Systematic Review of Skills-Based Active Labor Market Interventions on Intermediate and Final Migration Outcomes: Protocol

Daniela Anda
International Initiative for Impact Evaluation (3ie)

Shannon Shisler
3ie

Carolyn Huang
3ie

Miriam Berretta
3ie

Promise Nduku
South Africa Centre for Evidence, University of Johannesburg (SACE)

Andile Madonsela
SACE

Systematic Review Protocol

September 2023

Note to readers:

The protocol was developed by 3ie with funding by the International Organization for Migration (IOM) through the United States Agency for International Development (USAID) funded project, Addressing the Root Causes of Irregular Migration in Guatemala. The content of this report is the sole responsibility of the authors and does not represent the opinions of the IOM, USAID, 3ie, its donors or its Board of Commissioners. Any errors and omissions are also the sole responsibility of the authors. Please direct any comments or queries to the corresponding author, Maria Daniela Anda Leon at danda@3ieimpact.org.

Suggested citation: Anda, Daniela, Shannon Shisler, Carolyn Huang, Miriam Berretta, Promise Nduku, Andile Madonsela, 2023. Protocol | Systematic Review of Skills-Based Active Labor Market Interventions on Intermediate and Final Migration Outcome. New Delhi: International Initiative for Impact Evaluation (3ie).

Contents

1 Background

1.1 The problem, condition, or issue

In a globalized world, migration serves important development purposes. Besides being an internationally recognized human right as the "natural expression of people's desire to choose how and where to lead their lives, which is a fundamental component of human development" (UNSD, 2022, p. 12), some evidence shows that it may also improve development in countries of origin through remittances (Ghosh, 2006; Faist, 2008; Hossain, 2022). However, when individuals are forced to migrate out of necessity or survival and there are limited means outside of formal channels, migration can increase the vulnerability of already-disadvantaged populations.

Irregular migration affects millions of people around the world (Yayboke and Gallego, 2019), putting them at greater risk of financial and/or labor exploitation, physical harm, violence, or death (Vutha, Pide and Dalis, 2011; Yayboke and Gallego, 2019; United Nations Office on Drugs and Crime, 2021; ILO, 2022). This has induced governments and international organizations to invest significant resources in addressing the "root causes" of irregular migration that create unfavorable conditions in countries of origin (e.g., economic disparity, conditions exacerbated by climate change, political instability, insecurity and transnational crime) and humanitarian crises such as conflicts, wars, or persecution (Vutha, Pide and Dalis, 2011; Loschmann, Kuschminder and Siegel, 2014; Yayboke and Gallego, 2019; National Security Council, 2021; Rose *et al.*, 2021; UNHCR, 2022a).

Although migration behavior is often attributed to a single or primary reason, there are often multiple factors behind an individual's decision to migrate. These factors are jointly considered and may include broader drivers (Gent 2002). Several large-scale policies have been designed to address what have been identified as "root causes" (Table 1), where poor conditions and limited opportunities in countries of origin may make migration to destination countries attractive.

Table 1. Select current and salient policy responses

Program	Resources invested	Root causes addressed	Beneficiaries
The Netherlands 2016-2021 Addressing Root Causes of Conflict, Instability and Irregular Migration (ARC) program (ECORYS, 2020)	EUR €90 million	Security, rule of law, peace processes, political governance, and socioeconomic reconstruction	Burundi, Democratic Republic of Congo, Ethiopia, Jordan, Lebanon, Mali, Somalia, South Sudan, Sudan, and Syria
	EUR €37 million	Governance, rule of law, access to markets and employment, peace, and security	Pakistan and Afghanistan

Program	Resources invested	Root causes addressed	Beneficiaries
The EU Emergency Trust Fund for Africa (EUTF for Africa) (Knoll and Sheriff, 2017; European Commission. Directorate General for International Partnerships., 2022)	EUR €4.2 billion	Diverse causes of instability, irregular migration and forced displacement to support all aspects of stability, security and resilience.	Sub-Saharan Africa
The US Root Causes Strategy (Office of Management and Budget, 2022)	USD $987 million	Economic insecurity, inequality, governance, human rights and free press, and gender-based violence and trafficking	Central America

However, there is insufficient empirical research examining whether such "root cause" interventions effectively decrease irregular migration, despite the large programs that adopt these approaches (Berretta et al., forthcoming). Rather, the existing evidence base is primarily descriptive, describing why individuals choose to migrate, characteristics of who decides to migrate, and the broader development impacts of migration (Obokata, Veronis and McLeman, 2014; IMF, 2015; Goldin *et al.*, 2018; Pitoski, Lampoltshammer and Parycek, 2021). Systematic evidence on the effectiveness of programs addressing root causes of irregular migration is, therefore, still scant. The objective of this systematic review is to synthesize the evidence on the effectiveness of one such type of interventions addressing the root causes of irregular migration, by using the quantitative impact evaluations identified by Berretta and colleagues (2023).

One such line of root-cause programing are active labor market policy (ALMP) interventions that aim to create and improve employment opportunities for potential migrants. These include skills-based training or apprenticeships programs, job search assistance programs, employment pipelines/pathways, public works schemes and self-employment promotion efforts. The literature on ALMP has focused on the effects on earnings and employment in local or national labor markets (Card, Kluve and Weber, 2015; McKenzie, 2017), and is mostly concentrated in high-income countries (Dar and Tzannatos, 1999; Betcherman, Olivas and Dar, 2004). To our knowledge, this is the first systematic review of the literature on the effectiveness of such policies on migration outcomes for low- and middle-income countries (L&MICs).

1.2 The intervention

We will include interventions that address the root causes of migration related to economic instability through active labor market policies (ALMP). This includes demand-side intervention at countries of origin aimed to increasing individuals' access to employment and entrepreneurship opportunities. Further, we will focus on a specific type of ALMP: skills-based interventions.

1.3 Theory of Change

The theory of change linking skills-based ALMP and irregular migration assumes that the net benefits of migration are reduced when unfavorable systemic conditions at home, such as economic insecurity or lack of employment opportunities, are addressed. Such improvement in

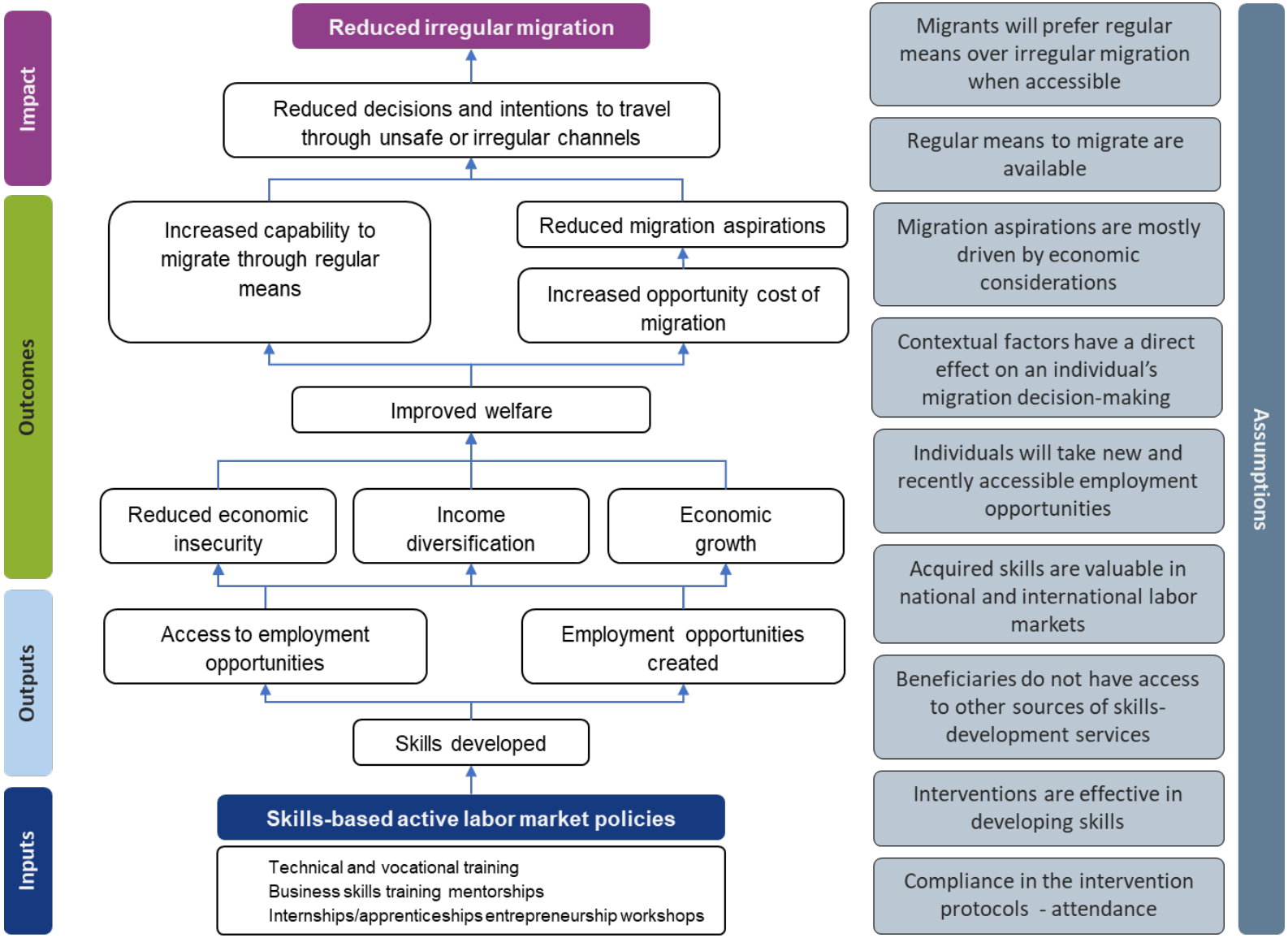
economic security increases the prospects of staying in the origin country the opportunity cost of migrating.

