LASER PULSE

PREVENTING/COUNTERING VIOLENT EXTREMISM SYSTEMATIC MIXED METHODS REVIEW: **METHODOLOGY**

Jessica Baumgardner-Zuzik | Shaziya DeYoung | Allyson Bachta

SUPPLEMENT TO AGREEMENT NO. AID-7200AA18CA00009

AOR Name: Brent Wells

January 2024

USAID FROM THE AMERICAN PEOPLE | PURDUE UNIVERSITY | ACLED Bringing clarity to crisis | AIR Advancing Evidence. Improving Lives. | Alliance for Peacebuilding

Authors

This publication was produced by the Alliance for Peacebuilding (AfP) under a sub-award funded by United States Agency for International Development (USAID) Long-term Assistance and Services for Research (LASER) Partners for University-led Solutions Engine (PULSE) - Co-operative agreement AID-7200AA18CA00009. It was prepared by Jessica Baumgardner-Zuzik (AfP), Principal Investigator (PI); Shaziya DeYoung (AfP), Lead Researcher; and Allyson Bachta (AfP), Researcher under the LASER PULSE program. This report was designed by Nicholas Gugerty (AfP), Senior Associate for Communications.

Suggested Citation

Baumgardner-Zuzik, Jessica, Shaziya DeYoung, and Allyson Bachta. 2023. Preventing/Countering Violent Extremism Systematic Mixed Methods Review: Methodology. West Lafayette, IN: Long-term Assistance and Services for Research—Partners for University-Led Solutions Engine (LASER Pulse Consortium).

About LASER PULSE

LASER (Long-term Assistance and Services for Research) PULSE (Partners for University-Led Solutions Engine) is a 10-year, $70M program funded by USAID's Innovation, Technology, and Research Hub, that delivers research-driven solutions to field-sourced development challenges in USAID partner countries.

A consortium led by Purdue University, with core partners Catholic Relief Services, Indiana University, Makerere University, and the University of Notre Dame, implements the LASER PULSE program through a growing network of 3,700+ researchers and development practitioners in 86 countries.

LASER PULSE collaborates with USAID missions, bureaus, and independent offices and other local stakeholders to identify research needs for critical development challenges, and funds and strengthens the capacity of researcher-practitioner teams to co-design solutions that translate into policy and practice.

Disclaimer

CONTENTS

ACRONYMS

AfP	Alliance for Peacebuilding
CVE	Countering Violent Extremism
CVP LAIT	Conflict and Violence Prevention Learning Agenda Implementation Team
DDRR	Disengagement, Deradicalization, Rehabilitation, and Reintegration
DEC	Development Experience Clearinghouse
IRR	Inter-rater Reliability
LASER PULSE	Long-term Assistance and Services for Research Partners for University-Led Solutions Engine
MEL	Monitoring, Evaluation, and Learning
MM	Mixed Methods
MMAT	Mixed Methods Appraisal Tool
OTI	Office of Transitions Initiatives
P/CVE	Preventing/Countering Violent Extremism
PI	Principal Investigator
PICO	Population, Intervention, Control, and Outcomes
PVE	Preventing Violent Extremism
QUAL	Qualitative
QUAN	Quantitative
RCT	Randomized Control Trial
ToC/s	Theory/ies of Change
TVE	Terrorist and Violent Extremism
USAID	United States Agency for International Development
VE	Violent Extremism

INTRODUCTION TO THE STUDY

As part of the Conflict and Violence Prevention Learning Agenda Implementation Team (CVP LAIT), the Alliance for Peacebuilding (AfP) carried out a Systematic Mixed Methods Review to map the evidence base for preventing/countering violent extremism (P/CVE) programming including what approaches work in which contexts and identify gaps that require greater investigation. The CVP LAIT was tasked with co-creating and implementing a bureau-wide learning agenda that established the evidence base for effective approaches to armed conflict and violence prevention; identified opportunities for CVP investments that would produce new knowledge to fill gaps in the existing literature; provided USAID staff with events, tools, resources, and/or guidance to incorporate learning agenda findings into their work; and conducted original research into armed conflict and violence prevention. Through an intensive, multi-stakeholder consultation process with USAID Washington and mission staff, P/CVE was identified as an effort that, if backed by sound evidence and guidance, could benefit program design, outcomes, policy, and knowledge generation.

Violent extremism (VE) stands as one of the most significant security threats facing the international community, with the frequency of violent acts and atrocities perpetrated by extremists escalating across the world. Despite the looming threats and known impacts of VE, universal agreement on how to define, discuss, and respond to it remains elusive. Over the past 20 years, the peacebuilding field has advanced its understanding of the drivers of VE. We now understand that radicalization is a fluid, nonlinear, highly individualized process, and the field has developed a series of approaches for P/CVE. Despite the surge in P/CVE programming, many interventions are inadequately substantiated, display a lack of rigor in both design and evaluation, and require further development and investment in more rigorous methods of measurement.

Evaluating P/CVE interventions introduces a myriad of methodological and logistical challenges, including: the absence of clearly articulated theories of change (ToCs); challenges in quantifying shifts in perceptions and ideologies; small sample sizes; the unavailability of comparative control groups; and potential stigmatization risks while focusing on susceptible individuals and communities. Additional hurdles comprise obtaining comprehensive data for evaluation and demonstrating causal effects. These complexities are further magnified due to the absence of uniform indicators and measures to gauge intervention outcomes and participant changes. Consequently, the actual impact of many interventions remains ambiguously documented, leaving the effectiveness of different approaches largely unassessed, especially in relation to VE goals. The lack of aggregated evidence of what has worked and what has not in P/CVE has hindered the field's ability to articulate cohesive programmatic and policy responses to VE. This deficiency also renders the field susceptible to a spectrum of practical, theoretical, and ethical problems.

To address these deficiencies, this research involved a systematic mixed method review of the relevant literature. It is important to note this was not a systematic review in the traditional sense, but has been adapted to include a greater breadth of mixed methods studies, particularly qualitative and non-randomized studies. However, best practices in systematic review methodology have been applied throughout to the improve rigor and transparency of the research. The objective was to collect and synthesize evidence related to P/CVE ToCs and their supporting rigorous, promising, and anecdotal evidence[1] across three primary programming responses: (1) *prevention (PV)*; (2) *containment/interdiction (CI)*; and (3) *disengagement, deradicalization, rehabilitation, and reintegration (DDRR)*. Articulating clear ToCs hypothesizing how change will occur is critical for testing and evaluating the impact of P/CVE interventions. This research applies an innovative ToC process, culminating in the development of 17 distinct, theoretically anchored, and empirically testable ToCs across the three primary P/CVE programming responses. These overarching ToCs serve three primary functions: to categorize

1 For the purposes of this research the following definitions were applied:
(1) Rigorous Evidence: Findings that are derived from research questions and hypotheses, backed by strong, methodologically sound research, and demonstrate clear, empirically validated results.
(2) Promising Evidence: Findings from approaches that, while not yet rigorously tested or of lower research quality, offer strong rationales and initial evidence suggesting effectiveness. These findings may come from innovative practices, pilot studies, or emerging research.
(3) Anecdotal Evidence: Findings that are unsupported with no demonstrated rationale that an intervention may be likely to improve outcomes in a different context and evidence collected in an informal manner that relies heavily or entirely on personal testimony.

programs with shared foundational logic and assumptions; to elucidate this logic and its underlying assumptions; and to create a framework for evidence-based mapping. This research applied the following learning agenda question and research questions to achieve these objectives.

Learning Agenda Question:

What are evidence-based ToCs and interventions to address P/CVE? What is working and what is failing?

