TECHNICAL NOTE No IDB-TN-3231

Evidence Gap Maps as a Tool for Evidence-Informed Decision-Making

Camilo Acosta
Minji Kang
Shannon Shisler

Inter-American Development Bank
Knowledge and Learning Division

October 2025

Evidence Gap Maps as a Tool for Evidence-Informed Decision-Making

Camilo Acosta
Minji Kang
Shannon Shisler

[1] Inter-American Development Bank
[2] International Finance Corporation
[3] International Initiative for Impact Evaluation (3ie)

Inter-American Development Bank
Knowledge and Learning Division

October 2025

Cataloging-in-Publication data provided by the Inter-American Development Bank Felipe Herrera Library

Acosta, Camilo.

Evidence gap maps as a tool for evidence-informed decision-making / Camilo Acosta, Minji Kang, Shannon Shisler.

p. cm. — (IDB Technical Note; 3231)

Includes bibliographical references.

1. Knowledge management-Latin America. 2. Knowledge management-Caribbean Area. 3. Evidence-based policy-Latin America. 4. Evidence-based policy-Caribbean Area. 5. Research-Evaluation-Latin America. 6. Research-Evaluation-Caribbean Area. I. Kang, Minji. II. Shisler, Shannon. III. Inter-American Development Bank. Knowledge and Learning Division. IV. Title. V. Series.

IDB-TN-3231

http://www.iadb.org

Copyright © 2025 Inter-American Development Bank ("IDB"). This work is subject to a Creative Commons license CC BY 3.0 IGO (https://creativecommons.org/licenses/by/3.0/igo/legalcode). The terms and conditions indicated in the URL link must be met and the respective recognition must be granted to the IDB.

Further to section 8 of the above license, any mediation relating to disputes arising under such license shall be conducted in accordance with the WIPO Mediation Rules. Any dispute related to the use of the works of the IDB that cannot be settled amicably shall be submitted to arbitration pursuant to the United Nations Commission on International Trade Law (UNCITRAL) rules. The use of the IDB's name for any purpose other than for attribution, and the use of IDB's logo shall be subject to a separate written license agreement between the IDB and the user and is not authorized as part of this license.

Note that the URL link includes terms and conditions that are an integral part of this license.

The opinions expressed in this work are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent.

EVIDENCE GAP MAPS AS A TOOL FOR EVIDENCE-INFORMED DECISION-MAKING

Camilo Acosta **Minji Kang** **Shannon Shisler**

ABSTRACT[1]

This note outlines the methodology for developing Evidence Gap Maps (EGMs) as adapted for the Inter-American Development Bank, positioning them as tools to enhance the quality and impact of development policies and operations. EGMs compile and structure existing studies to highlight evidence clusters and gaps around specific topics, guiding institutions to target resources where they can have the greatest effect. The note outlines the main steps involved in building an EGM and draws lessons from two EGMs recently developed by the IDB, including one co-produced with the International Initiative for Impact Evaluation. Beyond documenting available evidence, EGMs can inform policy dialogue and guide operational design, contributing to more effective use of public resources. Looking ahead, new technologies and software will allow EGMs to be produced more quickly and of higher quality, making them an increasingly powerful tool for policymakers and practitioners.

Keywords: evidence, EGM, impact evaluation, systematic review, evidence-informed policy.

JEL Codes: A10, A30, O22, Y90

1. We would like to thank Anil Thota, Daniela Anda, Diana Cordova, Mark Engelbert, Tomasz Kozakiewicz, Thomas Kelly, and Pradeep Singh from 3ie for all their work and the valuable knowledge they shared through the implementation of this project. We are also grateful to Lorena Rodríguez and Oscar Mitnik for supporting this work agenda, and to Indira Porto, Adela Barrio, Magdalena Rossetti, Cristina Querubín, Valentina Sequi, María Paula Gerardino, Rodrigo Serrano-Berthet, Sergio Britto Lima, Harold Villalba, Leonel Borja, Agustina Calatayud, María Eugenia Rivas, Vileydy González, Gonzalo Santamaría, Santiago Sánchez, Luisa Blanco, and other collaborators for their valuable contributions at different stages of the construction of the EGMs or other associated products.

1. INTRODUCTION

In the rapidly evolving landscape of international development, robust evidence and targeted knowledge are more critical than ever for enhancing the effectiveness and impact of projects. Evidence should inform every stage of the project cycle, from programming with governments to design, execution, closure, and final evaluation. This need is particularly critical at the Inter-American Development Bank (IDB) Group, given its importance as a development partner in Latin America and the Caribbean and its mandate as the Knowledge Bank of the region.

To meet this challenge, the IDB Group is strengthening mechanisms that encourage the generation, systematization, and use of knowledge at both portfolio and project levels. This means learning from both successes and failures while upgrading the tools and incentives that support knowledge management. The agenda is embedded in the Institutional Strategy 2024–2030 and operationalized through the 2024–2030 Impact Framework, which translates strategy into measurable outcomes that guide the work of the IDB Group (IDB Group, 2024, 2025).

At the same time, institutions like the IDB Group must remain actively engaged with knowledge generated outside their own walls — by universities, development partners, and intergovernmental organizations. Failing to do so risks creating an echo chamber, where lessons are drawn only from internal operations, drifting away from the knowledge frontier while potentially reinforcing biases or ineffective practices. The IDB's Development Effectiveness Policy Framework (DEPF) reinforces this approach by embedding knowledge and evidence-based practices into project design, monitoring, and evaluation—placing impact at the center of the IDB's work and promoting learning throughout the project cycle (IDB, 2025a, 2025b).

However, searching for high-quality and relevant literature is costly and complex. Every year, thousands of studies are published, often with conflicting findings about the effectiveness of similar interventions, making it difficult for teams to navigate and validate results. Development institutions also invest heavily in producing knowledge, yet much of it remains underutilized. In 2014, the World Bank reported that 87% of its policy reports had never been cited (Doemeland & Trevino, 2014). Similarly, an initial analysis at the IDB suggests that approximately 45% of products in the IDB publications catalog have never been cited in Google Scholar. These realities underscore the need for cutting-edge synthesis tools, such as Evidence Gap Maps (EGMs), to organize knowledge efficiently and make it actionable for policy and operations.

EGMs are among the most widely used tools in evidence synthesis. They identify areas where evidence concentrates and where evidence is lacking in a given field or topic, thereby directing research and knowledge generation efforts. Visually, EGMs map the literature in a two-dimensional matrix of interventions and outcome variables by following a systematic methodology to deliver curated evidence. This approach differs significantly from many evidence repositories and literature reviews that follow less structured and sometimes subjective processes.

By systematically identifying where knowledge is concentrated and where gaps persist, EGMs enable the design of better-informed and more impactful knowledge and development strategies (Villa–Turek et al., 2023). Ultimately, these maps strengthen the effectiveness of interventions and contribute to sustainable development outcomes. Recognizing their strategic value for strengthening the quality of public policy, this note outlines how EGMs can support the IDB Group's institutional goals, presenting both a methodological approach for their development and a roadmap for embedding a culture of evidence mapping into organizational practice. While tailored to IDB audiences, the guidance and experiences summarized here have broader relevance for any institution interested in developing EGMs and leveraging them as a tool to integrate evidence more effectively into policy and practice.

Knowledge investments, such as the EGMs, are crucial in identifying development challenges, informing policy dialogues, providing technical support, and fostering innovation. Knowledge is needed throughout the entire programming and operational cycles. To this end, reforms aimed at improving knowledge generation and systematization, both from analytical and operational work, should ensure that evidence—such as research findings and data—consolidates as a public good and supports the effectiveness of development operations. By enhancing governance structures, improving data quality, streamlining operational processes, and upgrading knowledge systems, these reforms will enable better learning from both successes and failures. Additionally, they will support the optimization of impact measurement and the effective use of artificial intelligence tools (Fielding & Briss, 2006; Ali et al., 2020).

Knowledge assets are central to international development, shaping how institutions identify challenges, inform policy dialogue, provide technical support, and attract private investment, among other activities. To strengthen effectiveness in its operational knowledge cycle, the IDB Group must ensure that its operations are grounded in robust evidence, consistently evaluated, and open to new approaches. In this context, EGMs are instrumental: by systematically mapping where evidence is concentrated and where it is scarce, EGMs foster a culture of evidence-based decision-making and enable the design of more impactful strategies and interventions. Embedding EGMs into the IDB Group's operational framework will enhance project design, execution, and evaluation, ensuring that resources are directed to the region's most pressing development challenges.

Beyond their methodological value, EGMs closely align with the IDB Group's reform agenda by establishing a transparent evidence base for project preparation and programming. They also promote *strategic selectivity* by showing where proven approaches can be scaled and where new ones should be tested. Finally, EGMs strengthen the Knowledge Bank mandate by turning scattered studies into accessible, actionable knowledge, underscoring the IDB Group's role as a bridge between global evidence and regional needs. These contributions align with the Group's emphasis on scale and impact, as highlighted in the Impact Report 2025 (IDB, 2025).

EGMs were developed and have been extensively promoted by the International Initiative for Impact Evaluation (3ie). In December 2024, the IDB and 3ie launched a collaboration to develop the Bank's first institution-wide EGMs. This partnership has three main components: (i) a thorough revision of the EGM on policing interventions (hereafter referred to as the *policing EGM*); (ii) the co-production of an EGM on transport interventions (hereafter referred to as the *transport EGM*); and (iii) a series of knowledge-sharing sessions to familiarize IDB audiences with the value of EGMs for decision-making, their practical use, and the steps involved in building them. This note is released together with these two EGMs.

This methodological note is meant to complement other articles and guides on how to use and build EGMs. In particular, we highlight the seminal paper by Snilstveit et al. (2016) and methodological contributions by Colombia's National Planning Department (Delgado et al., 2019; Villa-Turek et al., 2023). While written mainly for IDB audiences, this note could also be useful for other development banks and organizations interested in developing EGMs, with a view to linking these tools to strategic documents and the operational portfolio. This note is released alongside the policing[2] and transport EGMs,[3] together with a User Guide designed to facilitate the use of existing and future EGMs (Acosta et al., 2025). The visualizations of these two EGMs are presented in **Figures 1 and 2.**

The remainder of this note has the following structure. Section II discusses the current challenges faced by different IDB Group teams in terms of accessing frontier evidence. Section III presents a more precise definition of the EGMs and their intended audiences. Section IV lays out the detailed methodology to build Evidence Gap Maps. Section V presents the conclusions. This note also includes numerous annexes presenting templates and guides that are meant to help teams develop their own EGMs.

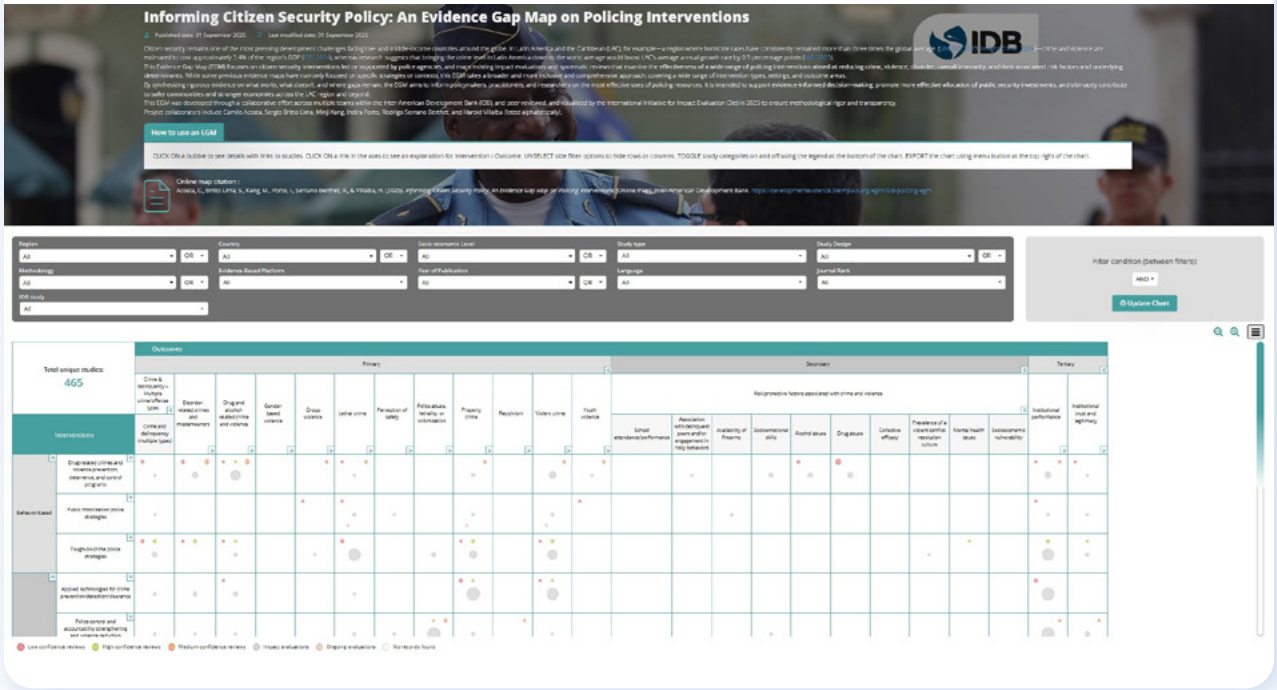
2. CURRENT CHALLENGES FOR EVIDENCE-BASED DECISION MAKING AT THE IDB

As highlighted above, the IDB Group Institutional Strategy 2024-2030 emphasizes the importance of consolidating the institution as the knowledge bank of Latin America and the Caribbean by producing cutting-edge research focused on the development issues in the region and serving as a bridge between development research made in the rest of the world that could be useful for its members. Achieving this vision requires strengthening the feedback between operations and knowledge by ensuring easy access to empirical evidence, insights from past projects, and tools to generate new knowledge from ongoing operations.