Our theory of change is adapted from Carling's (2002) theory of migration aspirations and abilities – further adapted by Carling and Talleraas (2016) and Carling and Schewel (2018) which present a framework of the individual decision-making process. According to this conceptual model, "voluntary" migration decision-making is driven by poor conditions, limited prospects, and/or individual perceptions in origin countries leading to a sense of stagnation or hopelessness and a desire for change, which may affect migration aspirations.

We hypothesize that skills-based ALMP create employment opportunities. This leads to improvements in local conditions and prospects such as greater economic security and income diversification through multiple livelihoods entrepreneurship, and economic growth. Further, such changes may reduce decisions and intentions to travel through unsafe or irregular channels that were based on a lack of better opportunities (Figure 1).

However, improving economic security may be insufficient to modify the life and migration aspirations of an individual, especially if other factors drive aspirations and perceptions (Soto, 2021). Therefore, we present a path through which irregular migrations is reduced as a result of an increased capability to migrate through regular means (Massey *et al.*, 1993; Kleemans, 2015). This represents the dichotomous effect of improvements in welfare: on one hand, individuals may possess better means to achieve migration aspirations; on another hand, greater economic opportunity may increase the opportunity cost of migrating.

FIGURE 1. THEORY OF CHANGE



1.4 Rationale for the review

This systematic review is expected to inform decisions regarding skills-based active labor market policies. Given the resources invested in intergovernmental programs addressing the root causes of irregular migration, key decision-makers have indicated interest in this area and can utilize the results of this review to inform interventions creating economic opportunities and developing skills in the workforce that aim to improve migration outcomes.

2 Research questions

1. What does the evidence indicate about the magnitude and direction of the effects of skills-based active labor market interventions on intermediate migration outcomes (intention to migrate, and knowledge, perceptions, attitudes and expectations) in low and middle-income countries (L&MICs)?
2. What does the evidence indicate about the magnitude and direction of the effects of skills-based active labor market interventions on final migration outcomes (any migration, international migration, migration flow, migration stock) in low and middle-income countries (L&MICs)?
3. Are there any unintended consequences of such interventions?
4. Do effects vary by context, intervention type, design or population characteristics (e.g., age, sex, SES, etc.)?
5. What are contextual barriers to and facilitators of intervention effectiveness?
6. How can future research enrich the evidence on the effects of active labor market interventions designed to improve migration in L&MICs?
7. What is the cost-effectiveness of these interventions?

3 Methodology

To respond to these research questions, we will conduct a theory-based mixed-methods systematic review using best practices outline by Snilstveit (2012) as well as by Cochrane and the Campbell Collaboration (Shemilt *et al.*, 2013; Kugley, Wade, Thomas, Mahood, A. K. Jørgensen, *et al.*, 2017; Higgins *et al.*, 2019) . The evidence included in this review will be based on the systematic literature search of key academic databases and grey literature sources conducted for the Evidence Gap Map (EGM) addressing root causes and drivers of irregular migration, see (2023) for search details. The studies identified by the EGM which evaluate the effects of skills-based ALMP will be assessed for quality and summarized visually and in a narrative format. Whenever the number of studies and levels of heterogeneity in intervention, outcomes and context suggest that it is reasonable to pool effect sizes together, we will also perform a meta-analysis to estimate an average effect size. We will implement a search for linked publications to the programs evaluated in included studies to identify documents that can inform a qualitative synthesis of the evidence and address research questions related to unintended consequences, the intervention context and barriers and facilitators of change (research questions 3 & 4).

3.1 Criteria for including and excluding studies in the review (PICOS)

Criteria	Included	Excluded
Participants	People of any age and gender residing in low- and middle-income countries (L&MICs)	High-income countries
Intervention(s)	Skills based active labor market policies including classroom and on-the-job training interventions	Other ALMP such as job search assistance programs, employment pipelines/pathways, public works schemes, self-employment promotion efforts and all else
Comparison	Business as usual, including pipeline and waitlist controls An alternate intervention	No comparator
Outcome(s)	Intermediate migration outcomes: • Intention to migrate • Knowledge, perceptions, attitudes, and expectations Final migration outcomes: • Attempted migration • Any migration measure unspecified as to international and/or irregular • International migration flow • International migration stock	All else
Study designs	Experimental and quasi-experimental impact evaluations; cost evidence, descriptive studies, process evaluations, and other qualitative studies linked to programs in included impact evaluations	Efficacy trials, before-after with no control group, feasibility/ acceptability studies, reviews.

3.1.1 Types of study participants

We will only include studies which consider populations in low- and middle-income countries (L&MIC;as defined using the) in the first year of intervention, if not available, then publication year will be considered. The exception to this is if a country held high-income status for only one year before reverting to L&MIC status. These will be included even if the intervention began in the high-income year. As of the writing of this protocol, this applies to Argentina (2014, 2017), Venezuela (2014), Mauritius (2019), and Romania (2019). If the study is conducted in a high-income country but measures impact on people, firms, or institutions in an L&MIC, it can be included. For example, we would not exclude a study that measures the impact of New Zealand's immigration visa lottery on residents of Tonga, or the Netherlands 2016-2021

Addressing Root Causes of Conflict, Instability and Irregular Migration (ARC) program on potential migrants from Ethiopia.

3.1.2 Types of interventions

Eligible interventions were identified during the development of the Addressing the Root Causes and Drivers of Irregular Migration Evidence Gap Map (Berretta et al., 2023). The map defined active labor market interventions as “Demand-side interventions aimed to increase individuals' access to employment and entrepreneurship opportunities. This may include skills-based interventions such as technical and vocational education training (TVET), business skills training, mentorships, internships/apprenticeships, entrepreneurship workshops; job placement centers and matching programs, employment pipelines/pathways within communities; wage subsidies; or public works schemes.” After completing the map, we found that these interventions were primarily related to skills development through training or apprenticeships. This systematic review will focus on such skills-based ALMP implemented through either classroom or on-the-job training interventions.

3.1.3 Types of outcome measures

The table below outlines outcomes that will be extracted. These outcomes can be measured using a variety of indicators such as rates, proportions, occurrence, etc. Whenever available, we will prefer outcomes associated with irregularity (e.g., irregular migration, intention to migrate irregularly), but based on the limited evidence on irregular migration (Berretta et al., forthcoming), we will also extract alternate outcomes such as intention to and final migration through regular channels, or migration unspecified as to whether it is regular or irregular.

Outcome	Indicators

Intermediate migration outcomes	- Intention to migrate
	  o Unspecified as to regular or irregular migration
	  o Regular migration
	  o Irregular migration
	- Knowledge, perceptions, attitudes and expectations
	  o Perception/psychosocial condition of current situation
	  o Expectations, awareness, knowledge, or attitudes on risks, benefits, costs, and/or consequences of movement through irregular channels
	  o Knowledge or awareness of legal pathways, legalization processes, or asylum seeking processes
	  o Knowledge or awareness of migrant labor rights

Final migration outcomes	- Any migration measure
	  o Any migration unspecified as to international and/or irregular
	  o International – unspecified as to regular or irregular
	  o International – regular
	  o International – irregular
	  o Forced displacement – unspecified as to international

	○ Forced displacement – international - International migration flow ○ International Migration flow – unspecified ○ International migration flow – regular ○ International migration flow – irregular - International migration stock ○ International migration stock – unspecified ○ International migration stock – regular ○ International migration stock – irregular

3.1.4 Types of Comparators

- Business as usual, including pipeline and waitlist controls
- An alternate intervention
- Studies with no comparator are excluded

3.1.5 Types of study design

Experimental and quasi-experimental impact evaluations were considered. The following study designs were included in the EGM from which the studies will be drawn.

- Randomized controlled trial
- Regression discontinuity design
- Controlled before-and-after studies, including:
    - Propensity-weighted multiple regression
    - Instrumental variable
    - Fixed-effects models
    - Difference-in-differences (and any mathematical equivalents)
    - Matching techniques
    - Synthetic control
- Interrupted time series

Further, for included impact evaluations, sister publications of process evaluations, qualitative evaluations, descriptive studies and cost evidence will also be included.

3.1.6 Date, language, and form of publication

All restrictions related to publication date, language and publication status are from the EGM, as that search was the basis for this review.

- Publication date: 1990 or later.
- Language: Search terms were in English, however, we screened records of studies in other languages (i.e., Spanish, Portuguese, Italian).
- Publication status: Academic and grey literature were included.

3.2 Search strategy

We will not perform any new searches for this systematic review given that the search for the EGM was conducted less than a year ago (between December 2022 and April 2023). The EGM search strategy was developed through a comprehensive and systematic process that adhered to the gold standard methodologies utilized in a systematic review (Kugley *et al.*, 2017; Berretta *et al.*, 2023; Appendix 3). Ultimately, the EGM includes ten impact evaluations looking at skills-based ALMPs.

In addition to qualitative evidence from the included studies, to assess factors that determine or hinder the effectiveness of interventions using a combination of qualitative synthesis, we will undertake targeted searching for qualitative studies, process evaluations, and project documents for those interventions evaluated in the ten included studies. We will conduct citation tracking of included studies to identify relevant sister papers and conduct internet and database searches using the names of programs from the ten studies. We will search Google and Google Scholar as well as the funder and implementer websites of the identified programs, looking for the following relevant information (Snilstveit, Oliver and Vojtkova, 2012): qualitative studies, descriptive quantitative studies and process evaluation.

3.3 Selection and coding of studies

3.3.1 Screening

Because we are utilizing the results of the EGM, there is no additional search for quantitative impact evaluations and thus no additional screening process to select the included impact evaluations. The EGM screening processes included independent duplicate screening at title and abstract and independent duplicate screening of the full text of potentially includable studies.

