Voices for Peace' Impact Evaluation of a Radio Drama to Counteract Violent Extremism in the Sahel Region in Burkina Faso—Endline Report (AID-OAA-M-13-00013; p. 116). USAID and NORC at the University of Chicago. https://pdf.usaid.gov/pdf_docs/pa00w4g3.pdf
- Bilazarian, T.** (2016). *Countering Violent Extremism: Lessons on Early Intervention from the United Kingdom's Channel Program* (p. 11). The Program on Extremism at George Washington University. <https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/downloads/Channel%20CVE%20UK.pdf>
- Björge, T., & Horgan, J.** (2009). *Leaving terrorism behind: Individual and collective disengagement* (37177830; 3892825). Routledge; ProQuest Sociology Collection. <http://search.proquest.com/docview/37177830?accountid=28979>
- Boucek, C.** (2008). Counter-Terrorism from Within: Assessing Saudi Arabia's Religious Rehabilitation and Disengagement Programme. *The RUSI Journal*, 153(6), 60-65. <https://doi.org/10.1080/03071840802670106>
- Boucek, C.** (2009). Extremist re-education and rehabilitation in Saudi Arabia. In T. Björge & J. Horgan (Eds.), *Leaving Terrorism Behind* (p. 230-241). Routledge. <https://doi.org/10.4324/9780203884751-22>
- Boulton, A.** (2010). Education for development, CD for Peace: Producing the "globally competitive" child. *Geoforum*, 41(2), 329-336. <https://doi.org/10.1016/j.geoforum.2009.09.013>
- Bou Zeid, M. R.** (2019). Countering extremism through service learning: Narratives learned from journalism students. *Journal of Applied Research in Higher Education*, 11(4), 686-697. <https://doi.org/10.1108/JARHE-11-2018-0246>

- Bowie, R., & Revell, L.** (2018). How Christian Universities Respond to Extremism. *Education Sciences*, 8(3), 140. <https://doi.org/10.3390/educsci8030140>
- Boyd-MacMillan, E.** (2016). Increasing Cognitive Complexity and Collaboration Across Communities: Being Muslim Being Scottish. *Journal of Strategic Security*, 9(4), 79-110. <https://doi.org/10.5038/1944-0472.9.4.1563>
- Boyle, P., Bouasla, E., & Abderebbi, M.** (2016). *Mid-Term Evaluation Favorable Opportunities to Reinforce Self-Advancement for Today's Youth (FOR SATY)* (AID-608-M-15-00002; p. 126). USAID.
- Brett, J., & Kahlmeyer, A.** (2017). *Strengthening Resilience to Violent Extremism – STRIVE (Horn of Africa) Evaluation Report* (p. 51). TANA, Conflict Management Consulting, The European Union. <https://ct-morse.eu/wp-content/uploads/2017/04/170124-STRIVE-evaluation-Report-Final.pdf>
- Briggs, R.** (2010). Community engagement for counterterrorism: Lessons from the United Kingdom. *International Affairs [London]*, 86(4), 971-981. International Bibliography of the Social Sciences (IBSS). <https://doi.org/10.1111/j.1468-2346.2010.00923.x>
- Broadbent, R.** (2013). Using Grass Roots Community Programs as an Anti-Extremism Strategy. *Australian Journal of Adult Learning*, 53(2), 187-210. ERIC.
- Brooks, M. C., & Ezzani, M. D.** (2017). "Being Wholly Muslim and Wholly American": Exploring One Islamic School's Efforts to Educate Against Extremism. *Teachers College Record*, 119, 1-32.
- Brottsförebyggande rådet (Sweden) (Éd.).** (2001). *Exit: A follow-up and evaluation of the organisation for people wishing to leave racist and nazi groups*. Brottsförebyggande rådet. https://www.bra.se/download/18.cba82f7130f475a2f1800028108/1371914734840/2001_exit_a_follow-up_and_evaluation.pdf
- Bryan, H.** (2017). Developing the political citizen: How teachers are navigating the statutory demands of the Counter-Terrorism and Security Act 205 and the Prevent Duty. *Education, Citizenship and Social Justice*, 12(3), 213-226. <https://doi.org/10.1177/1746197917717841>
- Busher, J., Choudhury, T., & Thomas, P.** (2019). The enactment of the counter-terrorism "Prevent duty" in British schools and colleges: Beyond reluctant accommodation or straightforward policy acceptance. *Critical Studies on Terrorism*, 12(3), 440-462. <https://doi.org/10.1080/17539153.2019.1568853>
- Busher, J., Choudhury, T., Thomas, P., & Harris, G.** (2017). *What the Prevent duty means for schools and colleges in England: An analysis of educationalists' experiences* (p. 68). Centre for Trust, Peace and Social Relations, Coventry University, Durham University, University of Huddersfield. <https://pdfs.semanticscholar.org/8371/af62c44e0eb9e0f35a3a24e9767d182f7299.pdf?ga=2.124265677.614667292.1610595013-1125023608.1610595013>
- Carthy, S. L., Doody, C. B., Cox, K., O'Hora, D., & Sarma, K. M.** (2020). Counter-narratives for the prevention of violent radicalisation: A systematic review of targeted interventions. *Campbell Systematic Reviews*, 16(3). <https://doi.org/10.1002/cl2.1106>
- Chatellier, S.** (2012). *Pakistani Women Moderating Extremism A Coalition-Building Case Study* (p. 18). The Institute for Inclusive Security. <https://www.inclusivesecurity.org/wp-content/uploads/2012/09/Pakistani-Women-Moderating-Extremism-A-Coalition-Building-Case-Study.pdf>
- Cherney, A.** (2020). Evaluating interventions to disengage extremist offenders: A study of the proactive integrated support model (PRISM). *Behavioral Sciences of Terrorism and Political Aggression*, 12(1), 17-36. <https://doi.org/10.1080/19434472.2018.1495661>
- Cherney, A., & Belton, E.** (2019). Evaluating Case-Managed Approaches to Counter Radicalization and Violent Extremism: An Example of the Proactive Integrated Support Model (PRISM) Intervention. *Studies in Conflict & Terrorism*, 1-21. <https://doi.org/10.1080/1057610X.2019.1577016>
- Chowdhury Fink, N., Romaniuk, P., & Barakat, R.** (2013). *Evaluating countering violent extremism programming: Practice and progress*. Center on Global Counterterrorism Cooperation.
- Christiaens, E., Hardyns, W., & Pauwels, L.** (2018). *Evaluating the BOUNCEUp Tool: Research Findings and Policy Implications* (p. 95). Ghent University, Faculty Law and Criminology, Institute for International Research on Criminal Policy (IRCP), Federal Public Service Home Affairs.
- Christmann, K., Rogerson, M., Hirschfield, A., Wilcox, A., & Sharratt K.** (2012). *Process Evaluation of Preventing Violent Extremism: Programmes for Young People*. <http://rgdoi.net/10.13140/2.1.3117.9042>
- Cifuentes, R., Whittaker, G. R., & Lake, L.** (2013). The Think Project: An Approach to Addressing Racism and Far-Right Extremism in Swansea, South Wales. *Democracy and Security*, 9(3), 304-325. Sociological Abstracts. <https://doi.org/10.1080/17419166.2013.802985>

- Cipaku, J.** (2013). *Mid-Term Evaluation Reducing Recidivism: A Process for Effective Disengagement of High-Risk Prisoners in Indonesia* (N° 7; p. 22). Search for Common Ground and New Zealand's International Aid & Development Agency. https://www.sfcg.org/wp-content/uploads/2014/07/INA_MT_Dec13_SCGF_NZL_MTR_Report_Revisi.pdf
- Clemens-Hope, O. M.** (2015). *USAID Peace Through Development II* (p. 56). USAID. https://blumont.org/wp-content/uploads/2016/01/PDev-II-Jan-Mar-2015-Qly-Rept_FINAL-4-29-2015.pdf
- Clement, P.-A., Madriaza, P., & Morin, D.** (2021). *Les intervenants et l'évaluation en prévention de l'extrémisme violent: Entre contraintes et opportunités*. Chaire UNESCO en prévention de la radicalisation et de l'extrémisme violents (Chaire UNESCO-PREV).
- Cockayne, J., O'Neil, S., Felbab-Brown, V., Chowdhury Fink, N., & Oswald, B. "Ossie".** (2015). *UN DDR in an Era of Violent Extremism: Is It Fit for Purpose ?* (p. 164). United Nations University. https://collections.unu.edu/eserv/UNU:5532/UN_DDR_in_an_Era_of_Violent_Extremism_2018.pdf
- Colibaba, A., Colibaba, S., Gheorghiu, I., Colibaba, C., Dinu, C., & Ursa, O.** (2017). The Xeno-Tolerance Project- A Useful Tool in Doing Quality Research. *Lucrări Științifice*, 60(1), 141-144.
- Connell, J. P., & Kubisch, A. C.** (1998). Applying a theory of change approach to the evaluation of comprehensive community initiatives: Progress, prospects, and problems. *New approaches to evaluating community initiatives*, 2(15-44), 1-16.
- Court, D.** (2006). Foolish Dreams in a Fabled Land: Living Co-Existence in an Israeli Arab School. *Curriculum Inquiry*, 36(2), 189-208. <https://doi.org/10.1111/j.1467-873X.2006.00352.x>
- CPN-PREV.** (2020). *Train the Trainers Manual: A toolkit to facilitate training on the prevention of violent radicalization for practitioners across Canada*. Canadian Practitioners Network for the Prevention of Radicalization and Extremist Violence (CPN-PREV).
- Cragin, K., & Chalk, P.** (2003). *Terrorism & development: Using social and economic development to inhibit a resurgence of terrorism* (1st éd.). Rand. https://www.rand.org/content/dam/rand/pubs/monograph_reports/2005/MR1630.pdf
- Davey, J., Birdwell, J., & Skellett, R.** (2018). *Counter Conversations A model for direct engagement with individuals showing signs of radicalisation online* (p. 32). Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/wp-content/uploads/2018/03/Counter-Conversations_FINAL.pdf
- Davey, J., Tuck, H., & Amarasingam, A.** (2019). *An imprecise science: Assessing interventions for the prevention, disengagement and de-radicalisation of left and right-wing extremists*. Institute for Strategic Dialogue.
- Demant, F., Wagenaar, W., & van Donselaar, J.** (2009). *Racism & Extremism Monitor Deradicalisation in practice*. Amsterdam, Netherlands: Leiden University, Anne Frank House. http://www.annefrankdagboek.nl/ImageVaultFiles/id_12097/cf_21/Deradicalisation_ebook.PDF
- Dhungana, S. K., Ismanbaeva, R., & Aisakhunova, A.** (2016). *Reducing Violent Religious Extremism and Preventing Conflict in Kyrgyzstan and Central Asia 2013-2016* (p. 69). Search for Common Ground. https://www.sfcg.org/wp-content/uploads/2017/04/CSSF_KGZ501-Review-Final-Report_13052016.pdf
- Dietrich, K.** (2018). *The Way Forward, Assessing the Impact of the "White Dove" CVE Radio Project in Northern Nigéria* (p. 33). Equal Access International. <https://www.equalaccess.org/wp-content/uploads/2018/11/2018-EAI-Nigeria-White-Dove-Final-Assessment.pdf>
- Dunn, K. M., Atie, R., Kennedy, M., Ali, J. A., O'Reilly, J., & Rogerson, L.** (2015). Can you use community policing for counter terrorism ? Evidence from NSW, Australia. *Police Practice and Research*, 17(3), 196-211. <https://doi.org/10.1080/15614263.2015.1015126>
- Dwyer, C. D.** (2010). "Sometimes I wish I was an 'ex' ex-prisoner": release and reintegration: the experience of 'politically motivated' former prisoners in Northern Ireland [PhD Thesis, Queen's University]. https://pureadmin.qub.ac.uk/ws/portalfiles/portal/178891096/Dwyer_Sometimes_I_69438803.pdf
- Dwyer, C. D., & Maruna, S.** (2011). The Role of Self-Help Efforts in the Reintegration of 'Politically Motivated' Former Prisoners: Implications from the Northern Irish Experience. *Crime, Law and Social Change*, 55(4), 293-309. <https://doi.org/10.1007/s10611-011-9284-7>
- Education Development Center (EDC) & USAID.** (2019). *USAID's Mindanao Youth for Development (MYDev) Program FY17 Impact Evaluation Report & FY18/19 (Extension) Performance Evaluation Report Measuring Youth's Employment, Perceptions and Engagements, and Skills* (p. 33). USAID and Education Development Center (EDC).

- El-Said, H.** (2015). *New approaches to countering terrorism: Designing and evaluating counter radicalization and de-radicalization programs*. <http://www.palgraveconnect.com/doi/10.1057/9781137449979>
- Elwick, A., & Jerome, L.** (2019). Balancing securitisation and education in schools: Teachers' agency in implementing the Prevent duty. *Journal of Beliefs & Values*, 40(3), 338-353. <https://doi.org/10.1080/13617672.2019.1600322>
- Eriksson, A.** (2008). Challenging cultures of violence through community restorative justice in Northern Ireland. In *Sociology of Crime Law and Deviance* (Vol. 11, p. 231-260). Emerald (MCB UP). [https://www.emerald.com/insight/content/doi/10.1016/S1521-6136\(08\)00410-7/full/html](https://www.emerald.com/insight/content/doi/10.1016/S1521-6136(08)00410-7/full/html)
- Feddes, A. R.** (2015). *Socio-psychological factors involved in measures of disengagement and deradicalization and evaluation challenges in Western Europe*. <http://www.mei.edu/sites/default/files/Feddes.pdf>
- Feddes, A. R., & Gallucci, M.** (2015). A literature review on methodology used in evaluating effects of preventive and de-radicalisation interventions. *Journal for Deradicalization*, 5, 1-27.
- Feddes, A. R., Huijzer, A., van Ooijen, I., & Doosje, B.** (2019). Fortress of Democracy: Engaging Youngsters in Democracy Results in More Support for the Political System. *Peace and Conflict: Journal of Peace Psychology*, 25(2), 158-164. <https://doi.org/10.1037/pac0000380>
- Feddes, A. R., Mann, L., & Doosje, B.** (2015). Increasing self-esteem and empathy to prevent violent radicalization: A longitudinal quantitative evaluation of a resilience training focused on adolescents with a dual identity: Increasing self-esteem and empathy to prevent violent radicalization. *Journal of Applied Social Psychology*, 45(7), 400-411. <https://doi.org/10.1111/jasp.12307>
- Finkel, S. E., Belasco, C. A., Gineste, C., Neureiter, M., & McCauley, J.** (2018). *Peace Through Development II Burkina Faso, Chad, and Niger Impact Evaluation Endline Report* (p. 149). USAID. https://pdf.usaid.gov/pdf_docs/PA00SWPK.pdf
- Finkel, S. E., Belasco, C. A., Neureiter, M., McCauley, J., Hoepers, B., & Corrigan, C. C.** (2017). *USAID/WEST Africa (USAID/WA) Evaluation & Analytical Services (EAS) Project for the Regional Peace and Governance Programs, Midline Report for Impact Evaluation of the Peace Through Development Phase II (PDEV II) Project in Chad, Niger, and Burkina Faso* (p. 163). USAID.
- Finkel, S. E., Rojo-Mendoza, R. T., Schwartz, C. L., Belasco, C. A., & Kreft, A.** (2015). *Evaluation & Analytical Services (EAS) Project for The Regional Peace and Governance Programs Impact Evaluation of Peace through Development II (P-DEV II) Radio Programming in Chad and Niger Final Report* (p. 50). USAID and University of Pittsburgh. https://pdf.usaid.gov/pdf_docs/PA00KTF3.pdf
- Finn, M., Momani, B., Opatowski, M., & Opondo, M.** (2016). Youth Evaluations of CVE/PVE Programming in Kenya in Context. *Journal for Deradicalization*, 7, 164-224.
- Franssen, A., Dal, C., & Rinschbergh, F.** (2019). *Rapport d'évaluation du Réseau de prise en charge des radicalismes et extrémismes violents* (p. 186). Centre d'Études Sociologiques (CES) de l'Université Saint-Louis - Bruxelles (USL-B). https://extremismes-violents.cfwb.be/fileadmin/sites/RAR/uploads/Documents_Reseau/Rapport_final_Evaluation_Reseau-FWB_11-11-2019.pdf
- Frenett, R., & Dow, M.** (2015). *One to One Online Interventions A pilot CVE methodology* (p. 28). Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/wp-content/uploads/2016/04/One2One_Web_v9.pdf
- Garaigordobil, M.** (2012). Evaluation of a program to prevent political violence in the Basque conflict: Effects on the capacity of empathy, anger management and the definition of peace. *Gaceta Sanitaria*, 26(3), 211-216. <https://doi.org/10.1016/j.gaceta.2011.06.014>
- Gatewood, C., & Boyer, I.** (2019). *Building Digital Citizenship in France Lessons from the Sens Critique project* (p. 36). Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/wp-content/uploads/2019/03/Sens-Critique-uk-screen_E3.pdf
- Gielen, A.-J.** (2017). Countering Violent Extremism: A Realist Review for Assessing What Works, for Whom, in What Circumstances, and How? *Terrorism and Political Violence*, 31(6), 1149-1167. <https://doi.org/10.1080/09546553.2017.1313736>
- Gill, P., Clemmow, C., Hetzel, F., Rottweiler, B., Salman, N., Van Der Vegt, I., Marchment, Z., Schumann, S., Zolghadriha, S., & Schulten, N.** (2020). Systematic Review of Mental Health Problems and Violent Extremism. *The Journal of Forensic Psychiatry & Psychology*, 1-28.

