Building Stronger Transport Policy: An Evidence Gap Map

Camilo Acosta
Inter-American Development Bank (IDB)

Leonel Borja
IDB

Indira Porto Gutierrez
IDB

María Daniela Anda León
International Initiative for Impact Evaluation (3ie)

Diana Belén Córdova-Aráuz
3ie

Shannon Shisler
3ie

Evidence Gap Map Protocol

April 2025

About IDB

The Inter-American Development Bank (IDB) is the leading source of development financing for Latin America and the Caribbean. Since 1959, it has been dedicated to improving lives across the region by providing financial and technical support to governments and partners. Together with IDB Invest and IDB Lab, the IDB Group promotes sustainable and inclusive growth by funding projects, generating cutting-edge research, and developing innovative solutions to address the region's most pressing challenges.

About 3ie

The International Initiative for Impact Evaluation (3ie) develops evidence on how to effectively transform the lives of the poor in low- and middle-income countries. Established in 2008, we offer comprehensive support and a diversity of approaches to achieve development goals by producing, synthesizing and promoting the uptake of impact evaluation evidence. We work closely with governments, foundations, NGOs, development institutions and research organizations to address their decision-making needs. With offices in Washington DC, New Delhi and London and a global network of leading researchers, we offer deep expertise across our extensive menu of evaluation services.

Evidence gap maps

An evidence gap map (EGM) is a thematic collection of information about impact evaluations or systematic reviews that measure the effects of international development policies and programs. The EGMs provide a visual display of completed and ongoing systematic reviews and impact evaluations in a sector or sub-sector, structured around a framework of interventions and outcomes.

About this EGM protocol

This protocol provides all the supporting documentation for the production of the EGM, including thematic background information, and details of the methods that will be applied to systematically search and screen the evidence base, as well as extract data from included studies.

The content of this report is the sole responsibility of the authors and does not represent the opinions of the IDB, 3ie, their donors, their respective Boards, or the countries IDB and its Board of Directors represent. Any errors and omissions are also the sole responsibility of the authors. Please direct any comments or queries to the corresponding author, Camilo Acosta (camiloac@iadb.org).

Suggested citation: Acosta, C., Porto, I., Borja, L., Anda-León, M.D., Córdova-Aráuz, D., & Shisler, S. (2025). Protocol: Building Stronger Transport Policy, An Evidence Gap Map.

Table of Contents

1. Background

1.1. The problem, condition, or issue

The IDB Group Institutional Strategy 2024-2030 highlights the importance of consolidating the institution as the knowledge bank of Latin America and the Caribbean by producing cutting-edge research on development issues in the region and serving as a bridge to development research from the rest of the world that could be useful to its members. For this issue, improving the flow of feedback between operations and knowledge is of utmost importance by providing easy access to empirical evidence, insights from past projects, and tools to generate new knowledge from operations.

The Transport Division at the IDB executes operations across all 26 borrowing members of the IDB Group, and its portfolio is among the largest at the institution. To ensure impact, it is vital that the operations in this category are formulated around a solid theory of change based on the best available knowledge, including frontier empirical evidence.

1.2. State of affairs and policy responses

The Transport Division currently has a team of economists dedicated to supporting its operational portfolio by providing empirical evidence to inform its operations and by participating in the respective impact evaluations. To provide frontier evidence, the knowledge team from the Transport Division hosts a database that includes state-of-the-art research with its relevant metadata. The current database contains around 200 articles across 8 intervention categories (institutional capacity, rural roads, roads and tunnels, diversity, logistics, urban mobility, road security, and technology) and 10 outcome categories (transportation, pollution, employment, health, education, productivity, housing prices, poverty, citizen security, and others).

1.3. Importance of developing this evidence gap map

Given the institutional mandate of promoting evidence-based decision making and leveraging the current efforts of the Transport Division's knowledge team, 3ie and the IDB signed an agreement in 2024 to co-produce an evidence gap map for this division with the support of the IDB's Knowledge and Learning Division.

This map is intended mainly for sectoral specialists and teams working in operations and knowledge-oriented tasks in headquarters and country offices. However, everyone inside the IDB Group will have access to this resource. In particular, it could also be helpful for country economists and their teams, development effectiveness specialists, and research teams, among others.

This EGM will focus on a similar type of interventions to those already captured in the existing database but grouped into three macro categories: urban mobility infrastructure; roads, regional transportation, and logistics; and law, regulations, and policy. Also, outcome variables will be grouped into five categories: access, quality, affordability, service management, and socioeconomic results. An important remark is that the first four categories

are highly aligned with the existing Indicators' Catalog for this division.

This map intends to complement other existing synthesis efforts, such as the *Evidence and gap map-studies of the effectiveness of transport sector intervention in low and middle-income countries* produced by the Centre of Excellence for Development Impact and Learning (Malhotra et al, 2021).[1] Even though this EGM is a valuable resource, the map produced under this protocol has a broader scope in terms of interventions. Moreover, the interventions and outcomes set in this map are highly aligned with the IDB's portfolio of transport operations and with the indicators being used to measure their effectiveness and impact.

2. Study objectives and questions

2.1. Objectives

Provide state-of-the-art empirical evidence to support the operations of the IDB's Transportation Division and its knowledge agenda, promoting strategic selectivity and strengthening impact.

2.2. Research Questions

- What are the primary causal evidence concentrations and gaps in the transportation literature focusing on the construction of roads, ports, and urban mobility infrastructure, and the policies surrounding them?

- What are the main characteristics of the empirical evidence on the effects of interventions in the transportation sector? In particular, how is this evidence distributed geographically and over time, and what study designs and methods have been used to evaluate the effectiveness of these interventions?

3. Methods

3.1. Framework development and scope

The Intervention-Outcome (I-O) framework from the Transportation EGM leverages existing knowledge from the Transport Sectoral Framework Document (SFD), the historical portfolio of IDB loan operations, and the SPD Indicator Catalog. Aligning with these sources is essential to facilitate navigation within the EGM, enhance usability, and ensure consistency with the IDB's taxonomy and operations. Definitions and more details of each source are explained in the methodological note, but some relevant details are presented here.

As a first step, the knowledge team from the Transportation division provided a list of intervention and outcome categories based on the division's current literature synthesis efforts. This categorization was

[1] This map is available at https://cedilprogramme.org/publications/evidence-and-gap-map-transport-sector-intervention

complemented using the Transport SFD. SFDs are knowledge documents that should provide a synthesis of the main development challenges and the best and most relevant evidence on a topic. These documents should summarize which interventions work in which particular contexts and identify knowledge gaps to guide future research efforts. All the SFDs are available at https://www.iadb.org/en/who-we-are/about-idb/operational-policies. In the case of the Transport SFD, after a complete read, the relevant interventions and outcomes were manually identified and included in the document.

To further improve the intervention framework, the IDB Knowledge team revised the approved loan proposals from the IDB Transportation division in the last 15 years to identify potential intervention categories that were not in the initial mapping. For example, the category "bridges" was added in this step. To give more structure to the outcome's framework, the teams leverage the recently released indicators' catalogue. This catalogue is a tool developed by the Office of Strategic Planning and Development Effectiveness (SPD) to standardize, facilitate, and enhance the IDB's measurement of project impact. Aligning outcome categories with the catalogue is vital to ensure the usability and consistency of different tools.

Throughout the process, the framework was constantly validated, revised, and approved by the Transportation Division's knowledge coordinator and other team members.

3.2. Criteria for including or excluding studies (PICOS)

3.2.1. Population

The EGM will cover all countries (low-income, lower-middle-income, upper-middle-income and high-income). Although this is a broad population focus, finding causal evidence on transportation interventions is challenging due to the difficulties in identifying interventions with experimental or quasi-experimental variation that enable a causal evaluation of a policy or intervention. However, a filter will be available to allow users to select studies focusing on countries from Latin America and the Caribbean, as well as other regions.

3.2.2. Interventions

The Transportation EGM covers interventions grouped into three main categories: (a) urban mobility infrastructure, (b) roads, regional transportation, and logistics, and (c) laws, regulations, and policies. Table 1 outlines the definitions and provides examples of interventions under each category. Given the current large scope of the map, interventions related to road safety and airports were excluded from this version. Their potential inclusion will be assessed in future updates of the EGM by the Transport Division team.