Records obtained from the targeted search for qualitative evidence linked to the programs evaluated in the ten included impact evaluations, will be single screened using the following definitions for inclusion:

- *A qualitative study* using mixed- methods or qualitative methods to collect and analyze primary data on all of the following: the research question, procedures for collecting data, procedures for analyzing data, and information on sampling and recruitment.
- *A descriptive quantitative study* using quantitative methods to collect primary data, and descriptive quantitative analysis on all of the following: the research question, procedures for collecting data, procedures for analyzing data, and information on sampling and recruitment, including at least two sample characteristics.
- *A process evaluation* assessing whether an intervention is being implemented as intended and what is felt to be working well, and why. Process evaluations may include the collection of qualitative and quantitative data from different stakeholders to cover subjective issues, such as perceptions of intervention success or more objective issues, such as how an intervention was operationalized. They might also be used to collect organizational information.

3.3.2 Data extraction and coding procedures

We will modify data extraction templates from 3ie's repository coding protocol and the coding protocols typically used for systematic reviews (Appendix 1). This includes bibliographic, geographic information and substantive data, as well as standardized methods information. In addition, two reviewers will independently extract data on outcomes, population (including gender/age disaggregation, when available), and effect sizes corresponding to the outcomes indicated above. Any discrepancies will be reconciled through consensus, or with a third team member if necessary. We will also extract qualitative information on barriers and facilitators to implementation, sustainability and equity implications, and other considerations for practitioners.

While the identification of qualitative evidence is limited to studies linked to the included impact evaluations, the process of data extraction and evidence synthesis is independent (Noyes et al., 2019). That is, coding of additional information is performed by a different researcher and using separate data extraction tools.

3.3.3 Critical appraisal

Two members of the team will independently appraise all of the included quantitative impact evaluations using 3ie's critical appraisal tool (Appendix 2.1 and 2.2), to assess the internal validity of experimental and quasi-experimental impact evaluation designs. The 3ie's tool, expands the bias domains of the Cochrane's ROBINS-I tool and RoB2.0 (Higgins *et al.*, 2016; Sterne *et al.*, 2019) and covers potential risks of selection bias, confounding bias, deviations from intended interventions, performance bias, outcome measurement bias and reporting bias.

For each study in our sample, we will report the results of the assessment for each criterion. In addition, we will produce an overall rating for each study as either "High risk of bias", "Some concerns" or "Low risk of bias", drawing on the decision rules in RoB2.0 (Higgins et al. 2016):

- "High risk of bias": if any of the bias domains were assessed as "No" or "Probably No."
- "Some concerns": if one or more domains were assessed as "No Information", and none were "No" or "Probably No."
- "Low risk of bias": if all the bias domains were assessed as "Yes" or "Probably Yes."

Whenever we are able to pool together effect sizes for an outcome of interest, we will attempt to explore whether the results are moderated by the overall rating, that is, whether there are systematic differences in outcome effects between primary studies with different risk of bias. Further, if meta-analysis is feasible, we will conduct sensitivity analysis to assess the robustness of the results to the risk of bias in included studies.

3.4 Analytical approach for quantitative and qualitative data

If sufficient data are available, we will conduct meta-analysis to provide summary effect estimates. We will choose the appropriate formulae for effect size calculations in reference to, and dependent upon, the data provided in included studies. We will conduct random effects meta-analyses when we identify two or more studies that we assess to be sufficiently similar.

We will assess heterogeneity by calculating the *Q* statistic, I^2, and τ^2 to provide an estimate of the amount of variability in the distribution of the true effect sizes (Borenstein, 2009). We will explore heterogeneity through the use of moderator analyses if the data allows. We will also test for the presence of publication bias when at least 10 studies are included in the analysis. The meta-analysis conducted with the quantitative data will be complemented by a thematic synthesis utilizing the extracted qualitative data.

3.4.1 Standard effect sizes

For pooling the effects of included studies, we need to standardize the effect sizes. For this purpose, we will use the formulae provided by Borenstein and colleagues (2009) to compute standardized mean differences (SMDs) for continuous outcomes, known as Cohen's d.

$$d = \frac{x_{Tp+1} + x_{Cp+1}}{SD}$$

Where, x is the reported mean for treatment (T) and control (C) groups at follow up ($p+1$), and SD is the pooled standard deviation[1]. For studies reporting regression results, we will follow the approach suggested by Keef and Roberts (2004) using the regression coefficient and the pooled standard deviation of the outcome. Where the pooled standard deviation of the outcome is unavailable, we will use regression coefficients and standard errors, t-statistics or significance levels, in that order upon availability of the data.

Where outcomes are reported in proportions of individuals, we will calculate the Cox-transformed log odds ratio effect size following Sánchez-Meca et al. (2003).

We will then adjust SMDs using Hedges' method to deal with potential biases in cases where sample sizes are small using the formula by Ellis (2010):

$$g \cong d\left(1 - \frac{3}{4(n_T + n_C) - 9}\right)$$

In all cases after synthesis, we will convert pooled effect sizes to commonly used metrics such as percentage changes and mean differences in outcome metrics typically used (e.g., weight in kg) whenever feasible.

3.4.2 Unit of analysis issues

Unit of analysis errors can arise when the unit of allocation of a treatment is different to the unit of analysis of effect size estimate, and this is not accounted for in the analysis (e.g., by clustering standard errors at the level of allocation). We will assess studies for unit of analysis errors (The Campbell Collaboration 2019), and where they exist, we will correct for them by

[1] If the study does not report the pooled standard deviation, it is possible to calculate it using the following formula:

$$SD_{p+1} = \sqrt{\frac{(n_{Tp+1} - 1)SD_{Tp+1}^2 + (n_{Cp+1} - 1)SD_{Cp+1}^2}{n_{Tp+1} + n_{cp+1} - 2}}$$

Where the intervention was expected to change the standard deviation of the outcome variable, we used the standard deviation of the control group only.

adjusting the standard errors according to the following formula (Higgins et al. 2020; Waddington et al. 2012; Hedges 2009):

$$SE(d)' = SE(d) * \sqrt{1 + (m-1)c}$$

where m is the average number of observations per cluster and c is the intra-cluster correlation coefficient. Where included studies use robust Huber-White standard errors to correct for clustering, we will calculate the standard error of d by dividing d by the t-statistic on the coefficient of interest.

3.4.3 Dealing with missing data

In cases of relevant missing or incomplete data in studies identified for inclusion, we will contact study authors to obtain the required information. If we are unable to obtain the necessary data, we will report the characteristics of the study but state that it could not be included in the meta-analysis or reporting of effect sizes due to missing data.

3.4.4 Assessment of heterogeneity

We will assess heterogeneity by calculating the Q-statistic, I^2, and Tau^2 to provide an estimate of the amount of variability in the distribution of the true effect sizes (Borenstein et al. 2009). We will complement this with an assessment of heterogeneity of effect sizes graphically using forest plots. Additionally, we will explore heterogeneity using moderator analysis in meta-regression specifications where there are at least four studies and significant heterogeneity. While some have suggested 10 studies as a minimum for moderator analysis (Higgins et al., 2019), as Borenstein and colleagues (2009) note, there are no hard and fast rules. However, we will ensure that for categorical moderators, there are a minimum of two effects per cell.

3.4.5 Quantitative data synthesis

We will conduct meta-analyses of studies that we assess to be sufficiently similar with respect to both the type of intervention being evaluated and the type of outcomes being measured. We will work with independent effect sizes, prioritizing outcomes based on comparability among studies when authors report more than one impact estimate for each of our intended analysis. The inclusion criteria for the review are narrow and we anticipate including studies that report on a limited set of interventions and outcomes. However, given the small sample of studies, it is difficult to predict how meta-analysis would be used in the review prospectively. The minimum criteria will be to only combine studies using meta-analysis when we identify two or more effect sizes using a similar outcome construct and where the comparison group state is judged to be similar across the two, similar to the approach taken by (Wilson et al. 2011). If pooling effect sizes for each outcome is not possible, we will check if we can conduct separate analyses for the major outcome categories (i.e., intermediate migration outcomes and final migration outcomes).

Moderator analyses can take into account multiple interventions as moderator variables, allowing us to also examine the impact of different intervention types by outcome. Where there are too few studies, or included studies are considered too heterogeneous in terms of interventions or outcomes, we will present a discussion of individual effect sizes along the causal chain. As heterogeneity exists in theory due to the variety of intervention characteristics

and contexts included, we will use inverse-variance weighted, random effects meta-analytic models (Higgins et al. 2020).

We will use the metafor package (version 2.4.0; Viechtbauer 2010) in R software to conduct the meta-analyses (version 4.0.4; R Core Team 2020). The amount of heterogeneity (i.e., $\tau\tau 2$), will be estimated using the DerSimonian-Laird estimator (DerSimonian & Laird, 1986).

Finally, we will conduct sensitivity analysis to assess whether the results of the meta-analysis are sensitive to the removal of any single study. We will do this by removing studies from the meta-analysis one-by-one and assessing changes in results. We will also assess sensitivity of results to inclusion of high risk of bias studies by removing these studies from the meta-analysis and comparing results to the main meta-analysis results. Studentized residuals and Cook's distances will be used to examine whether studies may be outliers and/or influential in the context of the model (Viechtbauer & Cheuuer and Cheung, 2010). Studentised residuals express the difference between the predicted effect size (based on the entire body of evidence in the analysis) and the observed effect size for any given study. These are standard diagnostic tools for outliers in meta-analysis. Studies with a studentized residual larger than the $100 \times (1 - 0.05/(2 \times k))$ the percentile of a standard normal distribution are considered potential outliers (i.e., using a Bonferroni correction with two-sided $\alpha = 0.05$ for k studies included in the meta-analysis). Studies with a Cook's distance larger than the median plus six times the interquartile range of the Cook's distances are influential.