2. Informing Citizen Security Policy: An Evidence Gap Map on Policing Interventions. Available at https://developmentevidence.3ieimpact.org/egm/IDB-policing-egm

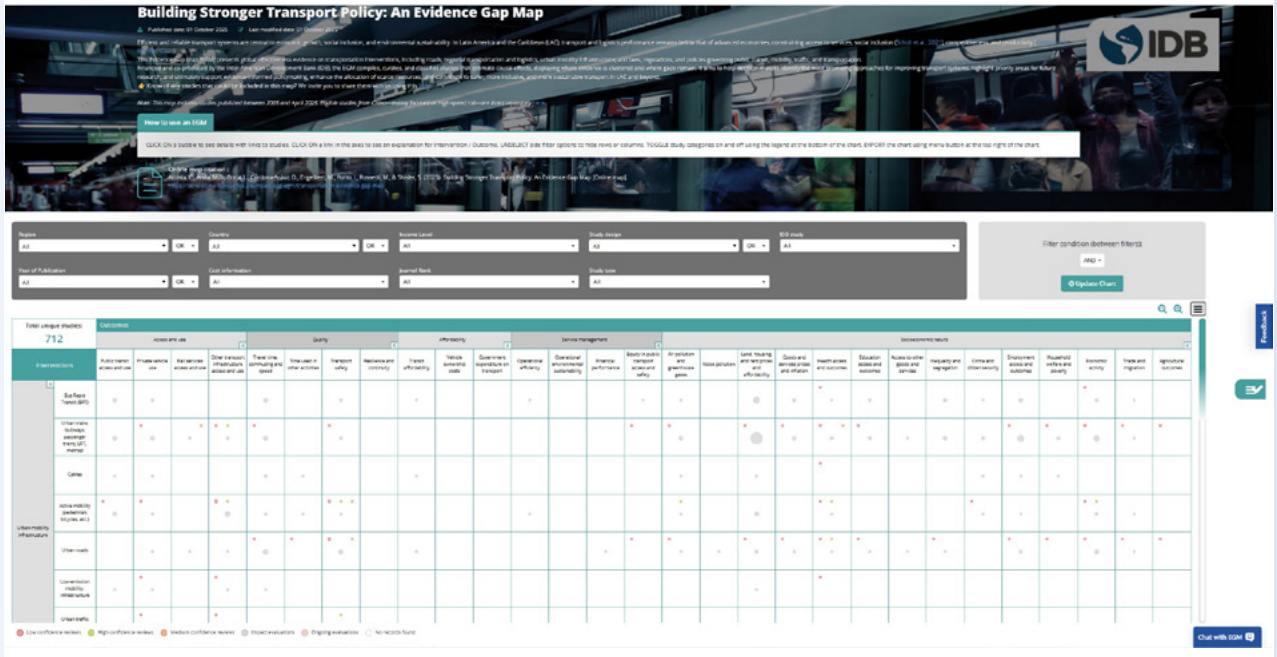
3. Building Stronger Transport Policy: An Evidence Gap Map. Available at https://developmentevidence.3ieimpact.org/egm/transportation-evidence-gap-map

Figure 1. Informing Citizen Security Policy: An Evidence Gap Map on Policing Interventions



Source: *https://developmentevidence.3ieimpact.org/egm/IDB-policing-egm*

Figure 2. Building Stronger Transport Policy: An Evidence Gap Map



Source: *https://developmentevidence.3ieimpact.org/egm/transportation-evidence-gap-map*

Against this backdrop, a central question arises: where is knowledge most needed along the programming and operational cycles? The short answer is at every stage. Specifically, during the programming cycle, knowledge underpins diagnostic assessments of country needs, informing and shaping the project pipeline, and guiding the prioritization of interventions most aligned with societal demands, government objectives and IDB Group's strategic priorities. Thus, evidence that complements the specialists' experience and contextual knowledge is highly valuable in the construction of Country Strategies and Country Development for Impact (CDI) documents, as well as when establishing dialogues with government counterparts. Importantly, these country-level strategies cascade into project development by shaping the pipeline, informing the selection of sectors and interventions, and ensuring that individual operations are aligned with broader strategic objectives. In this way, evidence strengthens not only upstream policy dialogue but also the downstream design and execution of projects, enhancing their relevance, quality and impact.

In addition, during the design stage of an operation, high-quality evidence is essential to validate the internal and external validity of proposed interventions, enabling the construction of robust Theories of Change, realistic targets, and measurable results frameworks. As projects move into execution, reformulations should be informed by actionable insights derived from previous findings. Finally, the closure stage focuses on evaluating project effectiveness through rigorous impact evaluations that should draw on existing knowledge to select the most appropriate methodologies, position new findings within the broader evidence base, and generate lessons for future interventions. Importantly, EGMs can guide which outcomes to prioritize for evaluation and whether specific operations merit an impact evaluation, helping the IDB strategically allocate limited evaluation resources to maximize learning and impact.

Beyond operations, cutting-edge evidence also plays a critical role in shaping the Bank's knowledge agenda. EGMs and other synthesis tools help identify key evidence gaps, guiding where new research investments are most needed and where limited resources can have the greatest effect. Moreover, existing literature should inform the development of all knowledge products, from initial concept through publication, ensuring that each output builds on the current state of evidence rather than duplicating it.

Operational and knowledge teams across all IDB units employ multiple strategies to access evidence. Teams can access evidence through Thematic Framework Documents (TFDs), previously known as Sectoral Framework Documents (SFDs). TFDs are knowledge documents that should provide a synthesis of the main development challenges and the most relevant evidence on a specific topic. These documents should summarize which interventions work in which particular contexts and identify knowledge gaps to guide future research efforts. TFDs are intended to be a flexible guide and reference for the IDB Group's work on a particular topic. Although TFDs and SFDs cover a wide array of topics, there is no SFD for every theme.

Currently, SFDs are available only on the Bank's website and are not included in the IDB publications catalog, which may hinder their visibility and limit their use. As a result, project teams often resort to alternative strategies to systematize and access evidence. In some cases, teams invest significant time and resources conducting literature reviews for a single project. In other cases, teams have also resorted to more dynamic, tailored and publicly available synthesis solutions, such as the *Evidence Platform on Security and Justice* (led by the Citizen Security Division), the *Skills Bank* (led by the Research Department and the Education Division), and the *What Works to Improve Lives series* (led by the Strategic Development Effectiveness Division).[4]

3. EVIDENCE GAPS MAPS

The use of evidence in public policy and development interventions has grown considerably in recent decades. Within the IDB Group, demand for knowledge is evident across multiple units, prompting teams to implement diverse solutions to meet this need. Evidence can take many forms—scientific articles, databases, reports, websites, and more—yet this multiplicity of formats creates important challenges for operational teams and policymakers. Specifically: (1) ensuring material access to the relevant evidence for policy formulation, design, implementation, and evaluation; (2) determining whether and where evidence exists on specific topics of interest, including the impacts and results of related interventions; and (3) validating the quality of the evidence identified.

A myriad of tools have been developed to synthesize and present evidence that supports policymaking. These possibilities include systematic reviews and meta-analyses, rapid evidence assessments, and EGMs, among others (Snilstveit et al., 2016). In essence, EGMs are instruments that allow for the mapping and synthesizing of large volumes of relevant evidence to offer a general visual overview of existing knowledge about specific public policy interventions on larger outcomes of interest.

EGMs have been developed and promoted extensively by 3ie and were conceived as a response to this challenge and the need to facilitate evidence-based decision-making (Snilstveit et al., 2016). In particular, 3ie EGMs are thematic collections of information about impact evaluations and systematic reviews. At the same time, these maps provide a visual representation structured around a framework of interventions and outcomes. An EGM can be accompanied by a report that provides all the supporting documentation for the map, including background information for its theme, the methods, protocols, and the analysis of results. 3ie's first EGM was produced

4. These collections can be found at (i) https://plataformadeevidencias.iadb.org , (ii) https://skillsbank.iadb.org/en, and (iii) https://publications.iadb.org/en/Advanced-Search?a%5B0%5D%5Bf%5D=field_title&a%5B0%5D%5Bv%5D=%22what%20works%20to%20improve%20lives%22.

in 2010 (Gaarder, 2010), focusing on agriculture interventions and health and nutrition outcomes. Since then, 3ie has produced more than 40 EGMs.[5]

The usefulness of the EGMs could be summarized in four ways. **First**, EGMs are helpful to facilitate informed policymaking and evidence-based decisions in both development institutions and governments. **Second**, EGMs clearly identify which policy aspects have not been evaluated, or which aspects are over-evaluated. They can signal the existence of absolute gaps when few primary studies exist in a given intervention-outcome intersection, and synthesis gaps when multiple studies exist but few recent or high-confidence reviews are available (Villa-Turek et al., 2023). These mappings allow for a better allocation of resources for both operational and knowledge investments.

Third, EGMs use a systematic methodology that attempts to be more objective than traditional literature reviews. They draw on broad searches rather than research from specific authors or institutions, resulting in a more comprehensive body of evidence. **Fourth**, and more internal to the IDB Group, EGMs can complement important sectoral documents—such as TFDs—by facilitating access to cutting-edge knowledge and serving as a valuable input for their construction. At the same time, EGMs can act as a visualization layer for TFDs, making them more dynamic and interactive, while their complementarity would also allow for easier updates of both.

Importantly, EGMs should be seen as complementary intelligence. Their value lies in complementing the tacit and contextual knowledge of specialists with a systematic overview of the evidence base. When used together, they make decisions more credible and relevant, turning research into actionable insights for operations and policy dialogue. Therefore, the IDB's EGMs are designed for the following users:

- Sectoral specialists and teams, both in headquarters and in country offices, working on operations and knowledge-oriented tasks.
- Country economists and their teams.
- Development Effectiveness specialists.
- Research teams, including specialists and their research assistants.
- Country Representatives, Division Chiefs, and Sector Managers, who oversee strategic alignment

5. The complete list of EGMs produced by 3ie can be found in their Development Evidence Portal (DEP), available at: https://www.3ieimpact.org/evidence-hub/evidence-gap-maps

Beyond these internal users, EGMs could also serve as a source of knowledge and public goods for specialists from other development organizations, policymakers, and other external stakeholders.[6]

However, it is essential to clarify what EGMs are not. First, EGMs are not primary research, as they do not generate new data or findings through original studies. Second, mapping alone does not indicate what the evidence says about the effectiveness of an intervention. Producing such insights would require a higher reliance on high-confidence systematic reviews or a more profound analysis of included evaluations, including extracting precise statistical information from the included articles and applying appropriate econometric techniques. Third, EGMs are not generative AI assistants, although the information they produce can inform such systems. For instance, the Transport EGM includes an AI chatbot that helps users navigate the map. Finally, a lack of evidence in some regions of an EGM does not prove an intervention is ineffective; it simply shows that more evidence is needed.

4. IDB ADAPTED METHODOLOGY

In this section, we describe the methodology used to create IDB-oriented EGMs. To expand the usefulness of EGMs and adapt them to the IDB context, the Knowledge and Learning Division (KLD) team collaborated with 3ie and different units across the Institution, including the Strategic Development Effectiveness Division (SPD/SDV) and sectoral teams, to adapt 3ie's established methodology. This process is organized into seven stages: defining the scope, establishing a protocol, conducting a literature search, screening studies, extracting data, critically appraising systematic reviews, and visualizing the map. The remainder of this section discusses each of these stages in detail.