Research Questions:

1. What interventions focused on P/CVE can be identified in the existing literature across each response?
2. What are the primary ToCs, outcomes, activities, target groups, and indicators across interventions?
3. Where are there similarities and differences in programming design, implementation, and results across different geographic, cultural, and extremism contexts?
4. Which ToCs are supported by research and evidence of impact to support the effectiveness of P/CVE interventions?

DEFINITIONS

Term	Definition
Violent Extremism (VE)	Advocating, engaging in, preparing, or otherwise supporting ideologically motivated violence to further social, economic, political, or religious objectives.[2]
Countering Violent Extremism (CVE)	Proactive actions to preempt or disrupt efforts by violent extremists to radicalize, recruit, and mobilize followers to violence, and to address specific factors that facilitate recruitment and radicalization to violence. CVE encompasses policies and activities to increase peaceful options for political, economic, and social engagement available to communities and local governments, and their abilities to act on them.[3]

Elucidating the differences between prevention and CVE presents its own conceptual problems, but this report distinguishes interventions aimed at Preventing Violent Extremism (PVE) based on their alignment with the level of involvement of individuals and groups across the spectrum of VE, in the following ways:

PVE	Proactive, upstream interventions aimed at the general population, at-risk and vulnerable communities, and individuals in the early stages of radicalization and recruitment to violence. PVE focuses on efforts to prevent or minimize recruitment and/or radicalization and the development of sympathy/alignment with the goals of VE actors.[4]
P/CVE	Actions that address the drivers of conflict and implement conflict transformation and reconciliation programming; create resilient communities by building immunity to recruitment by violent extremists by catalyzing community-based programming; and deter and disrupt recruitment and mobilization and assist with reintegration of former violent extremists.[5]

2 United States Agency for International Development. 2020. Policy for Countering Violent Extremism Through Development Assistance. https://www.usaid.gov/policy/countering-violent-extremism.

3 Ibid.

4 Alliance for Peacebuilding. 2022. CVP LAIT Learning Agenda Research Protocol: Preventing/Countering Violent Extremism. Available upon request.

5 Ibid.

Radicalization	A process by which a person or group adopts extreme ideas or beliefs and comes to view violence as a justified means to advance them.[6]
Disengagement	The process of shifting one's behavior to abstain from violent activities and withdraw from a violent extremist group.[7]
Deradicalization	The process of countering and undermining the ideology related to violent extremism and suggesting an alternative ideology by degrees. It refers to a change in beliefs, where an individual no longer holds extremist views or intentions, even if they might have once supported or participated in violent activities.[8]
Rehabilitation	The process where practitioners in a community or detention centers are involved in rehabilitating individuals after they have been deradicalized and/or disengaged from violent extremist ideologies.[9]
Reintegration	The process where practitioners help the transition of the completely rehabilitated individual back to society. Practitioners also work at the same time on society to ensure there is a positive response to the rehabilitated, and to mitigate social stigma. The ultimate goal of reintegration is to foster the social inclusion of the individual and prevent recidivism.[10]

Table 1: Definitions

FINALIZED CORPUS

A total of 2,677 records were identified through the initial targeted search. Once duplicates were removed and initial screening completed, a total of 2,285 records were retained for review based on criteria defined through AfP's population, intervention, control, and outcomes (PICO) framework.[11] In total, 605 full-text resources were assessed for eligibility, resulting in 129 included resources that contained 136 individual studies.[12] From the original 2,285 resources selected for abstract review, 6% were retained for the analysis, which is consistent with social science systematic review exclusion rates.[13]

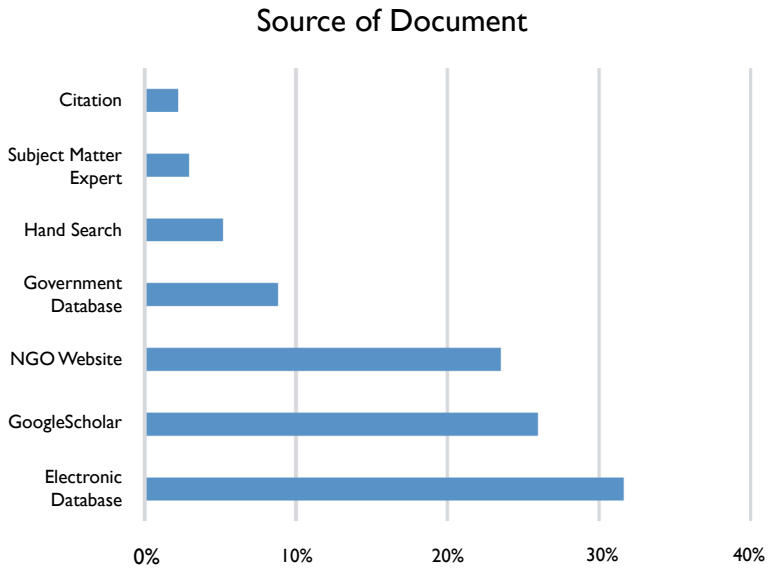


Figure 1: Breakdown of sources of included studies

The greatest number of resources were accessed through electronic databases, Google Scholar searches, and individual organizational websites. Additional sources were located through government websites (particularly USAID Development Experience Clearinghouse and the United Kingdom Home Office Research Database), hand searches through specific journals, subject matter experts' recommendations (received

6 United States Agency for International Development. 2020. *Policy for Countering Violent Extremism Through Development Assistance.*

7 Hedayah and Search for Common Ground. 2019. *Countering Violent Extremism: An Introductory Guide to Concepts, Programming, and Best Practices.*

8 Ibid, and Dalgaard-Nielsen, A. 2017. "Patterns of Disengagement from Violent Extremism: A Stocktaking of Current Knowledge and Implications for Counterterrorism." Expressions of Radicalization (Winter): 273-293.