Table 1: Intervention framework

Category	Intervention	Useful definitions	Detailed examples	Specific examples
Urban mobility infrastructure	Bus Rapid Transit	A Bus Rapid Transit (BRT) system is a bus-based public transport solution. Key characteristics include dedicated bus lanes separated from general traffic, off-board fare collection to expedite boarding, level boarding platforms for easy and accessible entry, traffic signal priority to reduce delays at intersections, and enhanced stations equipped that may include amenities like real-time information displays and shelters.	Construction of dedicated bus lanes, building level boarding platforms at stations; implementing off-board fare collection systems; installing traffic signal priority infrastructure; constructing depots and maintenance facilities; upgrading stations with real-time information displays and shelters; reinforcing or resurfacing dedicated lanes; integrating BRT with other transit modes; conducting regular inspections and repairs of infrastructure; maintaining BRT vehicles; updating fare collection and signal systems; improving service frequency; and displaying estimated wait times at stations. Installation of lighting and CCTV to enhance passenger security. Installation of air conditioning systems.	Transmilenio (Bogotá), Rede Integrada de Transporte (Curitiba), Metroplus (Medellín),
	Urban Trains (subways, passenger trains, LRT, metros)	Urban trains include rail-based public transport systems such as subways, light rail transit (LRT), passenger trains, and metros. Key characteristics include dedicated rail tracks separated from road traffic, (usually) electric-powered train sets, high-capacity service, and fixed stations or terminals often integrated with other modes of transport	Construction of an entirely new urban train system; construction of underground or elevated rail lines; building or renovating train stations with level platforms and accessibility features; electrification of lines; installation of signaling and control systems; procurement of electric train cars; upgrading station facilities with real-time information systems; installing fare collection infrastructure; building train depots and maintenance facilities; retrofitting existing lines for higher capacity or automation; track and system maintenance; integration with bus or BRT terminals for multimodal connectivity; and upgrading trains or stations with features such as air conditioning, improved lighting, and noise reduction systems. Installation of platform screen doors, intrusion prevention systems, lighting and CCTV to enhance passenger security	Line 2 of the Panama Metro, Line 3 of the Guadalajara Light Rail, Metro de Bogotá

Category	Intervention	Useful definitions	Detailed examples	Specific examples
	Cables	Cables refer to aerial cable car or gondola lift systems used for urban or rural public transport, especially in areas with difficult terrain. They typically operate with cabins suspended from a continuously moving cable, with fixed stations for boarding and alighting.	Construction of a new cable or gondola lift system; construction or extension of cable lines; installation of towers, cables, and cabins; building or renovating boarding stations with accessibility features; installation of propulsion and control systems; construction of maintenance and storage facilities; upgrading stations with real-time information and fare collection systems; integration with other modes of public transport; reinforcing structural components; and retrofitting cabins or stations with features such as air conditioning or improved accessibility.	Metrocable (Medellín), Mi Teleférico (La Paz), Cablebús (Mexico, D.F)
	Active Mobility (pedestrian, bicycles, etc.)	Active mobility refers to human-powered modes of transportation such as walking, cycling, and micro-mobility (e.g., scooters). These modes depend on dedicated infrastructure that ensures safety, accessibility, and integration with other transport systems.	Construction of new sidewalks, pedestrian paths, and bike lanes; renovation or widening of existing sidewalks and cycling infrastructure; installation of protected bike lanes and pedestrian crossings; construction of pedestrian bridges or tunnels; development of bike parking and storage facilities; creation of public bike-sharing stations; implementation of traffic calming measures such as speed bumps, raised crossings, and curb extensions, raised crossings, bollards, traffic-calming zones, pedestrian refuge islands, visibility enhancements (e.g., reflective paint, lighting), and buffer-protected bike lanes; installation of signage and wayfinding systems; improvement of lighting and drainage on active mobility corridors; and integration of pedestrian and cycling routes with public transport stations.	Tembici Urban Electric Bike-Sharing, Pedestrianization of downtown streets

Category	Intervention	Useful definitions	Detailed examples	Specific examples
	Urban roads	Urban roads refer to vehicular roadways within cities, including streets, avenues, boulevards, and parkways, primarily intended for motorized transport. This category includes infrastructure for the construction, maintenance, and physical improvement of roadways and their characteristics, excluding standalone interventions targeting active mobility.	Construction of new roadways or expansions of existing roads; resurfacing or repaving of urban streets and avenues; road widening and lane reconfiguration (adding turning lanes, reversible lanes, or bus/HOV lanes); construction of grade-separated or at-grade intersections, roundabouts, or flyovers; installation or upgrade of street lighting (including smart LED systems) and storm-water drainage; deployment of adaptive traffic-signal control, variable-message signs, and electronic speed-feedback displays; construction of dedicated toll lanes and electronic toll-collection infrastructure; development of retaining walls, medians, guardrails, sound or wind barriers, and other roadside safety barriers; installation of automated enforcement devices such as speed cameras and red-light cameras; implementation of traffic-calming works such as speed humps, raised tables, chicanes, curb extensions, and pedestrian refuge islands; construction of tunnels or elevated road segments for vehicular traffic; roadbed reinforcement, slope stabilization, or other structural rehabilitation for climate resilience; installation of additional road-safety elements such as crash cushions, rumble strips, high-visibility lane markings, and vehicle containment systems; and deployment of CCTV and incident-detection sensors for real-time traffic monitoring.	Anillo vial periférico (Lima), Buenos Aires to Puerto Madero connection.

Category	Intervention	Useful definitions	Detailed examples	Specific examples
	Low Emission Mobility: infrastructure	Low-emission mobility infrastructure refers to public investments in infrastructure or equipment that support the transition to cleaner, lower-emission transport systems. This includes systems for electric, hybrid, or other non-fossil-fuel-based vehicles. Public-sector vehicle purchases can be included when they are part of a broader low-emission transport intervention. Private individual vehicle purchases are excluded.	Installation of electric vehicle charging stations in public or fleet areas; construction of electric or hydrogen bus depots with fueling and maintenance infrastructure; deployment of low-emission zones with enforcement systems; installation of charging or fueling docks for electric or hydrogen ferries and boats; development of infrastructure for electric bike or scooter charging; grid or fuel supply upgrades to support transport electrification or hydrogen distribution; procurement of electric, hydrogen, or CNG buses, municipal service vehicles, or ferries as part of fleet electrification or conversion; construction of fueling stations for hydrogen or CNG vehicles in public fleets; and deployment of solar-powered or renewable energy-based charging infrastructure for light electric vehicles.	E-Mobility Program for Sustainable Cities, Hydrogen-Powered Trains in
	Urban traffic management systems	Urban traffic management systems refer to technologies and systems implemented to regulate, monitor, or optimize vehicular traffic flow within urban areas. These systems aim to improve traffic efficiency, reduce congestion, and enhance road safety. This category excludes physical road construction unless directly tied to traffic control systems.	Installation of smart traffic lights with adaptive signal control; implementation of electronic toll collection systems; deployment of traffic cameras for monitoring or enforcement; installation of variable message signs and real-time traffic information displays; development or upgrade of traffic control centers operated by transit or road authorities; deployment of vehicle detection sensors and automatic incident detection systems; integration of traffic data platforms for real-time monitoring and decision-making; implementation of geolocation-based or AI-driven traffic optimization technologies; pedestrian crossing sensors, and integration of accident-prone zone alerts in control centers.	Buenos Aires Sustainable Mobility Plan, Traffic Management plans (e.g., Mexico, Perth).
Roads, Regional	Rural roads	Rural roads refer to roadways located in non-urban areas, typically serving low-density populations and	Construction of new tertiary or feeder roads in rural areas; upgrading rural tracks to all-weather roads; rehabilitation or resurfacing of existing rural	Rural road improvement pro-