3.4.6 Treatment of qualitative research

We will include a distinct review component to synthesize qualitative evidence on review questions 3, 4 and 5. Qualitative data coding will contemplate factors related to the context of the intervention, its design and implementation, and population characteristics.

Whenever we identify sufficient in-depth qualitative studies and empirical primary data reported across the evidence-base, linked to groups of interventions and outcomes along the review's theory of change, thematic synthesis will be conducted to address research questions 3 through 5. The objective of this approach is to identify analytical themes on intervention mechanisms and contexts that mitigate or reinforce intervention effects.

Following Thomas and Harden's (2008) thematic synthesis, we will use inductive coding techniques to first identify common descriptive themes based on the reported findings of the primary studies. We will use EPPI-Reviewer's coding software to illustrate the link between the inductive codes in the primary studies and the identified descriptive themes. In a second step, following the identification of descriptive themes, we will configure these into higher level analytical themes, which present the results of the thematic synthesis.

We will use four analytical lenses for the process of generating inductive codes, descriptive themes, and final analytical themes that refer to the interplay of context, intervention design, intervention implementation, and population characteristics as outlined in more detail below:

I. Context: Any variable related to external factors beyond the program's control that affect its impact. This can refer to political factors such as types of governance, societal factors such as norms, economic factors such as a recession, and cultural factors such as beliefs.

II. Intervention design: any variable that is related to the design and planning of the applied intervention. Design and planning of an intervention refer to the blueprint or schedule of the intervention and will typically outline what components the intervention consists of and in what sequence they will be applied. Examples of design variables refer to: size or type of cash transfer; outreach strategy, posters; reminders; type of training.
III. Intervention implementation: any variable that is related to the implementation of the intervention in practice. This refers to variables that emerge while the intervention is applied and are usually not known in advance. Examples of implementation variables refer to the lack of attendance or uptake, payment difficulties, corruption, elite capture.
IV. Population characteristics: any variable related to the population targeted by the intervention or the population in which the effects are measured (in cases where these differ). This can refer to the socio-economic status of the population, its educational status, and asset ownership.
V. Interplay with program effect: We will extract data and codes needed to relate to the program effect, outcome, or impact. That is, we will not extract descriptive data on the intervention design, implementation, context, and population—we are only interested in data that reports on how variables in these four categories are affecting program effects.

3.5 Data presentation

We will provide a narrative summary of the papers identified. This will include an overall description of the available literature and a general synthesis of findings. Key information from each study, such as intervention type, study design, country, outcomes, measurement type, effect sizes, and confidence rating will be summarized in a table. Results from meta-analyses and their associated forest plots will be presented when the data is sufficient. Qualitative information will be summarized narratively in a practitioner's brief to support project design and implementation and will be utilized to explain and contextualize quantitative findings.

3.6 Limitations

The small number of studies which are addressing the research questions for this review may restrict the possibility of synthesizing the evidence using meta-analysis and our ability to draw generalizable conclusions. We will highlight the caveats of our analyses and interpretation of findings in the final report.

4 References

Berretta, M. (2022) ‘Mapping evidence of what works to strengthen resilience to shocks and stressors’. Available at: https://www.3ieimpact.org/blogs/mapping-evidence-what-works-strengthen-resilience-shocks-and-stressors (Accessed: 21 December 2022).

Berretta, M., Lee, S., Kupfer, M., Huang, C., Ridlehoover, W., Frey, D., Ahmed, F., Song, B., Marie Edwards, K., Porciello, J., Eyers, J., and Snilstveit, B. (2022). Strengthening resilience against shocks, stressors, and recurring crisis in low- and middle-income countries: an evidence gap map. New Delhi: International Initiative for Impact Evaluation (3ie).

Berretta, M. *et al.* (Forthcoming) ‘Addressing root causes and drivers of irregular migration – an Evidence Gap Map’. New Delhi: International Initiative for Impact Evaluation (3ie).

Berretta, M., Huang, C., Anda, D., Shisler, S. (2022). Strengthening resilience against shocks, stressors, and recurring crisis in low- and middle-income countries: an evidence gap map. New Delhi: International Initiative for Impact Evaluation (3ie).

Betcherman, G., Olivas, K. and Dar, A. (2004) ‘‘Impacts of Active Labor Market Programs: New Evidence from Evaluations with Particular Attention to Developing and Transition Countries’, *Social Protection Discussion Paper Series* [Preprint].

Borenstein, M. (ed.) (2009) *Introduction to meta-analysis*. Chichester, U.K: John Wiley & Sons.

Card, D., Kluve, J. and Weber, A. (2015) ‘What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations’, *IZA Discussion Paper*, July.

Carling, J. (2002) ‘Migration in the age of involuntary immobility: Theoretical reflections and Cape Verdean experiences’, *Journal of Ethnic and Migration Studies*, 28(1), pp. 5–42. Available at: https://doi.org/10.1080/13691830120103912.

Carling, J. and Schewel, K. (2018) ‘Revisiting aspiration and ability in international migration’, *Journal of Ethnic and Migration Studies*, 44(6), pp. 945–963. Available at: https://doi.org/10.1080/1369183X.2017.1384146.

Carling, J. and Talleraas, C. (2016) *Root causes and drivers of migration Implications for humanitarian efforts and development cooperation*. Peace Research Institute Oslo (PRIO).

Dar, A. and Tzannatos, Z. (1999) ‘Active Labor Market Programs: A Review of the Evidence from Evaluations.’ World Bank Social Protection Working Paper.

ECORYS (2020) *Addressing Root Causes (ARC) Programme Final Report*. Rotterdam: The Netherlands. Available at: https://www.kpsrl.org/publication/addressing-root-causes-arc-programme-final-report.

European Commission. Directorate General for International Partnerships. (2022) *EU emergency trust fund for Africa: 2022 annual report.* LU: Publications Office. Available at: https://data.europa.eu/doi/10.2841/9748 (Accessed: 3 July 2023).

Faist, T. (2008) ‘Migrants as transnational development agents: an inquiry into the newest round of the migration-development nexus: Migrants as Transnational Development Agents’, *Population, Space and Place*, 14(1), pp. 21–42. Available at: https://doi.org/10.1002/psp.471.

Ghosh, B. (2006) *Migrants’ remittances and development: myths, rhetoric and realities*. Geneva, Switzerland: International Organization for Migration.

Goldin, I. *et al.* (2018) *Migration and the Economy Economic Realities, Social Impacts & Political Choices*. Available at: https://www.citivelocity.com/citigps/migration-and-the-economy/.

Higgins, J.P. *et al.* (2016) ‘A revised tool for assessing risk of bias in randomized trials’, *Cochrane Database of Systematic Reviews*, 10(Suppl 1), pp. 29–31.

Higgins, J.P.T. *et al.* (eds) (2019) *Cochrane Handbook for Systematic Reviews of Interventions*. 1st edn. Wiley. Available at: https://doi.org/10.1002/9781119536604.

Hossain, M.I. (2022) ‘Impacts of social remittances on economic activities: labour migration from a village of Bangladesh to Malaysia’, *Migration and Development*, 11(3), pp. 273–290. Available at: https://doi.org/10.1080/21632324.2020.1753962.

ILO (2022) *Global estimates of modern slavery forced labour and forced marriage*. Geneva: International Labour Office.

IMF (2015) *International Migration: Recent Trends, Economic Impacts, and Policy Implications*. Available at: https://www.imf.org/external/np/g20/pdf/2015/111515background.pdf.

Kleemans, M. (2015) ‘Migration Choice under Risk and Liquidity Constraints’. Available at: https://doi.org/10.22004/AG.ECON.200702.

Knoll, A. and Sheriff, A. (2017) *Making Waves: Implications of the Irregular Migration and Refugee Situation on Official Development Assistance Spending and Practices in Europe*. Expertgruppen för biståndsanalys (EBA). Available at: https://www.oecd.org/derec/sweden/201701-ECDPM-rapport.pdf.

Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A.K., *et al.* (2017) ‘Searching for studies: a guide to information retrieval for Campbell systematic reviews’, *Campbell Systematic Reviews*, 13(1), pp. 1–73. Available at: https://doi.org/10.4073/cmg.2016.1.

Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A.-M.K., *et al.* (2017) ‘Searching for studies: a guide to information retrieval for Campbell systematic reviews’, *Campbell Systematic Reviews*, 13(1), pp. 1–73. Available at: https://doi.org/10.4073/cmg.2016.1.

Loschmann, C., Kuschminder, K. and Siegel, M. (2014) *The root causes of movement: Exploring the determinants of Irregular Migration from Afghanistan*. Maastricht Graduate School of Governance (MGSoG) | UNU-MERIT. Available at: https://www.homeaffairs.gov.au/research-and-stats/files/irregular-migration-afghanistan.pdf.

Massey, D.S. *et al.* (1993) ‘Theories of International Migration: A Review and Appraisal’, *Population and Development Review*, 19(3). Available at: https://www.jstor.org/stable/2938462.

McKenzie, D. (2017) ‘How Effective Are Active Labor Market Policies in Developing Countries? A Critical Review of Recent Evidence’, *The World Bank Research Observer*, 32(2), pp. 127–154. Available at: https://doi.org/10.1093/wbro/lkx001.

National Security Council, U. (2021) *U.S. Strategy for Addressing the Root Causes of Migration in Central America*. National security council. Available at: https://www.whitehouse.gov/wp-content/uploads/2021/07/Root-Causes-Strategy.pdf.

Noyes, J. *et al.* (2019) ‘Synthesising quantitative and qualitative evidence to inform guidelines on complex interventions: clarifying the purposes, designs and outlining some methods’, *BMJ Global Health*, 4(Suppl 1), p. e000893. Available at: https://doi.org/10.1136/bmjgh-2018-000893.