9 Hedayah and Search for Common Ground, Ibid.

10 Ibid.

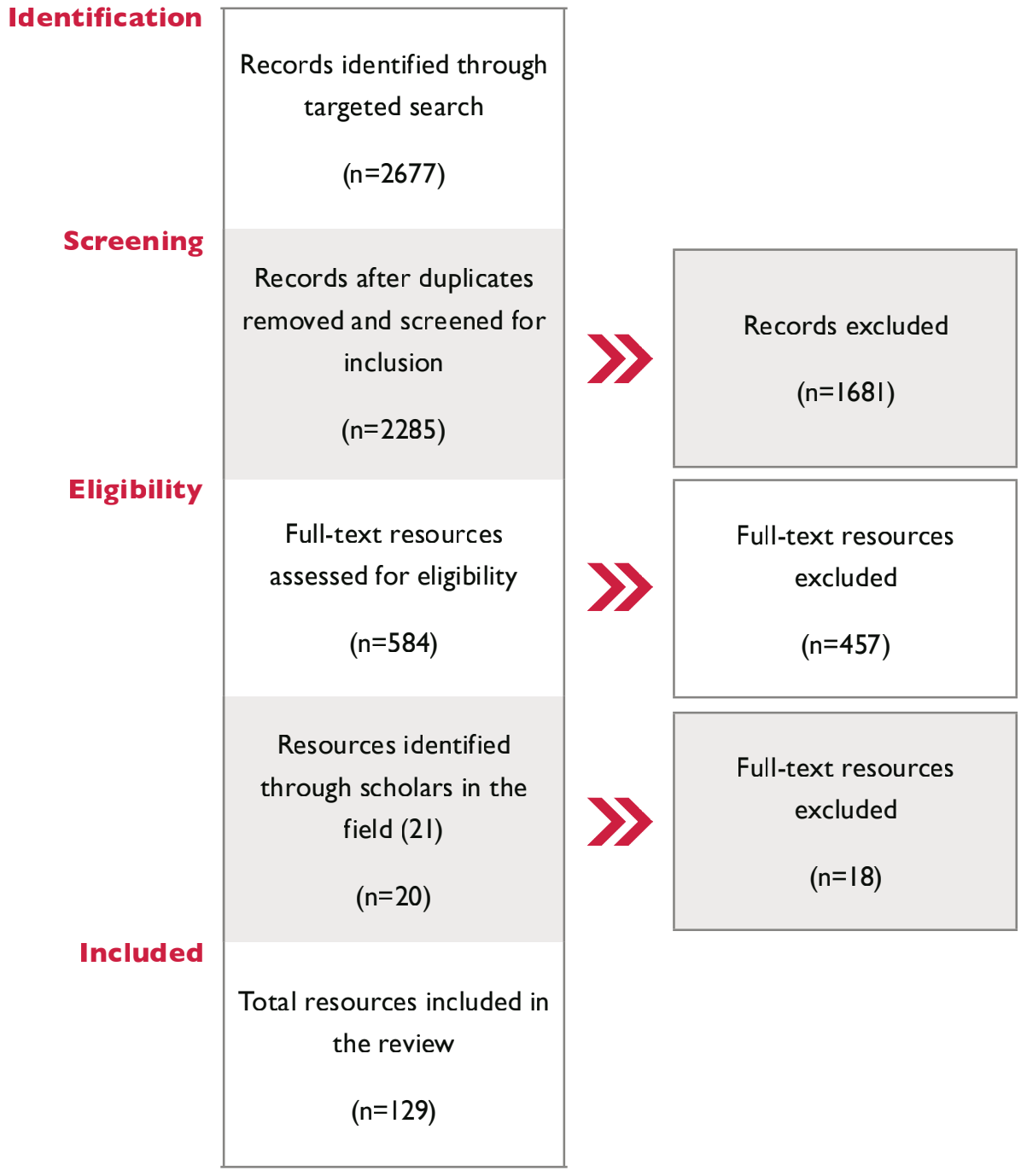
11 A detailed description of the Methodological Approach is provided in the next section.

12 See page 7 for additional detail.

13 Meline, Timothy. 2006. "Selecting Studies for Systemic Review: Inclusion and Exclusion Criteria." *Contemporary Issues in Communication Science and Disorders* 33 (Spring): 21–27..

during Phase I key informant interviews and focus group discussions with USAID, as well as open evidence calls), and citations extracted from included resources.

The finalized resources were split between journal submissions (26%) and self-published evaluations (74%). Journal submissions were identified as resources that had an ISSN/ISBN/DOI or similar journal serial number. Self-published evaluations referred to program final, mid-term, or endline evaluations. Mid-term reports were only used when final or endline evaluations were not available.



Following resource identification and collection, screening, and eligibility review, the 129 included resources were assigned to their respective programming response/s: 116 studies within PV, 50 studies within CI, and 25 studies within DDRR. In practice, many of the studies overlap, attempting to address key aspects of PV, CI, and sometimes even DDRR within a single program. For example, a program could have activities related to capacity building for prison staff (a PV approach), as well as disengagement and deradicalization activities with incarcerated populations (a DDRR approach). Therefore, the total N of each approach (190) is greater than the N of resources (129). Furthermore, the N of included studies (136) is larger than the N of resources (129) because some resources featured more than one study.

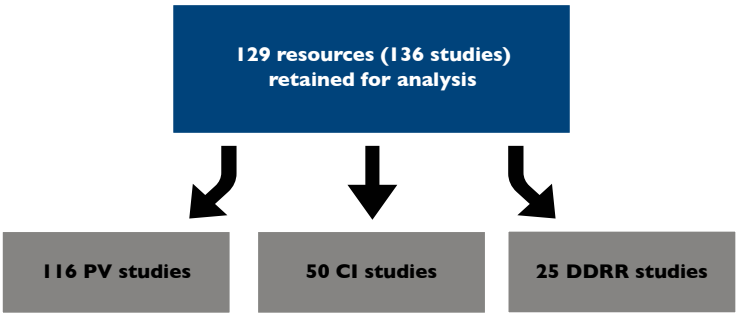


Figure 2: Number of studies included

METHODOLOGICAL APPROACH

In order to address the research questions highlighted above, this research employs a systematic mixed methods approach focusing on PV, CI, and DDRR programming responses. These responses capture interventions that align with the level of involvement of individuals and groups across the cycle of VE. Each response may capture sub-interventions that share a similar aim—i.e., population-level prevention activities compared to activities to disrupt radicalization or recruitment all aim to prevent radicalization.

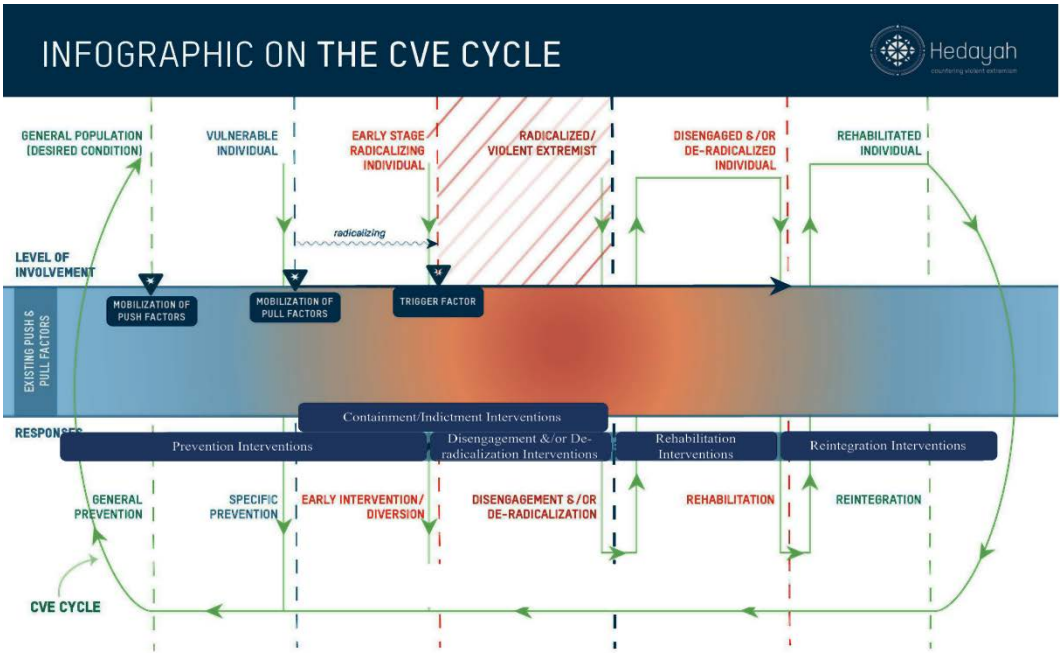


Figure 3: Adapted Hedayah CVE Cycle illustrating the three primary P/CVE programming responses[14]

14 The infographic is adapted from Hedayah's *The CVE Cycle: An individual trajectory* that is overlaid with our proposed target responses within P/CVE programming.

The approach to this research involved: (1) resource identification and collection using the PICO criteria identified for the research; (2) determining eligibility of collected resources; (3) thematic coding for ToC analysis across studies; (4) scoring studies for quality using a Mixed Methods Assessment Tool (MMAT);[15] (5) scoring studies for strength of evidence using a Strength of Evidence Continuum; (6) Evidence Base Mapping and Maturity Analysis; and (7) conducting Thematic Analysis and Evidence Synthesis of studies.