Category	Intervention	Useful definitions	Detailed examples	Specific examples
Transportation and Logistics		connecting communities to markets, services, or main transport corridors. This category includes the construction, rehabilitation, or maintenance of rural roads, with a focus on tertiary or feeder roads. It excludes urban roads or major highways.	roads; road stabilization works (e.g., drainage, culverts, slope protection); maintenance of unpaved or gravel roads; construction of rural bridges or fords as part of rural road networks.	grams in El Salvador, Ecuador or Colombia
	Highways and national roads	Highways and national roads refer to major interurban or regional roadways that form part of a country's primary transport network. This category includes the construction, expansion, rehabilitation, and maintenance of highways, freeways, and other national or trunk roads. It excludes urban roads, rural tertiary roads, and standalone bridge or tunnel projects.	Construction of new highways or national roads; expansion or widening of existing corridors; resurfacing or rehabilitation of major roadways; installation of median barriers, guardrails, rumble strips, and high-visibility lane markings; deployment of automated speed-management equipment (speed cameras, red-light cameras, speed-detection zones); slope-stabilization and rockfall-protection works; addition of emergency stopping bays and arrester beds; variable speed-limit systems and dynamic signage; road-weather monitoring systems; climate-resilience measures such as elevating roadbeds, using permeable pavements, and reinforcing embankments; construction of grade-separated crossings for pedestrians or wildlife; and development of dedicated rest areas and service plazas.	East-West Link (Suriname), Ruta del Sol (Colombia), San José-Caldera (Costa Rica)
	Railways and Intermunicipal Trains	Railways and intermunicipal trains refer to train systems that operate over long distances, connecting cities, towns, and regions. These systems are used to move either people (passenger trains) or goods (cargo trains) between municipalities. This category includes the construction, maintenance, or improvement of the	Construction of new railway lines for passengers or cargo; rehabilitation or upgrading of existing tracks; installation of signaling and communication systems; electrification of rail corridors; building or renovating intercity train stations and maintenance depots; improvements to rail safety infrastructure such as fencing and level crossing removals; and preparation of technical studies to	Central Railroad Project (Uruguay), São Paulo Regional Rail Project, Tren Maya (Mexico)

Category	Intervention	Useful definitions	Detailed examples	Specific examples
		infrastructure needed for these services to operate, such as tracks, stations, and control systems. It does not include urban trains like subways or metros.	assess feasibility, costs, and risks of railway investments.	
	Tunnels	Tunnels are underground or underwater passages constructed to enable road vehicles or pedestrians to pass through obstacles such as mountains, cities, or bodies of water. This category includes the construction, maintenance, or improvement of road and pedestrian tunnels, along with associated infrastructure. It excludes rail tunnels and aboveground road works.	Construction of new roads or pedestrian tunnels; rehabilitation or resurfacing of tunnel interiors; installation or upgrade of tunnel lighting, ventilation, and fire safety systems; drainage and waterproofing improvements; structural repairs or reinforcements; seismic retrofitting of tunnels	Agua Negra International Tunnel, Tunel de Oriente (Colombia)
	Bridges	Bridges are structures built to span physical obstacles such as rivers, roads, valleys, or railways, enabling the movement of vehicles, pedestrians, or cargo. This category includes the construction, maintenance, or improvement of bridge infrastructure, including related safety and structural systems. It covers feasibility studies and financing but excludes tunnels or elevated urban roads not classified as bridges.	Construction of new road or pedestrian bridges; structural rehabilitation or reinforcement of existing bridges; resurfacing of bridge decks; installation or upgrade of guardrails, drainage systems, or lighting; replacement of expansion joints or bearings; and implementation of seismic or climate resilience upgrades.	Takutu River Bridge (Guyana-Brasil), Puente Binacional Rio Sixaola (Costa Rica-Panama).
	Ports	Ports are facilities that support the movement of goods and passengers by ship, typically located along coastlines, rivers, or lakes. This category includes the construction, maintenance, and improvement of	Construction or expansion of cargo or passenger terminals; rehabilitation of piers, docks, and wharves; dredging of navigation channels; installation or upgrade of cargo handling equipment such as cranes, conveyors, or scanners; construction of	Port Expansion Project – Trinidad and Tobago, Terminal Zárate project in Argentina,

Category	Intervention	Useful definitions	Detailed examples	Specific examples
		infrastructure for both cargo and passenger ports. It also includes investments in logistics, equipment, and digital systems that improve port operations and cargo handling, and ancillary facilities that help streamline port operations.	storage infrastructure including refrigerated warehouses or silos; electrification and shore power systems for docked ships; deployment of digital logistics platforms for customs, cargo tracking, or scheduling; and construction or modernization of dry ports.	Puerto de Manzanillo – Mexico
Law, regulations and policy	Subsidies for public transit	A subsidy is financial support provided by governments or institutions to reduce the cost of public transit for users. Subsidies may be directed either to users (e.g., through discounted or free fares) or to operators (e.g., through compensation for fare reductions or revenue loss). The focus is on making public transit more affordable and accessible. This category excludes investment in infrastructure or general operational funding not tied to fare reduction.	Implementation of free or discounted fares for specific groups (e.g., students, seniors, low-income populations); government compensation to operators for offering reduced fares; public transit voucher or pass programs; universal basic mobility pilots offering fare-free access; and employer- or government-subsidized bulk purchase of transit passes, are free transit pilots linked to universal basic mobility initiatives.	Brazil's Vale Transporte, Abono gratuito para viajeros frecuentes (Spain), Subsidio Nacional al Transporte Público (Santiago de Chile)
	Low Emission Mobility: policy and regulations	Low emission mobility refers to transportation systems and technologies that significantly reduce greenhouse gas or air pollutant emissions compared to conventional fossil fuel-based mobility. This policy and regulation category includes the implementation or reform of laws, standards, and public programs that support this transition. It includes financial incentives such as subsidies or tax exemptions, as well as regulatory measures targeting emissions reduction. It excludes infrastructure	Subsidies or tax incentives for the adoption of electric, hybrid, or alternative-fuel vehicles; vehicle-emission standards and fuel-economy regulations; zero-emission fleet targets for public and private operators; rules for the public procurement of low-emission vehicles; regulations for hydrogen, biofuel, or other clean-energy technologies; phase-out deadlines for internal-combustion engines; low-emission transition strategies within national or municipal transport policies; CO_2-based registration or circulation taxes on internal-combustion vehicles; city-wide cycling-promotion programmes	Proconve P7 (Brazil), Mandate Euro 5 (Chile), National Green Hydrogen Strategy (Chile)

Category	Intervention	Useful definitions	Detailed examples	Specific examples
		development and direct vehicle purchases.	designed to shift mode share as an emissions-reduction strategy; and green traffic-promotion campaigns such as “Car-Free Day” initiatives.	
	Price-based traffic restrictions and tolls	Price-based traffic restrictions refer to the implementation or modification of systems that regulate vehicle access or road use through monetary charges. This includes congestion pricing, tolls, and distance-based or time-based fees applied to influence travel behavior, reduce traffic, or generate revenue. This category excludes physical construction of toll-related infrastructure, which is classified separately.	Introduction of congestion pricing zones in city centers; adoption of time-variable or dynamic pricing schemes; implementation of distance-based road user charges; policy changes adjusting toll rates or vehicle class exemptions; design of pricing mechanisms for express lanes; enforcement rules for electronic toll collection; integration of toll systems with broader urban mobility pricing strategies; and legal frameworks enabling or revising pricing-based access control.	Annual Congestion Fee (Buenos Aires), New York’s or London’s congestion pricing
	Circulation restrictions, non-price based	Non-price-based circulation restrictions refer to regulatory measures that limit vehicle access or use based on criteria other than emissions levels or monetary charges. These restrictions aim to manage congestion, improve traffic flow, or address trade and operational concerns, without using tolls or emission-based exclusions.	Implementation of license plate-based restrictions (e.g., odd-even schemes); time-based driving bans for certain vehicle types or categories; area-based restrictions for freight vehicles or motorcycles; access limitations for vehicles registered outside a jurisdiction; circulation rules linked to traffic reduction policies; and trade-related vehicle restrictions based on vehicle origin or route.	Pico y Placa (multiple Colombian cities), Hoy no Circula (Mexico), Rodízio Veicular (Sao Paulo)

Category	Intervention	Useful definitions	Detailed examples	Specific examples
	Transportation network companies (ride-sharing firms)	Transportation Network Companies (TNCs) refer to app-based ride services, such as ride-sharing or ride-hailing platforms, that connect passengers with drivers using private vehicles. This category includes interventions related to the implementation, regulation, or policy development surrounding TNCs. It covers both the introduction or expansion of such services and the legal, institutional, or operational frameworks that govern their functioning, including aspects like safety, labor standards, pricing, data use, and coordination with urban transport systems.	Introduction or expansion of ride-sharing services in a city; implementation of licensing or permitting systems for TNCs; establishment or revision of regulatory frameworks; development of labor protections or standards for app-based drivers; enforcement of safety and vehicle inspection requirements; policies on dynamic pricing, data sharing, or trip reporting; introduction of low-emission or accessibility requirements for TNC fleets; and integration of ride-sharing platforms into public transport or city mobility planning tools.	Entry of firms or regulatory measures in Mexico, Brazil, and Colombia.