Before starting an EGM, certain criteria should be in place. The interested division or sector should be willing to lead its development and demonstrate readiness to oversee the process, as EGMs require significant time, financial, and human resources. Moreover, it is recommended to conduct a preliminary online search to verify if there are existing EGMs or other synthesis tools that may be useful to the interested parties.[7] The Knowledge and Learning Division adds value as a strategic, active partner by providing the methodological framework, quality assurance, and accompanying the process.

6. This characteristic is not unique to the EGMs. For instance, the Evidence Platform on Security and Justice is intended for both operational and knowledge teams of the Citizen Security Division and for external stakeholders, such as policymakers and researchers, in Latin America and the Caribbean.

7. It is recommended to start the search in specialized organizations that produce Evidence Gap Maps, such as 3ie (https://www.3ieimpact.org/evidence-hub/evidence-gap-maps), the Campbell Collaboration (https://www.campbellcollaboration.org/tag/evidence-maps/), or the World Health Organization (https://www.who.int/initiatives/decade-of-healthy-ageing/evidence-gap-map), or by conducting a series of searches in a web browser. Colombia's National Planning Department has also worked on developing its own EGMs (https://sinergia.dnp.gov.co/evaluaciones/paginas/mbe.aspx)

It is worth highlighting that, currently, EGMs are not self-updating products. To maintain their relevance, it is recommended that they be updated every three to five years, depending on the pace of new evidence emerging in the sector and the resources available for updating.

4.1. Define the Scope

The first step in creating an EGM is to define its scope. This step is one of the most critical stages in the construction of an EGM—or any evidence synthesis effort—because it sets the foundation for all subsequent steps. At this stage, the team should articulate the **research question** the map or review seeks to answer, framing it around the key policy issue of interest. For example: "*What is the extent of the empirical evidence on the effects of interventions to address [a policy issue] in [the context of interest]?*" Defining the question also helps determine whether an EGM is the most appropriate tool for the team's objectives.

With the question defined, another tool frequently used to determine the EGM scope is the so-called **PICOS framework**. PICOS, which stands for Population, Intervention, Comparison, Outcome, Study Design, is a structured approach commonly used in evidence-based research to define research questions and guide systematic reviews. In the context of EGMs, PICOS can help systematically categorize interventions and outcomes and identify the target population in an orderly fashion. Regarding the **population of interest (P)**, 3ie's EGMs tend to focus on low and middle-income countries. However, if evidence from high-income countries is valuable to inform operations and public policy in the region, mapping available studies and reviews from any region is valid. The final scope of the population covered in each map depends on the interests and resources of the respective divisions, and it must be the same for all the interventions and outcomes covered on the map.[8]

Once the population is defined, an **Interventions-Outcome (I-O) Framework** must be developed. This framework encompasses all significant interventions and outcomes within the sector or topic being analyzed. In this framework, the rows should represent all relevant interventions for the specific topic, while the columns should capture all outcomes of interest. These outcomes should be organized along the causal chain, ranging from direct topic-specific outcomes to variables analyzing the impact and cost-effectiveness of the interventions, involving more comprehensive outcomes showing the long-term impact of the intervention. Both interventions and outcomes should be included in more aggregated categories that facilitate the use of and navigation within the EGM. To improve user experience, it is recommended that an

8. For example, in the case of the transport EGM, it was essential to consider worldwide evidence for several reasons. First, finding causal studies with experimental or even quasi-experimental variation to evaluate transportation interventions is not straightforward. Second, some policies (e.g., urban tolls and congestion charges) have been applied in only a handful of cities or territories.

EGM does not have more than 30 interventions or outcomes. If the sectoral teams would like to add more intervention categories, they could also consider breaking the framework and creating two separate EGMs.

For example, in the transport EGM, one intervention category is "Urban Mobility," which includes specific interventions such as the construction of new Bus Rapid Transit systems or the implementation of public transportation subsidies. Among the outcomes, some categories measure direct results like "Access and Use," and others measure broader impacts such as "Socioeconomic Indicators." Among others, "Access and Use" includes outcomes related to the access and use of public transit systems, while "Socioeconomic Indicators" includes different labor market indicators and outcomes measuring productivity, income, or household consumption.

Handling multicomponent interventions is a recurring challenge in the construction of EGMs. These programs often combine several stand-alone activities—for example, a public transit program that includes both the construction of a light rail system and the provision of transit subsidies—which may correspond to different intervention categories in the framework. To ensure comparability and avoid double-counting, studies should generally be coded into a single intervention, following the principle of mutually exclusive categories, unless the paper separately reports the effects of each component. The chosen approach must be documented in the protocol. In the transport EGM, this issue was addressed by creating a "multi-intervention" category within each main intervention group, as well as an additional category for studies spanning more than one group. Annex 1 includes a more detailed description of this problem along with some proposed solutions.

Defining an I-O framework for interventions and outcomes involves two steps. The first step focuses on constructing an initial framework, and the second step is a validation process with IDB specialists. In this stage, the I-O framework should leverage existing knowledge from (i) the TFDs/SFDs, (ii) the portfolio of IDB loan operations, and (iii) the SPD Indicator Catalog (when available). Aligning with these sources is essential to facilitate navigation within the EGM, enhance usability, and ensure consistency within the institution. The use of each of these sources in the construction of the framework will be briefly presented below.

A. Thematic or Sectoral Framework Documents: As mentioned in Section II, the TFDs review the most prevalent development challenges and evidence around one topic. These documents provide strategic guidance on priority areas, including *lines of action* to address pertaining challenges, discussing the work the IDB has done (or is thinking about) to deal with them. These documents are also meant to guide future knowledge agendas. [9]

9. All the Sectoral Framework Documents are available at https://www.iadb.org/en/who-we-are/about-idb/operational-policies

Therefore, TFDs help align the EGMs with the IDB's sectoral strategies, ensuring relevance to the Bank's goals and facilitating the identification of evidence that informs its operations. TFDs also offer context-specific insights into regional challenges and opportunities, cross-sectoral linkages, and stakeholder engagement. TFDs should also include relevant literature in the knowledge frontier and could be an important first input for the EGMs. Moreover, an updated and aligned EGM could also serve as the primary input for future TFDs and other synthesis products.

In addition, TFDs can inform the I-O framework by identifying the relevant interventions and outcomes included in the document. This identification can be achieved manually or by utilizing AI-powered systems, such as the IDB Group Knowledge Platform, to identify key challenges, lines of action, and priority areas. Both approaches ensure that the I-O framework is aligned with IDB's strategic goals and responsive to the actual needs and opportunities in the region. Annex 2 presents a comparison between the results from the human and the AI-assistant approaches in identifying intervention categories from the Transportation SFD.

B. Loan Operations: The EGM team should meticulously identify and curate a comprehensive list of all approved operations in the division(s) related to the EGM theme. Revising all relevant operations ensures that the interventions and outcomes devised for the EGM align with the IDB's main operational agenda and may suggest additional types of interventions or outcomes. In general, this approach ensures alignment with the institution's strategic goals and enhances the EGM framework's relevance and effectiveness.

C. SPD's Catalog of Results Indicators: Since 2023, the Office of Strategic Planning and Development Effectiveness (SPD) has been working on consolidating a catalog of results indicators to standardize, facilitate, and enhance how the IDB measures project impact. These indicators are based on "roots," which comprise both fixed components and attributes built using AI and ML, streamlining the creation of new indicators. These new indicators are also part of aggregated outcome categories and aligned with different intervention categories. In the Fall of 2024, SPD released an initial version with indicators for the three divisions in the Infrastructure and Energy Department. By the end of 2026, it is expected that all sectors will be included in the Catalog.[10]

By strategically aligning the TFDs and Bank's portfolio with these top-line indicators and metrics, the approach promotes coherence and consistency with the overarching institutional priorities and strategic goals of the IDB. This linkage helps to provide clear, measurable outcomes for each intervention category, facilitates better monitoring and analysis of project impacts, and supports the broader development effectiveness efforts by enhancing data systems within the IDB and in partner countries.

10 . The Catalog of Results Indicators is available for IDB Employees at https://indicators-catalog.iadb.org/catalog

Ultimately, there are two additional advantages to using this Catalog. First, as these catalogs were built with collaboration from the sectoral teams, the presented categorization already has a vital degree of validation from them. Second, these indicators provide a robust, evidence-based approach to assessing and improving the understanding of operational/research causality and effectiveness of development interventions.

D. Additional Sources: The I-O Framework could also involve policy documents from major local or international policy organizations in the field of interest. Furthermore, it could include existing systematic reviews, EGMs, and other literature sources. For instance, for constructing the policing EGM, 3ie's previous EGM on "Building Peaceful Societies" was a frequent source of information.

Once an initial version of the I-O framework is consolidated, a validation exercise is conducted with the division's specialist in charge and the supporting team to verify the framework's validity. This step is critical as it ensures ownership and commitment of each participating division. Once this validation occurs, a second version of the framework is consolidated, and the respective specialist decides whether to undergo an additional round of validation by their peers or external experts.

Following the PICOS framework, defining the **Comparison (C)** element can be challenging for EGMs, since they map existing evidence rather than conduct direct comparative or causal analyses. However, in practice, it simply answers the question of whether the team wants to include studies that compare an intervention to the population's status quo or to a specific alternative.

More relevant for an EGM is the definition of the **Study Design (S)**, which specifies the types of studies to be included in the map. Given the IDB's emphasis on impact and effectiveness, it is recommended to include only studies that employ strong causal inference methodologies (e.g., randomized controlled trials and quasi-experimental designs), while excluding purely descriptive or qualitative research.

Additional inclusion criteria:

Finally, teams could also establish additional inclusion criteria to tailor the EGM to their needs. Some examples are:

- **Type of studies:** Teams should incorporate impact evaluations and systematic reviews. Impact evaluations help determine the effect of a particular intervention and policy, and their absence in a specific I-O combination may inform the existence of important knowledge gaps. However, the results of most impact evaluations tend to be very local with little external validity, making

it hard to generalize conclusions about the effectiveness of some interventions. Thus, including systematic reviews can also be rendered appropriate. In general, it is recommended to include both types of articles and show them on the map as different objects.

- **Languages:** Although English dominates academic research, restricting inclusion to only English-language documents could introduce bias. At the same time, including all possible languages may create significant challenges for screening studies and extracting data. A pragmatic recommendation for IDB EGMs is therefore to include studies written in any of the four official languages: English, Spanish, Portuguese, and French.

- **Time frame:** It is recommended to define a time frame to limit the number of studies included in the search strategy and to ensure that the evidence remains up to date and relevant. Acceptable starting points include setting the initial date 20 years before the search or using 1995, which aligns with the beginning of the credibility revolution in economics.

4.2. Establish a Protocol

Establishing a protocol is a critical step in the development of any EGM and is considered part of the best practices in synthesis tools. A protocol ensures that the entire process is transparent, systematic, and replicable, giving credibility to the map and confidence to those who will rely on it. It provides a clear record of the criteria defined above. Beyond guiding the immediate construction of the map, a protocol plays an essential role for future users. Teams other than the original authors, or even the same team years later, can refer back to the protocol to understand the scope and methodology, interpret the results properly, and update the map when new evidence becomes available. Without a protocol, updating or replicating a map risks inconsistency, duplication of effort, or even misinterpretation of its findings.

The protocol also serves as a practical project management tool. It sets out the databases to be searched, the screening criteria to be applied, and the data to be extracted, ensuring that all steps follow a high-rigor methodology. A protocol makes the process not only more credible but also more efficient, as teams can work from a shared understanding of the tasks and quality standards. Annex 3 of this document provides an outline of the minimum recommended elements that every protocol should include to safeguard quality and comparability across EGMs developed at the IDB.

4.3. Search the Literature

The search stage is a critical step that involves conducting a comprehensive and systematic search for relevant studies. This search should be as thorough and unbiased as possible to ensure that the resulting EGM provides a reliable and accurate overview of the existing evidence. By meticulously searching across multiple databases, repositories, and sources, researchers aim to minimize biases that could skew the findings or lead to incomplete conclusions.

To ensure high-quality evidence, searches should cover major academic databases and, where possible, thematic websites relevant to the study area. Annex 4 lists recommended repositories for scholarly research, including Elsevier's *ScienceDirect,* the National Bureau of Economic Research, and the CESifo Network, among others. Systematic reviews should be identified through specialized repositories such as 3ie, Cochrane, or Campbell, as well as discipline-specific libraries. In addition, searches should extend to specialized institutions and repositories to capture relevant grey literature.[11] For example, the transport EGM drew on sources including the IDB, World Bank, RAND Transportation, and UN Trade and Development.

The selection of keywords is a critical step in developing the search strategy for an EGM. Keywords should capture the main topic, the interventions and outcomes of interest, and the types of studies and methods to be considered. They should also reflect the specific context of interventions, such as population groups or geographic regions and, when appropriate, include additional terms that broaden or refine the evidence base to ensure comprehensive coverage of relevant literature.