- Glazzard, A., & Reed, A.** (2018). *Global Evaluation of the European Union Engagement on Counter-Terrorism* (p. 35). Royal United Services Institute and International Center for Counter-Terrorism - The Hague (ICCT). <https://icct.nl/app/uploads/2018/10/eu-ct-evaluation-v7-final.pdf>
- Goaziou, V. L.** (2018). L'éducatif au prisme de la radicalisation—La Cellule d'écoute et d'accompagnement des familles (CEAF) de l'ADDAP13. *Recherches et pratiques pour le Groupe addap13*, 2, 52.
- Government of the United Kingdom.** (2011). *Prevent Strategy: Equality Impact Assessment* (p. 17). Government of the United Kingdom. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/97979/prevent-review-eia.pdf
- Greiner, K.** (2010). *Applying Local Solutions to Local Problems: Radio Listeners as Agents of Change* (IQC DFD-I-00-05-00244-00; p. 57). USAID and University of South Florida. https://pdf.usaid.gov/pdf_docs/PDAC538.pdf
- Harahap, H. I., Irmayani, T., & Lubis, F. H.** (2019). The Rationality of De-radicalization Efforts for the Children of Terrorists at Al-Hidayah Islamic Boarding School. *International Journal of Islamic Thought*, 16(1), 38-50. <https://doi.org/10.24035/ijit.16.2019.004>
- Harris-Hogan, S., Barrelle, K., & Smith, D.** (2019). The role of schools and education in countering violent extremism (CVE): Applying lessons from Western countries to Australian CVE policy. *Oxford Review of Education*, 45(6), 731-748. <https://doi.org/10.1080/03054985.2019.1612343>
- Hassan, G., Brouillette-Alarie, S., Ousman, S., Kilinc, D., Varela, W., Lavoie, L., Fetiu, A., Harris-Hogan, S., Madriaza, P., Borokhovski, E., Pickup, D., Boivin, M., Srimathi Narayana, M., Rousseau, C., Gill, P., Thompson, S., McCoy, J., Venkatesh, V., & Morin, D.** (2021). *A Systematic Review on the Outcomes of Primary and Secondary Prevention Programs in the Field of Violent Radicalization*. Canadian Practitioners Network for the Prevention of Radicalization and Extremist Violence (CPN-PREV).
- Hassan, G., Brouillette-Alarie, S., Ousman, S., Savard, É., Kilinc, D., Madriaza, P., Varela, W., Pickup, D., Danis, E., & CPN-PREV team.** (2021). *A Systematic Review on the Outcomes of Tertiary Prevention Programs in the Field of Violent Radicalization*. Canadian Practitioners Network for the Prevention of Radicalization and Extremist Violence (CPN-PREV). <https://cpnprev.ca/wp-content/uploads/2021/11/Intervention-report-FINAL-2021.pdf>
- Hassan, G., Ousman, S., Madriaza, P., Fetiu, A., Boily, L.-A., Levesque, F., Squalli, Z., Ajrouche, K., El-Tahry, N., Lampron-De Souza, S., Desmarais, L., Duong, E., & Moyano, R.** (2020). *D'un océan à l'autre : Cartographie des initiatives de prévention secondaire et tertiaire oeuvrant dans un contexte de radicalisation et d'extrémisme violent au Canada*. Réseau des Praticiens Canadiens pour la Prévention de la Radicalisation et de l'Extrémisme Violent (RPC-PREV). <https://cpnprev.ca/wp-content/uploads/2020/12/FR-Rapport-mapping-final-1.pdf>
- Heath-Kelly, C., & Strausz, E.** (2018). The banality of counterterrorism “after, after 9/11”? Perspectives on the Prevent duty from the UK health care sector. *Critical Studies on Terrorism*, 12(1), 89-109. <https://doi.org/10.1080/17539153.2018.1494123>
- Helmus, T., & Klein, K.** (2018). *Assessing Outcomes of Online Campaigns Countering Violent Extremism: A Case Study of the Redirect Method* (p. 19). RAND Corporation. https://www.rand.org/pubs/research_reports/RR2813.html
- Heydemann, S.** (2014). Countering violent extremism as a field of practice. *Insights*, 1-4.
- Hiariej, E., Rachmawati, A. D., Taek, A. M., Kurniasari, M., & Alvian, R. A.** (2017). *Final Evaluation Reducing the Recruitment and Recidivism of Violent Extremists in Indonesia* (p. 122). Search for Common Ground. https://www.sfcg.org/wp-content/uploads/2018/01/INA029_DOS_BC_external_Evaluation_Report_FINAL_2017.pdf
- Hirschi, C., & Widmer, T.** (2012). Approaches and challenges in evaluating measures taken against right-wing extremism. *Evaluation and program planning*, 35(1), 171-179.
- Holmer, G.** (2013). Countering Violent Extremism: A Peacebuilding Perspective. *USIP Special Report*, 336. <http://www-preview.usip.org/sites/default/files/SR336-Countering%20Violent%20Extremism-A%20Peacebuilding%20Perspective.pdf>
- Holmer, G., Bauman, P., & Aryaeinejad, K.** (2018). *Measuring Up: Evaluating the Impact of P/CVE Programs* (p. 2018-09). United States Institute of Peace.
- Hong, Q. N., & Pluye, P.** (2019). A conceptual framework for critical appraisal in systematic mixed studies reviews. *Journal of Mixed Methods Research*, 13(4), 446-460.

- Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., & Nicolau, B.** (2018). *Mixed Methods Appraisal Tool (MMAT) Version 2018—User guide*. McGill University. http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_criteria%C3%A2%E2%82%AC%20manual_2018%C3%A2%E2%82%AC%2008%C3%A2%E2%82%AC%2001_ENG.pdf
- Horgan, J., & Braddock, K.** (2010). Rehabilitating the Terrorists?: Challenges in Assessing the Effectiveness of De-radicalization Programs. *Terrorism and Political Violence*, 22(2), 267-291. International Bibliography of the Social Sciences (IBSS). <https://doi.org/10.1080/09546551003594748>
- Iacopini, G., Stock, L., & Junge, K.** (2011). *Evaluation of Tower Hamlets Prevent Projects* (p. 63). The Tavistock Institute. https://www.tavistockinstitute.org/wp-content/uploads/2012/12/Tavistock_Projects_Tower-Hamlets-PVE_evaluation_final_report.pdf
- Ipp, O., Prado, A., & Fourati, Y. A.** (2014). *Tunisia Transition Initiative (TTI) Final Evaluation Report* (AID-OAA-I-13-00007; p. 140). USAID and The QED Group. https://pdf.usaid.gov/pdf_docs/PA00JZD7.pdf
- Istiqomah, M.** (2011). *De-radicalization program in Indonesian prisons: Reformation on the correctional institution*. 29-34. <https://doi.org/10.4225/75/57a4200e2b5a3>
- i-works research ltd.** (2013). *The Think Project Interim Evaluation* (p. 26). Ethnic Youth Support Team (EYST).
- Jackson, B., Rhoades, A., Reimer, J., Lander, N., Costello, K., & Beaghley, S.** (2019). *Practical Terrorism Prevention: Reexamining U.S. National Approaches to Addressing the Threat of Ideologically Motivated Violence*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2647.html
- Jailobaeva, K., & Asilbekova, G.** (2017). *Social Media for Deradicalization in Kyrgyzstan: A model for Central Asia Final project evaluation* (p. 58). Search for Common Ground. <https://www.dmeformpeace.org/resource/final-evaluation-social-media-for-deradicalization-in-kyrgyzstan-a-model-for-central-asia/>
- Jerome, L., & Elwick, A.** (2016). *Evaluation report on the ACT Building Resilience Project Full Report* (p. 64). Middlesex University London and Association for Citizenship Teaching (ACT). <https://eprints.mdx.ac.uk/22857/1/The%20ACT%20Building%20Resilience%20Project%20Full%20Evaluation%20Report%20FINAL%2031.10.16.pdf>
- Jerome, L., & Elwick, A.** (2019). Identifying an Educational Response to the Prevent Policy: Student Perspectives on Learning about Terrorism, Extremism and Radicalisation. *British Journal of Educational Studies*, 67(1), 97-114. <https://doi.org/10.1080/00071005.2017.1415295>
- Johns, A., Grossman, M., & McDonald, K.** (2014). "More Than a Game": The Impact of Sport-Based Youth Mentoring Schemes on Developing Resilience toward Violent Extremism. *Social Inclusion*, 2(2). ProQuest Sociology. <https://doi.org/10.17645/si.v2i2.167>
- Johnston, D., Hussain, A., & Cataldi, R.** (2008). *Madrassa Enhancement and Global Security A Model for Faith-Based Engagement*. <https://icrd.org/wp-content/uploads/2018/01/Madrassa-Enhancement-and-Global-Security.pdf>
- Joyce, C. A.** (2018). *Exploring teachers' beliefs, values and attitudes towards radicalisation, extremism and the implementation of anti-radicalisation strategies* [The University of Sheffield School of Education]. <https://etheses.whiterose.ac.uk/21452/1/Craig%20Joyce%20Thesis%20August%202018.pdf>
- Khalil, J., Brown, R., Chant, C., Olowo, P., & Wood, N.** (2019). *Deradicalisation and Disengagement in Somalia—Evidence from a Rehabilitation Programme for Former Members of Al-Shabaab* (p. 50). Royal United Services Institute for Defence and Security Studies. https://rusi.org/sites/default/files/20190104_whr_4-18_deradicalisation_and_disengagement_in_somalia_web.pdf
- Khalil, J., & Ipp, O.** (2016). *USAID/OTI PDQIII Task Order #10, Activity #3 Mali Transition Initiative: Final Evaluation* (Q011OAA1500012; p. 62). USAID.
- 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 & Development*, 3(1), 12. <https://doi.org/10.5334/sta.ee>
- Khurshid, K., Bhatti, A. J., & Hussain, B.** (2018). Education for Social Justice: Commitments and Achievements. *Pakistan Journal of Social Sciences*, 38(1), 199-219.
- Knox, C., & Hughes, J.** (1996). Crossing the Divide: Community Relations in Northern Ireland. *Journal of Peace Research*, 33(1), 83-98.

- Kollmorgen, J.-C., & Barry, C.** (2017). *Evaluation Report Ex-Post Performance Evaluation of USAID/RDMA Sapan Program* (AID-486-I-14-00001/ AID-486-TO-16-00007; p. 116). USAID. https://pdf.usaid.gov/pdf_docs/PA00MXJ8.pdf
- Kollmorgen, J.-C., Ogada, M., Korir, S., & Dena, E.** (2019). *Strengthening Community Resilience Against Extremism (SCORE) Mid-Term Performance Evaluation—Final Report* (AID-OAA-M-13-00011; p. 114). USAID and Social Impact, Inc.
- Kruglanski, A. W., Gelfand, M. J., Bélanger, J. J., Sheveland, A., Hetiarachchi, M., & Gunaratna, R.** (2014). The Psychology of Radicalization and Deradicalization: How Significance Quest Impacts Violent Extremism: Processes of Radicalization and Deradicalization. *Political Psychology*, 35, 69–93. <https://doi.org/10.1111/pops.12163>
- Kundnani, A.** (2009). *Spooked! How not to prevent violent extremism*. Institute of Race Relations. <https://www.kundnani.org/wp-content/uploads/spooked.pdf>
- Kundnani, A.** (2012). Radicalisation: The journey of a concept. *Race & Class*, 54(2), 3-25. International Bibliography of the Social Sciences (IBSS); Sociological Abstracts. <https://doi.org/10.1177/0306396812454984>
- Kurtz, J.** (2015). *Does Youth Employment Build Stability? Evidence from an Impact Evaluation of Vocational Training in Afghanistan* (p. 40). Mercy Corps. https://www.mercycorps.org/sites/default/files/2020-01/MercyCorps_AfghanistanINVEST_ImpactEvaluation_2015.pdf
- Kyriacou, C., Szczepek Reed, B., Said, F., & Davies, I.** (2017). British Muslim university students’ perceptions of Prevent and its impact on their sense of identity. *Education, Citizenship and Social Justice*, 12(2), 97-110. <https://doi.org/10.1177/1746197916688918>
- Lakhani, S.** (2012). Preventing violent extremism: Perceptions of policy from grassroots and communities. *Howard Journal of Criminal Justice*, 51(2), 190-206. ProQuest Sociology Collection. <https://doi.org/10.1111/j.1468-2311.2011.00685.x>
- Lamhaidi, N.** (2017). *Women’s Caravan for Peace Final Evaluation* (p. 18). Search for Common Ground. https://www.sfcg.org/wp-content/uploads/2018/07/WC4P-Evaluation_EN-FINAL.pdf
- Letsch, L.** (2018). Countering Violent Extremism in Tunisia – Between Dependency and Self-Reliance. *Journal for Deradicalization*, 17, 163-195.
- Levy, D., Jamankulov, K., & Sartbay, T.** (2019). *Project Evaluation #JashStan: Youth as Agents of Peace and Stability in Kyrgyzstan* (p. 73). Evidence Research Institute. <https://www.dmeformpeace.org/resource/final-evaluation-jashstan-youth-as-agents-of-peace-and-stability-in-kyrgyzstan/>
- Liht, J., & Savage, S.** (2013). Preventing Violent Extremism through Value Complexity: Being Muslim Being British. *Journal of Strategic Security*, 6(4), 44-66. <https://doi.org/10.5038/1944-0472.6.4.3>
- Lindekilde, L.** (2012a). Neo-liberal Governing of “Radicals”: Danish Radicalization Prevention Policies and Potential Iatrogenic Effects. *International Journal of Conflict and Violence*, 6(1), 109-125. ProQuest Sociology Collection.
- Lindekilde, L.** (2012b). Value for Money? Problems of Impact Assessment of Counter-Radicalisation Policies on End Target Groups: The Case of Denmark. *European Journal on Criminal Policy and Research*, 18(4), 385-402. ProQuest Central; ProQuest Sociology. <https://doi.org/10.1007/s10610-012-9178-y>
- Lindekilde, L.** (2014). Refocusing Danish counter-radicalisation efforts: An analysis of the (problematic) logic and practice of individual de-radicalisation interventions. In *Counter-Radicalisation* (p. 223-241). Routledge. <https://doi.org/10.4324/9781315773094-14>
- Lobnikar, B., Cajner Mraović, I., & Prisljan, K.** (2019). Preventing Radicalisation and Extremism – The Views of Police Students in Croatia. *VARSTVOSLOVJE*, 2, 161-183.
- Lum, C., Kennedy, L. W., & Sherley, A. J.** (2006). The Effectiveness of Counter-Terrorism Strategies: Campbell Systematic Review Summary. *Campbell systematic reviews*, 2(1), 1-50.
- Mackenzie, M., & Blamey, A.** (2005). The Practice and the Theory: Lessons from the Application of a Theories of Change Approach. *Evaluation*, 11(2), 151-168. <https://doi.org/10.1177/1356389005055538>
- Madriaza, P., Morin, D., Ousman, S., Autixier, C., Hassan, G., & Venkatesh, V.** (2021). *Improving evaluations of programs for prevention of radicalization and violent extremism: An exploratory international study*. UNESCO Chair in Prevention of Radicalization and Violent Extremism (UNESCO-PREV Chair).