3.2.3. Comparators

We will include any type of comparator but exclude studies that only use simulation or forecast models, ex-ante impact assessments, or scenario analyses.

3.2.4. Outcomes

The EGM will include outcome measures grouped into five main categories: access, quality, affordability, service management, and socio-economic results. Table 2 outlines the definition of outcomes for each category.

Table 2: Outcome framework

Category	Outcome	Description
Access	Public Transit Access and Use	Outcomes related to access and usage of public transit systems, such as total passengers, ridership by line or station, frequency of use, bike-sharing dock utilization.
	Private Vehicle Use	Outcomes related to private vehicle usage, such as vehicle ownership, mode choice, and trip frequency.
	Rail Services Access and use	Outcomes related to the use of rail services (excluding urban public transit), including passenger and freight transport.
	Other transport infrastructure access and use	Outcomes related to the access and use of transport infrastructure not covered by public transit, private vehicles, or traditional rail services. This includes, for example, rural roads, highways, tunnels, bridges, ports, fluvial transport systems, and active mobility infrastructure such as bike paths and pedestrian walkways. This category is especially relevant when the use of a transport infrastructure is directly linked to the intervention being evaluated. Indicators may refer to the frequency or volume of users, or to evidence that the infrastructure is being used as intended.
Quality	Travel Time and Speed	Outcomes related to travel speed, commuting time, and commuting flows across modes or road segments, including time savings from new infrastructure (e.g., bike lanes, bus lanes, HOV lanes, navigation systems). This also covers variability between peak and off-peak periods or across transport modes. These outcomes capture congestion relief, network efficiency, and user convenience. When speed indicators serve mainly to assess safety measures (e.g., reduced speeds in school zones), they belong under Transport Safety.
	Time used in other activities	Outcomes related to reallocating time to non-travel activities, such as leisure, sport (including walking or cycling), school, or family time.

Category	Outcome	Description
	Safety of the Infrastructure	Outcomes related to crashes and their consequences, including accident frequency, injuries, and fatalities. Indicators may also capture roadway safety standards, pedestrian and cyclist protection, vehicle speeds in sensitive areas (e.g., school zones), and compliance with traffic rules. This category also includes systemic aspects such as emergency-response times and post-crash survival rates. In general, any outcome linked to the safety of transport users or infrastructure belongs here.
	Resilience and continuity	Outcomes related to infrastructure capacity to maintain operations, adapt to disruptions, and recover from shocks such as natural disasters, extreme weather, or political and economic crises.
Affordability	Transport affordability	Outcomes related to the cost of using transport and its affordability for different population groups, especially vulnerable households. Indicators may include household spending on fares, fuel, tolls, or other travel costs; the share of income devoted to transport; or fare levels relative to median or household income. This category also considers how cost burdens or savings are distributed across income, age, or geographic groups, and whether households in rural or peripheral areas have at least one reasonably priced transport option available.
	Vehicle Ownership Costs	Outcomes related to the costs associated with owning and maintaining a vehicle (e.g., purchase price, insurance, maintenance, fuel).
	Government Expenditure on Transport	Outcomes related to government spending on transport, including infrastructure, operations, and maintenance costs.
Service Management	Operational Efficiency	Outcomes related to efficiency changes associated with the operation of a transportation system derived from new infrastructure, an intervention, an initiative, or regulation compliance. Indicators can include passengers per hour, delays, routes per hour, operational time savings, etc.
	Operational Environmental Sustainability	Outcomes related to transport system sustainability, such as emissions, air quality inside transit, or adoption of clean energy.

Category	Outcome	Description
	Financial Performance	Outcomes related to the financial performance of transport systems, such as fares, fleet costs, tolls, revenues, and profitability.
	Equity in Public Transport Access and Safety	Outcomes related to disparities in access, use, or safety of public transport across population groups, including those defined by gender, race, income, disability, or other characteristics.
Socioeconomic Results	Air Pollution and Greenhouse Gases	Outcomes related to air quality (e.g., PM2.5, PM10) and greenhouse gas emissions (e.g., CO2, O3) in a given area or population.
	Noise Pollution	Outcomes related to noise levels (e.g., dB) in a given area or population.
	Land, Housing, and Rent Prices and Affordability	Outcomes related to the value or price of land, housing, and rents, measured at the level of individual plots, properties, or broader areas such as neighborhoods or municipalities. This category also includes measures of housing affordability, such as the ratio of housing costs to income, etc.
	Goods and Services Prices and Inflation	Outcomes related to prices or inflation of goods and services, such as groceries, oil, energy, etc, in a defined market or area
	Health Access and Outcomes	Outcomes related to healthcare access and broader health conditions influenced by transport. Indicators may include healthcare utilization (e.g., physician visits, hospitalizations), availability of facilities, and travel times to reach them. This category also covers population health outcomes linked to environmental exposure or accessibility, such as respiratory illnesses from air pollution or improved health from greater access to care. Transport-related injuries are excluded and are addressed under Transport Safety.
	Education Access and Outcomes	Outcomes related to access to educational services (e.g., classes, learning opportunities) and infrastructure (e.g., schools, universities), including indicators such as attendance and enrollment. This category also covers educational performance outcomes, such as test scores, completion rates, and graduation rates for specific populations.
	Access to other goods and services	Outcomes related to access to markets and services other than education and health. Indicators

Category	Outcome	Description
		may include accessibility to grocery stores, financial institutions, government services, food and agricultural markets, and similar essential services.
	Inequality and segregation	Outcomes related to inequality, segregation, and social disparities regarding wealth, income, education, health, or other relevant dimensions. Indicators could include indicators like the Gini index, differences in income between groups of populations, literacy rate gaps, life expectancy differences by income or region, gender pay gaps, etc. Can include variables measuring economic inclusion, such as employment rate, financial inclusion, or representation for different marginalized groups like women, ethnic minorities or sexual minorities.
	Crime and Citizen Security	Outcomes related to crime and public safety, including crime rates by type, arrests, and police presence.
	Employment Access and Outcomes	Outcomes related to access to labor markets and employment opportunities, such as access to employment in certain locations, likelihood of attending an interview, labor participation rates, and more traditional labor market indicators, such as unemployment rate, number of jobs, labor participation rates, youth unemployment, wages, etc.
	Household Welfare and Poverty	Outcomes related to household material well-being, including changes in income, expenditure, or poverty status. This category includes absolute and relative measures of economic resources and living standards, as well as poverty and extreme poverty rates. Indicators may capture both average household conditions and disparities among households but exclude aggregate macroeconomic growth measures.
	Economic Activity	Outcomes capturing aggregate or regional economic performance and growth, such as GDP, GNP, economic growth rates, productivity (labor, firm, land, TFP), sectoral output, trade flows, nightlight intensity, and population growth. Indicators can be defined for specific geographic areas, industries, or groups of firms, but exclude household-level income, expenditure, or poverty measures. This category also includes firm level outcomes, such as, revenue, productivity, plant opening, investment, etc.

Category	Outcome	Description
	Trade and Migration	Outcomes related to different variables measuring the flow of goods, services, workers and other inputs. It can include different variables used in international trade, such as exports, imports, trade balance, number of destinations, number of products traded, etc., or used in migration studies, such as the number of immigrants, emigrants, net flows, etc. They can be defined for individual firms, groups of firms, or firms in a geographic area.
	Agricultural Markets	Outcomes related to agricultural production, market equilibrium (e.g., prices, quantities, crop yields, inputs), and the functioning of agricultural institutions. This category also covers frictions such as land allocation, access to capital, and information gaps. Indicators can be defined by crop, at the farm level, or for a broader geographic area.