Once identified, core keywords should be combined into structured search strings using Boolean operators ("AND," "OR," "NOT"), which are supported by most bibliographic databases. These search strings must align closely with the EGM's scope and inclusion criteria, reducing irrelevant results while capturing relevant studies efficiently. Annex 5 provides an example of the search strategy developed for the transport EGM. As emphasized by Snilstveit et al. (2016), the involvement of an Information Specialist can be highly valuable in designing and refining the search strategy, ensuring both comprehensiveness and precision. In addition, users can leverage large language models (LLMs) to generate a set of relevant keywords and Boolean search strings, complementing traditional approaches and improving efficiency.

11. Grey literature usually refers to "information produced on all levels of government, academia, business and industry in electronic and print formats not controlled by commercial publishing," i.e., where publishing is not the primary activity of the producing body." (Rothstein & Hopwell, 2009)

After conducting a comprehensive search, techniques such as snowballing or citation tracking can be employed to expand the evidence base and ensure a comprehensive representation of available knowledge. Snowballing refers to tracking back and pursuing references of relevant papers from both primary studies and reviews to increase the body of evidence (Greenhalgh & Peacock, 2005). Forward citation tracking refers to the identification of newer studies or reports that cite a specific article that has already been published and is highly relevant for the EGM. All these searches and processes should be documented in the EGM's protocol.

To avoid duplication and redundancy, it is recommended that searches be documented, including the search date, the search string(s) used, and the number of studies found and downloaded. Moreover, the information from relevant studies should be stored in **reference management** software, such as Zotero or Mendeley, which offer comprehensive, low-cost solutions for managing bibliographic data. Current versions of these programs are easy to use. First, creating most new references only requires the article's Digital Object Identifier (DOI), which is issued by most publishers nowadays, and the software will then complete the rest of the article's information using available metadata from the Internet. Second, once all the references are stored, easy-to-use RIS or BibTeX files can be exported to Word, LaTeX, or other text editors. The use of this software is highly recommended, as proper storage of references facilitates the construction of EGMs.

4.4. Screen and Select Relevant Studies

A comprehensive literature search may yield an overwhelming number of results. For example, the search strategy for the transport EGM yielded nearly 39,000 records. The natural question that follows is how to filter this volume of information to identify the studies that are truly relevant to the EGM's scope. This requires a systematic screening process, designed to ensure that only studies meeting the agreed inclusion criteria move forward to data extraction and mapping.

Before starting this process, a deduplication of studies is recommended if the search strategy involved multiple databases and repositories. For example, a relevant study with coauthors from the IDB and World Bank could be found in at least two institutional catalogs. This deduplication can be easily done by leveraging AI tools or statistical matching algorithms. For the transport EGM, deduplication and the screening process were done using **EPPI-reviewer**, a web-based software developed by the Evidence for Policy & Practice (EPPI) Centre to support all aspects of literature review and evidence synthesis, including deduplication, screening, and extracting data (Thomas et al., 2023).[12] For the transport EGM, this process led to the deletion of 4,000 records.[13]

12. More information on EPPI-Reviewer can be found at: https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2914

13. In the case of the deduplication, EPPI suggests a highly accurate similarity score between studies. For the transport EGM, all studies with a similarity score below 0.7 will be considered non-duplicates, while scores of 0.95 and above will be considered duplicates. Bibliographic information from studies with scores between 0.71 and 0.94 was reviewed to complete this deduplication.

After deduplication, the recommended screening process follows two steps: first, a screening of titles and abstracts only (TAS), and second, a full-text screening (FTS). During TAS, the screening team reviews only titles and abstracts to assess inclusion according to the criteria defined in the protocol. These criteria could be made explicit in a checklist. TAS is efficient and practical, and it reduces the cognitive load of the more demanding FTS. To ensure consistency, the initial rounds of TAS should be conducted in duplicate: two reviewers screening the same set of studies to align definitions and resolve potential misunderstandings. Once agreement levels are high and conflicts become minor, TAS can proceed independently. Once mastered, TAS of a single study can take less than one minute. Annex 6 includes an example of a TAS screening checklist.

When the number of search results is very large and a substantial set of studies has already been screened, AI classifiers can be trained to accelerate the process. In the transport EGM, the classifier embedded in EPPI Reviewer was iteratively trained and refined until the proportion of relevant studies fell below an inclusion threshold of 20% was sufficiently low. Using this approach, the team screened about 6,000 studies at TAS and flagged around 2,500 for FTS.[14] It is important to note, however, that classifiers trained on too few studies can generate high rates of false positives and false negatives, which undermines both the reliability and the usefulness of the EGM.

During FTS, teams should begin by retrieving and downloading the full texts of the studies flagged for this stage. Reviewers then assess these studies using a checklist similar to the one applied at TAS, but with greater detail regarding study methods and evaluated outcomes. Annex 7 includes an example of an FTS screening checklist. To ensure rigor, it is recommended that FTS be conducted in duplicate, with differences reconciled between reviewers. When resources are limited, more flexible strategies may be adopted, such as conducting duplicates until a minimum level of consistency is reached or using AI classifiers. However, full-text classifiers tend to be more costly and time-consuming to develop than title-and-abstract classifiers, which makes them less common in practice. For the transport EGM, out of the approximately 2,500 that were screened in duplicates at FTS, around 900 studies were deemed eligible for data extraction.

At this stage, main and linked publications should be identified to avoid double counting or over-weighting evidence for a particular intervention. Linked studies can often be detected through identical trial numbers, study names or acronyms, or overlapping authors. The team should establish and document in the protocol a clear decision rule for selecting the main study, for example, prioritizing the most recent publication, peer-reviewed work over working papers, the most complete reporting, or the longest follow-up period, or leave all versions.

14. The high number of studies reflects the broad scope of the transport EGM, which was set to include evidence from high- and upper-middle-income countries (e.g., the US, China, Europe) where transportation research is abundant.

4.5. Extract Data

After selecting the relevant studies, the team must extract the information required to build the EGM. Data extraction can be conducted in EPPI Reviewer, Excel, or through custom-built platforms such as 3ie's Development Evidence Portal. This process goes beyond collecting basic bibliographic details (e.g., authors, URL/DOI, journal, year, country, and language, most of which are already available in reference management software) to include methodological and contextual variables such as study design, population of interest, and, most importantly, the type of intervention and outcomes evaluated. These variables allow each study to be accurately positioned within the intervention–outcome matrix, enable the creation of map filters, and capture additional information that may be valuable to the team. Annex 8 provides an example of the data extraction template, with asterisks marking the mandatory fields. Upon completing this step—and before proceeding to visualization—it is recommended that the team conduct a quality assurance review to ensure that all fields are complete and consistent across studies.

EGMs do not typically provide data on the direction or magnitude of effects, as this would require additional resources to extract estimates, standard errors, and sample sizes from all included studies. Consequently, these variables are usually excluded from standard EGM data extraction. However, if a team anticipates conducting a meta-analysis or a more detailed synthesis in the future, it is advisable to collect this information during the data extraction stage to avoid duplication later. That said, displaying effectiveness information directly within an EGM carries important risks. Simplified summaries, such as vote counting based on the number of positive or negative effects reported, can be misleading because they fail to account for study quality, heterogeneity, or statistical significance (McKenzie & Brennan, 2019).

4.6. Critically Appraise Systematic Reviews

In parallel with data extraction, teams may carry out the critical appraisal of systematic reviews. In general, critical appraisals help users distinguish high-quality studies and reviews from weaker ones, even after the screening process. Given the large number of studies typically included in an EGM, it is recommended to prioritize the appraisal of systematic reviews. Appraising individual impact evaluations is generally not undertaken, as it requires substantial subject-matter expertise and resources beyond the scope of most EGMs.

For the appraisal of systematic reviews, teams can use standardized tools such as the SURE checklist (see Annex 9), which assigns an overall rating of high, medium, or low confidence in the findings. This appraisal focuses on the quality of the review's methodology rather than the individual studies it covers. Depending on available resources and the desired level of rigor, appraisals can be conducted using a "fatal flaw" approach or through a complete review, either individually or in duplicate. High- and medium-confidence systematic reviews can then be reliable sources of credible, decision-ready knowledge for answering questions about the effectiveness of interventions.

4.7. Visualize the Map

The eligible studies and their extracted data should be displayed in an interactive EGM tool. Visualization makes evidence more accessible and actionable by showing where knowledge is strong and where gaps remain. In these tools, the concentration of studies is represented by larger circles within each intervention–outcome cell, while the results of critical appraisals are typically shown through traffic-light color coding to indicate the confidence level of systematic reviews. EGMs can also display protocols or ongoing studies in another color. Because individual studies may cover multiple interventions or outcomes, they can appear in more than one cell if they present the effects of different interventions separately (see the discussion regarding multicomponent interventions in Section IV.1). The interactive visualization also displays different filters that users can leverage to navigate the EGM, which depend on the information extracted (e.g., methodology, countries, languages, etc.).

Most organizations currently use their own visualization platforms, though providers such as EPPI are developing shared solutions. The IDB's first two EGMs (on policing and transport) were visualized using 3ie's DEP, while an in-house visualization tool is planned for launch in 2026. At the same time, advances in AI are expanding possibilities for user interaction. For example, the transport EGM piloted a generative AI chatbot that allows users to navigate studies by asking questions directly about the evidence available within the EGM framework, making evidence more accessible and actionable.

Finally, every EGM should be accompanied by a summary report that synthesizes the findings, describes the methodology, and highlights key evidence gaps, policy implications, and directions for future research (Snilstveit et al., 2016). This report, ideally written by subject-matter experts on the EGM's topic (such as IDB sectoral specialists), is essential to ensure that the evidence presented is not only visualized but also interpreted, contextualized, and translated into actionable insights and recommendations.

5. CONCLUSION

This note outlines the methodology and importance of developing Evidence Gap Maps (EGMs) at the IDB. EGMs provide a structured way to synthesize and visualize the existing body of evidence, clarifying where evidence is clustered and where gaps remain. By following a systematic process—from defining the scope and establishing a protocol to screening, data extraction, and visualization—EGMs strengthen the Bank's capacity to link cutting-edge knowledge with operations and inform policy dialogue.

The IDB has already advanced this agenda with two pioneering EGMs. The *policing EGM*, developed with contributions from multiple units and refined through collaboration with 3ie, provides a systematic mapping of evidence on policing strategies. The *transport EGM*, co-produced with 3ie, is one of the largest sectoral EGMs to date, drawing from over 35,000 initial search results and resulting in nearly 900 included studies. Both EGMs were guided by detailed protocols that ensured rigor and transparency, and their underlying data are publicly accessible through the IDB Open Data Catalog, reinforcing accountability and openness.

Looking forward, the Bank plans to expand its portfolio of EGMs across priority areas, strengthening the evidence base for operational and policy decisions. The Knowledge and Learning Division (KLD) will continue to play a central role in supporting divisions, offering methodological guidance, templates, and technical expertise. Collaboration with 3ie or similar organizations remains key for both quality assurance and capacity building within the institution. Just as importantly, building and maintaining these tools requires significant resources and effort, highlighting the need to work together with other development institutions. When developed collaboratively, EGMs can serve as a public good, accessible to a wide range of stakeholders while multiplying their value across sectors and institutions.

The future of EGMs at the IDB should also involve greater use of emerging technologies. Advances in artificial intelligence promise to accelerate the production of high-quality EGMs by automating parts of the search, screening, and data extraction processes, leveraging software integration, APIs, and other tools. As noted by Villa-Turek et al. (2023), AI has the potential to support the development of novel synthesis tools, enabling more efficient, accurate, and scalable evidence mapping. The transport EGM has already piloted a generative AI chatbot to facilitate interaction with studies and data, offering a glimpse of the opportunities ahead.

EGMs represent a critical lever for consolidating the IDB Group as the go-to knowledge bank in Latin America and the Caribbean. Far more than methodological tools, they embed evidence into decision-making and transform research into actionable development outcomes, reaching high levels of stringency, innovation, and impact. By combining methodological rigor, institutional commitment, and new technologies, the IDB is positioning itself as the forefront of evidence synthesis— setting a global benchmark on how development institutions harness knowledge to deliver lasting impact.