1. Resource Identification and Collection and 2. Eligibility

PICO-Defined Search Criteria and Eligibility Criteria

AfP defined the parameters of this study using the PICO criteria, which is the standard used in Cochrane and Campbell Collaborative systematic reviews.

Criteria	PICO Criteria Particulars	Eligibility Criteria
Population/ Problem	VE; terrorism/targeted violence; P/CVE programming	Evaluations, systematic reviews, meta-reviews of evaluations, and program evaluation reports focused on VE; terrorism/targeted violence; P/CVE programming.
Intervention	This research focuses on P/CVE interventions focused on PV, CI, and/or DDRR	Evaluations, systematic reviews, meta-reviews of evaluations, and program evaluation reports on VE; terrorism/targeted violence; P/CVE programming interventions focused on PV, CI, and/or DDRR.
Control	No restrictions	
Outcome	No restrictions	
Countries	All countries except the U.S.	Evaluations, systematic reviews, meta-reviews of evaluations, and program evaluation reports on P/CVE programs present in all countries except the U.S.[16]
Language	English, French, Spanish – other languages with included translations	
Year	Post-9/11/2001	Evaluations, systematic reviews, meta-reviews of evaluations, and program evaluation reports on P/CVE programs globally post-9/11/2001.
Publication	Academic, government, private, and scholarly literature	Evaluations, systematic reviews, and meta-reviews to be included must have an outlined methodology.

Table 2: PICO criteria and inclusion/exclusion criteria

15 Hong, Q. N. et al. 2018. "Mixed Methods Appraisal Tool (MMAT) Version 2018—User guide." *Education for Information* Vol 34, No 4: 285-291. https://doi.org/10.3233/EFI-180221.

16 The U.S. employs distinct legal and operational definitions for "domestic terrorism" compared to international P/CVE contexts, uniquely influencing how activities are prosecuted and addressed. Additionally, domestic approaches to terrorism in the U.S. are heavily intertwined with concerns related to civil liberties, particularly around surveillance, profiling, and potential infringement on First Amendment rights. Incorporating U.S. domestic terrorism into a systematic review of global P/CVE could introduce significant heterogeneity due to these varied legal, operational, and cultural contexts, potentially complicating the extraction of cohesive findings or actionable recommendations from the results. Following in-depth consultations with P/CVE experts and given these differences, the U.S. was excluded from this research.

Resource Search Strategy

To develop a comprehensive corpus of relevant resources, both published and unpublished, AfP leveraged well-known P/CVE knowledge hubs, prior experience in P/CVE research, and its own membership network and communities of practice to identify a multi-track data collection process with four distinct sources for resource identification and collection:

1. AfP developed databases and corpora including the P/CVE Digest, the Eirene Peacebuilding Database, and internally maintained resources in relation to AfP's previous P/CVE systematic scoping review.

2. An open call for evidence.

3. Internet hand searches of online databases, journals, and development and peacebuilding organizations including, but not limited to: Ministry of Foreign Affairs - IOB Publications, The USAID DEC, RAND Database of Worldwide Terrorism Incidents; UK Home Office Research Database, and The Directory of Open Access Journals. Structured searches were performed using primary terms, secondary terms, and logical operators. Primary: PVE, CVE, P/CVE, VE, TVE, prevention, violent extremism, preventing violent extremism, countering violent extremism, transforming violent extremism. Secondary: evaluate [evaluating/evaluate/evaluation], impact, intervention/s, program/s, review/s, evidence. Logical operators: and/or.

4. Key Informant Interviews and Focus Group Data: during the multi-stakeholder consultation process, the LAIT met with 90+ individuals who were vetted through the CVP team and the Peace and Security Council, alongside 16 USAID Missions. As part of this process, individuals were asked to identify any resources they considered useful for subsequent research. AfP employed structured searches in this database for resources relevant to P/CVE.

5. Additional material through snowballing: using the references and bibliographies of collected resources, any relevant resources that were omitted from the initial search were identified and collected for inclusion.

AfP used Microsoft Excel to track references and code key characteristics documented for each resource. This method allowed researchers to quickly access information in one place, check each other's work to avoid duplication, and efficiently evaluate characteristics of each resource against the inclusion criteria when deciding whether to include for full text coding and review.

Two independent researchers constructed the search queries, as well as identified and collected relevant resources using the PICO criteria. Following the initial relevance assessment, the two researchers assessed the curated corpus for eligibility criteria. Eligibility status per resource was validated by a third researcher (the PI). Any disagreements in the codes were resolved by discussion.

3. Theory of Change Analysis

Once the included studies for this research were finalized, each study was assigned applicable broad level ToC/s. As part of this process, a wide range of P/CVE resources, including proposals, reports, research projects, and evaluations were reviewed with the goal of generating a working list of the primary ToCs extent in the field. Although the exact ToCs described or implied across P/CVE literature are often highly specific to individual programs and contexts, this research aimed to develop a more general list that can be used to classify programs that share underlying logic and assumptions and to assist in making these logic and assumptions more explicit. These ToCs were originally organized across five P/CVE programming responses: (1) *prevention*; (2) *containment/interdiction*; (3) *disengagement/deradicalization*; (4) *rehabilitation*; and (5) *reintegration*.

A systematic initial review of the P/CVE literature focused on identifying unique ToCs. Using an adapted first and second

cycle coding framework,[17] a technical expert used the initial review to generate a list of ToCs by programming response, identifying unique causal relationships between P/CVE interventions and response-specific outcomes: (1) *prevention (PV)*; (2) *containment/interdiction (CI)*; (3) *disengagement/deradicalization (DD)*; (4) *rehabilitation (RB)*; and (5) *reintegration (RI)*. In the second cycle, this longer list of program-specific ToCs was reduced to the smallest number of unique causal relationships that could effectively "fit" the overall first list underneath them. Through this process, 19 distinct and field-wide ToCs were identified across these five programming responses, each outlining identifiable and theoretically-informed approaches that lay the foundation to test cause/effect assumptions. The abstracts of all included studies were reviewed and coded into these 19 ToCs by the technical expert. As part of full-text coding, two independent researchers assessed and verified this ToC coding. Any disagreements in the codes were resolved by discussion.

As part of the evidence mapping and synthesis, the PI completed a review of all 136 studies and verified the TOC coding. While in theory, each of the five programming responses are seen as discrete and one can make a distinction between outcomes and program logic, in practice, many of the included studies applied a combination of approaches with little to no distinction between their outcomes. As such, a decision was made to combine the P/CVE programming responses between *disengagement/deradicalization*; *rehabilitation*; and *reintegration* into one programming response of DDRR. One additional ToC was also added to PV to account for a unique group of interventions working on prison reform and prison staff capacity building that was not initially captured in the ToC analysis. Following this reclassification, a total of 17 ToCs (6 for PV; 4 for CI; 7 for DDRR) were finalized and each of the 136 studies were coded into three primary programming responses and their relevant ToCs.

Given the overlap in practice of the studies between programming responses and ToCs, many studies were classified into multiple ToCs. Only 40% (n=55) of P/CVE studies included in the final corpus were assigned to a single ToC. The remaining 60% of studies had more than one applicable ToC. Within these 60%, a single program could be applicable to anywhere from two different ToCs to 10 different ToCs.

Many P/CVE programs articulate ToCs focused on training and providing resources to key stakeholders, with the assumption (sometimes explicit, often implied) that these resources and techniques will eventually impact communities and individuals targeted by P/CVE activities. Where applicable, this research collapsed these ToCs into their unique activities. For instance, a ToC arguing that training prison staff in trauma-informed care will produce more trauma-informed care programming for VE offenders collapsed into a more general ToC about the effect of trauma-informed and mental health care on recidivism. The implications of this choice are reflected in the limitation section below.