3.2.5. Study designs

We define the study design eligibility criteria below, drawing on commonly accepted standards for impact evaluations (Gertler et al., 2016) and systematic reviews (Waddington et al., 2012).

We will include only quantitative effectiveness literature, focused on impact evaluations and systematic reviews that use attributional, causal designs to evaluate the effects of a clearly defined development intervention delivered in a real-world setting, rather than on natural or market-based occurrences or on controlled laboratory experiments without a discernible intervention component. Therefore, we will exclude studies primarily designed to determine the extent to which a specific technique, technology, treatment, procedure, or service works under ideal conditions rather than to answer a question relevant to the roll-out of a large program (i.e., a lab-in-the-field).

We will only include studies that implement at least one of the following study designs widely used to evaluate intervention effectiveness (Aloe et al. 2017; Reeves, Wells, and Waddington 2017):

A. Prospective studies that allocate participants to treatment and control groups using random assignment or quasi-experimental methods:
    a. Randomised controlled trials (RCTs), with assignment at individual, household, community, or other cluster level, and quasi-RCTs using prospective methods of assignment (such as alternation).
    b. Natural experiments with clearly defined intervention and comparison groups, which exploit natural randomness in implementation assignment by decision makers (e.g., public lottery) or random errors in implementation.

B. Quasi-experimental designs where treatment arms are created without random assignment:
    a. Regression discontinuity designs (RDD), either sharp or fuzzy designs, and other derived methods (i.e., kink RDD, differences in discontinuity).
    b. Instrumental variables (IV). This category may include Arellano-Bond and Arellano-Bover estimators if they are included as instruments in the econometric specification.
    c. Endogenous treatment-effects models, endogenous switching regression, and other methods synonymous to the Heckman two-step model.
    d. Difference-in-differences (DID), two-way fixed-effects (TWFE), high-dimensional fixed effects, and two-way Mundlak regressions (TWM).
    e. Interrupted time series (ITS) models, with or without a contemporaneous comparison group. An ITS model should include pre-intervention outcome data for at least 3 time points.
    f. Weighting and matching approaches which control for observable confounding, including non-parametric approaches (e.g., statistical matching, covariate matching, coarsened-exact matching, propensity score matching) and parametric approaches (e.g., propensity-weighted multiple regression analysis).
    g. Synthetic control methods, including their extensions: synthetic differences in differences, and generalized or augmented synthetic control methods.

Note that natural experiments where the assignment to intervention and control groups was not part of a planned experiment could use different inclusion criteria (e.g., RCT, RDD, ITS). These cases will be categorized as RCT, RDD, ITS, etc.

In panel datasets, additional estimation strategies are often employed to address time dynamics, autocorrelation, and endogeneity, particularly when outcomes are persistent over time. These strategies include random effects models, feasible generalized least squares, and dynamic panel estimations. While these methods support inference, they only yield causal effects when combined with exogenous variation, valid instruments, or a robust identification strategy. The same applies to gravity models. **Therefore, unless a clear identification strategy is given by one of the methodologies above, these studies should be excluded** (but not necessarily during the title and abstract screening stage).

The same rationale should apply to other theoretical or emerging methods, such as machine learning, Bayesian estimation, or simulation-based methods. Data-driven simulations and simulations of general equilibrium models sometimes incorporate causal study designs, as described above. Therefore, these approaches should not be dismissed outright at the title and abstract screening stage if the screener is uncertain about their methodology.

Observational studies, evaluations, and case studies that do not meet the methodological conditions described above, such as before-after studies without a comparison group or cross-sectional studies that do not adequately address selection bias or confounding, will not be included. Finally, we will also exclude the following study types: qualitative studies, feasibility studies, acceptability studies, and studies that examine willingness-to-pay for goods, services, processes, and business models. We acknowledge that the study types excluded from this map may contain valuable information; however, the focus of this EGM is to map existing rigorous evidence of intervention effectiveness.

A systematic review is a synthesis of research evidence on a particular topic, obtained through an exhaustive and transparent search across multiple academic databases and other relevant sources. The search process is systematic and reproducible, often including studies in different languages to avoid bias, while maintaining a clear focus on a specific intervention and pre-defined outcomes of interest. Systematic reviews apply widely accepted scientific strategies to minimize bias at every stage and, indeed, assess the quality and reliability of included studies by classifying them according to their design and the credibility of their findings.

The reviews included on this map go beyond a mere “state of the art” synthesis, since they aim to evaluate the effectiveness of interventions and to estimate their causal impact on relevant outcomes. Reviews that include study designs or methods not eligible for this map will be retained if at least one eligible study design is included and if the review reports results for at least one relevant intervention and one relevant outcome. When the study design of the included evaluations in a review is unclear, the review will be eligible if it aims to answer an effectiveness question. In addition, systematic reviews do not need to include a meta-analysis to be included in the map, since meta-analysis is often unsuitable when interventions are highly heterogeneous.

3.2.6. Other inclusion and exclusion criteria

Language

Studies published in English, Spanish, Portuguese, or French will be included, although the search terms will be in English. For grey literature, the four languages will be considered for the search if the institutions of interest include language filters.

Publication date

Studies will be included if they were published in 2005 or later. Given the field's evolution, it is widely recognized that the late 1990s and early 2000s marked a turning point in economics, commonly referred to as the “Credibility Revolution.” During this period, there was a growing emphasis on enhancing the reliability of empirical research through the adoption of more rigorous research designs and the increased use of experimental and quasi-experimental methods to evaluate policies and interventions. In light of this, the 2005 or later threshold has a low likelihood of missing eligible studies, while also limiting the overall breadth of the evidence mapping

project, ensuring that the exercise remains manageable and within our current resource constraints.

3.3. Search strategy

The EGM will employ a systematic search strategy, adhering to the guidelines for literature searching outlined by MacDonald et al. (2024). The search will avoid publication bias by covering academic bibliographic databases and grey literature sources, such as websites of organisational and international agencies. An information specialist will perform electronic searches in two subject-specific databases (EconLit and the 3ie Development Evidence Portal) and one general database (Web of Science). We will search the AEA RCT Registry and 3ie's Registry of International Development Impact Evaluations (RIDIE) to identify relevant ongoing studies. We will also perform backward and forward citation tracking and check the list of included studies for all systematic reviews if resources and time allow, considering the project timeframe and the number of included studies. When feasible, we will use Google Scholar for forward citation tracking and Web of Science for backwards citation tracking for each included study, as these resources have the most complete citation data (Martín-Martín et al. 2018). Alternatively, we will use the CitationChaser tool to automate the citation tracking process (Haddaway et al., 2021).[2]

To identify relevant grey literature, the team will manually search the databases and websites of organizations identified by the Transport Division and the Knowledge and Learning Division of the IDB. The full list of these institutions is provided in Appendix A.

In addition, approximately 200 publications manually collected and curated by the IDB Transport Division in 2017 and updated in 2024 will be incorporated into the search results. These studies will be processed under the same screening protocol as all other records to ensure consistency and minimize selection bias.

3.4. Reference management and screening protocol

We will document each step in the screening process in detail and graphically present the process in a flow chart to facilitate replication of the findings. We will manage the selection of studies for data extraction as part of the map using EPPI-Reviewer 6 software (Thomas et al. 2023) by implementing the following steps:

3.4.1. Import study records and remove duplicates

We will import all output files (e.g., RIS or .txt files) of the search strategy into EPPI Reviewer. We will use an automated process within EPPI Reviewer to remove duplicate references. In particular, 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

[2] CitationChaser relies on data from Lens.org to find citation relationships among papers. This allows for an automated process, but Lens.org citation data are incomplete relative to Google Scholar and Web of Science, so some papers citing/cited by included studies may be missed.

from studies with scores between 0.71 and 0.94 will be reviewed to manually complete this deduplication.

3.4.2. Training of screeners

Training will focus on understanding the subject matter and the screening process. Initially, all screeners will screen the same set of studies, two batches of 50 records, and will be evaluated for consistency. Screeners will achieve an 85% level of consistency before proceeding to independent title and abstract screening.