REFERENCES

Acosta, C., Rossetti, M., Porto, I., Anda, M. D., & Shisler, S. (2025). *Evidence Gap Maps: User Guide.* https://doi.org/10.18235/0013738

Delgado, C., Aragón, J., Celis, M., López, D., Uribe, L., & Sáenz, M. (2019). *Guía metodológica para la construcción de mapas de evidencia.* Departamento Nacional de Planeación, Colombia. https://colaboracion.dnp.gov.co/CDT/Sinergia/Documentos/2_Guia_metodologica_para_la_construccion_de_Mapas_de_Evidencia.pdf

Doemeland, D., & Trevino, J. (2014). Which World Bank reports are widely read? *World Bank Policy Research Working Paper,* **(6851).** http://documents.worldbank.org/curated/en/387501468322733597

Gaarder, M. (2010). Evidence on links in the causal chain: Agriculture, nutrition and health. *International Initiative for Impact Evaluation.*

Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ, 331*(7524), 1064-1065. DOI: 10.1136/bmj.38636.593461.68

Fielding, J. E., & Briss, P. A. (2006). Promoting evidence-based public health policy: can we have better evidence and more action? *Health Affairs, 25*(4), 969-978. DOI: 10.1377/hlthaff.25.4.969

Inter-American Development Bank (2025a). New Development Effectiveness Policy Framework. https://idbdocs.iadb.org/wsdocs/getdocument.aspx?docnum=EZIDB0000578-308104095-506

Inter-American Development Bank (2025b). Driving Impact: Foundations of Development Effectiveness Reforms at the IDB. https://www.iadb.org/document.cfm?id=EZIDB0000578-1350314980-832

IDB Group (2024). IDB Group Institutional Strategy: Transforming for Scale and Impact. https://www.iadb.org/en/who-we-are/institutional-strategy

IDB Group (2025). *Impact Report 2025: Transforming for Scale and Impact.* Washington, DC: IDB. http://dx.doi.org/10.18235/0013629

McKenzie, J. E., & Brennan, S. E. (2024). *Chapter 12: Synthesizing and presenting findings using other methods.* **In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.),** *Cochrane Handbook for Systematic Reviews of Interventions* (version 6.5). John Wiley & Sons. https://training.cochrane.org/handbook

Rothstein, H. R., & Hopewell, S. (2009). Grey literature. *The handbook of research synthesis and meta-analysis,* 2, 103-125.

Sabet, S.M., Miranda, J., & Engelbert, M. (2020). Impact Evaluation Repository Search and Screening Protocol: 2017 Update. *International Initiative for Impact Evaluation, Washington, DC.*

Snilstveit, B., Vojtkova, M., Bhavsar, A., Stevenson, J., & Gaarder, M. (2016). Evidence & Gap Maps: A tool for promoting evidence-informed policy and strategic research agendas. *Journal of Clinical Epidemiology, 79*, 120-129. DOI: 10.1016/j.jclinepi.2016.05.015

Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O'Driscoll, P., Bond, M., & Koryakina, A. (2023). *EPPI-Reviewer: Advanced software for systematic reviews, maps and evidence synthesis.* EPPI Centre, UCL Social Research Institute, University College London.

Villa-Turek, E., Insuasti-Ceballos, H. D., Ruiz-Saenz, J. A., & Campo-Robledo, J. (2023). Evidence Gap Maps as Critical Information Communication Devices for Evidence-based Public Policy. *arXiv preprint arXiv:2304.10576.* https://doi.org/10.48550/arXiv.2304.10576

ADDITIONAL MATERIAL

ANNEX 1. DEALING WITH STUDIES OF MULTICOMPONENT PROGRAMS WITHIN EGMS

There are five approaches for handling studies that span multiple interventions (see table below). The specific approach needs to be tailored to suit a particular topic, the nature of that evidence base, and the framework and breadth of intervention categories. However, there should be consistency within and across EGMs.

The following guidelines are suggested:

1. The Principal Investigator makes the final call on the coding approach for an EGM, based on the objective of the EGM, the relevant literature, and the opinion of the substantive expert and EGM executive editors before implementing the coding protocol.

2. Consider how to approach multicomponent interventions from the outset. Aim to identify any common packages of interventions as soon as possible, while developing the framework.

3. Include a binary question of whether a study evaluates a multicomponent program in the data extraction codebook. In cases where this is not possible, teams may also consider indicating that studies include multiple components by including an asterisk or a special filter in the interactive map.

4. Justify and describe the approach within the EGM report.

Approaches to coding studies of 'multiple interventions' adopted in 3ie EGMs

OPTION	MAY WORK BEST WHEN	LIMITATIONS
1. Code a multicomponent study into a "multicomponent interventions" category only	Evidence base contains very few multicomponent interventions	May underrepresent the amount of evidence in a map. The user might be interested in a particular intervention type but unable to identify the full amount of evidence when it is not specified under the multicomponent row. A catch-all row provides very little information on those studies.
2. Create categories representing packages of interventions	The set of activities represents a common package of approaches OR the set of activities being evaluated cannot be feasibly broken apart (e.g., integrated management of childhood illnesses (IMCI)	This is only applicable to some scenarios, as there are not always standard packages on an evidence base.

OPTION	MAY WORK BEST WHEN	LIMITATIONS
3. As far as possible, code studies according to the 'principal' intervention component. Code additional components for additional descriptive analysis (not on the map)	Principal intervention components are easily identified.	It can be challenging to identify principal intervention components.
4. Code a study into multiple intervention categories only (Approach not recommended)	The evidence base contains a large number of multicomponent interventions that vary greatly.	May overrepresent the amount of evidence in the map and distort the theories of change. Does not accurately represent what was evaluated. It is difficult to understand which intervention types are paired with others.
5. Code a study into both multiple categories and into a multicomponent category row (Approach not recommended)	The evidence base contains a moderate number of multicomponent interventions.	May overrepresent the amount of evidence in the map and distort the theories of change. Does not accurately represent what was evaluated. It is difficult to understand which intervention types are paired with others. Multicomponent row loses its value as numbers increase.

ANNEX 2. INTERVENTION CATEGORIES IN THE TRANSPORTATION SFD

I. Interventions Identified by a Human (Fall 2024)

1. Transit

- Mass transit systems: metro
- Mass transit systems: BRTs
- Mass transit systems: cables
- Access to transit information

2. Roads

- Investment in highways
- Investment in rural roads
- Investment in urban roads
- Fare and funding structures: fuel taxes, parking fees, congestion charges, valorization taxes
- Vehicle restrictions
- Road maintenance programs
- Concessions and APPs

3. New Technologies

- Digital transformation in transportation (ITS)
- Electric mobility: standards, circulation, supply, environments, and regulations
- Automation technology

4. Other types of transportation

- Investment in ports, airports, and border crossings
- Investment in railways
- Maintenance programs for other infrastructure types
- Consolidation of fluvial navigation networks
- National and regional logistics plans
- Other regulations (cargo, technical, climate, etc.)
- Special loading and unloading areas
- Electric trucks, freight vehicles, and freight bicycles
- Fleet renewal
- Elimination of freight tables and other quotas
- Open skies agreements

II. Interventions Identified by an AI Assistant (Fall 2024)

Prompt: Please list the interventions included in the transportation Sectoral Framework Document and group them into around five categories.

Response: The Transportation Sector Framework Document outlines various interventions aimed at improving the transportation sector in Latin America and the Caribbean. These interventions can be grouped into five main categories:

1. Infrastructure Investments

- Construction, Rehabilitation, and Maintenance: Investments to improve transportation services through the construction, rehabilitation, and maintenance of infrastructure.
- Regional and Global Integration: Support for infrastructure that promotes regional and global integration.

2. Policy and Regulatory Frameworks

- Technical and Economic Regulations: Implementation of regulations to address market failures, such as social and environmental standards.
- Institutional and Regulatory Strengthening: Enhancing institutional and regulatory frameworks to support sector management.

3. Sustainability and Inclusivity

- Environmental and Social Sustainability: Support for the construction and maintenance of environmentally and socially sustainable infrastructure.
- Inclusive Transportation: Policies aimed at achieving greater inclusion through transportation infrastructure and services, with special emphasis on women, people with disabilities, and other population groups.

4. Technological Transformation

- Adoption of New Technologies: Encouraging the adoption of technologies in the transportation sector to improve productivity, competitiveness, safety, and sustainability.
- Regulatory Sandboxes and Pilot Projects: Development of regulatory sandboxes and pilot projects to test new technologies.

5. Multisectoral Approaches

- Comprehensive Programs: Implementation of comprehensive programs, particularly in rural and isolated locations and for small countries, to meet diverse needs.
- Synergies Among Infrastructure Sectors: Expanding multisector approaches to exploit synergies among different infrastructure sectors.

ANNEX 3. SUGGESTED EGM PROTOCOL OUTLINE

- **EGM title: Protocol**
- **Authors**
- **Date**

1. Background: Ensure that the purpose of developing the EGM is clearly understood for future updates. Some of the content in the background section of an Evidence Gap Map (EGM) protocol might be optional (as suggested in brackets after each title), particularly as the EGM will be developed primarily for internal use.

1.1. The problem, condition or issue (OPTIONAL): Provide a description of the problem, condition, or issue that this map addresses.

1.2. State of affairs and policy responses (OPTIONAL): Outline the policy or development problem addressed by the EGM. Provide an up-to-date overview of the issue, incorporating the latest findings and recent developments to contextualize the current state of affairs. A second section may include actions taken to address the issue, such as policy responses or private initiatives by research groups, NGOs, international aid organizations, and other stakeholders.

1.3. The intervention (SUGGESTED): Describe the intervention(s) of interest. Since this will also be covered in the methods section, it is not critical to include it here, but an introductory paragraph may be helpful to orient the users of the protocol.

1.4. Importance of developing this evidence gap map (CRITICAL): Clearly describe the justification for developing this EGM and who the intended users are. This section can also highlight policy relevance and potential contributions to decision-making.

2. Study objectives and questions: State the EGM's objectives and outline the research questions it aims to address. These questions should guide the mapping process and align with the policy or development problem identified.

2.1. Objectives (SUGGESTED): State the objectives of the EGM explicitly. EGMs can be undertaken for a number of reasons. For example, EGMs can be constructed to:

a. Provide a structured and accessible guide to users in a specific policy domain.
b. Identify gaps in the available evidence.
c. Identify clusters of evidence suitable for a systematic review.
d. Contribute to scoping a research program of primary studies and systematic reviews.
e. Build connections between related areas of research.

2.2. Research Questions (SUGGESTED): State the research questions that the EGM will answer, considering that EGMs can inform the extent and characteristics of the evidence rather than answering what that evidence says. Figure 1 provides examples of research questions that an EGM can address and the specific focus each conveys.

Figure 1: Examples of research questions

RESEARCH QUESTION	TYPE
1. What is the extent and what are the characteristics of empirical evidence on the effects of interventions to strengthen good governance through improved government effectiveness in L&MICs?	Coverage
2. What are the major primary and synthesis evidence gaps in the literature?	Gaps
3. What intervention/outcome areas should be prioritized for primary research and/or evidence synthesis?	Research needs

3. Methods: Describe the methods you will use for the EGM. If, during the development of the EGM you face challenges that make you deviate from what is described here, it should be noted in the final report with justification.

3.1. Framework development and scope (SUGGESTED): Describe how the framework was developed (e.g., consulted strategic plans, policy docs, stakeholders, funders, adapting existing frameworks, etc.), and describe the scope of the EGM. This context will help future users of the protocol understand the process used to determine relevant interventions and outcomes.

3.2. Conceptual framework (OPTIONAL): In an Evidence Gap Map (EGM), the conceptual framework serves as the foundation for organizing and understanding the evidence. Define the key concepts, relationships, and structure that guide the mapping process.

3.3. Criteria for including or excluding studies (PICOS) (CRITICAL): Define the PICOS framework. This is essential for determining what is includable or excludable in the review, and it will drive all future decision inclusions.

a. Population: Define the countries, groups, or individuals targeted by the interventions being studied (e.g., if the EGM only targeted studies of adolescents or only studies of L&MICs, that would be reflected here).

b. Intervention: Describe the types of interventions targeted by the EGM. This section should include a table listing the interventions considered for inclusion. Table 1 provides a template for structuring an intervention framework.

Table 1: Template for the intervention framework

DOMAIN	CATEGORY	DEFINITION	EXAMPLES
Intervention Domain 1	Category 1 of Intervention Domain 1	Definition of the category	Examples of the category. This section can also include examples of excluded cases.
	Category 2 of Intervention Domain 1	Definition of the category	Examples of the category. This section can also include examples of excluded cases.

c. Comparators: Specify what the interventions will be compared against, such as no intervention, business as usual, alternative interventions, or standard practices.

d. Outcomes: Describe the key outcome categories of interest. As with the intervention section, it is advisable to include the table of the outcomes included here. Table 2 presents a template illustrating the structure of an outcome framework.

Table 2: Template for the outcome framework

DOMAIN	CATEGORY	DESCRIPTION
Outcome Domain 1	Category 1 of Intervention Domain 1	Description of the category
	Category 2 of Intervention Domain 1	Description of the category

e. Other eligibility criteria: State any additional eligibility criteria applied to the EGM (e.g., restrictions on publication date, language restrictions, other).