Obokata, R., Veronis, L. and McLeman, R. (2014) ‘Empirical research on international environmental migration: a systematic review’. Available at: https://pubmed.ncbi.nlm.nih.gov/25132701/.

Office of Management and Budget (2022) ‘Budget of the U.S. Government Fiscal Year 2023’. U.S. Government Publishing Office. Available at: https://www.whitehouse.gov/wp-content/uploads/2022/03/budget_fy2023.pdf.

Pitoski, D., Lampoltshammer, T.J. and Parycek, P. (2021) ‘Drivers of Human Migration: A Review of Scientific Evidence’. Available at: https://www.mdpi.com/2076-0760/10/1/21.

Rose, S. *et al.* (2021) *Addressing the “Root Causes” of Irregular Migration from Central America: An Evidence Agenda for USAID*. 243. Centre for Global Development. Available at: https://www.cgdev.org/publication/addressing-root-causes-irregular-migration-central-america-evidence-agenda-usaid.

Shemilt, I. *et al.* (2013) ‘Issues in the incorporation of economic perspectives and evidence into Cochrane reviews’, *Systematic Reviews*, 2(1), p. 83. Available at: https://doi.org/10.1186/2046-4053-2-83.

Snilstveit, B., Oliver, S. and Vojtkova, M. (2012) ‘Narrative approaches to systematic review and synthesis of evidence for international development policy and practice’, *Journal of Development Effectiveness*, 4(3), pp. 409–429. Available at: https://doi.org/10.1080/19439342.2012.710641.

Sonnenfeld, A. *et al.* (2023) ‘Rule of Law and Justice: an evidence gap map’. International Initiative for Impact Evaluation (3ie). Available at: https://www.3ieimpact.org/evidence-hub/publications/evidence-gap-maps/rule-law-and-justice-evidence-gap-map (Accessed: 11 August 2023).

Soto, A.G.R. (2021) ‘Charting a New Regional Course of Action: The Complex Motivations and Costs of Central American Migration’.

Sterne, J.A.C. *et al.* (2019) ‘RoB 2: a revised tool for assessing risk of bias in randomised trials’, *BMJ*, p. l4898. Available at: https://doi.org/10.1136/bmj.l4898.

UNHCR (2022a) *Budget and Expenditure*, *Global Focus*. Available at: http://reporting.unhcr.org/financial (Accessed: 4 November 2022).

United Nations Office on Drugs and Crime (2021) *COVID 19 and the Smuggling of Migrants*. United Nation Office on Drugs and Crime. Available at: https://www.unodc.org/documents/human-trafficking/SOM_and_COVID-19_Publication_final_EN_final.pdf.

UNSD (2022) 'Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development'. Available at: https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework%20after%202022%20refinement_Eng.pdf.

Vutha, H., Pide, L. and Dalis, P. (2011) 'Irregular Migration from Cambodia: Characteristics, Challenges, and Regulatory Approach', (2011–26). Available at: https://www.econstor.eu/handle/10419/126861.

Yayboke, E.K. and Gallego, C.G. (2019) *Out of the Shadows Shining a Light on Irregular Migration*. CSIS Project on Prosperity and Development. Available at: https://reliefweb.int/report/world/out-shadows-shining-light-irregular-migration.

Appendix 1: Data extraction tools

1. Quantitative data extraction tool for effect size calculation

VARIABLE LABEL	EXPLANATION
Study ID	This is the study ID - it should match the study ID from the Outcome Mapping Sheet (e.g., 946578)
Estimate ID	The estimate ID will provide a specific number for each effect size extracted and should include the original study number, underscore, then the unique ID number (e.g., 946578_1, 946578_2 and so on)
STUDY DESCRIPTIVE INFORMATION	
Author	Author last name For 1 author: leading author last name (e.g. Gomez) For 2 authors: both author last names with ampersand in between (e.g. Smith & Bahn) For 3 or more authors: leading author last name followed by et al. (e.g. Gupta et al.)
Year	Year published
Location	Country of intervention. If it is an intervention for an specific location within a country, like city, write down the city instead.
Design	0=Experimental Design, 1=Quasi-Experimental Design
How Counterfactual is Chosen	Free text (e.g., RCT, Cluster RCT, propensity score matching, Instrumental variable, Fixed effects, etc.) - Multiple codes are ok
Estimate Type	Type of data for this effect size: 1 = Continuous - means and SDs, 2 = Continuous - mean difference and SD, 3 = Dichotomous outcome - proportions, 4 = Regression data - dichotomous outcome, 5 = Regression data - continuous outcome
Population	Drop down menu
Subgroup	Is this analysis of a subgroup? 0=no, 1=yes
If yes to subgroup, describe	Free text, describe the subgroup if applicable (e.g., boys, girls). If no subgroup, type N/A
Source	Note the page number, table number, column, and row you used to extract the data

Intervention description	Provide detailed description of the intervention such that a reader could easily understand what happened. Avoid copying text directly from the article as it is likely to be verbose. Summarize in your own words but include page numbers for quick reference. If more than two or more interventions are being evaluated, please provide descriptions for each intervention arm under separate rows, e.g. description of cash transfer (in all rows where estimate id's evaluate the cash transfer), description of cash transfer + community mobilization (in all rows where estimate id's evaluate the multicomponent intervention).
Intervention code	Dropdown menu with intervention codes
Exposure to intervention (in months)	How long is the intervention exposure itself? If time series is used, indicate the length of the period covering data points when the intervention was going on.
Evaluation period (in months)	The total number of months elapsed between the end of an intervention and the point at which an outcome measure is taken post intervention, or as a follow-up measurement. If less than one month, use decimals (e.g., measurement immediately after the intervention end would be coded as 0, one week would be .25, etc.)
Post-intervention or change from baseline?	0 = Post-intervention, 1 = Change from baseline
OUTCOMES	
Outcome description	Record the outcome for the corresponding effect size. Use this open answer field to enter, in the author's own words, a description of the outcome. Be selective and concise with the excerpts being transcribed here to ensure accurate and precise descriptions of the outcome. To the extent possible, be sure to include numbers, units, population, and comparators. Include page numbers with every excerpt extracted.
Outcome codes	Dropdown menu with outcome codes
Dataset	Record if data for this outcome comes from an identified dataset
EFFECT SIZE DATA EXTRACTION	
Reverse Sign (i.e., decrease is good)	Record no if an increase is good, record yes if a decrease is good and the sign needs to be reversed.
Unit of analysis	What is the unit of analysis? UOA for this effect size: 1= Individual, 2= Household, 3= Group (e.g. community organization), 4= Village, 5 = Other, 6 = Not clear
mean_t	Outcome mean for the treatment group
sd_t	Outcome standard deviation for treatment group
mean_c	Outcome mean for the comparison group
sd_c	Outcome standard deviation for control group
mean_overall_diff	Overall mean difference (treatment - control)
diff_se	Standard error of the overall mean difference

diff _t	t-statistic of mean difference
diff_p-value	p-value of mean difference
Odds ratio	Odds ratio reported in the study
OR_se	Odds ratio standard error reported in the study
Risk ratio	Risk ratio reported in study
RR_se	Risk ratio standard error
reg_coeff	Report the regression coefficient of the treatment effect
reg_SE	Report the associated standard error of the regression coefficient.
reg_t	Report the associated t statistic of the effect size (coefficient/SE)
reg_CI_LB	Report the associated Lower bound of the 95% Confidence interval of the effect size. If CI is reported for a different confidence level, indicate that in the notes section.
reg_CI_UP	Report the associated Upper bound of the 95% Confidence interval of the effect size. If CI is reported for a different confidence level, indicate that in the notes section.
Exact p value	Exact p value if given, if not, record as written in the manuscript (e.g., p < .001, or p > .05)
clust_t	Number of clusters - treatment group
clust_c	Number of clusters - control group
clust_T	Number of clusters - total sample
n_t	Sample size - treatment group
n_c	Sample size - control group
n_T	Sample size - total sample
periods (1 if cross sectional)	Record how many periods of evaluation there are (e.g., cross section is 1, panel data with 3 measurements is 3)
Does the sample size need to be corrected?	Often in panel data, models will report the number of observations rather than number of participants. In this column you will indicate 1="Yes" if the sample size needs to be divided by the number of periods, and 0="No" if either it is cross-sectional data, or if the authors have already divided the number of observations by the number of panel assessments and thus no correction is necessary.
Treatment Variable	Record the treatment variable as written in the model (e.g., the variable name the author uses, such as ("Intervention x Time")
CODING RECORDS	
coder	Record your name
Notes	Record any notes important for the team

Appendix 2: Critical appraisal tools

2.1 Full Appraisal of Risk of Bias for Impact Evaluations using RCT Designs

The following table provides a provisional tool to guide the risk of bias assessment for quantitative impact evaluations. If necessary, we could amend the tool to better inform the appraisal of primary studies.