- Madriaza, P., & Ponsot, A.-S.** (2015). *Preventing radicalization: A systematic review*. International Centre for the Prevention of Crime.
- Madriaza, P., Ponsot, A.-S., Marion, D., Monnier, C., Ghanem, A., Zaim, B., Nait-Kac, W., Hassani, N., & Autixier, C.** (2017). *The prevention of radicalization leading to violence: An international study of front-line workers and intervention issues* (p. 137). International Centre for the Prevention of Crime. http://www.crime-prevention-intl.org/fileadmin/user_upload/Publications/2017/EN_Rapport_Radicalisation_Final_Aout2017.pdf
- Madriaza, P., Valendru, F., Stock-Rabbat, L., Ponsot, A.-S., & Marion, D.** (2018). *Dispositif d'intervention sur la radicalisation violente en milieu ouvert (SPIP) en France* (p. 146). International Centre for the Prevention of Crime.
- Manby, M.** (2009a). *Evaluation of Kirklees Youth Offending Team (PVE) Pilot Parenting Project*. Ravensthorpe Community Centre (p. 25). Nationwide Children's Research Centre.
- Manby, M.** (2009b). *Kirklees Youth Offending Team Prevent (previously PVE) Programme (3) Evaluation of Theatre Project* (p. 21). Nationwide Children's Research Centre.
- Manby, M.** (2009c). *Kirklees Youth Offending Team. Prevent (previously PVE) Programme—(2) Evaluation of Film Project* (p. 41). Nationwide Children's Research Centre.
- Manby, M.** (2009d). *Kirklees Youth Offending Team. Prevent (previously PVE) Programme—Evaluation of Diversity Group* (p. 25). Nationwide Children's Research Centre.
- Manby, M.** (2010a). *Evaluation of Kirklees Youth Offending Team Prevent (previously PVE) Project Evaluation of Citizenship Programme* (p. 25). Nationwide Children's Research Centre.
- Manby, M.** (2010b). *Kirklees Youth Offending Team Prevent (previously PVE) Project Evaluation of Pathways into Adulthood Programme* (p. 28). Nationwide Children's Research Centre.
- Mansour, S.** (2017). *The Morocco Transforming Violent Extremism Media Training Program Final External Evaluation* (p. 57). Search for Common Ground. https://www.sfcg.org/wp-content/uploads/2018/06/MAR039_Final_Evaluation.pdf
- Marret, J.-L., Bellasio, J., van den Berg, H., van Gorp, A., van Hemert, D., Leone, L., Meijer, R., Warnes, R., & Van Wonderen, R.** (2017). *Final report providing background for using and further developing the validated toolkit Impact* (FP7 GA no. 312235). Impact Europe.
- Mastroe, C.** (2016). Evaluating CVE: Understanding the Recent Changes to the United Kingdom's Implementation of Prevent. *Perspectives on Terrorism*, 10(2), 50-60.
- Mastroe, C., & Szmania, S.** (2016). *Surveying CVE Metrics in Prevention, Disengagement and Deradicalization Programs*. Report to the Office of University Programs, Science and Technology Directorate, Department of Homeland Security.
- McDonald, B., & Mir, Y.** (2011). Al-Qaida-influenced violent extremism, UK government prevention policy and community engagement. *Journal of Aggression, Conflict and Peace Research*, 3(1), 32-44. ProQuest Sociology.
- McDowell-Smith, A., Speckhard, A., & Yayla, A. S.** (2017). Beating ISIS in the Digital Space: Focus Testing ISIS Defector Counter-Narrative Videos with American College Students. *Journal for Deradicalization*, 10, 50-76.
- McGlynn, C., & McDaid, S.** (2016). Radicalisation and Higher Education: Students' Understanding and Experiences. *Terrorism and Political Violence*, 31(3), 559-576. <https://doi.org/10.1080/09546553.2016.1258637>
- McRae, D.** (2009a). DDR and Localized Violent Conflict: Evaluating Combatant Reintegration Programs in Poso, Indonesia. *Indonesian Social Development Paper*, 14.
- McRae, D.** (2009b). *Reintegration and localized conflict: Promoting police-combatant communication*. The World Bank.
- McRae, D.** (2010). Reintegration and localised conflict: Security impacts beyond influencing spoilers. *Conflict, Security & Development*, 10(3), 403-430. <https://doi.org/10.1080/14678802.2010.484204>
- Meringolo, P., Bosco, N., Cecchini, C., & Guidi, E.** (2019). Preventing violent radicalization in Italy: The actions of EU project PROVA. *Peace and Conflict: Journal of Peace Psychology*, 25(2), 165-169. <https://doi.org/10.1037/pac0000375>

- Mitts, T.** (2017). Do Community Engagement Efforts Reduce Extremist Rhetoric on Social Media? *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2940290>
- Moffett, K., & Sgro, T.** (2016). School-Based CVE Strategies. *The ANNALS of the American Academy of Political and Social Science*, 668(1), 145-164. <https://doi.org/10.1177/0002716216672435>
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group.** (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, 6(7), e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Monzani, B., Sarota, A., & Venturi, B.** (2018). *Evaluation Report Inuka! Community-Led Security Approaches to Violent Extremism in Coastal Kenya* (p. 35). Search for Common Ground and Agency for Peacebuilding. <https://www.dmeforpeace.org/resource/final-evaluation-inuka-community-led-security-approaches-to-violent-extremism-in-coastal-kenya-october-2018/>
- Muncy, D., David, R., & Saleh, B.** (2015). *Training of Leaders on Religious and National Co-Existence (TOLERANCE) Project Mid-term Evaluation Report* (p. 113). USAID. https://pdf.usaid.gov/pdf_docs/PA00KMKD.pdf
- Murtaza, N., Sohail, A., Perveen Shaikh, R., Ahmed, S., Anver, S., Ahmad, A., Asghar, M., Ashraf, J., & Yar Khan, U.** (2018). *Punjab Youth Workforce Development Project Midterm Evaluation* (AID-391-C-15-00004; p. 116). USAID.
- Neumann, P.** (2008). Introduction. In J. ICST (Éd.), *Perspectives on Radicalisation and Political Violence: Papers from the First International Conference on Radicalisation and Political Violence*. International Centre for the Study of Radicalisation and Political Violence. <https://www.nonviolent-conflict.org/wp-content/uploads/2016/11/Perspectives-on-Radicalisation-Political-Violence.pdf>
- Neumann, P.** (2011). *Preventing violent radicalization in America*. Bipartisan Policy Center.
- Neumann, P., & Kleinmann, S.** (2013). How rigorous is radicalization research? *Democracy and Security*, 9(4), 360-382. International Bibliography of the Social Sciences (IBSS). <https://doi.org/10.1080/17419166.2013.802984>
- Nicolls, M., & Hassan, A.** (2014). *Evaluation Report Mid-Term Performance Evaluation of the USAID Somali Youth Leaders Initiative (SYLI)* (AID-RAN-I-00-09-00016; p. 98). USAID. https://pdf.usaid.gov/pdf_docs/PA00K3XD.pdf
- Octavia, L., & Wahyuni, E.** (2014). *Final Evaluation Report For the Project: Countering & Preventing Radicalization in Indonesian Pesantren* (p. 75). Search for Common Ground. https://www.sfcg.org/wp-content/uploads/2014/08/DUT_Evaluation_Report_FINAL.pdf
- Onyima, J. K.** (2017). Sub-Saharan Africa: Societal Reintegration of Ex-Militant Youths. *Conflict Studies Quarterly*, 21, 76-100. <https://doi.org/10.24193/csqs.21.4>
- Orban, F.** (2019). Le programme de mentorat norvégien pour détenus radicalisés: Premier bilan, premiers enseignements « La prison au-delà des frontières ». *Les presses de l'ENAP*, 13.
- O'Toole, T., DeHanas, D. N., & Modood, T.** (2012). Balancing tolerance, security and Muslim engagement in the United Kingdom: The impact of the 'Prevent' agenda. *Critical Studies on Terrorism*, 5(3), 373-389. <https://doi.org/10.1080/17539153.2012.725570>
- O'Toole, T., Meer, N., DeHanas, D. N., Jones, S. H., & Modood, T.** (2016). Governing through prevent? Regulation and contested practice in State-Muslim engagement. *Sociology*, 50(1), 160-177.
- O'Toole, T., DeHanas, D., Modood, T., Meer, N., & Jones, S.** (2013). *Taking Part Muslim Participation in Contemporary Governance* (p. 72). Centre for the Study of Ethnicity and Citizenship. <http://www.bristol.ac.uk/media-library/sites/ethnicity/migrated/documents/mpcgreport.pdf>
- Parker, D., & Lindekilde, L.** (2020). Preventing Extremism with Extremists: A Double-Edged Sword? An Analysis of the Impact of Using Former Extremists in Danish Schools. *Education Sciences*, 10(4), 1-19. <https://doi.org/10.3390/educsci10040111>
- Parker, L., Boyer, I., & Gatewood, C.** (2018). *Young Digital Leaders Impact Report* (p. 44). Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/wp-content/uploads/2018/10/YDL_Impact-Report_Final_October_2018.pdf
- Peracha, F. N., Khan, R. R., & Savage, S.** (2016). Sabaoon: Educational methods successfully countering and preventing violent extremism. In *Expanding Research on CVE* (p. 84-104).

- Peterson, A.** (2012). Legitimacy and the Swedish Security Service's Attempts to Mobilize Muslim Communities. *International journal of criminology and sociology*, 1, 109-120.
- Piasecka, S.** (2019). Performing PREVENT: Anti-extremist theatre-in-education in the service of UK counter-terrorism, a Freirean analysis. *Critical Studies on Terrorism*, 12(4), 715-734. <https://doi.org/10.1080/17539153.2019.1615660>
- Pickering, S., McCulloch, J., & Wright-Neville, D.** (2008). Counter-terrorism policing: Towards social cohesion. *Crime, Law and Social Change*, 50(1-2), 91-109. <https://doi.org/10.1007/s10611-008-9119-3>
- Pipe, R., Egal, J., Malla, L., Billow, Z., & Abdi, A.** (2016). *Somalia Program Support Services Final Performance Evaluation of the Transition Initiatives for Stabilization Project* (IDIQ AID-623-I-14-00009; p. 156). USAID and International Business & Technical Consultants, Inc. (IBTCI). <https://static1.squarespace.com/static/5db70e83fc0a966cf4cc42ea/t/5f491ed3020a2654cb8d19b7/1598627541959/1344.pdf>
- Pistone, I., Eriksson, E., Beckman, U., Mattson, C., & Sager, M.** (2019). A scoping review of interventions for preventing and countering violent extremism: Current status and implications for future research. *Journal for Deradicalization*, 19, 1-84.
- Powers, S. T.** (2015). Expanding the Paradigm: Countering Violent Extremism in Britain and the Need for a Youth Centric Community Based Approach. *Journal of Terrorism Research*, 6(1), 19-26. <https://doi.org/10.15664/jtr.1074>
- Pratchett, L., Thorp, L., Wingfield, M., Lowndes, V., & Jabbar, R.** (2010). *Preventing Support for Violent Extremism through Community Interventions: A Review of the Evidence-Rapid Evidence Assessment Full Final Report*. Department for Communities and Local Government.
- Ranstorp, M.** (2010). *Understanding violent radicalisation: Terrorist and jihadist movements in Europe* (889171749; 4230150). Routledge; International Bibliography of the Social Sciences (IBSS). <http://search.proquest.com/docview/889171749?accountid=28004>
- Reeves, J., & Crowther, T.** (2019). Teacher feedback on the use of innovative social media simulations to enhance critical thinking in young people on radicalisation, extremism, sexual exploitation and grooming. *Pastoral Care in Education*, 37(4), 280-296. <https://doi.org/10.1080/02643944.2019.1618377>
- Reichardt, C. S.** (2009). Quasi-Experimental Design. In *The SAGE Handbook of Quantitative Methods in Psychology* (p. 47-72). SAGE Publications Ltd. <https://doi.org/10.4135/9780857020994>
- Reynolds, L.** (2017). *Internet Citizens Impact Report* (p. 50). Institute for Strategic Dialogue (ISD). <https://www.isdglobal.org/wp-content/uploads/2017/12/Internet-Citizens-ISD-Impact-Report-Dec-2017.pdf>
- Reynolds, L., & Parker, L.** (2018). *Digital Resilience Stronger Citizens Online* (p. 42). Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/wp-content/uploads/2018/10/Digital-Resilience-Project-Report-FINAL_web.pdf
- Ris, L., & Ernstorfer, A.** (2017). *Borrowing a wheel: Applying existing design, monitoring and evaluation strategies to emerging programming approaches to prevent and counter violent extremism*. Peacebuilding evaluation consortium.
- Rodon, C.** (2018). *Rapport d'évaluation d'un dispositif / d'un programme en milieu ouvert de déradicalisation et désengagement de l'idéologie violente pour la prévention et la réduction du risque terroriste* (p. 120). ITG-Consultant. https://www.researchgate.net/publication/337495623_Rapport_d%27evaluation_d%27un_dispositif_d%27un_programme_en_milieu_ouvert_de_deradicalisation_et_desengagement_de_l%27ideologie_violente_pour_la_prevention_et_la_reduction_du_risque_terroriste_Paris_France
- Romaniuk, P.** (2015). *Does CVE work?: Lessons learned from the global effort to counter violent extremism*. Global Center on Cooperative Security.
- Rooke, A., & Slater, I.** (2010). *Prevent in Southwark – 2009-2010 Evaluation Report*. The Centre for Urban and Community Research.
- Rothstein, H., Sutton, A. J., & Borenstein, M. (Éds.).** (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Wiley.
- Rustan, E., Hanifah, N., & Kanro, B.** (2018). De-radicalization in the Implementation of Islamic Education Curriculum in SMA Masamba South Sulawesi. *Dinamika Ilmu*, 18(2), 271-283. <https://doi.org/10.21093/di.v18i2.1338>

- Sabir, R.** (2014). *Understanding Counter-Terrorism Policy and Practice in the UK since 9/11* [University of Bath]. https://purehost.bath.ac.uk/ws/portalfiles/portal/187958374/Rizwaan_Sabir_2014_Understanding_Counter_Terrorism_Policy_and_Practice_in_the_UK_since_9_11..pdf
- Salkind, N.** (2010). Quasi-Experimental Design. In *Encyclopedia of Research Design* (p. 1172-1176). SAGE Publications, Inc. <https://doi.org/10.4135/9781412961288>
- Saltman, E. M., Dow, M., & Bjornsgaard, K.** (2016). *Youth Innovation Labs A Model for Preventing and Countering Violent Extremism* (p. 40). Institute for Strategic Dialogue (ISD). <https://www.isdglobal.org/wp-content/uploads/2016/07/YouthCAN-Labs.pdf>
- Salyk-Virk, M. J.** (2018). Building Community Resilience? Community Perspectives of the Countering Violent Extremism Pilot Program in Minneapolis/St. Paul. *Studies in Conflict & Terrorism*, 43(11), 1011-1042. <https://doi.org/10.1080/1057610X.2018.1514054>
- Sarota, A.** (2017). *Baseline Evaluation of: Katika Usalama Tunategemeana and Pamoja! Strengthening Community Resilience in Tanzania* (p. 107). Search for Common Ground. https://www.sfcg.org/wp-content/uploads/2017/07/Baseline-Report.Final_Public.pdf
- Savage, S., Khan, A., & Liht, J.** (2014). Preventing violent extremism in Kenya through value complexity: Assessment of Being Kenyan Being Muslim. *Journal of Strategic Security*, 7(3), 1-26. International Bibliography of the Social Sciences (IBSS). <https://doi.org/10.5038/1944-0472.7.3.1>
- Savoia, E., Su, M., Harriman, N., & Testa, M. A.** (2019). Evaluation of a School Campaign to Reduce Hatred. *Journal for Deradicalization*, 21, 43-83.
- Savoia, E., Testa, M. A., Stern, J., Lin, L., Konate, S., & Klein, N.** (2016). *Evaluation of the Greater Boston Countering Violent Extremism (CVE) Pilot Program* (p. 57). Harvard T.H. Chan School of Public Health. https://www.dhs.gov/sites/default/files/publications/OPSR_TP_CVE-Formative-Evaluation-Greater-Boston-CVE-Pilot-Program-Report_161121-508.pdf
- Schanzer, D., & Eyerman, J.** (2019). *Engaging With Communities to Prevent Violent Extremism A Review of the Obama Administration's CVE Initiative* (p. 85). Duke University and RTI International. https://www.researchgate.net/publication/336914453_Engaging_with_Communities_to_Prevent_Violent_Extremism_A_Review_of_the_Obama_Administration's_CVE_Initiative
- Schanzer, D., Kurzman, C., Toliver, J., & Miller, E.** (2016). *The Challenge and Promise of Using Community Policing Strategies to Prevent Violent Extremism: A Call for Community Partnerships with Law Enforcement to Enhance Public Safety, Final Report* (N° 249674; p. 87). Triangle Center on Terrorism and Homeland Security Sanford School of Public Policy, Duke University. <https://www.ncjrs.gov/pdffiles1/nij/grants/249674.pdf>
- Schmid, A. P.** (2013). Radicalisation, de-radicalisation, counter-radicalisation: A conceptual discussion and literature review. *ICCT Research Paper*, 97. <http://www.academia.edu/download/31064974/ICCT-Schmid-Radicalisation-De-Radicalisation-Counter-Radicalisation-March-2013.pdf>
- Schorn, F., Moubayed, L., & Auten, S.** (2010). *Review of the Office of Middle East Programs Youth Initiatives* (RAN-1-00-09-00016-00; p. 176). USAID and Aguirre Division of JBS International, Inc.
- Schulze, K. E.** (2008). Indonesia's Approach to Jihadist Deradicalization. *CTC Sentinel*, 1(8), 2.
- Schumicky-Logan, L.** (2017). Addressing Violent Extremism with a Different Approach: The Empirical Case of At-Risk and Vulnerable Youth in Somalia. *Journal of Peacebuilding & Development*, 12(2), 66-79. <https://doi.org/10.1080/15423166.2017.1336467>
- Schuurman, B.** (2018). Research on Terrorism, 2007–2016: A Review of Data, Methods, and Authorship. *Terrorism and Political Violence*, 32(5), 1011-1026. <https://doi.org/10.1080/09546553.2018.1439023>
- Schuurman, B., & Bakker, E.** (2016). Reintegrating jihadist extremists: Evaluating a Dutch initiative, 2013–2014. *Behavioral Sciences of Terrorism and Political Aggression*, 8(1), 66-85. <https://doi.org/10.1080/19434472.2015.1100648>
- Search for Common Ground.** (2011). *Program Evaluation Report "Countering and Preventing Radicalization in Indonesian Prisons"* (p. 14). Search for Common Ground. https://www.dmfeforpeace.org/sites/default/files/INA_EV_Feb11_Countering%20and%20Preventing%20Radicalization%20in%20Indonesian%20Prisons.pdf