3.4. Search strategy (CRITICAL): Have a clearly defined search strategy; this is critical for future updates to be aligned with the original EGM's scope and purpose.

Here you will:

· List all sources that will be searched and state any restrictions (e.g., on the time period covered by the search).

· Present the search strategy (or strategies) used for each academic database in an Appendix, including any limits and filters used, so that it can be replicated.

· If other sources are used (e.g. review registers or repositories, grey literature), report search terms used with dates of searches.

· Indicate whether special searches were conducted for different types of evidence, or if reference lists of reviews or other maps were searched.

· Note whether you will use citation tracking, manual searches of bibliographies, open calls for submissions of relevant work, and consultations with authors or experts.

· Note whether relevant ongoing studies and reviews will be included and explicitly address how you will identify them.

3.5. Screening approach (CRITICAL): Describe how studies will be reviewed and selected for inclusion in the EGM. This should reflect the PICOS and the rest of the criteria in section 4.2. Clarify whether screeners will work independently (e.g., will duplicate independent screening be implemented at the title and abstract stage? At the full-text stage?). It is highly desirable to use two independent reviewers to screen for inclusion. However, there is increasing use of automation. If automation and text-mining are used, describe in detail and with information about the validity and testing of the method. When the burden of screening is too high, other methods may be justified and should be fully described with any validated checks. Include any details on the screening software tool that will be used.

a. Handling of Duplicates: Explain how duplicate studies from different databases will be identified and removed.

b. Training of Screeners: Outline how the screening team will be trained to ensure consistency and reliability.

c. Title and Abstract Screening: Describe the process for reviewing study titles and abstracts to determine initial eligibility.

d. Full-Text Screening: Detail how full-text articles will be reviewed for final inclusion.

e. Checks for Linked Publications: Explain how related publications (e.g., working papers, preprints, final versions) will be identified and handled.

3.6. Data extraction and critical appraisal (CRITICAL): State how data will be extracted from reports of included studies, clarifying how many people will be involved (and whether independently), whether machine learning will be used (and if so, how), and how disagreements will be handled. Any use of automation and text-mining should be described in sufficient detail for replication and with information about the validity and testing of the method. List the types of information that will be sought from reports of included studies. Describe any attempts to obtain or clarify data from individuals or organizations, if applicable.

a. Data extraction training and pilot (OPTIONAL): Describe how the data extraction team will be trained. The training usually consists of sharing practice batches of about 50 studies, which are then revised by the core team and checked with the trainees during a reconciliation session. This can be repeated until the trainees achieve a minimum reliability of 85%.

b. Main-stage extraction (CRITICAL): Define whether data extraction will be conducted in duplicate. All systematic reviews should be appraised for quality using an accepted and validated tool. State the tool(s) or coding strategies that will be used to assess study quality/risk of bias/confidence for included reviews, how the tool(s) or coding strategies will be implemented, and the criteria that will be used to assign studies (e.g., how will you define low-confidence, medium-confidence, and high-confidence reviews).

3.7. Quality checks (SUGGESTED): Explain how the core team will assess the quality of the data extracted. Quality checks are particularly important when extraction is performed by a team of consultants not part of the core EGM team.

4. Analysis and reporting (OPTIONAL): Describe the fields on which the analysis will be based. For example, whether the analysis will examine the evolution of evidence over time (publication year), geographic distribution (regions, countries), interventions or outcomes (e.g., which have been more researched, which are emerging, which have stagnated, which have increased), and methods (e.g., trends in the use of RCTs versus quasi-experimental methods over time). It can include an announcement that the EGM will be an interactive tool, the website where it will be stored and if it will be shared with an external or internal public (or both).

4.1 Provide the results of the systematic reviews' appraisal: Explain how the systematic review's analysis will be conducted. Analyzing high- and medium-quality reviews is one of the most valuable tools an EGM can provide in assessing "what works."

5. Engagement and communication plan (OPTIONAL): Describe any strategy for external communication and dissemination of the work. This includes a presentation to stakeholders during and after the report is ready (for instance, a presentation of preliminary results can be done to the advisory group).

ANNEX 4. SEARCH REPOSITORIES

NAME	WEBSITE
ArXiv	https://arxiv.org/archive/econ
CAF Publications	https://www.caf.com/en/topics/r/research-for-development/publications/
Center of Economic Performance (CEPR)	https://cepr.org/publications/discussion-papers
CESifo Network	https://www.cesifo.org/en/publications/cesifo-working-papers
Cochrane Library	https://www.cochranelibrary.com/
DOAJ	https://doaj.org/
Education Resources Information Center (ERIC)	https://eric.ed.gov/
Elsevier's Science Direct	https://www.sciencedirect.com/
Google Scholar	https://scholar.google.com/
IDB Publications	https://publications.iadb.org/en
IMF Working Papers	https://www.elibrary.imf.org/view/journals/001/001-overview.xml
Institute of Labor Economics (IZA) Discussion Paper Series	https://www.iza.org/publications/dp
International Initiative for Impact Evaluation (3ie)	https://www.3ieimpact.org/evidence-hub/publications
JSTOR	https://www.jstor.org/
National Bureau of Economic Research (NBER)	https://www.nber.org/research
OAS Publications	https://www.oas.org/ext/en/security/crime-prevention-network/Resources/Digital-Library
PsycINFO	https://www.apa.org/pubs/databases/psycinfo
PubMed	https://pubmed.ncbi.nlm.nih.gov/
Project MUSE	https://muse.jhu.edu./
Redalyc	http://www.redalyc.org/home.oa
RePEc/ EconPapers	https://econpapers.repec.org/
SageJournals	https://journals.sagepub.com/
SciELO	https://www.scielo.org/
ScienceDirect	https://www.sciencedirect.com
SSRN	https://www.ssrn.com/index.cfm/en/
SpringerLink	https://link.springer.com
UNICEF Publications	https://www.unicef.org/reports
World Bank Publications	https://www.worldbank.org/en/research

ANNEX 5. EXAMPLE OF SEARCH STRINGS

This annex presents the search strings used on March 27, 2025, to search for relevant literature for the Transport EGM in EconLit. The acronyms “ti” and “ab” denote that the search was done in titles and abstracts, while the number at the end of each line represents the number of hits after that search.

1 (transport* or transit or logistics or paratransit or "active mobility").ti,ab. **35764**

2 (train? or rail or railway? or railroad? or monorail or subway* or (metro adj3 (underground or subterranean or passenger? or system? or network? or track? or line? or station?))).ti,ab. **8810**

3 (tram? or trolley? or streetcar? or "street car?" or cablecar? or "cable car?" or gondola?).ti,ab. **287**

4 (infrastructure or road* or highway? or tunnel* or hyperloop? or bridge? or sidewalk? or footpath? or pedestrian* or toll?).ti,ab. **41886**

5 ((bicycle? or bike? or cycle or cycling) adj3 (path* or trail* or lane?)).ti,ab. **179**

6 (parking or (speed adj3 limit*) or (("traffic light?" or stoplight? or "stop light?") adj3 (smart or intelligent or control* or coordinat* or phas*)) or ((congest* or traffic) adj3 (charg* or pric* or zone? or area? or manag* or monitor*)) or ((traffic or "license number" or "licence number" or "license plate" or "licence plate") adj3 (ration* or restrict*)) or "number coding scheme?" or "Unified Vehicular Volume Reduction Program" or UVVRP or "Pico y Placa" or "hoy no circula").ti,ab. **3024**

7 (bus or buses or busing or bussing or "low emission*" or "low carbon" or (fuel adj3 efficien*) or ((charg* or "electric vehicle?" or "electric car?" or EV) adj3 (station? or network?)) or (fleet? adj3 (electri* or EV)) or (electri* adj3 (car? or vehicle? or automobile?) adj3 (incentiv* or bonus* or rebate? or credit? or subsid* or "tax break?")) or ((ride* or carpool*) adj3 (app? or application? or inventiv* or regulat*)) or ((autonomous* or self-driving or connected) adj3 (vehicle? or car? or automobile?)) or ((driving or driver?) adj3 (aid? or assist*))).ti,ab. **7901**

8 (port? or harbo?r? or canal?).ti,ab. **5486**

9 ((fare? adj3 (digital* or automat*)) or "integrated ticket*").ti,ab. **31**

11 (market? or labo?r or worker? or employ* or unemploy* or "gross domestic product" or GDP or "gross national product" or GNP or econom* or trade or trading or traded or import* or export* or (flow* adj3 (goods or services)) or expend* or consume or consumption or produc* or TFP or crop? or farm* or ((land or acre? or acreage or hectare? or hectarage) adj3 (allocat* or "use" or usage)) or (population adj3 (grow* or increas* or shrink* or decreas*))).ti,ab. **1174313**

12 (poverty or poor or impoverish* or hous* or rent* or wage? or salar* or "capital acquisition").ti,ab. **218538**

13 (migra* or immigra* or emmigra*).ti,ab. **36631**

14 (((passenger? or commut* or delay* or "mode choice" or ridership or (vehicle? or car? or automobile?)) adj3 (private or own*)) or ((urban or sustainable or service? or smart) adj3 mobil*) or (travel* adj3 time) or "average speed" or micromobility or e-mobility or (connectivity adj3 "last mile")).ti,ab. **6726**

15 (fare? or price? or pricing or cost?).ti,ab. **351112**

16 (leisure or recreation* or (time adj3 (family or families)) or gender or sex or minorit* or inequalit* or inequit* or equit* or unequal or gini or (access* adj3 service?) or inclusion or noise or noisy or decibel?).ti,ab. **154520**

17 (injur* or fatal* or surviv* or collision? or crash or crashes or accident? or safe* or ((physician? or doctor?) adj3 (visit? or trip?)) or hospital* or "life expectan*" or crime or assault* or homicide? or murder? or arrest* or "police per individual").ti,ab. **67108**

18 (emission* or emit* or pollut* or "carbon footprint?" or (fuel adj3 efficien*) or (air adj3 quality) or (clean* adj3 energy) or (((pm or particulat*) adj3 concentrat*) or (sustainab* adj3 urban adj3 develop*))).ti,ab. **40715**

19 (school* or educat* or universit* or testing or (test adj3 scor*) or literacy or attain* or achiev* or graduat* or attend*).ti,ab. **239300**

20 or/11-19. **1407757**

21 ((experiment$4 adj2 (design or study or research or evaluat* or evidence or vary or varies or variation)) or ((random or randomi#ed or randomly) adj2 (trial or assign* or treatment or control* or allocat* or experiment$2 or evaluat* or vary or varies or variation or choose or chose*))).ti,ab. **17835**

22 ((match* adj2 (propensity or coarsened or covariate or neighbo?r or method or data)) or "propensity score" or ("difference* in difference*" or "difference-in-difference*" or "differences-in-difference*" or "double difference*") or (quasi-experiment$2 or "quasi experiment$2") or (estimator and evaluat*) or ("instrumental variable*" or (IV adj2 (estimation or approach))) or (Heckman adj3 (model* or approach*)) or ((two-stage or "two stage" or three-stage or "three stage" or four-stage or "four stage") adj3 (control* or function* or regression* or "least squares")) or 2SLS or "regression discontinuity" or "time series" or counterfactual or "segment* regression" or (non adj2 participant*) or ((control or comparison) adj2 (group* or condition* or area* or village* or household* or intervention)) or (panel$1 adj2 (data or household* or model*)) or ((exploit* or "tak* advantage") adj3 (variation* or variety or exogen* or heterogen*)) or (econometric adj2 (model* or adjust*)) or (select* adj2 (bias* or self))).ti,ab. **118050**

23 ((impact? or effect*) adj2 (evaluat* or assess or assessing or assessment or analyze or analyse or analyzing or analysing or analysis or analytical or estimate or estimating or estimation or cause or causal or strong* or weak* or exert*)).ti,ab. **63217**

24 ("program* evaluation" or "project evaluation" or "evaluation research" or "natural experiment*" or "program* effectiveness" or "outcome assessment" or "evaluation study" or "field experiment").ti,ab. **10301**

25 or/21-24. **184366**

26 10 and 20 and 25. **9231**

27 (L62 or L91 or L92 or L98 or R4* or o18 or h54).cc. **99666**

28 26 and 27. **4480**

ANNEX 6. EXAMPLE TITLE AND ABSTRACT SCREENING CHECKLISTS

QUESTIONS		DECISION				NOTES
1	Is the study in English, Spanish, Portuguese, or French?	NO	EXCLUDE on publication language	YES	Continue to the next question	If not sure about the language of the full text (e.g., title and abstract shown in multiple languages), continue to the next question. In such cases, this criterion should be verified during full-text screening.
2	Was the study published after 2005?	NO	EXCLUDE on publication year	YES	Continue to the next question	If the record does not indicate the year, continue to the next question. In such cases, this criterion should be verified during full-text screening.
3	Does the study evaluate an intervention listed in the EGM framework?	NO	EXCLUDE on intervention	YES	Continue to the next question	Refer to the protocol for a detailed description of includable interventions in this EGM, covering ....
4	Does the study use quantitative experimental or quasi-experimental designs? OR if a review, does it address effectiveness questions?	NO	EXCLUDE on study design	YES	Choose an option from below	Includable designs: randomized studies, matching (incl. PSM), FE (incl. DID), IV, RDD, synthetic control, interrupted time series, and other attributional methods that account for selection bias and confounding. Excluded studies: descriptive and regression studies without a clear identification of causal effects. If unclear at this point, screen at FT stage. Effectiveness systematic reviews and meta-analysis. Excluded studies: qualitative, descriptive, literature reviews and those failing to describe the methodology adequately.
5	Are you unsure about inclusion?	YES	EXCLUDE as known duplicate			Choose only one option.
6	Is the study a duplicate?	YES	INCLUDE second opinion			
7	Does the study meet all eligible criteria?	YES	INCLUDE on title and abstract			

Whenever the response to the question is UNCLEAR, continue to the next question.