Provisional risk of bias assessment tool (RCT)

General	ID	EPPI ID		
General	Study first author	Open answer		
General	Time taken to complete assessment	Minutes		
General	Design type: What type of study design is used?	1= Randomized controlled trial (RCT) (random assignment to households/individuals) or quasi-RCT 2= Cluster-RCT (quasi-RCT)	-	
General	Methods used for analysis: Which methods are used to control for selection bias and confounding?	1 = Statistical matching (PSM, CEM, covariate matching) 2 = Difference in differences (DID) estimation methods 3 = IV-regression (2stage least squares or bivariate probit) 4 = Heckman selection model 5 = Fixed effects regression	-	

		6 = Covariate adjusted estimation 7 = Propensity weighted regression 8 = Comparison of means = Other (please state)		
General	Design and analysis method description	Open answer	Briefly describe the study design and analysis method undertaken by the authors.	
General	Study population	Open answer	Provide any details in the paper that describe how the study population was selected, covering: a) How is the population selected? What is the sampling strategy to recruit participants from that population into the study? b) What are the characteristics of that study participants? Was this a pilot program aimed at being scaled up? d) Were there specific factors of success or failure in the implementation?	
General	Type of comparison group	1=No intervention (service delivery as usual) 2=Other intervention 3=Pipeline (wait-list) control (still service delivery as usual)	Indicate type of comparison group	
General	Type of comparison group (if other)	Open answer		

General	Ethical clearance	Open answer	Provide any details of ethical research clearances granted. Report unclear if this information is not available.	
General	Study registration	Open answer	Provide any details of study registration, including registry IDs, etc.	
1: Assignment mechanism - Assessment	Assignment mechanism: Was the allocation or identification mechanism random or as good as random?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) The authors describe a random component in sequence generation/ randomization method (e.g. lottery, coin toss, random number generator) and assignment is performed for all units at the start of the study centrally or using a method concealed from participants and intervention delivery b) If public lottery is used for the sequence generation, authors provide detail on the exact settings and participants attending the lottery. c) If a special randomization procedure is used to ensure balance, it is well described and justified given the study setting (stratification, pairwise matching, unique random draw, multiple random draws etch). d) A balance table is reported suggesting that allocation was random between all groups including subgroup receiving different treatment within control or treatment groups (if the comparison is relevant for this assessment).	Score “Yes” if all criterion a), b), c) and d) are satisfied. Score "Probably Yes" if only criterion a) and b) are not satisfied OR if only criteria c) is not satisfied. Score “Unclear” if d) is not satisfied because no balance table is reported. Score "Probably No" if d) is not satisfied because there is no balance table reported and there is evidence suggesting a problem in the randomization, such as baseline coefficients in a diff-in-diff regression table are very different or sample size is too small

				for the procedure used (using stratification when there are less than two units for each intervention and control group in each strata can lead to imbalance). Score “No” if d) is not satisfied because there are large imbalances concerning a large number of variables, providing evidence that the assignment was not random. If this is scored as no, use the NRS tool.
1: Assignment mechanism - Justification	Assignment justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
2: Unit of analysis - Assessment	Unit of analysis: Is unit of analysis in cluster allocation addressed in standard error calculation?	1=Yes 2=No 3=Not reported/unclear 4=Not applicable	Score "Yes" if UoA = UoR OR if UoA ≠ UoR and standard errors are clustered at the UoR level OR data is collapsed to the UoR level Score "Not reported/unclear" if not enough information is provided on the way the standard errors were calculated or what the unit of analysis is. Score "Not applicable" if it is not a cluster RCT. Score "No" otherwise.	

3: Selection bias - Assessment	Selection bias Was any differential selection into or out of the study (attrition bias) adequately resolved?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	Score "Yes" if there is no attrition or attrition falls into the green zone and the study establishes that attrition is randomly distributed (e.g. by presenting balance by key characteristics across groups) AND if survey respondents were randomly sampled. Score "Probably yes" if attrition falls into the green zone AND if survey respondents were randomly sampled. Score "Unclear" if there is an attrition problem but no information provided on the relationship between attrition and treatment status, OR if there is not enough information on how the population surveyed was sampled. Score "Probably no" if there is attrition which is likely to be related to the intervention OR there is some indication that the survey respondents were purposely sampled in a way that might have led the sampling to be different between treatment and control groups, or attrition falls into the yellow zone. Score "No" if attrition falls into the red zone.	
3: Selection bias - Justification	Selection bias justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	

4: Confounding - Assessment	Confounding and group equivalence: Was the method of analysis executed adequately to ensure comparability of groups throughout the study and prevent confounding	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) Baseline characteristics are similar in magnitude; b) Unbalanced covariates at the individual and cluster level are controlled in adjusted analysis; c) Adjustments to the randomization were taken into account in the analysis (stratum fixed effects, pairwise matching variables)? (Bruhn and McKenzie 2009)	Score “Yes” if criterion a) and b) are satisfied; Score "Probably yes" if a) is not satisfied but b) is satisfied and imbalances are small in magnitude OR if only a) is satisfied. Score “Unclear” if no balance table is provided or if imbalances are controlled for but they are very large in magnitude and assignment mechanism is not coded as "Yes" or "Probably yes" Score "Probably no" if a) and b) are not satisfied and the magnitude of imbalances are small. Score “No” if a) and b) are not satisfied and the magnitude of imbalances are large, and covariates are clear determinant of the outcomes.
4: Confounding - Justification	Confounding justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	

5: Deviations from intended interventions - Assessment	Deviations from intended interventions: Spillovers, crossovers and contamination: was the study adequately protected against spill-overs, crossovers and contamination?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) There were no implementation issues that might have led the control participants to receive the treatment (implementer's mistake). b) The intervention is unlikely to spillover to comparisons (e.g., participants and non-participants are geographically and/or socially separated from one another and general equilibrium effects are not likely) or the potential effects of spill overs were measured (e.g. variation in the % of unit within a cluster receiving the treatment). There is no risk of contamination by external programs: the treatment and comparisons are isolated from other interventions which might explain changes in outcomes. d) There is nothing in the surveys that might have given the control participants an idea of what the other group might receive OR they did but there is no risk that this has changed their behaviors; AND the survey process did not reveal information to the control group that they did not have before (e.g. the study aims to measure increase in take up of a service or product that participants might not know about) Authors might put something in place in the design of the study that allows to control for that survey effect (e.g. a pure control with no monitoring except baseline end line)	Score “Yes” if criterion a), b), c) and d) are satisfied. Score "Probably yes" if there is no obvious problem but there is no information reported on potential risks related to spill overs, contamination, or survey effects in the control group OR if there were issues with spillovers but they were controlled for or measured. Score “Unclear” if spillovers, crossovers, survey effects and/or contamination are not addressed clearly. Score "Probably no" if any of the criterion a), b), c) or d) are not satisfied but the scale of the issue is not clear. Score “No” if any of the criterion a), b), c) or d) are not satisfied and happened at a large scale in the study.

5: Deviations from intended interventions - Justification	Deviations justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages). For example, intervention groups are geographically separated, authors use intention to treat estimation or instrumental variables to account for non-adherence, and survey questions are not likely to expose individuals in the control group to information about desirable behaviors ('survey effects').	
6. Performance bias - Assessment	Performance bias: Was the process of monitoring individuals unlikely to introduce motivation bias among participants?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) The authors state explicitly that the process of monitoring the intervention and outcome measurement is blinded and conducted in the same frequency for treatment and control groups, or argue convincingly why it is not likely that being monitored could affect the performance of participants in treatment and comparison groups in different ways (such as resulting in Hawthorne or John Henry effects). b) The outcome is based on data collected in the context of a survey, and not associated with a particular intervention trial, or data are collected from administrative records or in the context of a retrospective (ex post) evaluation.	Score "Yes" if either criterion a) or b) are satisfied. Score "Probably yes" if the study is based on data collected during a trial and there is no obvious issue with the monitoring processes, but authors do not mention potential risks. Score "Unclear" if it is not clear whether the authors use an appropriate method to prevent Hawthorne and John Henry Effects (e.g., blinding of outcomes and, or enumerators, other methods to ensure

				consistent monitoring across groups). Hawthorne effects may result where participants know that they are being observed and John Henry Effects may result from participant knowledge of being compared. Score "Probably no" if there was imbalance in the frequency of monitoring in intervention groups, which might have influenced participants' behaviors. Score "No" if neither criterion a) or b) are satisfied.
6. Performance bias - Justification	Performance bias justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
7. Outcome measurement bias - Assessment	Outcome measurement bias: Was the study free from biases in outcome measurement?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) Outcome assessors are blinded, or the outcome measures are not likely to be biased by their judgement. b) For self-reported outcomes: respondents in the intervention group are not more likely to have accurate answers due to recall bias. c) For self-reported outcomes: respondents do not have incentives to over/under report something related to their	Score “Yes” if criterion a), b), c) and d) are satisfied: Score "Probably yes" if there is a small risk related to any of a), b), c) or d) and there is no more information provided to justify the absence of bias

			performance or actions, OR researchers put in place mechanisms to reduce the risk of reporting bias (researchers not strongly involved in the implementation of the program and it is clear that their answers to the survey will not affect what they receive in the future) OR authors have measured the risks of bias through falsification tests or measuring the effect on placebo outcomes in cases where there was a risk of reporting bias. d) Timing issue: the data collection period did not differ between intervention and comparison group; the baseline data is not likely to be affected by the beginning of the intervention or affects a small percentage of the study participants.	OR if there was a high risk of bias, but authors have either controlled it in their design or measured it with a placebo outcome. Score “Unclear” if there is a high risk related to any of a), b), c) or d) and there is no more information provided to justify the absence of bias. Score "Probably no" if there are high risk related to a), b), c) or d) and it is clear that authors were not able to control this bias. Score “No” if there is evidence of bias.
7. Outcome measurement bias - Justification	Outcome measurement justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
8. Reporting bias - Assessment	Analysis reporting: Was the study free from selective analysis reporting?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) A pre-analysis plan or trial protocol is published and referred to or the trial was preregistered, or the outcomes were preregistered. b) Authors report results corresponding to the outcomes announced in the method section (there is no outcome reporting bias);	Score "Yes" if all the criterion a), b), c), d), and e) are satisfied; Score "Probably yes" if all the conditions are met except a), or if all the conditions are met but there is some

			c) Authors report results of unadjusted analysis and intention to treat (ITT) estimation, alongside any adjusted and treatment-on-the treated/complier average-causal effects analysis.) d) Authors use the appropriate analysis method (use baseline data when available), and different treatment arms are differentiated in the analysis e) Authors have reported all the analysis which could help understand the results and no other bias is assessed as unclear due to the lack of an important analysis (e.g., a balance table or a subgroup analysis)	element missing that could have helped understand the results better (e). Score "Unclear" if there is not enough information to determine that there is an analysis missing; Score "Probably no" if any of the criterion b), c) or d) are not satisfied; Score "No" if any of the criterion b), c) or d) are not satisfied and there is evidence that the analysis results would be different because large imbalances were not controlled for, compliance was very low and ITT estimation was not reported or different treatment arms were pooled.
8. Reporting bias - Justification	Analysis reporting justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
9. Other bias - Assessment	Other risks of bias Is the study free from other sources of bias?	1= Yes, 4 = No		