- SecDev.Foundation.** (2016). *Extreme Dialogue: Social media Target Audience Analysis and Impact Assessments in support of countering violent extremism An abridged summary report of findings and lessons learned* (p. 41). SecDev.Foundation and The Institute for Strategic Dialogue. [https://preventviolentextremism.info/sites/default/files/Kanishka-Secdev %20Extreme %20Dialogue- %20 Social %20media %20Target %20Audience %20Analysis %20and %20Impact %20Assessments %20in %20 support %20of %20countering %20violent %20extremism.pdf](https://preventviolentextremism.info/sites/default/files/Kanishka-Secdev%20Extreme%20Dialogue-%20Social%20media%20Target%20Audience%20Analysis%20and%20Impact%20Assessments%20in%20support%20of%20countering%20violent%20extremism.pdf)
- Sheikh, S., Sarwar, S., & King, E.** (2012). *Evaluation of the Muslim Council of Wales' Prevent work Final report* (N° 23/2012; p. 90). Welsh Government Social Research. [https://gov.wales/sites/default/files/ statistics-and-research/2019-08/120719muslimcouncilen.pdf](https://gov.wales/sites/default/files/statistics-and-research/2019-08/120719muslimcouncilen.pdf)
- Sian, K. P.** (2015). Spies, surveillance and stakeouts: Monitoring Muslim moves in British state schools. *Race Ethnicity and Education*, 18(2), 183-201. International Bibliography of the Social Sciences (IBSS). <https://doi.org/10.1080/13613324.2013.830099>
- Silke, A.** (2001). The devil you know: Continuing problems with research on terrorism. *Terrorism and political violence*, 13(4), 1-14.
- Silke, A.** (2006). The impact of 9/11 on research on terrorism. In *Mapping terrorism research* (p. 90-107). Routledge.
- Silverman, T., Stewart, C. J., Amanullah, Z., & Birdwell, J.** (2016). *The Impact of Counter-Narratives Insights from a year-long cross-platform pilot study of counter-narrative curation, targeting, evaluation and impact* (p. 54). Institute for Strategic Dialogue (ISD) and Against Violent Extremism (AVE). [https:// www.isdglobal.org/wp-content/uploads/2016/08/Impact-of-Counter-Narratives_ONLINE_1.pdf](https://www.isdglobal.org/wp-content/uploads/2016/08/Impact-of-Counter-Narratives_ONLINE_1.pdf)
- Sjøen, M. M., & Mattsson, C.** (2019). Preventing radicalisation in Norwegian schools: How teachers respond to counter-radicalisation efforts. *Critical Studies on Terrorism*, 13(2), 218-236. [https://doi.org/10.1080/175 39153.2019.1693326](https://doi.org/10.1080/17539153.2019.1693326)
- Spalek, B., & Davies, L.** (2012). Mentoring in relation to violent extremism: A study of role, purpose, and outcomes. *Studies in Conflict and Terrorism*, 35(5), 354-368. International Bibliography of the Social Sciences (IBSS).
- Speckhard, A., Shajkovci, A., & Ahmed, M.** (2019). Intervening in and Preventing Somali-American Radicalization with Counter Narratives: Testing the Breaking the ISIS Brand Counter Narrative Videos in American Somali Focus Group Settings. *Journal of Strategic Security*, 11(4), 32-71. [https://doi. org/10.5038/1944-0472.11.4.1695](https://doi.org/10.5038/1944-0472.11.4.1695)
- Speckhard, A., Shajkovci, A., Wooster, C., & Izadi, N.** (2018). Mounting a Facebook Brand Awareness and Safety Ad Campaign to Break the ISIS Brand in Iraq. *Perspectives on Terrorism*, 12(3), 50-66.
- Sullivan, H., & Stewart, M.** (2006). Who Owns the Theory of Change? *Evaluation*, 12(2), 179-199. [https:// doi.org/10.1177/1356389006066971](https://doi.org/10.1177/1356389006066971)
- Supratno, H., Subandiyah, H., & Raharjo, R. P.** (2018). *Character Education in Islamic Boarding School as a Medium to Prevent Student Radicalism*. 222, 405-410. <https://doi.org/10.2991/soshec-18.2018.86>
- Swedberg, J.** (2011). *Mid-Term Evaluation of USAID's Counter-Extremism Programming in Africa* (p. 110). USAID. https://pdf.usaid.gov/pdf_docs/pdacr583.pdf
- Swedberg, J., & Reisman, L.** (2013). *Mid-Term Evaluation of Three Countering Violent Extremism Projects* (p. 136). USAID. https://pdf.usaid.gov/pdf_docs/pdacx479.pdf
- Taylor, E. (Lily), Taylor, P. C., Karnovsky, S., Aly, A., & Taylor, N.** (2016). "Beyond Bali": A transformative education approach for developing community resilience to violent extremism. *Asia Pacific Journal of Education*, 37(2), 193-204. <https://doi.org/10.1080/02188791.2016.1240661>
- Taylor, L., & Soni, A.** (2017). Preventing radicalisation: A systematic review of literature considering the lived experiences of the UK's Prevent strategy in educational settings. *Pastoral Care in Education*, 35(4), 241-252. <https://doi.org/10.1080/02643944.2017.1358296>
- Tesfaye, B., McDougal, T., Maclin, B., & Blum, A.** (2018). "If Youth Are Given the Chance" *Effects of Education and Civic Engagement on Somali Youth Support of Political Violence* (p. 42). Mercy Corps. [https://www.mercycorps. org/sites/default/files/2019-11/If %20Youth %20Are %20Given %20the %20Chance_LR_FINAL.pdf](https://www.mercycorps.org/sites/default/files/2019-11/If%20Youth%20Are%20Given%20the%20Chance_LR_FINAL.pdf)
- Tesfaye, B., & Mohamud, A.** (2016). *Critical Choices Assessing the Effects of Education and Civic Engagement on Somali Youths' Propensity Towards Violence* (p. 34). Mercy Corps. [https://www.mercycorps.org/sites/ default/files/2020-01/CRITICAL_CHOICES_REPORT_FINAL_DIGITAL.pdf](https://www.mercycorps.org/sites/default/files/2020-01/CRITICAL_CHOICES_REPORT_FINAL_DIGITAL.pdf)

- Thomas, P., Miah, S., & Purcell, M.** (2017). *The Kirklee Prevent Young Peoples' Engagement Team: Insights and lessons from its first year* (p. 21). Kirklees Council and the University of Huddersfield, Huddersfield Centre for Research in Education and Society (HudCRES). <http://eprints.hud.ac.uk/id/eprint/32393/1/Kirklees%20Prevent%20Engagement%20Team%20Report%20June%202017.pdf>
- Tines, J., Haq Siddiqui, Noman ul, Akhtar, N., Sadiq, M., Tanveer, T., & Iqbal Zaidi, S. Z.** (2017). *Karachi Youth Workforce Development Project: Midterm Evaluation Report (AMANTECH)* (AID-391-C-15-00004; p. 184). USAID.
- Tropp, L. R., Bilali, R., & Flickinger, S.** (2019). *Healing Our Communities: Promoting Social Cohesion in Rwanda* (p. 111). USAID, Karuna Center for Peacebuilding, AEGIS Preventing Crimes Against Humanity, HROC, Institute of Research and Dialogue for Peace (IRDIP). <https://www.karunacenter.org/wp-content/uploads/2018/03/Healing-Our-Communities-Final-Report.pdf>
- Tsuroyya, T.** (2017). Media to Counter Radicalization: A Case Study at Islamic (Boarding) Schools. *Advanced Science Letters*, 23(12), 11649-11653. <https://doi.org/10.1166/asl.2017.10486>
- Uhlmann, M.** (2017). *Evaluation of the Advice Centre on Radicalisation Final Report* (p. 108). Federal Office for Migration and Refugees and Research Centre Migration, Integration and Asylum. https://www.beratungsstelle-radikalisierung.de/SharedDocs/Anlagen/EN/broschuere-fb31-evaluationsbericht-pdf.pdf?__blob=publicationFile&v=3
- UNEG.** (2016). "Norms and Standards for Evaluation", *United Nations Evaluation Group*.
- United Kingdom House of Commons, & Communities and Local Government Committee.** (2010). *Preventing Violent Extremism, Sixth Report of Session 2009-10* (Report, Together with Formal Minutes, Oral and Written Evidence HC 65; p. 310). House of Commons, Communities and Local Government Committee. <https://publications.parliament.uk/pa/cm200910/cmselect/cmcomloc/65/65.pdf>
- United States Government Accountability Office (GAO).** (2017). *Countering Violent Extremism, Actions Needed to Define Strategy and Assess Progress of Federal Efforts* (GAO-17-300; p. 62). United States Government Accountability Office (GAO). <https://www.gao.gov/assets/690/683984.pdf>
- University of Amsterdam.** (2013). *Empirical Study (revised)* (N° 241744; p. 127). Scientific Approach to Formulate Indicators & Responses to Radicalisation (SAFIRE) and University of Amsterdam.
- Upton, M., & Grossman, M.** (2019). The Dury's Out: Participatory drama and applied theatre processes as ways of addressing radicalized thinking – a pilot study. *Applied Theatre Research*, 7(1), 51-66. https://doi.org/10.1386/atr_00005_1
- van der Heide, L., & Schuurman, B.** (2018). Reintegrating Terrorists in the Netherlands: Evaluating the Dutch approach. *Journal for Deradicalization*, 17, 196-239.
- van Hemert, D., van der Berg, H., van Vliet, T., Roelofs, M., Huis in 't Veld, M., Marret, J.-L., Gallucci, M., & Feddes, A. R.** (2014). *Synthesis report on the state-of-the-art in evaluating the effectiveness of counter-violent extremism interventions* (D2.2). Impact Europe.
- Veldhuis, T. M.** (2015). *Captivated by fear An evaluation of terrorism detention policy* [University of Groningen].
- Veldhuis, T. M., Gordijn, E. H., Lindenberg, S. M., & Veenstra, R.** (2010). *Terrorists in Prison Evaluation of the Dutch terrorism wing* (p. 8). University of Groningen Faculty of Behavioural and Social Sciences.
- Vergani, M., Iqbal, M., Ilbahar, E., & Barton, G.** (2020). The Three Ps of Radicalization: Push, Pull and Personal. A Systematic Scoping Review of the Scientific Evidence about Radicalization Into Violent Extremism. *Studies in Conflict & Terrorism*, 43(10), 854-854. <https://doi.org/10.1080/1057610X.2018.1505686>
- Vermeulen, F.** (2014). Suspect Communities-Targeting Violent Extremism at the Local Level: Policies of Engagement in Amsterdam, Berlin, and London. *Terrorism and Political Violence*, 26(2), 286-306. Sociological Abstracts. <https://doi.org/10.1080/09546553.2012.705254>
- Vittum, K., Ombok, O., Odary, K., Mmoji, G., & Management Systems International.** (2016). *Evaluation Kenya Tuna Uwezo: Final Performance Evaluation USAID/Kenya and East Africa Office of Democracy, Governance and Conflict* (AID-623-TO-16-00004; p. 102). USAID and Management Systems International.
- Walsh, M., & Gansewig, A.** (2019). A former right-wing extremist in school-based prevention work: Research findings from Germany. *Journal for Deradicalization*, 21, 1-42.
- Warrington, A.** (2018). 'Sometimes you just have to try something'—A critical analysis of Danish state-led initiatives countering online radicalisation. *Journal for Deradicalization*, 14, 111-152.

- Waterhouse Consulting Group.** (2008). *Preventing Violent Extremism, An Independent Evaluation of the Birmingham Pathfinder* (p. 53). Waterhouse Consulting Group. <https://wallscometumblingdown.files.wordpress.com/2008/11/birmingham-pve-final-report-14-11-08.pdf>
- Webb, E.** (2017). *For Our Children: An Examination of Prevent in the Curriculum* (p. 32). Centre for the Response to Radicalisation and Terrorism (CRT) and The Henry Jackson Society. <https://henryjacksonsociety.org/wp-content/uploads/2018/12/For-Our-Children-An-Examination-of-Prevent-in-the-Curriculum-.pdf>
- Webber, D., Chernikova, M., Kruglanski, A. W., Gelfand, M. J., Hettiarachchi, M., Gunaratna, R., Lafreniere, M.-A., & Belanger, J. J.** (2018). Deradicalizing Detained Terrorists. *Political Psychology*, 39(3), 539-556. <https://doi.org/10.1111/pops.12428>
- Weeks, D.** (2017). Doing Derad: An Analysis of the U.K. System. *Studies in Conflict & Terrorism*, 41(7), 523-540. <https://doi.org/10.1080/1057610X.2017.1311107>
- Weinberg, L., Pedahzur, A., & Hirsch-Hoefler, S.** (2004). The Challenges of Conceptualizing Terrorism. *Terrorism and Political Violence*, 16(4), 777-794. <https://doi.org/10.1080/095465590899768>
- Weine, S., Eisenman, D., Glik, D., Kinsler, J., & Polutnik, C.** (2016). *Leveraging a Targeted Violence Prevention Program to Prevent Violent Extremism:* (p. 34). University of Illinois at Chicago (UIC) and University of California, Los Angeles (UCLA). https://www.dhs.gov/sites/default/files/publications/862_OPSR_TP_LA-Formative-Evaluation_180817-508.pdf
- Widmer, T., Blaser, C., & Falk, C.** (2007). Evaluating Measures Taken Against Right-Wing Extremism. *Evaluation*, 13(2), 221-239. ProQuest Sociology Collection. <https://doi.org/10.1177/1356389007075225>
- Wilchen Christensen, T.** (2015). *A Question of Participation: Disengagement from the Extreme Right. A case study from Sweden* [Kopicentralen, Roskilde Universitet]. https://rucforsk.ruc.dk/ws/portalfiles/portal/56384428/twc_fin_ny.pdf
- Williams, M. J., Horgan, J. G., & Evans, W. P.** (2016). *Evaluation of a Multi-Faceted, U.S. Community-Based, Muslim-Led CVE Program* (N° 249936; p. 167). Georgia State University and University of Nevada, Reno. <https://www.ncjrs.gov/pdffiles1/nij/grants/249936.pdf>
- Williams, M. J., & Kleinman, S. M.** (2014). A utilization-focused guide for conducting terrorism risk reduction program evaluations. *Behavioral Sciences of Terrorism and Political Aggression*, 6(2), 102. ProQuest Sociology Collection.
- Wilner, A., & Rigato, B.** (2017). The 60 Days of PVE Campaign: Lessons on Organizing an Online, Peer-to-Peer, Counter-radicalization Program. *Journal for Deradicalization*, 12, 227-268.
- Wilson, N. L., & Krentel, J.** (2018). *Lessons from Strengthening Capacity in Countering Violent Extremism* (p. 20). United States Institute of Peace. <https://www.usip.org/publications/2018/05/lessons-strengthening-capacity-countering-violent-extremism>
- Winston, J., & Strand, S.** (2013). Tapestry and the aesthetics of theatre in education as dialogic encounter and civil exchange. *Research in Drama Education: The Journal of Applied Theatre and Performance*, 18(1), 62-78. <https://doi.org/10.1080/13569783.2012.756178>
- Wolfowicz, M., Litmanovitz, Y., Weisburd, D., & Hasisi, B.** (2019). A Field-Wide Systematic Review and Meta-analysis of Putative Risk and Protective Factors for Radicalization Outcomes. *Journal of Quantitative Criminology*. <https://doi.org/10.1007/s10940-019-09439-4>
- Young, H., Rooze, M., Russell, J., Ebner, J., & Schulten, N.** (2016). *Evidence-based Policy Advice Final Report* (p. 54). Terrorism and Radicalisation (TerRa).
- Younis, T., & Jadhav, S.** (2019). Keeping Our Mouths Shut: The Fear and Racialized Self-Censorship of British Healthcare Professionals in PREVENT Training. *Cult Med Psychiatry*, 43, 404-424. <https://doi.org/10.1007/s11013-019-09629-6>
- Zeuthen, M.** (2021). *Reintegration: Disengaging Violent Extremists - A Systematic Literature Review of Effectiveness of Counter-Terrorism and Preventing and Countering Violent Extremism Activities*. Rusi. https://english.iob-evaluatie.nl/binaries/iob-evaluatie-eng/documenten/sub-studies/2021/02/01/literature-studies-%E2%80%93-counterterrorism-and-preventing-and-countering-violent-extremism/Rusi_Reintegration_disengaging_violent_extremists_202102.pdf