4. Quality of Studies

Quality assessment as part of systematic reviews contribute to definitions on how much "weight" to attribute to conclusions of included studies. Without assessing quality, there is a risk that the simple *existence* of studies will be used as the basis for conclusions, irrespective of their intrinsic quality. While many different quality appraisal techniques, standards, and guidelines exist, most are biased towards evaluation method and prioritize specific methods (particularly randomized control trials) over others, regardless of the quality of research implementation. Furthermore, quality appraisal techniques lack consensus and are still undeveloped, particularly for systematic mixed methods reviews—i.e., reviews that include qualitative, quantitative, multi-methods, and mixed methods studies.

Across the corpus, 16% of studies were quantitatively evaluated, 29% were qualitatively evaluated, 20% were evaluated using multi-methods, and 35% were evaluated using a mixed methods framework. Due to the diversity in studies across the corpus, the MMAT[18] was used for this research. Two initial screening questions were applied to verify that the MMAT can

17 Miles, Matthew B., A.M. Huberman, and Johnny Saldaña. 2020. *Qualitative Data Analysis: A Methods Sourcebook*. 4th ed. Los Angeles: Sage.
18 Hong, Q. N. et al. 2018

be used to assess each study. The MMAT can only be used to appraise the quality of empirical studies and cannot be used for non-empirical papers, such as reviews and theoretical papers. Unlike other evaluation tools, the MMAT can be used to assess five different types of studies: qualitative, quantitative descriptive, quantitative non-randomized studies (i.e., quasi-experimental), quantitative randomized controlled trials, and mixed method designs—all of which feature in the corpus. The choice to use the MMAT controls for the complexity of applying multiple quality appraisal techniques for each method and the potential non-comparability between their results.

MMAT Study Categorizations[19]

Study Categorization	MMAT Manual Definitions	MMAT Methodological Criteria
Qualitative Studies	Research concerned with exploring and understanding the meaning individuals or groups ascribe to a social or human problem.	**Is the qualitative approach appropriate to answer the research question?** *Consider whether the qualitative approach/methodology was the planned research approach, or if changes were made during the course of data collection that led to qualitative methods.* **Are the qualitative data collection methods adequate to address the research question?** *Consider whether the method of data collection (e.g., in-depth interviews and/or group interviews; and/or observations) and the form of the data (e.g., tape recording, video material, diary, photo, and/or field notes) are adequate. Is there sufficient saturation of type of response across qualitative data to have confidence in the findings, or would additional time in the field have produced different findings? If different methods are triangulated to produce the finding, credibility is higher. If there is no indication of the number of interviews or time spent observing, credibility is weakened.* **Are the findings adequately derived from the data?** *Rate the amount of descriptive information presented to support the findings. Is there evidence of careful qualitative analysis, such as using multiple coders, validation methods, qualitative software, or discussions of data validity?* **Is the interpretation of results sufficiently substantiated by data?** *Are the findings clearly connected with direct quotes or thick description of observations, rather than just the opinion of the researcher with little connection to the evidence?* **Is there coherence between qualitative data sources, collection, analysis, and interpretation?** *Are there clear links between data sources, collection, analysis, and interpretation?*

Table 3: MMAT Study Categorizations

19 Ibid.

Study Categorization	MMAT Manual Definitions	MMAT Methodological Criteria
Quantitative Descriptive Studies	Research concerned with and designed only to describe the existing distribution of variables without much regard to causal relationships or other hypotheses. They are used for monitoring the population, planning, and generating hypotheses.	**Is the sampling strategy relevant to address the research question?** *Is the source of sampling relevant to the target population? Does the study provide a clear justification of the sample frame used?* **Is the sample representative of the target population?** *Is there a match between respondents and the target population? Indicators of representativeness include: clear description of the target population and of the sample (e.g., respective sizes and inclusion and exclusion criteria); reasons why certain eligible individuals chose not to participate; and any attempts to achieve a sample of participants that represents the target population.* **Are the measurements appropriate?** *Does the study explicitly provide indicators of appropriate measurements related to their stated goal and outcomes? I.e.: the measurements are justified and appropriate for answering the research question; the measurements reflect what they are supposed to measure.* **Is the risk of nonresponse bias low?** *Are the respondents and nonrespondents different on the variable of interest? Are findings based on at least 85% of original sample (or sub-sample if this finding is based on a sub-sample), i.e., almost all the participants contributed to almost all measures? Some indicators of low nonresponse bias can be considered such as a low nonresponse rate, reasons for nonresponse (e.g., noncontacts vs. refusals), and statistical compensation for nonresponse (e.g., imputation).* **Is the statistical analysis appropriate to answer the research question?** *Did the study explicitly state its analysis methods? Were there any presented problems or limitations with data analysis that might limit the interpretation of the results?*

Table 3: MMAT Study Categorizations

Study Categorization	MMAT Manual Definitions	MMAT Methodological Criteria
Quantitative Non-Randomized Studies	Research involves any quantitative studies estimating the effectiveness of an intervention or studying other exposures that do not use randomization to allocate units to comparison groups.	**Are the participants representative of the target population?** *Is there a match between respondents and the target population? Indicators of representativeness include: clear description of the target population and of the sample (such as respective sizes and inclusion and exclusion criteria); reasons why certain eligible individuals chose not to participate; and any attempts to achieve a sample of participants that represents the target population.* **Are measurements appropriate regarding both the outcome and intervention (or exposure)?** *Does the study explicitly provide indicators of appropriate measurements related to their stated goal and outcomes? I.e.: the measurements are justified and appropriate for answering the research question; the measurements reflect what they are supposed to measure.* **Are there complete outcome data?** *Are findings based on at least 85% of original sample (or sub-sample if this finding is based on a sub-sample) i.e., almost all the participants contributed to almost all measures.* **Are the confounders accounted for in the design and analysis?** *Have researchers explicitly discussed if confounding is expected, or presented appropriate methods to control for confounders (such as stratification, regression, matching, standardization, and inverse probability weighting)?* **During the study period, is the intervention administered (or exposure occurred) as intended?** *Were participants treated in a way that is consistent with the planned intervention? Did the study exhibit the presence of contamination (e.g., the control group may be indirectly exposed to the intervention) or whether unplanned co-interventions were present in one group?*

Table 3: MMAT Study Categorizations

Study Categorization	MMAT Manual Definitions	MMAT Methodological Criteria
Quantitative Randomized Controlled Trials	Research involves a clinical study in which individual participants are allocated to intervention or control groups by randomization (intervention assigned by researchers).	**Is randomization appropriately performed?** *Have researchers described how the randomization schedule was generated? A simple statement such as "we randomly allocated" or "using a randomized design" is insufficient to judge if randomization was appropriately performed. Is assignment predictable? Using odd and even record numbers or dates is not appropriate. At minimum, a simple allocation (or unrestricted allocation) should be performed by following a predetermined plan/sequence. It is usually achieved by referring to a published list of random numbers or a list of random assignments generated by a computer. Also, restricted allocation can be performed, such as blocked randomization (to ensure particular allocation ratios to the intervention groups), stratified randomization (randomization performed separately within strata), or minimization (to make small groups closely similar with respect to several characteristics). Was allocation concealed to protect assignment sequence until allocation? Researchers and participants should be unaware of the assignment sequence up to the point of allocation.* **Are the groups comparable at baseline?** *Have researchers discussed any potential baseline imbalances and/or ways to address any imbalance? Baseline imbalance between groups suggests that there are problems with the randomization. Indicators from baseline imbalance include: "(1) unusually large differences between intervention group sizes; (2) a substantial excess in statistically significant differences in baseline characteristics than would be expected by chance alone; (3) imbalance in key prognostic factors (or baseline measures of outcome variables) that are unlikely to be due to chance; (4) excessive similarity in baseline characteristics that is not compatible with chance; (5) surprising absence of one or more key characteristics that would be expected to be reported" (Higgins et al. 2016).* **Are there complete outcome data?** *Are findings based on at least 85% of original sample (or sub-sample if this finding is based on a sub-sample)? For example, almost all the participants contributed to almost all measures.* **Are outcome assessors blinded to the intervention provided?** *Are clear risks of bias for findings minimized? Things to consider are: (1) post hoc nature of finding (i.e., possible data fishing); and (2) whether outcome assessors are unaware of who is receiving which interventions.* **Did the participants adhere to the assigned intervention?** *Did at least 85% of participants continue with their assigned intervention throughout follow-up?*