ANNEX 7. EXAMPLE FULL TEXT SCREENING CHECKLIST

	QUESTIONS	DECISION				NOTES
1	Are you confident you can exclude the paper based on the abstract alone (you should always begin FT screening by reading the abstract first)	NO	MARKER - Exclude on TA	NO	Continue to the next question	This marker will probably be used mostly for incomplete or absent abstracts that got included at the title and abstract screening and once full text is retrieved, screening the complete abstract allows us to confirm or deny its relevance. However, all studies should be checked. Use the notes section to indicate the reason for exclusion.
2	The study is NOT evaluating the effectiveness of an intervention by using quantitative methods to establish a causal link between the intervention and one or more outcomes.	YES	EXCLUDE - Not a quantitative effectiveness study	NO	Continue to the next question	Use this code to exclude lab/efficacy studies, cost analysis that comes from previous studies not reporting on a new impact evaluation, process evaluations, studies only describing the design or implementation of an intervention, feasibility/acceptability studies, literature reviews, protocols/ongoing studies, and purely qualitative studies.
3	Does the study evaluate an intervention that is NOT relevant for this EGM?	YES	EXCLUDE - Not a relevant intervention	NO	Continue to the next question	Refer to the protocol for a detailed description of the included interventions. [Briefly describe includable interventions.]
4	Does the study evaluate the effects on outcomes OTHER THAN the ones included in the protocol [broad categories for outcomes of interest]?	YES	EXCLUDE - Not a relevant outcome	NO	Continue to the next question	Refer to the protocol for a detailed description of the outcomes included. [Briefly describe the includable outcomes. Define what to do with aggregate measures.]

QUESTIONS		DECISION				NOTES
5	The study lacks a control group or uses an invalid impact evaluation method to assess the causal relationship between the intervention and the outcomes.	YES	EXCLUDE - No valid causal inference	NO	Choose an option from below	Includable designs: statistical matching, FE (incl. DID), IV, RDD, synthetic control, and other attributional methods that account for selection bias and confounding. Please refer to the protocol for a whole detailed list of methodologies.
6	Is the study a duplicate?	YES	EXCLUDE as known duplicate			Choose only one option.
7	Are you unsure about inclusion?	YES	INCLUDE second opinion			
8	Does the study meet all the eligible criteria?	YES	INCLUDE on full text			

Whenever the response to the question is UNCLEAR, continue to the next question.

ANNEX 8. DATA EXTRACTION TEMPLATE

CATEGORY	VARIABLE
Study Information	Study internal ID*
	Title name*
	Language
Author Information	Author Name*
	Author Affiliation Institution
	Author Affiliation Country
Publication Information	Publication Type
	DOI
	Study status
	Abstract
	Keywords
	Journal name
	Journal volume
	Journal issue
	Pages
	Year of Publication*
	URL*
	Open access
	IDB Study
Geographic and Project Information	Continent name
	Country name
	Project name
Methodological information	Evaluation Design
	Evaluation Method
	Additional Quantitative Methods
	Unit of Observation
Intervention Information	Intervention group/arm 1 *
	Intervention group/arm 2 *
	Create different Intervention options in case there is more than one intervention group *
Outcome Information	Outcome (multiple fields to cover all relevant outcomes) *
	Outcome description *

Note: the asterisk (*) denotes those fields that are mandatory for any EGM.

ANNEX 9. CHECKLIST FOR CRITICALLY APPRAISING SYSTEMATIC REVIEWS (ADAPTED VERSION OF SURE CHECKLIST)[I]

Section A: Methods used to identify, include and critically appraise studies

A.1 Were the criteria used for deciding which studies to include in the review reported? *Did the authors specify:* *A.1.1 Types of studies* *A.1.2 Participants/ settings/ population* *A.1.3 Intervention(s)* *A.1.4 Outcome(s)* *Note.* ***If the authors do not report the search period, check the publication date of the earliest included study. If the study was published before 1990 this can be coded*** Yes.	**Yes** **Partially** **No** ***Coding guide:*** ***YES:*** ***All four should be yes*** ***NO:*** ***All four should be no*** ***PARTIALLY:*** ***Any other***
A.2 Was the search for evidence reasonably comprehensive? *Were the following done:* *A.2.1 Language bias avoided (no restriction of inclusion based on language)* *A.2.2 No restriction of inclusion based on publication status* *A 2.3 Relevant databases searched: at least one database that includes grey/unpublished literature, as well as either: (a) for health, at least two relevant comprehensive subject databases (such as PubMed/MEDLINE, EMBASE and CENTRAL), or (b) for social sciences, at least two relevant comprehensive subject databases (such as IDEAS) and one comprehensive general database (such as EconLit, PsychInfo, Scopus)* *A.2.4 Reference lists in included articles checked* *A.2.5 Authors/experts contacted* *Notes.* When authors do not mention limitations on language or publication status, code *Yes.* The use of "published" often simply means released (e.g., "studies published between 1990 – 2010") and not necessarily that studies were excluded based on publication status; do not code *No* simply because the authors use "published" in this way. When authors do not mention that reference lists were searched or experts contacted, code *No*. If authors were only contacted for study results data, code *No*. Checking reference lists of review articles does not fully meet A.2.4 requirement (code *Partially*) but is a mitigating factor. Grey literature typically means research that is not published in sources such as books or journal articles. The following databases include grey literature: Academic Search Complete (includes many conference proceedings), CAB Abstracts, searches conducted using CADATH checklist, clinicaltrials.gov, Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Library, Embase (includes 3.6m+ conference abstracts), Google, Google Scholar, Healthcare Management Information Consortium (HMIC), IDEAS/RePEc, National Technical Information Service (NTIS), OpenSIGLE/OpenGrey, PsycEXTRA, Scopus (includes ~10m conference papers). If you identify additional sources, please notify the DEP team. Searching websites of relevant governmental agencies and non-governmental organizations can also identify grey literature. Note that MEDLINE/PubMed, a comprehensive data base of journals, does not include grey literature: "For indexing in MEDLINE, NLM currently selects publications that it considers to be journals."; see also Citrome L. Beyond PubMed: Searching the "Grey Literature" for Clinical Trial Results. Innov Clin Neurosci. 2014;11(7-8):42-46. EBSCO and OVID are platforms, not databases. If an author only reports searching "EBSCO" without identifying the databases searched, code *Partially*	**Yes** **Partially** **No** **Can't tell** ***Coding guide - check the answers above:*** ***YES:*** ***All five should be yes*** ***PARTIALLY:*** ***Relevant databases and reference lists are both reported*** ***NO:*** ***Any other***

A.3 Does the review cover an appropriate time period?

Is the search period comprehensive enough that relevant literature is unlikely to be omitted?

Note. If the authors do not report the search period, check the publication date of the earliest included study. If the study was published before 1990 this can be coded *Yes.*

Yes
Can't tell (only use if no information about time period for search)
No
Unsure

Coding guide:

YES: Generally, this means searching the literature at least back to 1990
NO: Generally, if the search does not go back to 1990
CAN'T TELL: No information about time period for search

Note: With reference to the above – there may be important reasons for adopting different dates for the search, e.g. depending on the intervention. If you think there are limitations with the timeframe adopted for the search which have not been noted and justified by the authors, you should code this item as a NO and specify your reason for doing so in the comment box below. Older reviews should not be downgraded, but the fact that the search was conducted some time ago should be noted in the quality assessment. Always report the time period for the search in the comment box.

A.4 Was bias in the selection of articles avoided?

Did the authors specify:

A.4.1 Independent screening of **full text** by at least 2 reviewers
A.4.2 List of included studies provided
A.4.3 List of excluded studies provided

Notes. For A.4.1, independent screening means that both screeners screened all full text without knowing what the other screener decided (that is, one screener and one verifier does not meet criterion). If the authors note two screeners and do not use the word "independent" but mention a third reconciler to resolve differences, assume independence. Other acceptable methods include (a) the use of machine learning approaches (e.g., priority classifiers), provided a portion of machine excluded studies are checked or (b) double screening until an acceptable level of reliability (at least .85) is reached, with a percentage of subsequent coding being checked to protect against coder drift. If authors report double screening a small portion of studies, but do not report their inter-rater reliability, code No. When authors do not mention whether independent screening was conducted by at least two reviewers, code *No*. **Single screening at title and abstract is acceptable.**

Yes
Partially
No

Coding guide:

YES: All three should be yes, although reviews published in journals are unlikely to have a list of excluded studies (due to limits on word count) and the review should not be penalized for this.

The list of excluded studies does not need to include studies whose abstracts were screened out as ineligible. Because journals often have word count limits, reviews published in journals do not need to have a list of excluded studies and are coded *Not Applicable.*	*PARTIALLY:* ***Independent screening and list of included studies provided are both reported*** *NO:* ***All other. If a list of included studies is provided, but the authors do not report whether or not the screening has been done by 2 reviewers, then this section is downgraded to NO.***
A.5 Did the authors use appropriate criteria to assess the quality and risk of bias in analysing the studies that are included?[ii] Were the following done: *A.5.1 The criteria used for assessing the quality/ risk of bias were reported* *A.5.2 A table or summary of the assessment of each included study for each criterion was reported* *A.5.3 Sensible criteria were used that focus on the quality/ risk of bias (and not other qualities of the studies, such as precision or applicability/external validity). "Sensible" is defined as a recognized quality appraisal tool/ checklist, or similar tool which comprehensively assesses bias (internal validity) in included studies Please see footnotes for details of the main types of bias such a tool should assess.* *Notes.* Identified tools with sensible criteria include: Academy of Nutrition and Dietetics Quality Criteria Checklist, Cochrane Handbook, The Delphi List, Effective Public Health Practice Project (EPHPP) Quality Assessment Tool, Guide to Community Preventative Services Study Quality tool, Joanna Briggs Institute Checklists for RCT/QED, National Institutes of Health's Quality Assessment Tool for Controlled Intervention Studies (sometimes labelled NHLBI tool). Child Health Epidemiology Reference Group (CHERG) study design & quality standards, Grading of Recommendations Assessment, Development and Evaluation (GRADE) RoB criteria, (CHERG and GRADE provide a set of guidelines for synthesizing evidence from multiple impacts on an outcome. As part of these multi-step processes, RoB is assessed, but other dimensions are also assessed (such as consistency of results across all studies). For A5.3, what needs to be reported is the individual ratings for each study on design/quality standards (CHERG) or risk of bias (GRADE). For case-control studies and cohort studies, the Newcastle-Ottawa Scale uses sensible criteria that are focused on risk of bias as does Methodological Index for Non-Randomized Studies (MINORS). Note that these designs typically are not as rigorous as RCTs or even QEDs.	**Yes** **Partially** **No** **Not Applicable** (to be used only if there were no eligible quantitative studies) ***Coding guide:*** *YES:* ***All three should be yes*** *PARTIALLY:* ***The first and third criteria should be reported. If the authors report the criteria for assessing risk of bias and report a summary of this assessment for each criterion, but the criteria may be only partially sensible (e.g. do not address all possible risks of bias, but do address some), we downgrade to PARTIALLY.*** *NO:* ***Any other***
A.6 Overall – how much confidence do you have in the methods used to identify, include and critically appraise studies? *Use the guidance below to determine the overall score for section A, based on your answers to each of the questions in this section.* ***High confidence*** *applicable when the answers to the questions in section A are all assessed as 'yes'* ***Low confidence*** *applicable when any of the following are assessed as 'NO' above: not reporting explicit selection criteria (A1), not conducting reasonably comprehensive search (A2), not avoiding bias in selection of articles (A4), not assessing the risk of bias in included studies (A5)* ***Medium confidence*** *applicable for any other – i.e. section A3 is assessed as 'NO' or can't tell and remaining sections are assessed as 'partially' or 'can't tell'*	**Low confidence** (limitations are important enough that the results of the review are not reliable) **Medium confidence** (limitations are important enough that it would be worthwhile to search for another systematic review and to interpret the results of this review cautiously if a better review cannot be found) **High confidence** (only minor limitations)