Appendix A: List of evaluation studies included in this systematic review

Table 31. Evaluation studies included in this systematic review

Author(s) (merged studies) ⁴⁸	Country	Prevention level	Type of violent extremism targetted	Scope of intervention evaluated	Evaluation type	Methodological design	Study type	Post or pre-post design	Number of participants	Control group
Abu-Nimer et Nasser, 2017	Niger	primary	Islamist	program or project	process	qualitative	observational	post	46	
Admo et al., 2018	Canada	targetted primary	all types	program or project	impact, output	mixed	observational	post	102	
Aldrich, 2012	Mali, Chad, Niger	primary	all types	program or project	impact	quantitative	quasi-experimental	post	>1000	
Aldrich, 2014	Mali	primary	Islamist	program or project	impact	quantitative	quasi-experimental	post	200	
Algristian et al., 2019	Indonesia	general	all types	program or project	impact	quantitative	quasi-experimental	pre-post	16	
Ali and Saragih, 2018	Indonesia	general	all types	part of national strategy or plan	process	qualitative	observational	post	Not given	
Al-Maqosi et al., 2019	Jordan	targetted primary	Islamist	program or project	impact	quantitative	experimental	pre-post	23	25
Amanullah and Harrasy, 2017	Kenya	secondary	Islamist	program or project	impact	mixed	observational	post	Impossible to determine	
Anindya, 2019	Indonesia	secondary	Islamist	part of national strategy or plan	process	qualitative	observational	post	21	
Audit Commission, 2008	United Kingdom	targetted primary	all types	entire national strategy or plan	impact, audit	qualitative	observational	post	Not given	
Awan, 2012	United Kingdom	targetted primary	Islamist	part of national strategy or plan	impact	qualitative	observational	post	6	
Azam and Bareeha, 2017	Pakistan	tertiary	Islamist	program or project	impact, process	qualitative	observational	post	67	
Badurdeen and Goldsmith, 2018	Kenya	general	Islamist	part of national strategy or plan	impact	qualitative	observational	post	249	
Bala and Deman, 2017	Tunisia	targetted primary	all types	program or project	impact, process, output, other	qualitative	observational	post	516	

⁴⁸ The reference for any merged articles is shown between parentheses.

Barkindo and Bryans, 2016	Nigeria	tertiary	all types	part of national strategy or plan	process	qualitative	n/a	post	Not given	
Basse, 2018	Niger	secondary	Islamist	program or project	impact, process, output	mixed	observational	post	201	
Bastug and Evlek, 2016	Turkey	secondary, tertiary	left-wing, right-wing, Islamist, other	part of national strategy or plan	impact	quantitative	observational	post	Not given	
Bean et al., 2011	Chad, Niger	primary	all types	program or project	impact, process	qualitative	observational	post	Impossible to determine	
Bilali, 2019	Burkina Faso	primary	all types	program or project	impact	mixed	experimental	pre-post	1 452	1452
Bilazarian, 2016	United Kingdom	secondary, tertiary	all types	part of national strategy or plan	process	qualitative	observational	post	4	
Bou Zeid, 2019	Lebanon	targetted primary	all types	program or project	process	qualitative	observational	post	5	
Boucek, 2008 (Boucek, 2009)	Saudi Arabia	tertiary	Islamist	part of national strategy or plan	process	qualitative	observational	post	5	
Boulton, 2010	Philippines	general	all types	program or project	other	qualitative	observational	post	4	
Bowie and Revell, 2018	United Kingdom	targetted primary	all types	part of national strategy or plan	process	qualitative	observational	post	8	
Boyd-MacMillan, 2016	Scotland	targetted primary	Islamist	part of national strategy or plan	impact	mixed	quasi-experimental	pre-post	21	
Boyle et al., 2016	Morocco	targetted primary, secondary	all types	program or project	impact, other	mixed	observational	post	6	
Brett and Kahlmeyer, 2017	Kenya / Somalia	general	all types	part of national strategy or plan	impact, process	qualitative	observational	post	Not given	
Briggs, 2010	United Kingdom	targetted primary, secondary	all types	part of national strategy or plan	impact	mixed	observational	post	77	
Broadbent, 2013	Australia	secondary	all types	program or project	impact, process	quantitative	observational	post	16	
Brooks and Ezzani, 2017	United States	targetted primary	Islamist	program or project	process, other	qualitative	observational	post	Impossible to determine	
Brottsförebyggande rådet (Sweden), 2001 ⁴⁹	Sweden	secondary, tertiary	right-wing	entire national strategy or plan	impact, process, output	qualitative	observational	post	Not given	
Bryan, 2017	United Kingdom	secondary	right-wing, Islamist	part of national strategy or plan	process, output	qualitative	observational	post	3	
Busher et al., 2017	United Kingdom	primary	all types	part of national strategy or plan	process	mixed	observational	post	303	
Chatellier, 2012	Pakistan	targetted primary	all types	program or project	impact, process	mixed	observational	post	Not given	
Cherney and Belton, 2019 (Cherney, 2020)	Australia	secondary, tertiary	all types	program or project	impact, process	mixed	observational	post	14-22	
Christiaens et al., 2018	Netherlands	general	all types	program or project	impact, process	mixed	quasi-experimental	post with follow-up	101	

⁴⁹ Summary in English.

Christmann et al., 2012	United Kingdom	general	Islamist	part of national strategy or plan	impact, process, output, monitoring	mixed	observational	post	33-77	
Cifuentes et al., 2013	United Kingdom	targetted primary, secondary	right-wing	program or project	impact	mixed	observational	post	Impossible to determine	
Cipaku, 2013	Indonesia	tertiary	Islamist	program or project	impact, output	qualitative	observational	post	45	
Clemens-Hope, 2015	Niger, Chad and Burkina Faso	primary	all types	program or project	output	other	observational	post	Not given	
Cockayne et al., 2015	Somalia and Kenya	tertiary	Islamist	part of national strategy or plan	process	qualitative	observational	post	67	
Colibaba et al., 2017	Romania	general	all types	program or project	impact	qualitative	observational	post	40	
Court, 2006	Israel	targetted primary	all types	program or project	process	qualitative	observational	post	Not given	
Cragin and Chalk, 2003a	Israel	targetted primary	all types	program or project	impact	qualitative	observational	post	Not given	
Cragin and Chalk, 2003b	Philippines	primary	other, all types	part of national strategy or plan	impact	qualitative	observational	post	Not given	
Cragin and Chalk, 2003c	Israel	primary	all types	part of national strategy or plan	impact, process	qualitative	observational	post	Not given	
Davey et al., 2018	n/a	tertiary	right-wing, Islamist	program or project	impact	mixed	observational	post	> 800	
Demant et al., 2009	Netherlands	secondary, tertiary	right-wing	program or project	impact	qualitative	observational	post	22	
Dhungana et al., 2016	Kyrgyzstan and Central Asia	primary	Islamist	part of national strategy or plan	processus	qualitative	observational	post	48	
Dietrich, 2018	Nigeria	primary, targetted primary	Islamist	program or project	impact, processus	mixed	observational	post	1282	
Dunn et al., 2015	Australia	targetted primary	Islamist	part of national strategy or plan	impact, processus, autre	mixed	observational	post	33	
Dwyer and Maruna, 2011 (Dwyer, 2010)	Northern Ireland	tertiary	other	program or project	processus	mixed	observational	post	104	
Education Development Center (EDC) and USAID, 2019a	Philippines	targetted primary	all types	program or project	impact	quantitative	quasi-experimental	pre-post	1657	41
Education Development Center (EDC) and USAID, 2019b	Philippines	targetted primary	all types	program or project	impact	quantitative	quasi-experimental	pre-post	789	
Eriksson, 2008	Northern Ireland	primary	other	program or project	process, other	qualitative	observational	post	Not given	
Feddes et al., 2019a	Netherlands	targetted primary	all types	program or project	impact	quantitative	quasi-experimental	pre-post	228	
Feddes et al., 2019b	Netherlands	targetted primary	all types	program or project	impact	quantitative	quasi-experimental	pre-post	225	
Finkel et al., 2015a	Chad	primary, secondary	all types	program or project	impact, output, other	mixed	quasi-experimental	pre-post (with interim evaluation)	450	15

Finkel et al., 2015b	Niger	primary, secondary	all types	program or project	impact, output, other	mixed	quasi-experimental	pre-post (with interim evaluation)	450	15
Finkel et al., 2018a (Finkel et al., 2017)	Niger	primary, secondary	all types	program or project	impact, output, other	mixed	quasi-experimental	pre-post (with interim evaluation)	18185	
Finkel et al., 2018b (Finkel et al., 2017)	Chad	primary, targetted primary	all types	program or project	impact	quantitative	experimental	post with follow-up	18185	
Finkel et al., 2018c (Finkel et al., 2017)	Burkina Faso	primary, targetted primary	all types	program or project	impact	quantitative	experimental	post with follow-up	18185	
Finn et al., 2016	Kenya	general	all types	entire national strategy or plan	impact	qualitative	observational	post	Not given	
Franssen et al., 2019	Belgium	primary, secondary, tertiary	all types	program or project	impact, process, output	mixed	observational	post	945	
Frenett and Dow, 2015	n/a	secondary, tertiary	right-wing, Islamist	program or project	impact	mixed	observational	post	154	
Garaigordobil, 2012	Spain	general	other	program or project	impact	quantitative	quasi-experimental	pre-post	191	85
Gatewood and Boyer, 2019	France	targetted primary	all types	program or project	impact, process	mixed	quasi-experimental	pre-post	22	
Glazzard and Reed, 2018	n/a	general	all types	part of national strategy or plan	process, output	qualitative	observational	post	21	
Goaziou, 2018	France	secondary, tertiary	Islamist, all types		process	qualitative	N/A	post	Impossible to determine	1
Government of the United Kingdom, 2011	United Kingdom	targetted primary, secondary	Islamist	entire national strategy or plan	impact, process	mixed	observational	post	1113	
Greiner, 2010	Niger and Chad	primary	all types	program or project	impact	qualitative	observational	post évaluation	182	
Harahap et al., 2019	Indonesia	secondary	Islamist	program or project	process	qualitative	observational	post	Not given	
Harris-Hogan et al., 2019	Australia	general	all types	program or project	impact	quantitative	quasi-experimental	pre-post	117	
Heath-Kelly and Strausz, 2018	United Kingdom	primary	all types	part of national strategy or plan	impact, process, other	mixed	observational	post	335	
Helmus and Klein, 2018	n/a	secondary	right-wing, Islamist	program or project	impact	quantitative	observational	post	Not given	
Hiariej et al., 2017	Indonesia	secondary, tertiary	Islamist	part of national strategy or plan	impact, process	mixed	observational	post	1170	
Hirschi and Widmer, 2012a	Switzerland	primary	right-wing	program or project	impact	quantitative	quasi-experimental	pre-post	Not given	
Hirschi and Widmer, 2012d	Switzerland	primary	right-wing	program or project	impact	quantitative	quasi-experimental	pre-post	747	12
Hirschi and Widmer, 2012e	Switzerland	targetted primary, secondary	right-wing	program or project	impact	quantitative	observational	post	115	
Hirschi and Widmer, 2012g	Switzerland	targetted primary, secondary primary,	right-wing	program or project	impact	quantitative	quasi-experimental	pre-post	Not given	

Iacopini et al., 2011	United Kingdom	targetted primary	Islamist	part of national strategy or plan	impact, process	qualitative	observational	post	37	
Ipp et al., 2014	Tunisia	secondary	Islamist	part of national strategy or plan	impact, other	qualitative	observational	post	14	
Istiqomah, 2011	Indonesia	tertiary	Islamist	program or project	impact	qualitative	observational	post	4	
i-works research ltd., 2013	Wales	secondary	right-wing	part of national strategy or plan	impact	mixed	quasi-experimental	pre-post	Not given	
Jackson et al., 2019	United States	general	all types	part of national strategy or plan	other	qualitative	observational	post	50	
Jailobaeva and Asilbekova, 2017	Kyrgyzstan	tertiary	Islamist	program or project	impact, process, output	mixed	observational	post	41	
Jerome and Elwick, 2016 (Jerome and Elwick, 2019 ; Elwick and Jerome, 2019)	United Kingdom	targetted primary	all types	program or project	impact, process	mixed	quasi-experimental	pre-post	232	
Johns et al., 2014	Australia	targetted primary	Islamist	program or project	impact	mixed	observational	post	39	
Johnston et al., 2008a	Pakistan	primary	Islamist	program or project	impact	mixed	observational	post	Not given	
Johnston et al., 2008b	Pakistan	primary	Islamist	program or project	impact	mixed	observational	post	Not given	
Joyce, 2018	United Kingdom	secondary	all types	part of national strategy or plan	process	mixed	observational	post	38	
Khalil and Ipp, 2016	Mali	primary	all types	program or project	impact, other	qualitative	observational	post	Impossible to determine	
Khalil and Zeuthen, 2014	Kenya	primary	all types	program or project	other	qualitative	observational	post	Not given	
Khalil et al., 2019	Somalia	tertiary	Islamist	part of national strategy or plan	process	qualitative	observational	post	102	
Khurshid et al., 2018	Pakistan	targetted primary	all types	part of national strategy or plan	impact	mixed	observational	post	500	
Kollmorgen and Barry, 2017	Thailand	general	all types	program or project	impact, process	qualitative	observational	post	Not given	
Kollmorgen et al., 2019	Kenya	primary	Islamist	part of national strategy or plan	impact, process, output	mixed	observational	post	528	
Kundnani, 2009	United Kingdom	targetted primary, secondary, tertiary	Islamist	entire national strategy or plan	process, output	qualitative	observational	post	56	
Kurtz, 2015 (Kurtz et al., 2016)	Afghanistan	secondary	all types	program or project	impact, process	mixed	quasi-experimental	post	1129	
Kyriacou et al., 2017	United Kingdom	primary, targetted primary	Islamist, all types	entire national strategy or plan	impact, other	mixed	observational	post	9	
L. Parker et al., 2018	Italy, Romania and Sweden	targetted primary	all types	program or project	impact, process	mixed	quasi-experimental	pre-post	192	135
Lakhani, 2012	United Kingdom	targetted primary	Islamist, all types	entire national strategy or plan	process	qualitative	observational	post	56	
Lamhaidi, 2017	Morocco	secondary	all types	program or project	impact, process, output	mixed	observational	post	Impossible to determine	

Letsch, 2018	Tunisia	general	Islamist, all types	entire national strategy or plan	process	qualitative	observational	post	25
Levy et al., 2019	Kyrgyzstan	secondary	Islamist, all types	program or project	impact, process	mixed	observational	post	1644
Liht and Savage, 2013	United Kingdom	targetted primary	Islamist	program or project	impact	quantitative	quasi-experimental	pre-post	81
Lindekilde, 2012	Denmark	targetted primary, secondary, tertiary	all types	entire national strategy or plan	process, other	qualitative	observational	post	17
Lindekilde, 2014	Denmark	targetted primary	Islamist	part of national strategy or plan	impact	qualitative	observational	post	Not given
Lobnikar et al., 2019	Croatia	general	all types	program or project	impact	quantitative	quasi-experimental	pre-post	108
Madriaza et al., 2018	France	primary, secondary, tertiary	all types	program or project	impact, process	mixed	quasi-experimental	pre-post	15-81
Manby, 2009a	United Kingdom	targetted primary	all types	program or project	impact	mixed	quasi-experimental	pre-post	5
Manby, 2009b	United Kingdom	secondary	all types	program or project	impact	mixed	quasi-experimental	pre-post	9
Manby, 2009c	United Kingdom	secondary	all types	program or project	impact	mixed	quasi-experimental	pre-post	7
Manby, 2009d	United Kingdom	secondary	right-wing, Islamist, all types	program or project	impact	mixed	quasi-experimental	pre-post	6
Manby, 2010a	United Kingdom	secondary, tertiary	all types	program or project	impact	mixed	quasi-experimental	pre-post	5
Manby, 2010b	United Kingdom	tertiary	all types	program or project	impact	mixed	quasi-experimental	pre-post	9
Mansour, 2017	Morocco — international	primary	all types	part of national strategy or plan	impact, output	mixed	observational	post	57
Mastroe, 2016	United Kingdom	primary	all types	entire national strategy or plan	impact, process	qualitative	observational	post	20
McDonald and Mir, 2011	United Kingdom	targetted primary	Islamist	program or project	process, other	qualitative	observational	post	48
McDowell-Smith et al., 2017	United States	targetted primary	Islamist	program or project	impact	quantitative	observational	post	75
McGlynn and McDaid, 2016	United Kingdom	primary	all types	part of national strategy or plan	impact, other	qualitative	observational	post	11
McRae, 2010 (McRae, 2009a; McRae, 2009b)	Indonesia	tertiary	Islamist	program or project	impact, process	mixed	observational	post	Not given
Meringolo et al., 2019a	Italy	targetted primary	all types	program or project	impact	qualitative	observational	post	18
Meringolo et al., 2019b	Italy	targetted primary	all types	program or project	impact	mixed	observational	post évaluation	19
Mitts, 2017	United States	targetted primary	Islamist	part of national strategy or plan	impact	quantitative	quasi expérimentale	pre-post avec séries temporelles	Not given