Table 3: MMAT Study Categorizations

Study Categorization	MMAT Manual Definitions	MMAT Methodological Criteria
Mixed Methods (MM)	Research involves combining qualitative (QUAL) and quantitative (QUAN) methods. In this tool, to be considered MM, studies have to meet the following criteria: a) At least one QUAL method and one QUAN method are combined; b) Each method is used rigorously in accordance with the generally accepted criteria in the area (or tradition) of research invoked; and c) The combination of the methods is carried out at the minimum through a MM design (defined a priori, or emerging) and the integration of the QUAL and QUAN phases, results, and data.	**Is there an adequate rationale for using a MM design to address the research question?** *Are the reasons for conducting a MM study clearly explained? Several reasons can be invoked, such as to enhance or build upon qualitative findings with quantitative results and vice versa; provide a comprehensive and complete understanding of a phenomenon; or develop and test instruments.* **Are the different components of the study effectively integrated to answer the research question?** *Does the study present information on how qualitative and quantitative phases, results, and data were integrated? Such information includes how data gathered by both research methods was brought together to form a complete picture (e.g., joint displays) and when integration occurred (e.g., during the data collection-analysis or/and during the interpretation of qualitative and quantitative results).* **Are the outputs of the integration of qualitative and quantitative components adequately interpreted?** *Does the study apply meta-inference during the interpretation of findings from the integration of the qualitative and quantitative components? Does the study show the added value of conducting a MM study rather than having two separate studies?* **Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?** *Are any divergencies and inconsistencies (conflicts, contradictions, discordances, discrepancies, and dissonances) that occurred when integrating the findings from the qualitative and quantitative components explained? Did the study apply any strategies to address the divergences, such as reconciliation, initiation, bracketing and exclusion?* **Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?** *To appraise, use criteria for the qualitative component and the appropriate criteria for the quantitative component. The quality of both components must be at least 3 or higher for the MM study to be considered of good quality. The premise is that the overall quality of a MM study cannot exceed the quality of its weakest component. For example, if the quantitative component is rated high quality and the qualitative component is rated low quality, the overall rating for this criterion will be of low quality.*

Table 3: MMAT Study Categorizations

Each study categorization has 5 corresponding variables assessing methodological quality criteria, resulting in a total of 25 variables across the entire checklist. Scores for methodological quality are entered into the checklist for the designated study categorization by selecting "Yes=1," "No=0," or "Can't tell=0." The highest score available for the quality of a study is 5, whereas the lowest score is 0.

Two different researchers assessed each study's categorization. The PI then assessed the quality of each study and assigned scores. To ensure the data's reliability and control for perception bias, the research team applied an inter-rater reliability (IRR)[20] method by having a second researcher code a random 10% of included studies using the following formula.

$$\text{Reliability} = \frac{\text{Number of Agreements}}{\text{Number of Agreements + Disagreements}}$$

Figure 4: Inter-rater reliability formula

The percentage of IRR for the coded categories resulted in 88% agreement, which was considered sufficient by the research team, and no further testing was required.

5. Strength of Evidence Continuum Analysis

AfP's unique approach to systematic mixed methods reviews allows for the inclusion of many more studies that would traditionally be excluded due to their non-statistical research methods (i.e., non-randomized control trials and/or experimental/quasi-experimental methods). Given that peacebuilding is a developing field, this method allows for a deeper review of a broader state of evidence with the intention of improving practice and informing future research to support professionalization of the field. However, the inclusion of studies that would traditionally be excluded based on methods requires a more *nuanced assessment of evidence strength* across a continuum rather than more finite methods. Any continuum must also account for *research quality* as a critical characteristic informing strength of evidence. To supplement the Quality Analysis conducted with the MMAT, AfP developed a Strength of Evidence Continuum using the criteria described below. Assessing both the quality of evidence, which evaluates research methodological rigor and validity, and the strength of evidence, which considers the cumulative weight of consistent findings and their practical significance, is essential for a comprehensive understanding of the evidence base in a research project. This dual assessment approach helps ensure that not only are the methods sound, but also that the body of evidence as a whole supports informed decision-making.

Each included study was assessed using three criteria focused on programmatic effect, the evaluation timeline, and the research design. The choice of using programmatic effect, evaluation timeline, and research design as criteria for the strength of evidence continuum is advantageous for several reasons. First, programmatic effect assesses the real-world impact of interventions, ensuring that the evidence reflects practical significance. Second, the evaluation timeline considers the temporal aspect, allowing for the examination of both short-term and long-term effects, which is crucial for understanding sustainability of results. Lastly, research design evaluation enhances the assessment of methodological quality, contributing to a more comprehensive evaluation of the evidence base by accounting for potential biases and study design strengths. These three criteria collectively provide a well-rounded and robust framework that considers both real-world significance and research quality.

Individual points were assigned based on each criterion for a final score ranging between 1-11 points (see table 4). Each criterion score was evaluated using individual variables coded from the full-text coding of all included studies. These variables were coded by three separate researchers and underwent full data cleaning to identify errors and potential outliers as part of the overall research process.

20 Miles, Matthew B. and A. Michael Huberman. *Qualitative data analysis: an expanded sourcebook*. Thousand Oaks, Cal: Sage, 1994.

Category	Strong Evidence	Moderate Evidence	Promising Evidence/ Strong Theory	Unsupported
Programmatic Effect	**Statistically significant** effect on outcomes (4 points)	**Lacks significant findings on outcomes,** but provides statistically significant effect on other outcomes in the study (3 points)	**Lacks significant findings,** but demonstrates a rationale that an intervention may be likely to improve outcomes in another context (2 points)	Intervention is found to be **unsupported** without rationale that an intervention may be likely to improve outcomes and evidence collected in an informal manner that relies heavily or entirely on personal testimony (1 point)
Evaluation Timeline	**Ex-post evaluation** (>1 year post implementation) (3 points)	**Endline evaluation** (last 2 months of programmatic implementation to 1 year post-implementation) (2 points)		**Concurrent evaluation** (occurring during the program implementation) (1 point)
Research Design	**Experimental** study (4 points)	**Quasi-experimental** study (3 points)	Well-designed and well-implemented **correlational and/or case study** to examine the effects of an intervention (2 points)	Evaluation includes **single method assessments**, informal quantitative and qualitative data collection approaches, lack of clear evaluation methodology, and non-systematic approaches (1 point)

Table 4: Strength of Evidence Continuum

Studies were then assigned an overarching strength of evidence classification of Strong Evidence (9-11 points), Moderate Evidence (7-8 points), Strong Theory (4-6 points), and Anecdotal Findings (1-3 points).[21]

21 Strength of evidence classifications are highlighted in summary form across the technical reports as aggregates for each ToC. Those identified as having strong evidence are highlighted with bolded font.