Section B: Methods used to analyse the findings

B.1 Were the characteristics and results of the included studies reliably reported? *Was there:* *B.1.1a Independent data extraction by at least 2 reviewers* *B.1.1b Independent risk of bias assessment by at least 2 reviewers* *B.1.2 A table or summary of the characteristics of the participants, interventions and outcomes for each included study.* *B.1.3 A table or summary of the results of all the included studies* *A.2.5 Authors/experts contacted* *Notes.* Independent extraction means that both extractors extracted all data without knowing what the other extractor decided (that is, one extractor and one verifier does not meet criterion). If the authors note two reviewers and do not use the word "independent" but mention a third reconciler to resolve differences, assume independence. When authors do not mention whether independent extraction was conducted by at least two reviewers, code No. Forest plots are an appropriate summary of the results, as is reporting that summarizes the findings by outcome domain.	**Yes** **No** **Partially** **Not applicable** (e.g. no included studies) ***Coding guide:*** ***YES:* All three should be yes** ***PARTIALLY:* Criteria B.1.1 and B.1.3 are yes, but some information is lacking on B.1.2.** ***No:* None of these are reported. If the review does not report whether data was independently extracted by 2 reviewers (possibly a reporting error), we downgrade to NO.** ***NOT APPLICABLE:* if no studies/no data**
B.2 Are the methods used by the review authors to analyse the findings of the included studies clear, including methods for calculating effect sizes if applicable? *Notes.* An example of acceptable reporting: "fixed effects meta-analysis, with standardized mean differences for continuous outcomes and response ratios for dichotomous outcomes"	**Yes** **Partially** **No** **Not applicable** (e.g. no studies or no data) ***Coding guide:*** ***YES:* Methods used clearly reported. If it is clear that the authors use narrative synthesis, they don't need to say this explicitly.** ***PARTIALLY:* Some reporting on methods but lack of clarity** ***NO:* Nothing reported on methods**
B.3 Did the review describe the extent of heterogeneity? *B.3.1 Did the review ensure that included studies were similar enough that it made sense to combine them, sensibly divide the included studies into homogeneous groups, or sensibly conclude that it did not make sense to combine or group the included studies?* *B.3.2 Did the review discuss the extent to which there were important differences in the results of the included studies? (Note, This item is not about which specific factors might explain differences in the results - that is covered in section B6)*	**Yes** **Partially** **No** **Not applicable** (e.g. no studies or no data) ***Coding guide:*** ***PARTIALLY:* B.3.1 is yes** ***NO:* Any other** ***NOT APPLICABLE:* if no studies/no data**

B.3.3 If a meta-analysis was done, was the I^2, chi square test for heterogeneity or other appropriate statistic reported? If no statistical test was reported, is a qualitative justification made for the use of random effects?

Notes. Code B.3.1 *No* if analyses includes studies with implausibly different interventions, comparisons, or populations. If a narrative analysis, the authors need to have a rationale for why studies were not combined (such as interventions were too different) or Code B.3.1 as *No*. For meta-analyses, reporting a metric for heterogeneity is sufficient for B.3.2. For non-meta-analysis, mentioning heterogeneity in results is enough (for example, The impacts varied from X to Y or Study A found X and Study B found Y).

B.4 Were the findings of the relevant studies combined (or not combined) appropriately relative to the primary question the review addresses and the available data?

B.4.1 How was the data analysis done?

Descriptive only
Vote counting based on direction of effect
Vote counting based on statistical significance
Description of range of effect sizes
Random effects meta-analysis
Fixed effects meta-analysis
Meta-regression
Bayesian approaches
Network meta-analyses (NMA)
Other: specify
Not applicable (e.g. no studies or no data)

B.4.2 How were the studies weighted in the analysis?
Equal weights (this is what is done when vote counting is used)
By quality or study design (this is rarely done)
Inverse variance (this is what is typically done in a meta-analysis)
Number of participants (sample size – this was standard practice in early meta-analyses)
Other: specify
Not clear
Not applicable (e.g. no studies or no data)

B.4.3 Did the review address unit of analysis errors?
Yes - took clustering into account in the analysis (e.g. used intra-cluster correlation coefficient)
No, but acknowledged problem of unit of analysis errors
No mention of issue
Not applicable - no clustered trials or studies included

Note on B.4.1 There should be a clear justification if fixed effects meta-analysis is used. A fixed effects model assumes one true effect size, and that the only differences are due to sampling error. This is highly unlikely in international development due to large variations in context, participants, implementation, etc., thus a random effects model is typically most appropriate when meta-analysis is used.

For network meta-analysis (NMA), the review must (1) discuss why NMA is appropriate, (2) present a network diagram where the thickness of the lines reflects the number of studies for each direct effect, (3) provide information on inconsistency factors and global test for inconsistency, (4) provide a table with the relative effect between each pair of interventions, and (5) provide a ranking of interventions using rankograms and cumulative ranking plots. Authors should interpret these graphs carefully if inconsistency in the network is detected.

Bayesian approaches can be used for both meta-analysis and NMA. Prior distributions are needed for the particular intervention being analyzed.

Yes
Partially
No
Not applicable (e.g. no studies or no data)
Can't tell

Coding guide:

YES: ***If appropriate table, graph or meta-analysis (or descriptive where meta-analysis not possible and authors report magnitude of effects for all included studies) AND appropriate weights AND unit of analysis errors addressed (if appropriate).***

PARTIALLY: ***If appropriate table, graph or meta-analysis AND appropriate weights AND unit of analysis errors not mentioned or not addressed (and should have been).***

NO: ***If descriptive OR vote counting (where quantitative analyses would have been possible) OR inappropriate reporting of table, graph or meta-analyses.***

NOT APPLICABLE: ***if no studies/no data***

CAN'T TELL: ***if unsure (note reasons in comments below)***

Note on B.4.3: Unit of analysis issues arise when the unit assigned is a cluster, such as a school, but the units analyzed are individual people, such as students. If the analysis does not account for this clustering, the standard errors will be too large and accordingly the estimated statistical significance will be too small. Studies can account for the clustering using an appropriate hierarchical linear model or a random effects econometric model (note that random effects meta-analysis does not fix this problem, which exists at the study level). A systematic review can address these errors by requiring that the study use the correct analysis or by adjusting results using an intra-class correlation (typically the ICC is given a default value)

B. 5 Does the review report evidence appropriately?

B.5.1 The review makes clear which evidence is subject to low risk of bias in assessing causality (attribution of outcomes to intervention), and which is likely to be biased, and does so appropriately

B.5.2 Where studies of differing risk of bias are included, results are reported and analyzed separately by risk of bias status

Note: Making clear which evidence is subject to low risk of bias can be accomplished in a table listing RoB for each study or by listing RoB for each study on each RoB criterion; that is, if A5.2 is Yes, then B5.1 is *Yes* (but the reverse is not true). Reporting only study design is not sufficient to meet B5.1. For B5.2, narrative analysis must group or report by RoB, it is not sufficient to simply report RoB of each study. If the SR does not use sensible criteria to assess RoB, then B5.1 is *No*.

Note on reporting evidence and risk of bias: For reviews of effects of 'large n' interventions, experimental and quasi-experimental designs should be included (if available). For reviews of effects of 'small n' interventions, designs appropriate to attribute changes to the intervention should be included (e.g. pre-post with assessment of confounders).

For B.5.1, This item examines whether the SR clearly identifies which studies have low/high RoB, so that the reader understands the strength of evidence supporting each impact (the reporting can be for individual studies or an outcome domain). This differs from A5.2 (which examines the reporting of RoB at the criterion level) and B5.2 (which requires overall analysis/reporting by RoB). An overall GRADE quality of evidence rating cannot be used to meet this requirement because the GRADE rating is based on RoB but also additional factors such as consistency of results, indirectness of evidence, imprecision, and reporting bias. However, if the SR reports the RoB dimension separately (typically labeled "study limitations" or "risk of bias") for each outcome domain, that fulfills this criterion. For similar reasons, the overall CHERG quality assessment does not fulfill this requirement.

Item B.5.2 applies only when there are low risk of bias studies included in analyses. If all studies in an analysis are deemed some concerns or high risk of bias, this point is not applicable.

Yes
Partially
No
Not applicable

Coding guide:

YES: ***Both criteria should be fulfilled (where applicable)***

NO: ***Criteria not fulfilled***

PARTIALLY: ***Only one criterion fulfilled, or when there is limited reporting of quality appraisal (the latter applies only when inclusion criteria for study design are appropriate)***

NOT APPLICABLE: ***No included studies***

B.6 Did the review examine the extent to which specific factors might explain differences in the results of the included studies?

B.6.1 Were factors that the review authors considered as likely explanatory factors clearly described?

B.6.2 Was a sensible method used to explore the extent to which key factors explained heterogeneity?

Descriptive/textual
Graphical
Meta-analysis by sub-groups
Meta-regression
Other

Yes
Partially
No
Not applicable

Coding guide:

YES: ***Explanatory factors clearly described and appropriate methods used to explore heterogeneity***

	PARTIALLY: ***Explanatory factors described but for meta-analyses, sub-group analysis or meta-regression not reported (when they should have been)*** ***NO:*** ***No description or analysis of likely explanatory factors*** ***NOT APPLICABLE:*** ***e.g. too few studies, no important differences in the results of the included studies, or the included studies were so dissimilar that it would not make sense to explore heterogeneity of the results***
B.7 Overall - how much confidence do you have in the methods used to analyse the findings relative to the primary question addressed in the review? *Use the guidance below to determine the overall score for section B, based on your answers to each of the questions in this section.* ***High confidence*** *applicable when all the answers to the questions in section B are assessed as 'yes'.* ***Low confidence*** *applicable when any of the following are assessed as 'NO' above: critical characteristics of the included studies not reported (B1), not describing the extent of heterogeneity (B3), combining results inappropriately (B4), reporting evidence inappropriately (B5).* ***Medium confidence*** *applicable for any other: i.e. the "Partial" option is used for any of the 6 preceding questions and/or B.2 and/ or B.6 are assessed as 'no'.*	**Low confidence** (limitations are important enough that the results of the review are not reliable) **Medium confidence** (limitations are important enough that it would be worthwhile to search for another systematic review and to interpret the results of this review cautiously if a better review cannot be found) **High confidence** (only minor limitations)

Section C: Overall assessment of the reliability of the review

C.1 Are there any other aspects of the review not mentioned before which lead you to question the results?	Additional methodological concerns (e.g., reviews by a single author) Robustness Interpretation Conflicts of interest (of the review authors or for included studies) – note issues in comment section Other No other quality issues identified
C.2 Are there any mitigating factors which should be taken into account in determining the reviews reliability?	Limitations acknowledged (note, this is not a sufficient reason to upgrade a score, but should be noted in the assessment summary if limitation are acknowledged) Strong policy conclusions drawn (including in abstract/ summary) in the absence of high-quality evidence Any other factors Note. A low confidence review cannot be upgraded by simply acknowledging the limitations.

C.3 **Based on the above assessments of the methods how would you rate the reliability of the review?**

Coding guide:

High confidence in conclusions about effects: high confidence noted overall for sections A and B, unless moderated by answer to C1

Medium confidence in conclusions about effects: medium confidence noted overall for both sections A and B or that you have assessed medium for A or B and high for the other section.

Low confidence in conclusions about effects: low confidence noted overall for sections A or B, unless moderated by answer to C1 or C2. For example, if there is only one reason A or B is low confidence and there is a relevant mitigating factor that makes that reason less problematic, this can be assessed as Medium Confidence (e.g., the screening/extraction was not independent (leads to low) but two people screened/extracted all studies (for example, one checked the other and they report an acceptable level of reliability)).

Notes. There are two cases where an SR can receive High Confidence even though was assessed Medium Confidence on Section A and the only reason for Medium is because (1) authors were not contacted to identify additional studies; however, the literature search involved multiple website searches, which serves an equivalent function, and (2) authors did not cross-checked references in all included studies; however, the authors did crosscheck all references in other review articles (at least two), which serves an equivalent function.

[i] Adapted from Supporting the Use of Research Evidence (SURE) Collaboration. SURE checklist for making judgements about how much confidence to place in a systematic review. In: SURE guides for preparing and using policy briefs. www.evipnet.org/sure