Moffett and Sgro, 2016	n/a	primary	all types	program or project	output	quantitative	observational	post évaluation	Not given	
Monzani et al., 2018	Kenya	general	Islamist, all types	program or project	impact, other	mixed	quasi expérimentale	pre-post	194	145
Muncy et al., 2015	Nigeria	primary	all types	program or project	impact	qualitative	observational	post évaluation	191	
Murtaza et al., 2018	Pakistan	secondary	all types	program or project	impact	qualitative	observational	post évaluation	102	
Nicolls and Hassan, 2014	Somalia	targetted primary	all types	program or project	process, output	qualitative	observational	post évaluation	357	
Octavia and Wahyuni, 2014	Indonesia	secondary, tertiary	left-wing, right-wing, Islamist	program or project	process	qualitative	observational	post évaluation	Impossible to determine	
Onyima, 2017	Nigeria	targetted primary	Islamist	entire national strategy or plan	impact, process	qualitative	observational	post évaluation	68	
Orban, 2019	Norway	targetted primary	Islamist	part of national strategy or plan	impact, process	qualitative	observational	post évaluation	17	
O'Toole et al., 2012	United Kingdom	targetted primary	Islamist	program or project	impact, process, output	mixed	quasi-experimental	pre-post	Not given	
O'Toole et al., 2013 (O'Toole et al., 2016)	United Kingdom	tertiary	Islamist	program or project	impact, process	qualitative	observational	post évaluation	112	
Parker and Lindekilde, 2020 ⁵⁰	Denmark	targetted primary	all types	program or project	impact	quantitative	expérimentale	pre-post	955	976
Peracha et al., 2016	Pakistan	tertiary	Islamist	program or project	impact, process	qualitative	observational	post évaluation	4	
Peterson, 2012	Sweden	targetted primary	Islamist	part of national strategy or plan	process	qualitative	observational	post évaluation	13	
Piasecka, 2019	United Kingdom	targetted primary	right-wing, Islamist	program or project	impact, other	qualitative	observational	post évaluation	Not given	
Pickering et al., 2008	Australia	targetted primary	Islamist	part of national strategy or plan	process	mixed	observational	post évaluation	601	
Pipe et al., 2016	Somalia	primary	all types	program or project	impact, other	mixed	observational	post évaluation	2789	
Powers, 2015	United Kingdom	primary,	non spécifique	plan ou stratégie nationale	impact, processus	qualitative	observational	post évaluation	95	
Ranstorp, 2010	targetted primary	all types	entire national strategy or plans	impact, process	processus	qualitative	observational	post évaluation	Not given	
Reeves and Crowther, 2019	Indonesia	general	Islamist	entire national strategy or plan	process	quantitative	observational	post	146	
Reynolds, 2017	United Kingdom	targetted primary	all types	program or project	impact	mixed	quasi-experimental	pre-post	441	
Reynolds and Parker, 2018	n/a	targetted primary	all types	program or project	impact, process	mixed	quasi-experimental	pre-post	54	51


⁵⁰ A manuscript was sent by the authors in 2019.

Rodon, 2018	United Kingdom	targetted primary	all types	program or project	impact, process	qualitative	observational	post	Not given	
Rooke and Slater, 2010	France	secondary, tertiary	all types	program or project	impact, process	qualitative	observational	post	Not given	
Rustan et al., 2018	United Kingdom	targetted primary	Islamist	part of national strategy or plan	impact, process, output	qualitative	observational	post	Not given	
Sabir, 2014	United Kingdom	targetted primary, secondary, tertiary	Islamist	entire national strategy or plan	other	qualitative	observational	post	20	
Saltman et al., 2016	n/a	targetted primary	all types	program or project	impact	mixed	quasi-experimental	pre-post	Impossible to determine	
Salyk-Virk, 2018	United States	general	all types	part of national strategy or plan	impact	qualitative	observational	post	26	
Sarota, 2017	Tanzania	secondary	Islamist	program or project	impact, process	mixed	observational	post	391	
Savage et al., 2014	Kenya	targetted primary	Islamist	program or project	impact	mixed	quasi-experimental	pre-post	24	
Savoia et al., 2016	United States	general	all types	program or project	process	qualitative	observational	post	52	
Savoia et al., 2019	United States	primary	all types	program or project	impact	quantitative	quasi-experimental	pre-post	767	326
Schanzer and Eyerman, 2019	United States	targetted primary, general	Islamist, all types	entire national strategy or plan	impact, process	qualitative	observational	post	Not given	
Schanzer et al., 2016	United States	targetted primary, secondary	all types	part of national strategy or plan	process	mixed	observational	post	382 Départements de police	
Schorn et al., 2010a	Egypt	targetted primary	all types	program or project	impact, process, other	mixed	observational	post	73	
Schorn et al., 2010b	Egypt	targetted primary	all types	program or project	impact, process, other	mixed	observational	post	48	
Schorn et al., 2010c	Egypt	targetted primary	all types	program or project	impact, process, other	mixed	observational	post	15	
Schulze, 2008	Indonesia	tertiary	Islamist	part of national strategy or plan	impact, process	qualitative	observational	post	Not given	
Schumicky-Logan, 2017	Somalia	secondary, tertiary	all types	program or project	impact	mixed	quasi-experimental	pre-post	392	
Schuurman and Bakker, 2016	Netherlands	tertiary	Islamist	program or project	impact, process	qualitative	observational	post	6	
Search for Common Ground, 2011	Indonesia	tertiary	all types	program or project	impact, process	mixed	quasi-experimental	pre-post	Impossible to determine	
SecDev.Foundation, 2016	n/a	primary	all types	program or project	impact	mixed	observational	post	Impossible to determine	
Sheikh et al., 2012	Wales	targetted primary, secondary, tertiary	Islamist	part of national strategy or plan	impact, process, output	mixed	observational	post	65	

Sian, 2015	United Kingdom	targetted primary	all types	part of national strategy or plan	impact	qualitative	observational	post	Not given	
Silverman et al., 2016a	United States	targetted primary	Islamist	program or project	impact	mixed	observational	post	Impossible to determine	
Silverman et al., 2016b	United States	targetted primary	Islamist	program or project	impact	mixed	observational	post	Impossible to determine	
Silverman et al., 2016c	Pakistan	tertiary	right-wing	program or project	impact	mixed	observational	post	Impossible to determine	
Sjøen and Mattsson, 2019	Norway	secondary	all types	part of national strategy or plan	process	qualitative	observational	post	16	
Spalek and Davies, 2012	United Kingdom	secondary	all types	program or project	process	qualitative	observational	post	16	
Speckhard et al., 2018	Iraq	primary, secondary	Islamist	program or project	impact	mixed	observational	post	Impossible to determine	
Speckhard et al., 2019	United States	secondary	Islamist	program or project	other	mixed	observational	post	Impossible to determine	
Supratno et al., 2018	Indonesia	targetted primary	Islamist	program or project	process	qualitative	observational	post	Not given	
Swedberg, 2011a	Niger	general	all types	program or project	impact	mixed	quasi expérimentale	post	217	117
Swedberg, 2011b	Chad	general	all types	program or project	impact	mixed	quasi expérimentale	post	368	152
Swedberg, 2011c	Mali	general	all types	program or project	impact	mixed	quasi expérimentale	post	100	1
Swedberg and Reisman, 2013	Kenya	targetted primary	Islamist	part of national strategy or plan	impact	mixed	quasi expérimentale	post	962	484
Taylor et al., 2016 (Aly et al., 2014)	Indonesia	targetted primary	all types	program or project	impact	mixed	observational	post	21	
Tesfaye and Mohamud, 2016	Somalia	targetted primary	Islamist	part of national strategy or plan	impact	mixed	quasi-experimental	post	504	298
Tesfaye et al., 2018	Somalia	primary	Islamist	part of national strategy or plan	impact	mixed	quasi-experimental	post	937	283
Thomas et al., 2017a	United Kingdom	targetted primary	all types	program or project	process	mixed	observational	post	11	
Thomas et al., 2017b	United Kingdom	targetted primary	all types	program or project	impact, process	qualitative	observational	post	Not given	
Tines et al., 2017	Pakistan	secondary	all types	program or project	impact	mixed	quasi-experimental	pre-post	801	
Tropp et al., 2019a	Rwanda	secondary	all types	program or project	impact	quantitative	quasi-experimental	post	26	27
Tropp et al., 2019b	Rwanda	secondary	all types	program or project	impact, output, other	quantitative	quasi-experimental	pre-post (with follow-up)	68	
Tsuroyya, 2017	Indonesia	secondary	Islamist	part of national strategy or plan	impact, other	qualitative	observational	post	4	
Uhlmann, 2017	Germany	secondary, tertiary	all types	part of national strategy or plan	process, output	qualitative	observational	post	Not given	

United Kingdom House of Commons and Communities and Local Government Committee, 2010	United Kingdom	targetted primary	Islamist	entire national strategy or plan	process	qualitative	observational	post	33	
United States Government Accountability Office (GAO), 2017	United States	general	all types	entire national strategy or plan	output, audit	other	observational	post	6	
University of Amsterdam, 2013 (Feddes et al., 2015)	Netherlands	targetted primary	Islamist	program or project	impact	mixed	quasi-experimental	pre-post (with follow-up)	46	
Upton and Grossman, 2019	Australia	targetted primary	Islamist	program or project	impact	mixed	observational	post	41	
Van der Heide and Schuurman, 2018	Netherlands	tertiary	all types	program or project	impact	qualitative	observational	post	72	
Veldhuis, 2015	Netherlands	tertiary	all types	part of national strategy or plan	process, other	mixed	observational	post	Not given	
Veldhuis et al., 2010	Netherlands	tertiary	all types	part of national strategy or plan	process	qualitative	observational	post	Not given	
Vermeulen, 2014a	United Kingdom	targetted primary	all types	part of national strategy or plan	impact, process	qualitative	observational	post	12	
Vermeulen, 2014b	Germany	targetted primary	all types	part of national strategy or plan	impact, process	qualitative	observational	post	12	
Vermeulen, 2014c	Netherlands	targetted primary	all types	part of national strategy or plan	impact, process	qualitative	observational	post	12	
Vittum et al., 2016	Kenya	secondary	all types	program or project	process	qualitative	observational	post	47	
Walsh and Gansewig, 2019	Germany	targetted primary	right-wing	program or project	impact, process	mixed	expérimentale	pre-post	564	
Warrington, 2018	Denmark	primary	all types	part of national strategy or plan	other	qualitative	observational	post	Not given	
Waterhouse Consulting Group, 2008	United Kingdom	targetted primary	Islamist	part of national strategy or plan	impact	qualitative	observational	post	Impossible to determine	
Webb, 2017	United Kingdom	primary	all types	entire national strategy or plan	process, other	mixed	observational	post	35	
Webber et al., 2018 (Kruglanski et al., 2014)	Sri Lanka	tertiary	other	program or project	impact	quantitative	quasi-experimental	pre-post (with follow-up)	669	255
Weeks, 2017	United Kingdom	tertiary	all types	part of national strategy or plan	impact, process	qualitative	observational	post	23	
Weine et al., 2016	United States	primary	all types	program or project	process	qualitative	observational	post	Not given	
Wilchen Christensen, 2015	Norway	secondary	right-wing	part of national strategy or plan	process	qualitative	observational	post	Not given	
Williams et al., 2016	United States	targetted primary, secondary	Islamist	program or project	impact, process	mixed	quasi-experimental	post	323	46
Wilner and Rigato, 2017	Canada	primary	all types	program or project	output, other	mixed	observational	post évaluation	Not given	
Wilson and Krentel, 2018	United Arab Emirates	targetted primary	all types	program or project	impact, process	mixed	quasi-experimental	post with follow-up	Not given	

Winston and Strand, 2013	United Kingdom	targetted primary	all types	program or project	impact, process	mixed	observational	post	260	
Young et al., 2016a	Germany	targetted primary	all types	entire national strategy or plan	process	qualitative	observational	post	Not given	
Young et al., 2016b	Germany	secondary, tertiary	all types	entire national strategy or plan	process	qualitative	observational	post	Not given	
Young et al., 2016c	Denmark	secondary, tertiary	Islamist	program or project	process	qualitative	observational	post	Not given	
Young et al., 2016d	United Kingdom	tertiary	right-wing	program or project	process	qualitative	observational	post	Not given	
Young et al., 2016e	Netherlands	tertiary	Islamist	program or project	impact, process	qualitative	observational	post	Not given	
Younis and Jadhav, 2019	United Kingdom	primary, targetted primary	all types	part of national strategy or plan	other	qualitative	observational	post	16	



Appendix B: Complete methodology of this systematic review

The methodology that we used to conduct this systematic review is based on the review methods of the Campbell Collaboration (<https://www.campbellcollaboration.org>). We adopted their definition of a systematic review as “a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyze data from the studies that are included in the review” (Moher et al., 2009, p. 1). To develop our review strategy, we used the Methodological Expectations of Campbell Collaboration Intervention Reviews (MECCIR) conduct standards and the PRISMA Statement checklist and flowchart.

B1 OBJECTIVES, RESEARCH QUESTIONS, AND KEY DEFINITIONS

a) Objectives

The overall objective of this systematic review was to inventory all evaluations of programs for prevention of violent extremism (PVE) as reported in publications through December 2019.

In addition to this overall objective, we had the following specific objectives:

1. Identify the methodologies used in evaluations of PVE programs
2. Identify the shortcomings in the literature on evaluation of PVE programs
3. Assess the methodological quality of the existing evaluation studies in this field
4. Make recommendations for evaluation of PVE programs.

b) Research questions

Our main research question was therefore, “On the basis of the literature, what are the main recommendations that can be made regarding evaluation of programs for prevention of violent extremism?” This main question involved sub-questions associated with specific key concepts.

Specific key questions:

- 1) What primary prevention programs have been evaluated?
- 2) What secondary prevention programs have been evaluated?
- 3) What tertiary prevention programs have been evaluated?
- 4) What other prevention programs, not classified as primary, secondary or tertiary, have been evaluated?
- 5) What recommendations might be made regarding evaluation of such programs, in light of the opinions expressed by the practitioners and researchers involved in the studies that we reviewed?

For each study that we reviewed, we attempted to answer the following specific sub-questions:

- 1) What theoretical evaluation approach was used in this study?
- 2) What evaluation method was used?
- 3) What strategies, tools and indicators were used to conduct the evaluation?
- 4) How were the findings for these programs defined and measured?
- 5) What was the target population of the evaluated program?
- 6) What method was used to assess the quality of the evaluation?

c) Key definitions

Drawing inspiration from Schmid (2013), in this systematic review we distinguish between radicalization and radicalization to violence. Radicalization is a dynamic process that arises out of the gradual polarization of political, economic, social or religious ideas and that seeks to reject or undermine the status quo. Radicalization can have positive or negative results for individuals and society. It can create opportunities for social change, but it can also aggravate a climate of confrontation between people or groups. When the methods advocated for achieving a radical solution involve legitimizing the use of violence or considering recourse to violent actions, then we can speak of radicalization to violence. Schmid believes that radicalization can in fact serve the cause of democracy, while “extremists can be characterised as political actors

who tend to disregard the rule of law and reject pluralism in society.” (Schmid, 2013, p. 8). There is no consensus definition of terrorism (Weinberg, Pedahzur and Hirsch-Hoefler, 2004). For the purposes of this systematic review, we defined “terrorism” as engaging in acts of violence for the purpose of constraining the government and/or frightening the public so as to achieve political, philosophical, ideological, racial, ethnic, religious or other ends. We used this definition to exclude from our review any evaluations of anti-terrorism programs designed to prevent terrorist attacks.

Radicalization is a process undergone by individuals or groups. When considering society as a whole, we instead use the concept of “social polarization”, meaning the gradual division of society and the social environment into different groups and sub-groups whose identity is based on the exacerbation of opposing characteristics related to basic concepts such as sex, race, religion or political opinions (CPN-PREV, 2020).

By “prevention”, we mean all efforts to reduce or eliminate risk conditions that may make an individual or group more vulnerable to violent extremism or to recidivism (among individuals who have previously engaged in violence or belonged to extremist groups). As in the field of public health, prevention programs may be aimed at primary prevention (targeting the general population not considered at risk), secondary prevention (targeting individuals or groups that are considered to be at risk or in the initial stages of the process of radicalization to violence), or tertiary prevention (targeting individuals or groups that are already engaged in the final stages of this process, or that belong to extremist groups, or that have committed acts associated with violent extremism). In the case of PVE programs, we make a further distinction between primary prevention programs and targeted primary prevention programs; the latter, though universal, target a specific community.