6. Evidence Base Mapping and Maturity Analysis

For the ToC maturity analysis, each of the 17 broad ToCs could be classified as exhibiting an immature, developing, maturing, or mature evidence base depending on the studies that fell within them. The following calculations and classifications depict how the evidence mapping and maturity analysis were conducted for the purposes of this research.

Size of the Body of Evidence

The size of the body of evidence was determined based on the number of studies that fell within each ToC using the following scale.

Body of Evidence Classification	Ranges	Size Descriptor
Very Small	Range: 1-10 studies	Limited
Small	Range: 11-20 studies	Moderate
Medium	Range: 21-40 studies	Substantial
Large	Range: 41+	Extensive

Table 5: Body of Evidence Size Scale

Variability of Studies

The variability of studies within each broad ToC was calculated using the Coefficient of Variation (CV) for both quality and strength of evidence to understand the dispersion or variability of studies around the mean. The greater the dispersion indicates the studies are spread very widely around the mean, indicating both a high degree of diversity and low consistency across strength and quality of evidence. The smaller the dispersion indicates a much higher level of consistency, meaning studies in the ToC are of similar strength of quality and evidence. For the purposes of this research, CV was used as an indication of greater maturity.

CV was calculated using the following formula $CV = \frac{\sigma}{\mu}$, where σ refers to the standard deviation and μ refers to the standard mean, for both the quality and strength of evidence. The following scale was then used to describe the CV for each.

Variability Classification	Ranges
Very Low Variability	Range: 0.00 - 0.20
Low Variability	Range: 0.21 - 0.40
Moderate Variability	Range: 0.41 - 0.60
High Variability	Range: 0.61 - 0.80
Very High Variability	Range: 0.81 - 1.00

Table 6: Variability Scale

Group of studies has a **higher** level of consistency related to strength of quality and evidence.

Group of studies has a **lower** level of consistency related to strength of quality and evidence.

Evidence Mapping and Maturity Analysis

Researchers mapped each study within a scatterplot to provide a visual of the evidence base for each of the 17 broad P/CVE ToC. Within this scatterplot, the x-axis plots each study's quality of evidence score (MMAT score), the y-axis depicts the strength of evidence (continuum score), and a linear trend line was calculated and applied to provide a visual of the ToC's evidence base.

Figure 5 below is an example of how evidence was mapped onto a scatterplot for a ToC.

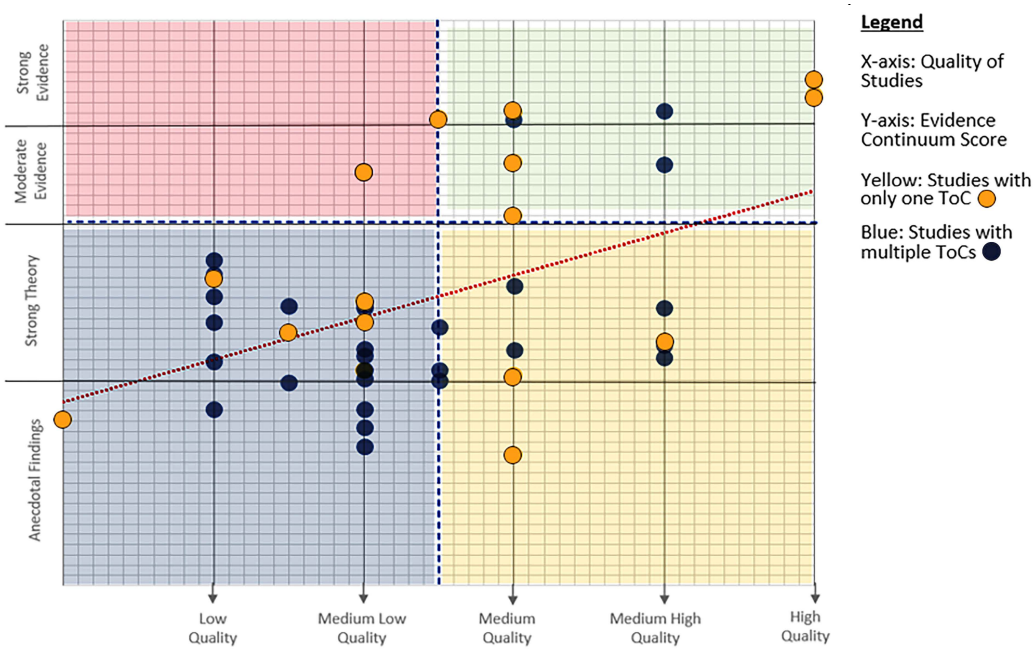


Figure 5: Scatterplot Example

Following the evidence mapping, maturity was then assessed for each ToC using the Maturity of Evidence Base Scale depicted in Table 6. The scale uses multiple variables including the size of the body of evidence; mean, standard deviation, and variance of the strength of evidence (using Strength of Evidence Continuum scores); mean, standard deviation, and variance of quality of evidence (using MMAT scores); and trends depicted from the evidence mapping for each ToC.

Each ToC's level of maturity could be described as either immature, developing, maturing, or mature evidence base depending on where the studies in the ToC fell within the Maturity of Evidence Base Scale.

Table 7: Maturity of Evidence Base Scale

Evidence Base Classification	Characteristics
Immature	1. *Majority of studies are most likely in the lower quadrants* (assessed using individual studies' MMAT and Strength of Evidence scores on scatterplot). 2. *Absence of studies in the upper quadrants* (assessed using individual studies' MMAT and Strength of Evidence scores on scatterplot). 3. Studies applicable to multiple broad ToCs, showing a general lack of focus (assessed using individual studies' use of singular or multiple ToCs as depicted on scatterplot in yellow or blue).

Evidence Base Classification	Characteristics
Developing	1. *Increasing presence in the lower right and occasionally upper quadrants* (assessed using individual studies' MMAT and Strength of Evidence scores on scatterplot). 2. *An upward trend in the mean quality and strength of evidence, but with evident variability* (assessed using individual studies' mean quality and strength of evidence scores to calculate linear trend line on scatterplot). 3. *Increasing empirical backing for some theoretical insights* (assessed using individual studies' MMAT and Strength of Evidence scores on scatterplot to show growth of studies falling within moderate and strong evidence).
Maturing	1. *Must include existence of high quality, strong evidence* (assessed using individual studies' MMAT and Strength of Evidence scores on scatterplot). 2. *Presence in the upper quadrants, but still sparse in the upper right quadrant* (assessed using individual studies' MMAT and Strength of Evidence scores on scatterplot). 3. *Reduced variability in the quality and strength of evidence, indicating improved consistency* (assessed using CV). 4. *Growing empirical validations for theoretical insights* (assessed using individual studies' MMAT and Strength of Evidence scores on scatterplot to show growth of studies falling within moderate and strong evidence). 5. *Clearer distinction between ToCs* (assessed using individual studies' use of singular or multiple ToCs as depicted on scatterplot in yellow or blue).
Mature	1. *Presence in the upper quadrants with increased studies in the upper right quadrant* (assessed using individual studies' MMAT and Strength of Evidence scores on scatterplot). 2. *Lower variability, indicating highly consistent and reliable studies* (assessed using CV). 3. *Existence of high quality, strong evidence* (assessed using individual studies' MMAT and Strength of Evidence scores on scatterplot). 4. *Trend line extends into strong evidence* (assessed using individual studies' mean quality and strength of evidence scores to calculate linear trend line on scatterplot). 5. *Existence of distinct and specialized research focus with little to no overlap between ToCs* (assessed using individual studies' use of singular or multiple ToCs as depicted on scatterplot in yellow or blue).