3.4.3. Title and abstract screening

Screeners will give a judgment of include, exclude, or unsure to records screened on the title and abstracts. Items marked unsure will be screened by a second screener (an approach demonstrated to produce comparable results to double screening at significantly lower cost; Shemilt et al. 2016). Several exclusion codes will be available to provide more information on the reasons for exclusion in each case. The researchers will apply screening codes in a hierarchical order to make consistent comparisons about why studies were excluded and at what stage in the screening process. The core team will hold periodic meetings to address studies flagged for a second opinion and make any refinements to the screening approach. The output of this process will be a set of screened studies that have been put forward for full-text screening. The inclusion and exclusion decision guide and screening tool are outlined in Appendix B, which also includes two additional markers for two intervention categories: road safety and airports and planes. However, the markers were created to identify studies that fall into these two topics and save the studies for potential future updates of the EGM.

We will explore the use of the machine-learning features of EPPI Reviewer, specifically Priority Screening, to accelerate the title and abstract screening process. We will begin by screening 200 random abstracts and the list of studies provided by the IDB's transportation division. We will then conduct a full-text screening of the studies from this sample that were included to determine which are eligible for the review. These screening data will serve as a training set for constructing a classifier that assigns a probability of inclusion to all remaining abstracts. We will screen all abstracts with a probability score of 0.3 or greater. We will then screen a random sample of 200 abstracts with lower probability scores to determine if any should be included for full-text screening. If more than 1 percent of this sample is found to be includable in the EGM, we will proceed to screen additional abstracts until this threshold is met (the process may also include building updated classifiers to ensure that we incorporate additional screening data).

3.4.4. Full-text screening

We will attempt to retrieve the full text for each study that meets the title and abstract inclusion criteria. Two coders will independently examine each full text in detail against the protocol. Again, we will apply a code to each study that indicates whether the study is included or why it is excluded. The output of this stage will be a set of studies deemed suitable to include in the EGM. The screening tool is outlined in Appendix C.

3.4.5. Checks for linked publications

The project team will attempt to group publications of the same study (i.e., an evaluation of an intervention on a specific population). This typically occurs when an author group publishes multiple papers on a single study. For each group of related publications, the research team will identify one main paper. We will extract descriptive information from the main paper. While each study will be represented only once in the EGM (regardless of the number of papers published on the study), all linked papers will be reviewed, and any additional information, particularly new outcome measures, will be incorporated into the dataset. This ensures that the extraction is as comprehensive as possible and prevents the evidence base from being artificially inflated. The identification of the main paper – the study that will appear in the map – will be consistent with the approach used by 3ie's Development Evidence Portal (DEP) team. Priority will be given to papers that already exist (and have their data extracted) in the DEP central database. If a potential main paper does not exist on the DEP, priority will be given to the most recent paper.

3.5. Data extraction and critical appraisal

The EGM team will systematically extract data from all included studies using the data extraction tool available in Appendix D. The data extraction will cover the following broad areas:

3.5.1. Basic study and publication information

This coding will focus on capturing the general characteristics of the study, including authors, publication date and status, study location, intervention type, outcomes reported, definition of outcome measures, population of interest, and study and program funders. Effect sizes for evidence synthesis will not be extracted.

3.5.2. Filters

The online map will display all the included studies in its default view. However, it will offer options to filter studies based on specific criteria, enabling users to view a subset of the evidence base. Table 3 shows the filters proposed for this map.

Table 3. List of variables that will be used as filters to classify included studies

Filter	Options for Dropdown	Explanation
Country	All countries*	The menu will allow identification of studies conducted in specific countries.
Region	• East Asia and Pacific (EAP) • Europe and Central Asia (ECA) • Latin America and the Caribbean (LAC) • Middle East and North Africa (MENA) • South Asia (SAR) • Sub-Saharan Africa (SSA) • North America	The menu will allow identification of studies conducted in specific regions based on the World Bank classification. This information is automatically completed by the DEP.
Income Level	• Low income • Lower middle income • Upper middle income • High income	The menu will allow identification of studies conducted in countries of a specific income category according to the World Bank classification. This information is automatically completed by the DEP.
Cost information	• Cost information (program costs and/or cost per participant) • Cost-benefit/cost-effectiveness analyses • None	The menu will allow for the identification of studies presenting cost evidence
Study design	• Randomized controlled trial • Regression discontinuity design • Instrumental variables • Fixed effects (including DiD) • Interrupted time series • Weighting and matching approaches • Synthetic control methods	The menu will allow for the identification of studies using a specific study design.
IDB Publication	• Yes/No	Teams will find IDB publications relatively easily
Year of publication	• 2005-2025	The menu will allow users to review evidence from specific time periods.
Journal Rank	• Q1 • Q2 • Q3 • Q4 • Not indexed/Other	The menu will allow users to review evidence from higher-quality journals ranked in Scopus's 2025 journal rankings.
Study type	• Impact evaluation • Systematic review	Users may want to focus on a specific type of publication.

The following processes will be implemented to collect this information:

3.5.3. Develop and refine data extraction tools and codebooks

The draft tools developed for this project will be reviewed and potentially refined in light of any feedback received by the IDB transportation division team, as the targeted main users of the EGM, and insights from project implementation.

3.5.4. Data extraction training and pilot

Coders assigned to each data extraction task will undergo theory- and practice-based training in using the tools provided. Each coder will code a 'training set' of studies, and detailed feedback will be provided.

3.5.5. Main-stage extraction

Coders will extract standardized information from each included study using the agreed-upon tool and codebook. This includes details on study design, population, intervention, comparator characteristics, and outcomes measured. Meetings will be held periodically with coders to provide support and resolve queries. For the extraction, the team will separate the studies into three distinct groups

a) Studies already in the DEP: for these studies, the team will review the bibliographic information and extract only the custom fields.
b) Studies not in the DEP but eligible: for these studies, the team will collect all the bibliographic, geographic, and methodological information, the custom fields, and all other required information for a study to belong to the DEP. DEP eligibility will be assessed by 3ie members to ensure consistency with the institution's products.
c) Studies not in the DEP and not eligible (for example, studies focusing on high-income countries): for these studies, the team will collect all the bibliographic, geographic, and methodological information, and the custom fields.

3.5.6. Quality checks

Throughout the data extraction process, the project team will check the extracted data. A core team member will check the consistency of the data extracted by coders. We will calculate measures of consistency and use them to inform the checking process. If additional review is warranted, targeted reviews will be conducted. This quality check process is put in place to ensure that the extracted data is accurate and does not assess the quality of the study itself or the evidence presented in the study.

3.5.7. Critical appraisal

For this EGM, we will critically appraise all included systematic reviews following the practices adopted by 3ie's systematic review appraisal tool, which draws on the SURE Checklist(Specialist Unit for Review Evidence (SURE) 2013). This appraisal assesses the extent to which each systematic review has used gold standard methodologies (Higgins et al. 2019; The Campbell Collaboration 2021), including criteria relating to the search, screening, data extraction, and analysis, and covers all the most common areas where biases are introduced. Each systematic review will be rated as low, medium, or high confidence, drawing on guidance provided in (Snilstveit et al. 2017). The tool used for this process is presented in

Appendix E. We will not critically appraise impact evaluations, as this is typically beyond the scope of EGMs. Critical appraisal assessments of systematic reviews will first be single-coded and then reviewed by a systematic review methods expert.

3.6. Dealing with multicomponent interventions

Multicomponent interventions are those that include several activities or components spanning different intervention categories, but whose effects are assessed collectively rather than separately. For example, an intervention could subsidize public bus fares to improve affordability for low-income commuters while also launching a public campaign to promote the environmental benefits of public transport. If the evaluation only quantifies the joint effect of providing the intervention components in combination, we consider the package of intervention components as a multicomponent intervention. If the effects of components are analyzed separately, we consider those to be separate interventions, and the study will be coded under each of those intervention types.