In the present systematic review, we regard the concepts of “prevention of radicalization to violence” and “prevention of violent extremism” as synonymous but use mainly the latter and its abbreviation, PVE, for convenience. But we do distinguish PVE measures from counterterrorism measures. The former target individuals who are vulnerable to becoming involved in violent extremism, while the latter are designed to address security threats and prevent or deter terrorist attacks. Arce and Sandler (2005) also distinguish between proactive and defensive counterterrorism measures. Proactive counterterrorism measures are often carried out directly by governments or their agents, against terrorists or their sponsors; examples of such measures would include destroying terrorist training camps, taking reprisals against sponsor states and infiltrating terrorist groups. In contrast, defensive counterterrorism measures are aimed at deterring terrorist attacks “by either making success more difficult or increasing the likely

negative consequences to the perpetrator“; examples would include building technological barriers, hardening potential targets, and securing borders (Arce and Sandler, 2005, p. 184).

Lastly, we adopt the definition of “evaluation” given by the United Nations Evaluation Group:

An evaluation is an assessment, conducted as systematically and impartially as possible, of an activity, project, program, strategy, policy, topic, theme, sector,

operational area or institutional performance. It analyses the level of achievement of both expected and unexpected results by examining the results chain, processes, contextual factors and causality using appropriate criteria such as relevance, effectiveness, efficiency, impact and sustainability. An evaluation should provide credible, useful evidence-based information that enables the timely incorporation of its findings, recommendations and lessons into the decision-making processes of organizations and stakeholders. (UNEG, 2016, p. 10).

B2 INCLUSION AND EXCLUSION CRITERIA

For this systematic review, we adopted maximally inclusive criteria so as to increase the likelihood of finding relevant studies despite variations in their methodological and theoretical frameworks. The following paragraphs summarize the criteria that we applied to determine whether a study was eligible for this review.

Our review targetted all studies published up to and including December 2019 in which primary, evidence-based data were used to evaluate PVE programs.⁵¹ The purpose of such programs is to reduce or eliminate risk conditions that may make an individual or group more vulnerable to becoming involved in violent extremism, or to recidivism.⁵² In keeping with the UNEG definition of evaluation, we included all studies whose purpose was to assess or judge a PVE program, project or strategy, even if they did not use the term “evaluation” explicitly. The target populations of the programs evaluated in these studies had to consist of adults. We thus targetted all evaluations of primary, secondary and tertiary PVE programs⁵³ that attempted to change the attitudes, emotions or behaviours of the target individuals or groups; of their families, friends and acquaintances; and of practitioners who work in this field. We excluded evaluations of programs that work with direct or indirect victims of terrorist actions,⁵⁴ evaluations of counterterrorism measures, and studies that evaluated continent-wide strategies or provided overall assessments of a continent-wide approach.

Because one publication can discuss more than one study, the unit of analysis for this review was the individual published study rather than the publication. We regarded a publication as discussing more than one study if it a) discussed more than one sample that had been analyzed independently and b) presented independent results for that sample.

Apart from distinguishing among the three levels of prevention, there were no other criteria that we could use to classify the programs. We therefore described the variables to be considered on the basis of a comparison among these three levels of prevention.

To be included in this review, the studies also had to have been written in English, French or Spanish (the languages read and spoken by the members of the research team).

As long as all of these conditions were met, we did not impose any further restrictions regarding the methodological characteristics of the studies.

⁵¹ Secondary data are data collected by someone other than the studies’ authors or their teams. Examples of secondary-data sources in the social sciences include population censuses, data collected by government departments, organizational records, and other data that were originally collected for purposes other than the research in question.

⁵² See the key definitions in the preceding section.

⁵³ Ibid.

⁵⁴ The families of the individuals who engaged in this process may be regarded as indirect victims of extremist groups. But here we understand “victims” to mean individuals and their families who were the target of attacks, attempted attacks or other violent acts by extremist groups.

B3 VARIABLES CODED

Each study included in this review was coded according to a global coding frame and a tool for appraising methodological quality.

a) Global coding frame

The following table shows the global coding frame that we developed for purposes of coding and then aggregating the data from the studies that we reviewed. The coding was done by a team of research assistants, using this tool.

Dimension	
Variable	Operational definition
General description of study	
Author	Author's name
Country	Country where the PVE program was delivered
Peer-reviewed	Whether the study was subjected to a blind peer review, as is typically the case for articles published in scientific journals
Funding sources	Whether the authors mention the sources of funding for their study (if yes, specify these sources)
Conflicts of interest	Whether the authors state their conflicts of interest
	List of stated conflicts of interest
	List of unstated conflicts of interest
Author(s) of study	
Gender	Author's gender
Country of origin	Author's country of origin
Discipline	Author's discipline
Profession	Author's profession
Number of publications as sole author	Number of publications as sole author, in the field of security studies
Number of publications as co-author	Number of publications as co-author, in the field of security studies
Number of publications in the database	Number of publications in the database for this systematic review
Region of first publication	Geographic region of the author's first publication
Prevention level	
Primary	All efforts that seek to reduce or eliminate risk factors or encourage protective factors and that target the general public not identified as being at risk. Primary prevention is a type of universal prevention; awareness campaigns are an example of primary prevention programs.
Targetted primary	All efforts that seek to reduce or eliminate risk factors or encourage protective factors and that target a specific community that is not identified as being at risk. Example: universal prevention programs in Muslim communities.
Secondary	All efforts that seek to reduce or eliminate risk factors or encourage protective factors and that target individuals or groups regarded as at risk and in the initial stages of the process of radicalization to violence.
Tertiary	All efforts that seek to reduce the factors that encourage recidivism among individuals or groups that are in the final stages of the process of radicalization, or who belong to extremist groups or have committed acts associated with violent extremism or with terrorism. Tertiary prevention programs also attempt to reintegrate such individuals and groups into society.
General	Prevention level not clearly indicated in the study

Type of violent extremism targeted

Left-wing (or synonyms)	The study clearly states that the program or project directly targets this type of extremism.
Right-wing (or synonyms)	The study clearly states that the program or project directly targets this type of extremism.
Islamist (or synonyms)	The study clearly states that the program or project directly targets this type of extremism.
Anarchist (or synonyms)	The study clearly states that the program or project directly targets this type of extremism.
Other	The program or project targets any other type of extremism that does not fit the other definitions.
All types	The study clearly states that the program or project targets all types of extremism. This is often the case for

Type of violent extremism targeted

Impact (summative)	An impact evaluation answers the question, “What worked?” In other words, it examines the effects that the intervention had on the participants and whether these effects matched the objectives that had been set. Impact evaluations assess how an intervention contributes to achieving a result or objective. That contribution may be intentional or unintentional, positive or negative, and long-term or short-term. Impact evaluations attempt to identify clear links between causes and effects and to explain how the intervention worked and for whom it worked.
Process (formative)	A process evaluation answers the questions “Why does it work?“, “How does it work?” and “How can we improve this process?” A process evaluation thus focuses on the factors that determine or influence the implementation of the program or project activities and provides insight into the changes that happen in the course of them. A process evaluation may start after the intervention begins (formative evaluation), or while it is under way (process evaluation) or in the middle of it (mid-course evaluation).
Output	Evaluation conducted after a program or a phase of a program is over, to determine to what extent the planned activities were carried out.
Audit	A quality-control evaluation, conducted objectively and independently, for the purpose of improving the operations of an organization and increasing their value. An audit helps the organization to achieve its objectives through a rigorous, systematic approach to observing and improving the effectiveness of risk management, control and governance processes.
Monitoring	An ongoing process of using selected indicators to systematically gather data about an action in progress, in order to let managers and stakeholders know what progress and objectives have been achieved and how the allocated funds are being spent.
Other	Any other type of evaluation

Evaluator type

Internal	Evaluation conducted by the people or department responsible for designing and implementing the program or project within the organization delivering it, or by its partner organizations or its funding agency.
Joint	Evaluation conducted by multiple funding agencies and/or their partners, but excluding program participants and practitioners.
Participatory	Evaluation in which all stakeholders (including program participants, practitioners and researchers) collaborate in designing it, conducting it and drawing conclusions from it.
External (independent)	Evaluation conducted by people and/or departments other than those responsible for designing and implementing the program or project, or from outside of the organization delivering it or the agency funding it.

Methodological design: according to overall approach

Quantitative	Studies that use quantifiable variables, gather quantitative data directly (through observations) or indirectly (through surveys), and perform statistical analyses of these quantitative data (numerically encoded observations, survey responses, etc.)
Qualitative	Studies that use qualitative methods for gathering and analyzing data (participants’ observations, ethnographies, interviews, focus groups, etc.)
Mixed (or mixed-methods)	Study that uses both quantitative and qualitative methods

Other	Any other overall approach
-------	----------------------------

Methodological design: according to manipulation of variables

Experimental (quantitative randomized controlled trials)	<p>A study that uses an experimental design actively manipulates the independent variable. In other words, the researcher arbitrarily selects the values of the independent variable (the intervention, for example) and applies them to various groups of subjects to test for a cause-and-effect relationship.</p> <p>Measurements are taken at a minimum of two points in time (before and after the intervention) and in more than one group. Normally, a study with an experimental design has a control group and an experimental group, and the subjects are randomly assigned to one group or the other.</p>
--	--

Quasi-experimental (quantitative non-randomized controlled trials)	A study with a quasi-experimental design also attempts to test for a cause-and-effect relationship between an intervention and measurements taken before and after it, but unlike in an experimental design, either there is no control group, or the groups tested are natural, intact or already formed, as opposed to being created randomly.
--	--

Other	
-------	--

Methodological design: according to program participants

Control group	A group of subjects who closely resemble the experimental group with regard to several demographic variables but do not receive the intervention and are thus used for purposes of comparison when the results of the intervention are evaluated.
---------------	---

Methodological design: according to whether measurements were taken repeatedly

Repeated measurements	In a program evaluation with repeated-measurement designs, measurements are taken on the same subjects at two or more points in time.
-----------------------	---

Post-evaluation	In a program evaluation with a post-evaluation design, measurements are taken at only one point in time, after the program ends or one of its cycles has been completed.
-----------------	--

Methodological design: according to number of independent variables

Simple	Only one independent variable
--------	-------------------------------

Complex or factorial	More than one independent variable
----------------------	------------------------------------

Methodological design or approach: according to number of dependent variables

Simple	Only one dependent variable
--------	-----------------------------

Complex or factorial	More than one dependent variable
----------------------	----------------------------------

Data-collection tools

Surveys	A survey is a method in which quantitative data are collected by means of a set of standardized questions that a sample of individuals are asked in order to determine various facts or their opinions on various matters.
---------	--

Interviews	An interview is a method of collecting qualitative data that is used in the social sciences to determine and examine an individual's opinions and attitudes about a specific subject through a conversational model.
------------	--

Focus groups	A focus group is a method of collecting qualitative data that is used in the social sciences to determine and examine the opinions and attitudes of a group of individuals with regard to a specific subject.
--------------	---

Observations	<p>Observations are a data-collection method that can be used in both qualitative and quantitative studies. In qualitative studies, researchers conduct observations to familiarize themselves with a particular group of individuals (such as a religious group, or a professional group, or a sub-culture or a particular community) and their practices. To conduct such observations, the researchers engage with the individuals intensively, in their own cultural environment, generally over a long period.</p> <p>In quantitative studies, researchers conduct observations by using a predesigned observation grid to collect data that will be quantified and analyzed statistically.</p>
--------------	--

Other	
-------	--

Scope of intervention evaluated

Entire national strategy or plan	Evaluation of all actions taken under a national strategy or plan
----------------------------------	---

Part of a national strategy or plan	Evaluation of some of the actions taken under a national strategy or plan, within a specific sample, sector or geographic area
-------------------------------------	--

Individual program or project	Evaluation of an individual prevention action designed to achieve specific objectives with predefined resources and a predefined work plan
-------------------------------	--

Sample	
Participants in the experimental group	Number of participants in the group receiving the intervention
Participants in the control group	Number of participants in the control group
Target population	
Individuals directly involved	Applies when interventions are directed at specific individuals and, in particular, when the goal is secondary or tertiary prevention, meaning that these individuals are already in the process of radicalization to violence or have already committed acts of violent extremism.
Families	Applies when a service is offered to the families of individuals who are already in the process of radicalization to violence or have already committed acts of violent extremism.
Community	Applies when the intervention involves working at the local level with community members other than families of individuals who are already in the process of radicalization to violence or have already committed acts of violent extremism (this is the case for most primary-prevention programs).
Societal group	Applies when the intervention involves working with a specific societal group (such as youth, Muslims, or women) but not with society as a whole
Society	Applies when the target of the intervention is the entire society, as in a primary or universal prevention program such as an awareness campaign.
Practitioners	Applies when another goal of the intervention is to work with everybody who has direct contact with the participants
Government	Applies when the intervention involves building prevention capacities within a government agency
Target setting	
Community	When the intervention involves working with the individual's broader community (excluding family) on the local level
Security	When the intervention targets law enforcement and the armed forces
Primary and secondary	When the intervention targets students, teachers and administrators in the primary and secondary education sector
Post-secondary education	When the intervention targets students, teachers and administrators in the post-secondary education sector
Justice	All agencies of the justice system (such as juvenile justice and the courts), excluding the correctional system and
Government	All institutions of government, excluding education, health and correctional settings
Cultural	
Correctional	When the intervention targets offenders in prisons, intermediate correctional settings and the probation system
Private sector	When the intervention targets employees of not-for-profit organizations
Health	All physical and mental health institutions
Other	
Type of indicators used or results obtained (quantitative or qualitative factors or variables that constitute simple, reliable means of measuring and reporting changes related to the intervention)	
Direct	Indicators that directly measure radicalization, violent extremism or sympathies for these phenomena
Indirect	Indicators not directly related to radicalization, violent extremism or sympathies for these phenomena—for example, self-esteem, leadership, etc.
Indicators used or results obtained	List of reported indicators
Types of effects	
List of positive and negative effects reported	Positive
	Negative
	Other

Limitations

Limitations	Do the authors report the limitations of the study?
Types of limitations reported	List of limitations reported

B4 Tool for appraising methodological quality

In addition to coding the preceding variables, we used the Mixed Methods Appraisal Tool (MMAT) (Hong et al., 2018; Hong and Pluye, 2019) to appraise the methodological quality of the evaluation studies included in this systematic review. Unlike other evaluation tools, the MMAT can be used to evaluate all of the different kinds of studies that we included in this review (qualitative, quantitative descriptive, experimental, quasi-experimental and mixed designs). Because we wanted to identify all methodologies that have been used to evaluate PVE programs, we did not use the MMAT as a criterion for including studies in

this review. We used it only to determine the quality of the methodologies used in the PVE evaluations that we did include.

The MMAT consists of 25 criteria divided into five groups representing the five types of designs just mentioned. This tool is used to assign each study a quality rating on a scale of 0 to 5. However, for studies that use mixed designs, the criteria associated with each design type must be coded. A study that uses mixed methodologies can thus potentially obtain a score of 0 to 25.

B5 Literature search strategies

The following Table shows the English and French keywords that we used to search the literature.