9. Other bias - Justification	Other bias justification	Open answer	Justification for coding decision	
10. Blinding - observers - Assessment	Blinding of participants?	1=Yes 2=No 8=unclear 9= N/A	If there is no information, code NO. If there is information but it is ambiguous, code UNCLEAR.	
10. Blinding - observers - Assessment	Blinding of outcome assessors?	1=Yes 2=No 8=unclear 9= N/A	If there is no information, code NO. If there is information but it is ambiguous, code UNCLEAR.	
10. Blinding - analysts - Assessment	Blinding of data analysts?	1=Yes 2=No 8=unclear 9= N/A	If there is no information, code NO. If there is information but it is ambiguous, code UNCLEAR.	
10. Blinding - method(s)	Method(s) used to blind	Open answer (including describe method of placebo control) No 9= N/A	Describe method(s) used to blind	
11. External validity - Assessment	External validity	Open answer	a) What do authors say about external validity?	Include all information that can help assess the external validity of the results.

			Summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	

2.2 Full Appraisal of Risk of Bias for Impact Evaluations using Quasi experimental designs

Provisional risk of bias assessment tool (QED)

Code	Question	Coding	Criteria	Decision-rules
General	ID	EPPI ID		
General	Time taken to complete assessment	Minutes		
General	Study first author	Open answer		
General	Outcomes assessed	Open answer		
General	Study design: What type of study design is used?		1= Natural experiment: randomized or as-if randomized 2= Natural experiment: regression discontinuity (RD) 3= CBA (non-randomized assignment with treatment and contemporaneous comparison group, baseline and end line data collection) – individual repeated measurement 4= CBA pseudo panel (repeated measurement for groups but different individuals) 5= Interrupted time series (with or without contemporaneous control group) 6= Panel data, but no baseline (pre-test) 7 = Comparison group with end line data only	
General	Methods used for analysis: Which methods are used		1 = Statistical matching (PSM, CEM, covariate matching)	

	to control for selection bias and confounding?		2 = Difference in differences (DID) estimation methods 3 = IV-regression (2-stage least squares or bivariate probit) 4 = Heckman selection model 5 = Fixed effects regression 6 = Covariate adjusted estimation 7 = Propensity weighted regression 8 = Comparison of means 9 = Other (please state)	
General	Study population	Open answer	Provide any details in the paper that describe how the study population was selected, covering: a) How is the population selected? What is the sampling strategy to recruit participants from that population into the study? b) What are the characteristics of that study participants? c) Was this a pilot program aimed at being scaled up? d) Were there specific factors of success or failure in the implementation?	
General	Ethical clearance	Open answer	Provide any details of ethical research clearances granted. Report unclear if this information is not available.	
1: Selection bias - Assessment	1 - Mechanism of assignment: was the allocation or identification mechanism able to control for selection bias?		1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	
1: Selection bias - Justification	For regression discontinuity designs	Open answer	a) Allocation is made based on a predetermined discontinuity on a continuous	Score “Yes” if criteria a), b), c) are all satisfied.

			variable (regression discontinuity design) and blinded to participants or; b) if not blinded, individuals reasonably cannot affect the assignment variable in response to knowledge of the participation decision rule; c) and the sample size immediately at both sides of the cutoff point is sufficiently large to equate groups on average.	Score "Probably Yes" if there are minor differences in between both sides of the cut-off point but authors convincingly argue that the differences are unlikely to affect the outcome, OR individuals are not blinded and there are low risk of them affecting the assignment, but the authors do not mention it. Score “Unclear” if it is unclear whether participants can affect it in response to knowledge of the allocation mechanism. Score "Probably No" if there are differences between individuals on both sides of the cut-off point, and there are doubts that the differences are due to individuals altering the assignment OR the participants are blinded but there is evidence that the decisions that determined the discontinuity is based on differences between the two groups or differences in time.

				Score “No” if the sample size is not sufficient OR there is evidence that participants altered the assignment variable prior to assignment. If the research has serious concerns with the validity of the assignment process or the group equivalence completely fails, we recommend assessing risk of bias of the study using the relevant questions for the appropriate methods of analysis (cross-sectional regressions, difference-in-difference, etc.) rather than the RDDs questions.
1: Selection bias - Justification	For assignment based nonrandomized program placement and self-selection (studies using a matching strategy or regression analysis, excluding IV)	Open answer	a) Participants and non-participants are either matched based on all relevant characteristics explaining participation and outcomes, or. b) all relevant characteristics are accounted for.** and the data set used contains relevant variables that are measured in a relevant ways (i.e. they were not collected for a different purpose initially and therefore are good proxy for some characteristics). **Accounting for and matching on all relevant characteristics is usually only feasible when the program allocation rule is known and there are no errors of targeting. It is unlikely	Score “Yes” if a) or b) and c) are satisfied. Score "Probably yes" if a) or b) are addressed for but there is some doubt related to c), OR authors combined statistical matching and difference-in-difference to cope with unobservable differences, OR they only did statistical matching and there were clear rules for selection into the program (no self-selection).

			that studies not based on randomization or regression discontinuity can score “YES” on this criterion. There are different ways in which covariates can be taken into account. Differences across groups in observable characteristics can be taken into account as covariates in the framework of a regression analysis or can be assessed by testing equality of means between groups. Differences in unobservable characteristics can be taken into account through the use of instrumental variables (see also question 1.d) or proxy variables in the framework of a regression analysis, or using a fixed effects or difference-in-differences model if the only characteristics which are unobserved are time-invariant	Score “Unclear” if · it is not clear whether all relevant characteristics (only relevant time varying characteristics in the case of panel data regressions) are controlled. Score "Probably no" if only a statistical matching was done and there was self-selection into the program. Score “No” if relevant characteristics are omitted from the analysis.
1: Selection bias - Justification	For identification based on an instrumental variable (IV estimation)	Open answer	Score “Yes” if an appropriate instrumental variable is used which is exogenously generated: for example, due to a ‘natural’ experiment or random allocation. Score "Probably yes" if there is less evidence (no balance table showing differences between the intervention and comparison group). Score “Unclear” if the exogeneity of the instrument is unclear (both externally as well as why the variable should not enter by itself in the outcome equation). Score "Probably no" if there is evidence that enrolment in the program is correlated with a variable that might also have an effect on outcome and on the instrumental variable.	

			Score “No” if it is clear that the instrument is not exogenous and affects the outcome through other channels than the program.	
2: Confounding - Assessment	2 - Group equivalence: was the method of analysis executed adequately to ensure comparability of groups throughout the study and prevent confounding?		1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	
2: Confounding - Justification	For regression discontinuity design	Open answer	a) The interval for selection of treatment and control group is reasonably small OR authors have weighted the matches on their distance to the cutoff point; and b) the mean of the covariates of the individuals immediately at both sides of the cut-off point (selected sample of participants and non-participants) are overall not statistically different based on t-test or ANOVA for equality of means. c) Significant differences in covariates of the individuals have been controlled in multivariate analysis; and for cluster-assignment, authors control for external cluster-level factors that might confound the impact of the program.	Score "Yes, if criterion a), b), c) and d) are addressed. Score "Probably yes" if b) is not addressed but c) is addressed and differences in means are not large. Score “Unclear” if insufficient details are provided on controls; or if insufficient details are provided on cluster controls. Score "Probably no" if b) is not addressed (absence of a difference test or balance table) and there are doubt regarding the continuity on

				both sides of the cut-off point (a). Score “No” otherwise.
2: Confounding - Justification	For non-randomized trials using difference-in-differences methods of analysis	Open answer	a) The authors use a difference-in-differences (or fixed effects) multivariate estimation method; b) the authors control for a comprehensive set of individual time varying characteristics, and for cluster assignment, authors control for external cluster-level factors that might confound the impact of the program**; c) and the attrition rate is sufficiently low and similar in treatment and control, or the study assesses that dropouts are random draws from the sample (for example, by examining correlation with determinants of outcomes, in both treatment and comparison groups); **Knowing allocation rules for the program – or even whether the non-participants were individuals that refused to participate in the program, as opposed to individuals that were not given the opportunity to participate in the program – can help in the assessment of whether the covariates accounted for in the regression capture all the relevant characteristics that explain differences between treatment and comparison groups.	Score "Yes, if a, b, c, d (if relevant) are addressed and baseline imbalances between groups were relatively low OR the method was combined by a statistical matching. Score "Probably yes" if all possible variables are controlled for and the selection into the program was done according to clear rules, but baseline imbalances between groups were very large. Score “Unclear” if insufficient details are provided; or if insufficient details are provided on cluster controls. Score "Probably no" if some time-varying characteristics are not controlled for and the program was self-selected by the intervention groups.