Multicomponent studies will be categorized based on the intervention components that have been bundled together. Bundled interventions that are evaluated five or more times will be added to the map as new intervention categories. We will create a "mixed" multicomponent bucket for all other combinations where there is no obvious pattern of specific components. This process will help us ensure that the map avoids artificially inflating the number of included studies and prevents double-counting of the same study. This coding adheres to common principles applied to 3ie EGMs:

A. All coding involves categorising studies into ideal types, so some simplification is necessary when describing studies in an EGM.
B. Coding of interventions to display studies in a typical EGM matrix should aim to describe the evaluative evidence (what the study is testing), rather than intervention components.
C. EGMs may have a secondary objective of describing program components based on the interventions included in EGMs, but the analysis should be clearly labelled as such.

The approach to deal with multicomponent interventions will be defined later on in the EGM construction process, and common combinations will be identified with the transportation division before the data extraction stage begins.

4. Analysis and reporting

The agreement between IDB and 3ie for this project does not include a technical report. The IDB's transportation division will be in charge of a posterior analysis of the resulting EGM, for which the Knowledge and Learning division could provide support, if needed.

5. Engagement and communication plan

The resulting Evidence Gap Map will be launched during the 2025 Knowledge Days. an event taking place inside the IDB in October 2025, which brings together the Vice Presidency for Sectors and Knowledge, the Vice Presidency for Countries, Sectoral and Regional Economic Advisors, Knowledge Coordinators, Country Economists, the Office of Strategic Planning and Development, IDB Invest, and IDB Lab.

The goal of this event is to identify common analytical priorities and opportunities for collaboration, discuss multisectoral knowledge gaps, and reflect on how to strengthen knowledge to improve the IDB Group's response to its main institutional objectives. During this event, senior management from 3ie will present the resulting Evidence Gap Map within a larger session on the importance of evidence for operations and policy.

The results of the map and its process will also be presented on the IDB's Abierto al Público blog, which is managed by the IDB's Felipe Herrera Library. The launch of this map will also be accompanied by a User Guide and a Technical Note to help users understand better how to use and build EGMs.

Finally, if there is interest from the IDB's transportation division, the map could also be presented at internal seminars from the division or the infrastructure department.

6. References

Aloe, A. M., Becker, B. J., Duvendack, M., Valentine, J. C., Shemilt, I., & Waddington, H. (2017). Quasi-experimental study designs series—Paper 9: Collecting data from quasi-experimental studies. *Journal of Clinical Epidemiology, 89*, 77–83. https://doi.org/10.1016/j.jclinepi.2017.02.013

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. (2016). *Impact evaluation in practice* (2nd ed.). Inter-American Development Bank and World Bank.

Haddaway, N. R., Grainger, M. J., & Gray, C. T. (2021). citationchaser: An R package and Shiny app for forward and backward citation chasing in academic searching. https://doi.org/10.5281/zenodo.4543513

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M., & Welch, V. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (Version 6.3). John Wiley & Sons.

MacDonald, H., Comer, C., Foster, M., Labelle, P. R., Marsalis, S., Nyhan, K., Premji, Z., et al. (2024). Searching for studies: A guide to information retrieval for Campbell systematic reviews. *Campbell Systematic Reviews, 20*(3), e1433. https://doi.org/10.1002/cl2.1433

Malhotra, S., White, H., de la Cruz, N., Saran, A., Eyers, J., John, D., Beveridge, E., & Blondal, N. (2021). *Evidence and gap map—Studies of the effectiveness of transport sector intervention in low- and middle-income countries (CEDIL/Campbell Gap Map 2021).* https://doi.org/10.51744/CSWP3

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics, 12*(4), 1160–1177. https://doi.org/10.1016/j.joi.2018.09.002

Reeves, B. C., Wells, G. A., & Waddington, H. (2017). Quasi-experimental study designs series—Paper 5: A checklist for classifying studies evaluating the effects on health interventions—a taxonomy without labels. *Journal of Clinical Epidemiology, 89*, 30–42. https://doi.org/10.1016/j.jclinepi.2017.02.016

Shemilt, I., Khan, N., Park, S., & Thomas, J. (2016). Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic Reviews, 5*(1), 1–13. https://doi.org/10.1186/s13643-016-0315-4

Snilstveit, B., Bhatia, R., Rankin, K., & Leach, B. (2017). *3ie evidence gap maps: A starting point for strategic evidence production and use* (Working Paper No. 28). International Initiative for Impact Evaluation (3ie). https://doi.org/10.23846/WP0028

Specialist Unit for Review Evidence (SURE). (2013). *Questions to assist with the critical appraisal of systematic reviews.* http://www.cardiff.ac.uk/specialist-unit-for-review-evidence/resources/critical-appraisalchecklists

The Campbell Collaboration. (2021). *Campbell Collaboration systematic reviews: Policies and guidelines*. https://doi.org/10.4073/cpg.2016.1

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.

Waddington, H., White, H., Snilstveit, B., Hombrados, J. G., Vojtkova, M., Davies, P., Bhavsar, A., Eyers, J., Perez Koehlmoos, T., Petticrew, M., & Valentine, J. C. (2012). How to do a good systematic review of effects in international development: A tool kit. *Journal of Development Effectiveness, 4*(3), 359–387. https://doi.org/10.1080/19439342.2012.711765

7. Appendices

Appendix A: List of relevant organizations identified for the search of grey literature

Name of the Organization	URL used in the search
IDB	https://publications.iadb.org/en
World Bank	https://documents.worldbank.org/en/publication/documents-reports/docadvancesearch
CAF	https://www.caf.com/en/action-areas/research-for-development/publications/
CEPAL	https://www.cepal.org/en/list/cepal_publication
Asian Development Bank	https://www.adb.org/publications
African Development Bank	https://www.afdb.org/en/documents/publications
European Investment Bank	https://www.eib.org/en/publications/all/index.htm
UN Trade and Development	https://unctad.org/publications
UN Habitat	https://unhabitat.org/knowledge/research-and-publications
International Road Federation	https://www.irf.global/irf-knowledge/
International Council on Clean Transportation	https://theicct.org/insight-analysis/publications/
World Resources Institute	https://www.wri.org/resources?query=&sort_by=created
C40 Cities	https://www.c40.org/research/
Institute for Transportation and Development Policy	https://itdp.org/publications/
Rand Transportation	https://www.rand.org/topics/transportation.html
Transformative Urban Mobility Initiative	https://transformative-mobility.org/knowledge-hub/multimedia-library/
International Transport Forum	https://www.itf-oecd.org/
International Energy Agency	https://www.iea.org/search/analysis?q=Publications
Transport Research Laboratory	https://www.trl.co.uk/publications

Appendix B: Title and abstract screening protocol

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.
4	Does the study evaluate a transportation 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 covering bus rapid transit, urban trains, cables, active mobility, urban roads, low emission mobility infrastructure or policy, urban traffic management systems, rural roads, highways, national roads, railways, trains, tunnels, bridges, ports, transit subsidies, traffic and circulation restrictions, and transportation network companies.
5	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

Questions		Decision				Notes
						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.
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 title and abstract			

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

Appendix C: Full-text screening protocol for impact evaluations[3]

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)	YES	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	Does the study only aim to describe the prevalence of a phenomenon or factors associated with/predictors of a phenomenon?	YES	EXCLUDE - No intervention	NO	Continue to the next question	Studies evaluating savings, credit, microfinance, and/or access to financial services, as predictor variables in a model describing empowerment as a phenomenon rather than aiming to establish a causal relationship between these should be excluded.
3	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

[3] For systematic reviews, a similar checklist will be used, with the necessary adaptations.