ANGLAIS	<p>(Extremi* OR Radicali* OR "Violent Extrem*" OR Indoctrinat* OR Terrori* OR "Homegrown Terror*" OR "Homegrown Threat*" OR "Radical Islam*" OR "Islamic Extrem*" OR "Religious Extrem*" OR Fundamentalis* OR Jihad* OR Islam* OR Salaf* OR "Lone wol*" OR "lone-wol*" OR "lone actor*" OR "foreign fight*" OR Returne* OR "White Supremacis*" OR "Neo-Nazi" OR "Right Wing" OR "Right-wing Extrem*" OR "far right" OR Fascis* OR "Left-wing Extrem*" OR "Left Wing" OR Anti-Semitis* OR Antifa* OR Anarch* OR "Eco-terror*" OR "Al Qaida-inspired" OR "ISIS-inspired" OR "Anti-Capitalis*" OR Incel* OR "Al Qaeda" OR ISIS OR ISIL)</p> <p>AND</p> <p>(Prevent* OR interven* OR respon* OR policy OR policies OR program* OR strategy* OR initiative* OR assess* OR eval* OR procedur* OR effect* OR *success* OR reduc* OR treat* OR counterterror* OR "counter-terror*" OR "de-radicali*" OR deradical* OR disengag* OR detect* OR "countering violent extrem*" OR CVE OR PVE OR Reint* OR Rehabilitat*)</p> <p>NOT</p> <p>(Cancer OR Disease OR hematoma OR "heart disease" OR "heart failure" OR cardiovascular OR "vortex generator*" OR "heat transfer" OR "bone" OR "fracture healing" OR "bone density" OR epilepsy OR "multiple sclerosis" OR Femin*)</p>
FRANÇAIS	<p>(Extremi* OR Radicali* OR "Extrem* Violent" OR Endoctrin* OR Terrori* OR " Terror* Domestique" OR "Islam* Radical" OR "Extrem* Islam*" OR "Extrem* Relig*" OR Fundamentalis* OR djihad * OR Islami* OR Salaf* * OR "Loup* solitaire*" OR "acteur solitaire *" OR (combattant* AND (étranger* OR terroriste*)) OR "Extrême droite" OR Suprémac* OR "Néo-Nazi" OR Néonazi* OR Fachis* OR "Extrem* Gauch" OR Antifa* OR Anti-Semitis* OR Anarch* OR "Eco-terror*" OR Incel* OR "Al Qaeda" OR ISIS OR ISIL)</p> <p>AND</p> <p>(Prevent* OR interven* OR repon* OR politique* OR program* OR stratégie* OR initiative* OR eval* OR procedur* OR effet* OR effect* OR succès OR réussi* OR résultat* OR reduc* OR traitem* OR contreterror* OR "contre-terror*" OR "de-radicali*" OR deradical* OR disengage* OR CVE OR PVE OR Reinteg* OR Rehabilitat* OR reinsert*)</p> <p>NOT</p> <p>(Cancer OR Maladi* OR Hématom* OR "Maladi* cardia*" OR "Insuffisan* cardia*" OR Cardiovasculair* OR "Générat* de tourbillon*" OR "Transfer* de chaleur*" OR Os OR "Consolid* de fractur*" OR "Densit* osseu*" OR Épileps* OR "Scléro*" OR femin*)</p>

Using the above inclusion criteria, exclusion criteria and keywords, we:

- searched the scientific literature
- searched the grey literature
- compared our findings with other frequently cited literature reviews, plus applied a “snowball” search strategy.

In addition, we consulted 14 experts by email to find out whether they knew of any other relevant studies.

a) Scientific literature

For the scientific literature, we had a librarian with expertise in the social sciences and humanities apply our search criteria to the following 21 databases.

ABI/Inform Global	International Political Science Abstracts
Academic Search Complete	Medline
ATLA Religion Database	OpenGrey.eu
Canadian Business et Current Affairs Complete	PAIS Index
Communication Abstracts	Political Science Complete
Canadian Public Policy Collection	ProQuest Dissertations et Theses Global
Canadian Research Index	PsycINFO
Education Source	Sociological Abstracts
ERIC	Sociological Index
Erudit / Persee	Web of Knowledge
FRANCIS	

These 21 databases contained not only published scientific articles and academic theses, but also a large volume of grey literature and conference papers. We also obtained access to the database from two recent systematic reviews by the Canadian Practitioners Network for the Prevention of Radicalization and Extremist Violence (CPN-PREV) (Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021; Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021) and merged this database with the 21 others.

b) Grey literature

To reduce “publication bias” (Rothstein et al., 2005) in our strategic review, we used Google to conduct an in-depth search of the grey literature. To identify additional documents, we also manually examined 228 websites of organizations involved in PVE. We selected these organizations from the UNESCO-PREV Chair’s map of centres of expertise in PVE (<https://chaireunesco-prev.ca/en/network/map/>). We also added other organizations in the course of this search. Table 32 is a complete list of the selected organizations.

Table 32. Organizations whose websites we searched manually

Academy on Human Rights and Humanitarian Law at American University's Washington College of Law	Center for Strategic and International Studies (CSIS)
Afghanistan Justice Organization	Center on Global Counterterrorism cooperation (CGCC)
AfPak programme Afghanistan/Pakistan (PSF)	Center on International Cooperation at New York University
Againstviolentextremism.org	Centre africain d'Etudes Internationales, Diplomatiques, Economiques et Stratégiques, en abrégé (CEIDES)
Akin Gump Strauss Hauer & Feld LLP	Centre for Environment, Human Rights and Development (CEHRD - Nigéria)
Alliance for Peacebuilding – Monitoring and Evaluation of CVE	Centre for Environmental Education and Development (CEED - Nigéria)
Alternative espaces citoyens (AEC - Niger)	Centre for Peace And Advancement (CEPAN - Nigéria)
Alternative to Violence Project	Centre for Research and Evidence on Security Threat (CREST)
American Bar Association Rule of Law Initiative (ABA ROLI)	Centre for the Advocacy of Justice and Rights (CAJR)
Amicale université populaire (Tchad)	Centre pour la Gouvernance Democratique
Amicus Legal Consultants	Century Foundation
AML Solutions International	Charity & Security Network
Amnesty International	Children and Young People Living for Peace (Nigéria)
Anti-Defamation League	Christian Foundation for Social Justice and Equity (CFSJE - Nigéria)
APO.org	Civipol
Asser Institute	Cleen Foundation (Nigéria)
Association burkinabé d'action communautaire (ABAC-ONG - Burkina Faso)	Clingendael – Netherlands Institute of International Relations
Association des jeunes juristes et sympathisants de Sikasso (AJJSS - Mali)	Club UNESCO de l'Université Abdou Moumouni (CUAM - Nigéria)
Association for Progressive Communications	CODE PAKISTAN
Association jeunesse pour la paix et la non-violence (AJPNV - Tchad)	Collectif des organisations de défense des droits de l'homme et de la démocratie (CODDHD - Niger)
Association of Francophone Supreme Courts (AHJUCAF)	Comité Interministériel de prévention de la délinquance et de la radicalisation (CIPDR)
Association pour l'enseignement coranique et la protection des enfants mouhadjirine (AECPEM - Tchad)	Commission Européenne
Association pour le dialogue entre les jeunes de diverses religions (ADJR - Tchad)	Community Motivation and Development Organization (CMDO)
Association rayons de soleil (Cameroun)	Community Policing Partners for Justice, Security & Democratic Reform (Nigéria)
Association tchadienne pour la promotion et la défense des droits de l'homme (ATPDH - Tchad)	Conflict Resolution Trainers Network (CROTINN - Nigéria)
Attah Sisters Helping Hand Foundation (ASHH - Nigéria)	Council of Europe
Baker & McKenzie	COWI
Bangladesh Enterprise Institute (BEI)	Danish Ministry of Defence (Broad Peace and Stabilisation Fund)
Bangladesh Institute of Peace and Security Studies (BIPSS)	Danish security and intelligence service
Better World Campaign	Defence, Australian Government
Bipartisan Policy Center	Design Monitoring and Evaluation for Peacebuilding
Blumont.org	Development Initiative of West Africa (DIWA - Nigéria)
Brennan Center for justice	Development, Education and Advocacy Resources for Africa (DEAR Africa - Nigéria)
Brookings Institution	Djamah-Afrik (Tchad)
Burkina Faso CRADHE	Dorwood Consultancy
Cadre africain de coopération civilo-militaire (CCCM- Niger)	East Africa Judges' and Magistrates' Association (EAJMA)
Care Fronting (Nigéria)	
Center for Evidence Based Crime Policy CEBCP	
Center for prevention of radicalization leading to violence	

Economic Community of West African States (ECOWAS)
 Educateagainsthat.com
 Education and Community Development
 EducommunicAfrik (Burkina Faso)
 Emergency Preparedness and Response Team (JDPC- Nigéria)
 Equal Access International
 EU Agency for Fundamental Rights
 European Counter-Radicalization and de-radicalization
 European Judges Training Network (EJTN)
 Exit Sweden
 Fantsuam Foundation (Nigéria)
 Federation burkinabé des associations, centres et clubs UNESCO (FBACU- Burkina Faso)
 Fondation Hirondelle (Niger et Mali)
 Ford Foundation
 Fourth Freedom Forum
 French Ministry of Interior Publications Database
 Friedrich Naumann Foundation (South Asia)
 Geneva Centre for Security Policy (GCSP)
 Georgetown University Center for Security Studies
 German National Center for Crime Prevention
 Global Center on Cooperative Security GCCS
 Global Community Engagement and Resilience Fund (GCERF)
 Global Counter Terrorism Forum (GCTF)
 Global Counter Terrorism Forum Violent Extremism (Hedayah)
 Global Initiative against Transnational Organized Crime
 Global Partnership for the Prevention of Armed Conflict
 Graduate Institute of International and Development Studies
 GW Program on Extremism
 Henry L. Stimson Center
 Hope for the Needy Association (HOFNA - Cameroun)
 Horn of Africa (HoA) programme (PSF)
 Human Rights First
 Human Rights Institute at Columbia University Law School
 Human security collective
 ICF
 IDP Goods (Cameroun)
 Impact Europe
 Inganta Rayuwa Peace Network (Nigéria)
 Insan Foundation
 Institut national de la statistique et des études économiques (INSEE)
 Institut of Security Studies
 Institute for Inclusive Security
 Institute for Justice and Reconciliation
 Institute for Social Policy and Understanding
 Institute for strategic dialogue (ISD)
 Integrity research and consultancy
 Integrityglobal.com
 Interfaith Council of Muslim and Christian Women's Associations (Nigéria)
 Intergovernmental Authority on Development (IGAD)
 International Association of Chiefs of Police (IACP)
 International Centre for Counter-Terrorism – The Hague (ICCT)
 International Centre for Peace, Charities and Human Development (INTERCEP - Nigéria)
 International Centre for the study of Radicalisation (ICSR)
 International Centre of Excellence for Countering Violent Extremism
 International Crisis Group
 International Institute for Justice and the Rule of Law (IIJ)
 International Monetary Fund (IMF)
 International Organization for Judicial Training (IOJT)
 International Peace Institute (IPI)
 International Republican Institute (IRI -Niger, Mali)
 Interpol
 Islamabad Policy Research Institute
 Islamic Counselling Initiatives of Nigeria (ICIN - Nigéria)
 Istituto Affari Internazionali
 Kecosce
 Kingsfaith Development and Youth Empowerment Initiative (Nigéria)
 Knowledge Platform Security & Rule of Law
 Leadership Initiative for Transformation and Empowerment (LITE- Africa - Nigéria)
 Leiden university
 Media Women for Peace (Cameroun)
 Ministry of Foreign Affairs of Denmark
 Moonshot
 Mouvement des jeunes pour le développement et l'éducation citoyenne (MOJEDEC - Niger)
 Nahdatul Ulama (NU)
 National Consortium for the Study of Terrorism and Responses to Terrorism (START)
 National Counterterrorism Center
 National Endowment for Democracy
 NATO Science for Peace and Security Program
 Neem Foundation (Nigéria)
 New Era Educational and Charitable Support Foundation (Nigéria)
 North East Youth Initiative for Development (Nigéria)
 Norwegian Ministry of Foreign Affairs

Observer Research Foundation (ORF)

Office of the United Nations High Commissioner on Human Rights (OHCHR)

ONG Adkoul (Niger)

ONG Jeunesse-enfance-migration-développement (JMED - Niger)

Open Society Foundation

Organisation for Security and Cooperation in Europe (OSCE)

Organisation pour la réflexion, la formation et l'éducation à la démocratie et au développement (ORFED - Mali)

OXFAM

PAIMAN Alumni Trust

Pak Institute for Peace Studies Pvt Ltd. (PIPS)

Peace and Stabilisation Fund (Danemark)

Peace Empowerment Foundation (Nigéria)

Peace Initiative Network (PIN) (Nigéria)

Prevention of and Fight against crime programme of the European union European commission

RAND Corporation

Regional Center for Strategic Studies

Réseau de Réflexion Stratégique sur la Sécurité au Sahel

Réseau panafricain pour la paix, la démocratie et le développement (REPPADD)

Royal Canadian Mounted Police (RCMP)

Royal United Services Institute (RUSI)

SaferWorld

Salesforce

Search for common Ground

South Asian Association for Regional Cooperation (SAARC)

Stockholm International Peace Research Institute (SIPRI)

Stop-djihadisme (France)

Stoppingviolentextremism.org

Strong Cities Network (SCN)

Tabara Youth Transformation Initiative (TYTI- Nigéria)

Taimako Community Development Initiative (Nigéria)

Tech Against Terror

The Campbell Collaboration

The Global Observatory

The John Sloan Dickey Center of International Understanding – Dartmouth University

The Prevention Project

The Unity Initiative (TUI)

U.S. Agency for International Development (USAID)

UiO C-REX - Center for Research on Extremism

UK College of Policing

UK Home Office Research Database

UK Ministry of Defence

UN Counter-Terrorism Implementation Task Force (CTITF)

UN Office of the High Commissioner for Human Rights

UN Office of the Special Adviser on Africa

UN Office on Drugs and Crime (UNODC)

UN Security Council Counter-Terrorism Committee Executive Directorate (CTED)

UN Women

UNESCO

Union Européenne

United Nations

United Nations Association – UK

United Nations Development Programme (UNDP)

United Nations Foundation

United Nations Office for Drugs and Crime's Terrorism Prevention Branch (UNODC)

United Nations Peacebuilding Support Office (PBSO), Niger et projet régional

United States Institute of Peace (USIP)

University Of Cambridge (institute of criminology)

US Department of Homeland Security

US National Criminal Justice Reference Service

Violence Prevention Network (Germany)

West Africa Network for Peacebuilding (WANEP)

Women Against Violent Extremism (WAVE - Nigéria)

Women and Girl Child Rescue and Development Initiative (Nigéria)

Women in International Security (WIIS)

World Affairs Council

World Bank

World Organization for Resource Development and Education (WORDE)

Youth Initiative Against Violence and Human Rights Abuse (YIAVHA - Nigéria)

Youth Justice Board

Youth Progressive Association in Taraba (TYPA - Nigéria)

Youths for Peace Building and Development in Africa (YOUPEDA - Nigéria)

c) Other frequently cited literature reviews

In addition to identifying documents by searching the scientific and grey literature as just described, we compared our findings with other frequently cited literature reviews (see Table 33).

Table 33. Systematic reviews and inventories of the literature on evaluating programs for preventing violent extremism

Literature review	Studies included	Studies excluded			
		CT*	NPD*	NE*	M*
Bellasio et al., 2018	28/48	7	3	2	8
Carthy et al., 2020	0/14	14			
Feddes et Gallucci, 2015	11/55	6	19	2	17
Gielen, 2017	25/73	4	38	3	3
Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021; Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021	44/48	2	1		1
Madriaza et al., 2017; Madriaza et Ponsot, 2015	12/23	6	3		2
Mastroe et Szmania, 2016	16/43	7	14	1	5
Pistone et al., 2019	17/38	5	12	2	2
Pratchett et al., 2010	1/18	4	6		7
Taylor et Soni, 2017	1/7	5		1	

CT: Studies classified as dealing with counterterrorism measures, not directly related to prevention or not dealing with any specific program

NPD: Studies with no primary data or with anecdotal data

NE: Non-evaluation studies

M: Publications inaccessible or merged with other publications that used the same sample and analysis

Every study that we thus found, that had been published in one of our three included languages, and that we had not previously identified, we added to our database. In addition to these reviews, we applied a snowball strategy using the bibliographies of the included studies.

d) Communications with experts

We also consulted 14 experts by email to find out whether they knew of any other relevant studies.

B6 PROCEDURE

Before starting this systematic review, we trained the five research assistants who were working with us, to clarify the concepts and work methodology. To search the scientific literature, we then used two bibliographic databases. One of them came from a similar systematic review done recently by the CPN-PREV team (Hassan, Brouillette-Alarie, Ousman, Kilinc et al., 2021; Hassan, Brouillette-Alarie, Ousman, Savard et al., 2021), with which our review had certain keywords in common. This database covered all existing publications to January 2018. Our librarian searched this database using the criteria previously mentioned and compiled a selection of scientific documents from it. Meanwhile, the research assistants reviewed the grey literature on the websites of the organizations mentioned above. Once collection of data from the grey literature had been completed, the databases were merged and any duplicates were eliminated. Also, the 14 experts were contacted during this period.

To eliminate any ineligible studies, the principal investigator and the research assistants screened the titles and abstracts of all of the documents identified in the above searches. During this first phase, to ensure

consistency, all team members coded the first 700 documents, analyzing and resolving any disagreements about how to code them. This phase also served as training for the team. Next, two coders reviewed each document. To ensure that there was sufficient agreement between the two coders, a Cohen's kappa coefficient was calculated. During this initial coding, we worked iteratively: each pair of coders worked on a limited number of items. Then Cohen's kappa was calculated. If its value fell below the minimum acceptable threshold of 0.6, the two coders reviewed their points of disagreement; if it was 0.6 or higher, they continued coding the next set of documents. The final kappa was 0.86.

The total number of publications selected was 211, but some publications discussed more than one study, so the total number of studies included in our systematic review was 219. (We regarded a publication as discussing more than one study if it discussed more than one sample that had been analyzed independently.)

We used the PRISMA model (<http://www.prisma-statement.org>) to record the results of our searches in the flow chart shown in Figure 1.



United Nations
Educational, Scientific and
Cultural Organization



"UNESCO Chair in Prevention of Radicalisation
and Violent Extremism", Université de Sherbrooke,
Concordia University, Université du Québec à Montréal