Table 7: Maturity of Evidence Base Scale

Returning to the Figure 5 scatterplot example and applying the Maturity of Evidence Base Scale, it can now be understood this map depicts a "mature evidence base" reflected by the large body of studies and resulting evidence included within this ToC. The increased presence of studies is witnessed in the upper right quadrant and the trend line extends into the strong evidence quadrant.[22] The existence of studies with distinct and specialized research focus are depicted in the visual through yellow dots, whereas the blue dots reflect studies that are applicable to multiple broad ToCs.

Two independent researchers assessed and verified this maturity coding. Any disagreements in the codes were resolved by discussion. Any evidence bases that fell distinctly between maturity levels were reviewed and a decision was reached based on consensus.

7. Thematic Analysis and Evidence Synthesis

Finally, researchers synthesized findings across relevant study characteristics for each of the 17 broad ToCs, including type of research methods, target groups, program beneficiary targeting strategy, and program activities and interventions, to assess effectiveness of programming and identify and understanding underlying causal mechanisms. Researchers employed a thematic analysis approach following full-text coding of 105 variables paired with computerized thematic[23] and descriptive analyses of the included studies.

Two coding teams separately conducted thematic analysis using a traditional card-sort theme extraction method[24] across relevant characteristics. Through this process, thematic categories relating to each characteristic were created inductively through a method of open coding. Once thematic categories were developed, the data was coded and restructured within relevant thematic categories for final category-based analysis. The two thematic analyses were compared and minor differences between the two were reconciled using cross-team discussion. The findings of these thematic analyses are reported across PV, CI, and DDRR evidence summaries.

To assess effectiveness of programming, the systematic mixed methods review went beyond a simple scoping and mapping to examine what programmatic ToCs have been put forth, where those ToCs are supported by evidence, what is the quality and strength of said evidence, and what are direct recommendations and challenges to improve practice. The PI applied a modified realist synthesis approach[25] of relevant evidence to the mechanisms by which interventions within each ToC work or not.

A realist approach involves identifying underlying causal mechanisms and exploring how they work and under what specific conditions—critical for the adaptive and complex environments in which P/CVE programming and evaluation occurs. The ToC Analysis established the underlying causal mechanisms being tested as part of this realist analysis and the MMAT and the Evidence Continuum extrapolated where evidence exists and of which quality. The PI then reviewed all full-text studies to extract what works and what does not across each ToC, including a particular focus on programmatic approaches, relevant target groups, shared challenges in implementation and evaluation, and practice recommendations. Special attention was given to vulnerable and historically marginalized populations, including persons with disabilities, youth, children, LGBTQI+ persons, indigenous communities, women, and girls.

---

22 The Figure does not depict the full trend line given limitations on space, but final end point given existence of strong evidence placed it within the strong evidence quadrant within the final scatterplot.

23 Computerized thematic content analysis used Computer Assisted Qualitative Data Analysis Software (CAQDAS) to assist in thematic coding identification across qualitative data to quickly identify and code specific emerging themes. For more information, reference Miles, Matthew B., A.M. Huberman, and Johnny Saldaña. 2020. *Qualitative Data Analysis: A Methods Sourcebook*. 4th ed. Los Angeles: Sage.

24 Card-sort theme extraction is a method for inductively analyzing qualitative data for the purposes of thematic analysis. Once data is organized into specific categories, a researcher physically or using CAQDAS, sorts the data into generally higher and higher groups to facilitate inductive reasoning. For more information, reference Miles, Matthew B., A.M. Huberman, and Johnny Saldaña. 2020. *Qualitative Data Analysis: A Methods Sourcebook*. 4th ed. Los Angeles: Sage.

25 A realist synthesis is the synthesis of a wide range of evidence that seeks to identify underlying causal mechanisms and explore how they work under what conditions, answering the question "what works for whom under what circumstances?" rather than "what works?" For more information, reference https://www.betterevaluation.org/methods-approaches/methods/realist-synthesis.

LIMITATIONS OF STUDY

Despite attempts at full transparency and the critical review of the research methodology by P/CVE subject matter experts, the team acknowledges that the scope and findings may demonstrate limitations.

The overall lack of P/CVE independent, peer-reviewed evaluations challenges the methodological rigor of this research and analysis. To complement traditional search methods, AfP did conduct multiple open calls for unpublished evaluations and grey literature from its network base, including to donors and research organizations. These search methods had an over-reliance on the English-language, biasing the scoping to Anglophone publications. While French and Spanish resources were included and other languages if a translation was available, this likely distorted the review's findings. It is possible that valuable resources may have been missed, leading to conclusions being drawn on partial data.

AfP's unique approach to systematic mixed methods reviews allows for the inclusion of many more studies that would traditionally be excluded due to their non-statistical research methods allowing for a deeper review of evidence base. If AfP had developed inclusion criteria based predominantly on research design, only 18 studies would have been included, severely limiting findings related to promising practices and strong theory. However, the inclusion of studies that would traditionally be excluded based on research methods requires a more nuanced assessment of evidence quality and strength. AfP did attempt to mitigate these limitations through the application of our MMAT and Strength of Evidence Continuum. While the MMAT is a peer-reviewed tool, it can present its own biases, particularly towards peer-reviewed literature, which may have impacted the overall quality appraisal of the predominantly grey literature found in this research. AfP developed the Strength of Evidence Continuum internally to provide a tool to better track evidence movement across studies, rather than simply pinpointing statistical evidence or not, which limits the intent of the research to both assess the evidence base alongside improving practice and informing future research to support professionalization of the field. While AfP made every effort to transparently apply these tools, the discretionary and subjective nature of grading can lead to inconsistencies and bias in its application.

Each study was assessed according to quality and strength using the MMAT framework and the Evidence Continuum as referenced above, although the individual studies and their associated grades have not been shared out of respect for the authors. While AfP had two coders independently code a relevant proportion of the corpus to mitigate bias for each of these tools, the researchers acknowledge this quality appraisal and strength of evidence process may still contain certain biases.

AfP's ToC analysis presents its own limitations to systematically distinguish between studies that included key activities and interventions that align with many ToCs, often even across programming responses. This was especially prevalent in endline evaluations of multi-year initiatives that influence drivers of VE, such as democracy, governance, and political stabilization. The combination of these programs complicates any evaluation attempt to assess the impact of P/CVE within more indirect programming and their categorization within the ToC analysis. There were also many programs that articulate ToCs focused on broad activities, like training, without explicit assumptions on how these approaches will impact P/CVE outcomes. The researchers attempted to collapse these individual programs into more general ToCs with direct P/CVE impacts, but this may have been erroneously classified when program logic was missing from the study. Both of these challenges can hyperinflate the number of studies for each ToC and potentially distort the evidence synthesis based on incorrect understanding of program logic. To mitigate this limitation, the evidence summary separates and focuses solely on the distinct activities and reported impacts relevant to each ToC, rather than including all reported findings.

Finally, AfP's PI employed a modified realist synthesis approach to distill the mechanisms by which interventions within each ToC work or not as part of the evidence synthesis. Narrative synthesis inherently has certain limitations. Primarily, it is more subjective compared to quantitative methods like meta-analysis and may be influenced by the PI's biases or interpretations, potentially leading to non-representative conclusions. While all efforts at transparency were made, the research recognizes

that without a standardized framework for conducting the evidence syntheses, consistency and reproducibility across the evidence synthesis may be limited.

Despite these limitations, this research effort provides valuable resources aimed at strengthening the knowledge base to improve P/CVE practice.

USAID FROM THE AMERICAN PEOPLE

PURDUE UNIVERSITY

ACLED Bringing clarity to crisis

AIR® Advancing Evidence. Improving Lives.

Alliance for Peacebuilding