Questions		Decision				Notes
						studies, and purely qualitative studies.
4	Does the study evaluate an intervention that does NOT have a transportation component?	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.]
5	Does the study evaluate the effects on outcomes OTHER THAN [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 included outcomes. [Briefly describe the includable outcomes. Define what to do with aggregate measures.]
6	The study does not have a control group or does not use a valid 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. Exclude studies that fail to adequately describe the methodology and those with insufficient clusters (assignment at the cluster level AND there is only one cluster in either the treatment or control condition).
7	Is the study a duplicate?	YES	EXCLUDE as known duplicate			Choose only one option.
8	Are you unsure about inclusion?	YES	INCLUDE second opinion			
9	Does the study meet all eligible criteria?	YES	INCLUDE on full text			

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

Appendix D: Data extraction template

Code	Subcode
Study Information	Study EPPI internal ID
	Coder name
	Title name
	Foreign Title
	Short title
	Language
Author Information	Author Name
	Author Affiliation Institution
	Author Affiliation Country
Publication Information	Publication Type
	DOI
	Study status
	Abstract
	Keywords
	Journal name
	Other journal name
	Journal volume
	Journal issue
	Pages
	Year of Publication
	URL
	Publisher location
	Open access
Sector Information	Sector name
	Sub-sector name
	DAC rank
	Primary DAC Code
	Secondary DAC Code
	CRS-Voluntary (tertiary) Code
	SDGs
	World Bank (WB) first theme
	WB first sub-theme
	WB second theme
	WB second sub-theme
	WB third theme
	WB third sub-theme
	Other topics
	Equity focus
	Equity dimension
	Equity description
Geographic Information	First year of intervention
	Continent name
	Country name
	Additional country
	Country income level
	Region name
	State/province name
	District name
	City/town name
	Location name

Code	Subcode
Target population and cost data	Age
	Sex
	Setting
	Sexual orientation
	Specific population group
	Cost data
	Type of cost data
Methodological information	Evaluation Design
	Evaluation Method
	Mixed Method
	Additional Quantitative Methods
	Unit of Observation
Program, Funding, and Implementation Information	Project Name
	Implementation Agency Category
	Implementation Agency Name
	Program Funding Agency Category
	Program Funding Agency Name
	Research Funding Agency Category
	Research Funding Agency Name
Intervention Information	Intervention group/arm 1
	Intervention group/arm 1 Description
	Intervention group/arm 2
	Intervention group/arm 2 Description
	Create 3 different Intervention options in case there is more than one intervention group.
Outcome Information	Outcome (multiple fields to cover all relevant outcomes)
	Outcome description

Appendix E: Critical appraisal tool for assessing Systematic Reviews

1. Title:

2. Author:

Checklist for making judgements about how much confidence to place in a systematic review of effects (adapted version of SURE checklist)[6]

Assessed by:
Date:

Overall Notes. When the primary publication does not provide the information needed to appraise, information in other version of the publication can be used (document additional source in the appraisal). Study protocols cannot be used as a source, as the plans described in the protocols might not have been implemented.

Provide documentation and page numbers in your justifications. If copying/pasting text from the manuscript, please use quotation marks.

We provide authors with the completed checklist and they can provide additional information.

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*. This information cannot be determined by looking at the types of studies included, because some eligible populations, designs, interventions, and outcomes might not have been examined in the studies.	☐ Yes ☐ Partially ☐ No *Coding guide - check the answers above* *YES: All four should be yes* *NO: All four should be no* *PARTIALLY: Any other*
Documentation/Comments (note important limitations or uncertainty – please provide documentation and page numbers for your justifications, and use quotation marks if	

copying/pasting text from the manuscript) *A.1.1 Types of studies* *A.1.2 Participants/ settings/ population:* *A.1.4 Intervention(s)* *A.1.4 Outcome(s)*	
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,[7] as well as either: (a) for health, at least two relevant comprehensive subject databases (such as PubMed/MEDLINE, EMBASE and CENTRAL),[8] 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.	☐ 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*

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 technical leader of the EGM. 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.	
Documentation/Comments (note important limitations or uncertainty – please provide documentation and page numbers for your justifications, and use quotation marks if copying/pasting text from the manuscript) *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* *A.2.4 Reference lists in included articles checked* *A.2.5 Authors/experts contacted*	
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	☐ Yes ☐ Can't tell (only use if no information about time period for search) ☐ No

included study. If the study was published before 1990 this can be coded *Yes*.	☐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.*
Documentation/Comments (note important limitations or uncertainty – please provide documentation and page numbers for your justifications, and use quotation marks if copying/pasting text from the manuscript) *A.3 Time Period for the search:*	

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	☐ Yes ☐ Partially ☐ No

☐ 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. 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*.	*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 penalised for this. 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.
Documentation/Comments (note important limitations or uncertainty – please provide documentation and page numbers for your justifications, and use quotation marks if copying/pasting text from the manuscript) *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*	
A.5 Did the authors use appropriate criteria to assess the quality and risk of bias in analysing the studies that are included?[9] ☐ A.5.1 The criteria used for assessing the quality/ risk of bias were reported	☐ Yes ☐ Partially ☐ No ☐ Not Applicable (to be used only if

☐ 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 recognised 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.	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
Documentation/Comments (note important limitations or uncertainty – please provide documentation and page numbers for your justifications, and use quotation marks if copying/pasting text from the manuscript)	

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*	
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)
Documentation/Comments (note important limitations or uncertainty)	

Section B: *Methods used to analyze 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.	☐ Yes ☐ No ☐ Partially ☐ Not applicable (e.g. no included studies) *Coding guide:*

☐ B.1.3 A table or summary of the results of all the included studies *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: All three should be yes* *PARTIALLY: Criteria one and three 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*
Documentation/Comments (note important limitations or uncertainty – please provide documentation and page numbers for your justifications, and use quotation marks if copying/pasting text from the manuscript) *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 the included studies* *B.1.3 A table or summary of the results of all the included studies*	
B.2 Are the methods used by the review authors to analyze the findings of the included studies clear, including methods for calculating effect sizes if applicable? *Note*. 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* *NOT APPLICABLE: if no studies/no data*
Documentation/Comments (note important limitations or uncertainty – please provide documentation and page numbers for your justifications, and use quotation marks if copying/pasting text from the manuscript) *B.2 Are the methods used by the review authors to analyze the findings of the included studies clear, including methods for calculating effect sizes if applicable*	
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? ☐ 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	☐ Yes ☐ Partially ☐ No ☐ Not applicable (e.g. no studies or no data) *Coding guide:* *YES: First two should be yes, and B.1.3 should be yes if applicable* *PARTIALLY: B.3.1 is yes* *NO: Any other* *NOT APPLICABLE: if no studies/no data*

impacts varied from X to Y or Study A found X and Study B found Y).	
Documentation/Comments (note important limitations or uncertainty – please provide documentation and page numbers for your justifications, and use quotation marks if copying/pasting text from the manuscript) *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?* *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?*	

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 analyses ☐ Network meta-analyses ☐ Other: specify ☐ Not applicable (e.g. no studies or no data)	☐ 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*

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. *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	*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)*

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).	
Documentation/Comments (note important limitations or uncertainty – please provide documentation and page numbers for your justifications, and use quotation marks if copying/pasting text from the manuscript) *B.4.1 How was the data analysis done?* *B.4.2 How were the studies weighted in the analysis?* *B.4.3 Did the review address unit of analysis errors?*	
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 analysed separately by risk of bias status *Notes.* 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	☐ Yes ☐ No ☐ Partially ☐ Not applicable *Coding guide:* *YES: Both criteria should be fulfilled (where applicable)* *NO: Criteria not fulfilled* *PARTIALLY: Only one criteria 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*

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.	
Documentation/Comments (note important limitations or uncertainty – please provide documentation and page numbers for your justifications, and use quotation marks if copying/pasting text from the manuscript) *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*	
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	☐ Yes ☐ Partially ☐ No ☐ Not applicable *Coding guide:* *YES: Explanatory factors clearly described and appropriate methods used to explore heterogeneity* *PARTIALLY: Explanatory*

☐ Meta-regression ☐ Other	*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*
Documentation/Comments (note important limitations or uncertainty – please provide documentation and page numbers for your justifications, and use quotation marks if copying/pasting text from the manuscript) *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?*	
B.7 Overall - how much confidence do you have in the methods used to analyze 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*	☐ 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)

"Partial" option is used for any of the 6 preceding questions and/or B.2 and/ or B.6 are assessed as 'no'.	
Documentation/Comments (note important limitations or uncertainty)	

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 the 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 limitations 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.
Documentation/Comments (note important limitations or uncertainty) *C.1 Are there any other aspects of the review not mentioned before which lead you to question the results?* *C.2 Are there any mitigating factors which should be taken into account in determining the review's reliability?*	

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

☐ Low confidence in conclusions about effects:

☐ Medium confidence in conclusions about effects:

☐ High confidence in conclusions about effects:

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)).

Note. 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.

Limitations should be summarized above, based on what was noted in Sections A, B and C.