				Score “No” if any of the criterion is not addressed.
2: Confounding - Justification	For statistical matching studies including propensity scores (PSM) and covariate matching** **Matching strategies are sometimes complemented with difference-indifference only uses in the estimation the common support region of the sample size, reducing the likelihood of existence of time variant unobservable differences across groups affecting outcome of interest and removing biases arising from time invariant unobservable characteristics. regression estimation	Open answer	a) Matching is either on baseline characteristics or time-invariant characteristics which cannot be affected by participation in the program; and the variables used to match are relevant (for example, demographic and socio-economic factors) to explain both participation and the outcome (so that there can be no evident differences across groups in variables that might explain outcomes); and, for cluster assignment, authors control for external cluster-level factors that might confound the impact of the program b) in addition, for PSM Rosenbaum’s test suggests the results are not sensitive to the existence of hidden bias; and, c) with the exception of Kernel matching, the means of the individual covariates are equated for treatment and comparison groups after matching; d) different matching methods including varying sample sizes yields the same results and authors take into account the use of control observations multiple times against the same treatment in their standard error calculation.	Score "Yes, if a, b, c, and d (if relevant) are addressed. Score "Probably yes" if the selection into the program was done according to clear rules, which are used for the matching but there are slight imbalances remaining after matching. Score “Unclear” if relevant variables are not included in the matching equation, or if matching is based on characteristics collected at end line; or if insufficient details are provided on cluster controls. Score "Probably no" if the program was self-selected by the intervention groups or participants OR if the selection into the program was done according to clear rules but there is no baseline data available to match the participants or groups on. Score “No” if matching was done based on variables

	methods. This combination approach is superior since it			that are likely to be affected by the program or any other scenario that affect a), b) c) or d).
2: Confounding - Justification	For regression-based studies using cross sectional data (excluding IV)	Open answer	a) The study controls for relevant confounders that may be correlated with both participation and explain outcomes (for example, demographic and socio-economic factors at individual and community level) using multivariate methods with appropriate proxies for unobservable covariates, and, for cluster-assignment, authors control particularly for external cluster-level factors that might confound the impact of the program; b) and a Hausman test with an appropriate instrument suggests there is no evidence of endogeneity**; c) and none of the covariate controls can be affected by participation; d) and either, only those observations in the region of common support for participants and non-participants in terms of covariates are used, or the distributions of covariates are balanced for the entire sample population across groups; **The Hausman test explores endogeneity in the framework of regression by comparing whether the OLS and the IV approaches yield significantly different estimations. However, it plays a different role in the different methods of analysis. While in the OLS regression framework the Hausman test mainly explores endogeneity and therefore is related with the	Score "Yes, if a, b, c and d are addressed. Score "Probably yes" if all criteria are addressed but authors did not report the Hausman test (b). Score “Unclear” if relevant confounders are controlled but appropriate proxy variables or statistical tests are not reported; or if insufficient details are provided on cluster controls. Score "Probably no" if any of the criterion other than b) is not addressed. Score “No" if none of the criterion are addressed.

			validity of the method, in IV approaches it explores whether the author has chosen the best available strategy for addressing causal attribution (since in the absence of endogeneity OLS yields more precise estimators) and therefore is more related with analysis reporting bias.	
2: Confounding - Justification	For identification based on an instrumental variable (IV estimation)	Open answer	a) The instrumenting equation is significant at the level of F≥10 (or if an F test is not reported, the authors report and assess whether the R-squared (goodness of fit) of the participation equation is sufficient for appropriate identification); b) the identifying instruments are individually significant (p≤0.01); for Heckman models, the identifiers are reported and significant (p≤0.05); c) where at least two instruments are used, the authors report on an overidentifying test (p≤0.05 is required to reject the null hypothesis); and none of the covariate controls can be affected by participation and the study, and authors convincingly assesses qualitatively why the instrument only affects the outcome via participation. If the instrument is the random assignment of the treatment, the reviewer should also assess the quality and success of the randomization procedure in part a). d) and, for cluster assignment, authors particularly control for external cluster level factors that might confound the impact of the program (for example, weather, infrastructure,	Score "Yes, if a, b, c, d (if relevant) are addressed. Score "Probably yes" if one of the test required for criterion a) or b) is not reported but the other is, and the rest of the criterion are addressed and the instrument is convincing. Score “UNCLEAR” if relevant confounders are controlled for but appropriate statistical tests are not reported; or if insufficient details are provided on cluster controls Score "Probably no" if exogeneity of the instrument is not convincing and appropriate tests are not reported. Score “No” otherwise if any of the tests required for

			community fixed effects, and so forth) through multivariable analysis.	criterion a), b) or c) are reported and not satisfied.
3: Performance bias - Assessment	3 - Performance bias: was the process of being observed free from motivation bias?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) For data collected in the context of a particular intervention trial (randomized or non randomized assignment), the authors state explicitly that the process of monitoring the intervention and outcome measurement is blinded, or argue convincingly why it is not likely that being monitored could affect the performance of participants in treatment and comparison groups in different ways (such as resulting in Hawthorne or John Henry effects). b) The study is based on data collected in the context of a survey, and not associated with a particular intervention trial, or data are collected from administrative records or in the context of a retrospective (ex post) evaluation.	Score “Yes” if either criterion a) or b) are satisfied; Score "Probably yes" if the study is based on survey data collected during a trial and there is no obvious issue with the monitoring processes but authors do not mention potential risks. Score “Unclear” if it is not clear whether the authors use an appropriate method to prevent Hawthorne and John Henry Effects (e.g. blinding of outcomes and, or enumerators, other methods to ensure consistent monitoring across groups). Hawthorne effects may result where participants know that they are being observed and John Henry Effects may result from participant knowledge of being compared. Score "Probably no" if there was imbalance in the frequency of monitoring in

				intervention groups, which might have influenced participants' behaviors. Score "No" if both criterion a) and b) are not satisfied.
3: Performance bias - Justification	Performance bias - Justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
4: Spill-overs, cross-overs and contamination - Assessment	4 - Spill-overs, cross-overs and contamination: was the study adequately protected against spill-overs, crossovers and contamination?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) There was no implementation issues that might have led the control participants to receive the treatment (implementer's mistake). The intervention is unlikely to spillover to comparisons (e.g. participants and non-participants are geographically and/or socially separated from one another and general equilibrium effects are not likely) or the potential effects of spill overs were measured (e.g. variation in the % of unit within a cluster receiving the treatment). c) There is no risk of contamination by external programs: the treatment and comparisons are isolated from other interventions which might explain changes in outcomes. b) There is nothing in the surveys that might have given the control participants an idea of what the other group might receive OR they did but there is no risk that this has changed their behaviors; AND the survey process did not reveal information to the control group that they did not have before (e.g. the study aims to measure increase in	Score “Yes” if criterion a), b), c) and d) are satisfied; Score "Probably yes" if there is no obvious problem but there is no information reported on potential risks related to spill overs, contamination, or survey effects in the control group OR if there were issues with spill-overs but they were controlled for or measured. Score “Unclear” if spillovers, cross-overs, survey effects and/or contamination are not addressed clearly. Score "Probably no" if any of the criterion a), b), c) or d) are not satisfied but the

			take up of a service or product that participants might not know about) Authors might put something in place in the design of the study that allows to control for that survey effect (e.g. a pure control with no monitoring except baseline end line)	scale of the issue is not clear. Score “No” if any of the criterion a), b), c) or d) are not satisfied and happened at a large scale in the study.
4: Spill-overs, cross-overs and contamination - Justification	Spill-overs, crossovers and contamination - Justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
5: Outcome measurement bias - Assessment	5 - Outcome measurement bias	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) Outcome assessors are blinded or the outcome measures are not likely to be biased by their judgement. b) For self-reported outcomes: respondents in the intervention group are not more likely to have accurate answers due to recall bias; c) For self-reported outcomes: respondents do not have incentives to over/under report something related to their performance or actions, OR researchers put in place mechanisms to reduce the risk of reporting bias (researchers not strongly involved in the implementation of the program and it is clear that their answers to the survey will not affect what they receive in the future) OR authors have measured the risks of bias through falsification tests or measuring the effect on placebo outcomes in cases where there was a risk of reporting bias. d) Timing issue: the data collection period did not differ between intervention and comparison group, the baseline data is not likely to be affected by the beginning of the	Score “Yes” if criterion a), b), c) and d) are satisfied: Score "Probably yes" if there is a small risk related to any of a), b), c) or d) and there is no more information provided to justify the absence of bias OR if there was a high risk of bias but authors have either controlled it in their design or measured it with a placebo outcomes. Score “Unclear” if it there is a high risk related to any of a), b), c) or d) and there is no more information provided to justify the absence of bias.

			intervention or affects a small percentage of the study participants.	Score "Probably no" if there are high risk related to a), b), c) or d) and it is clear that authors were not able to control for this bias. Score “No” if there is evidence of bias.
5: Outcome measurement bias - Justification	Outcome measurement bias - Justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
6: Reporting bias - Assessment	6 - Selective analysis reporting: was the study free from selective analysis reporting?	1= Yes, 2 = Probably Yes, 3 = Probably No, 4 = No, 8 = Unclear	a) a pre-analysis plan is published, especially for prospective NRS but it should also be for retrospective studies b) authors use ‘common’ methods of estimation (i.e. credible analysis method to deal with attribution given the data available) ; c) There is no evidence that outcomes were selectively reported (e.g. results for all relevant outcomes in the methods section are reported in the results section) ; d) Requirements for specific methods of analysis: - For PSM and covariate matching: (a) Where over 10% of participants fail to be matched, sensitivity analysis is used to re-estimate results using different matching methods (Kernel Matching techniques); (b) For matching with replacement, no single observation in the control group is matched with a large number of observations in the treatment group. - For IV (including Heckman) models, (a) The authors test and report the results of a Hausman test	Score “Yes” if a), b), c) and d) are satisfied OR if a) is not met and it is a retrospective NRS. Score "Probably Yes" if authors combined methods and reported relevant tests (d) only for one method OR if all the criteria are met except for a) and it is a prospective NRS Score "Unclear" if intended outcomes not specified in the paper OR if any of the requirements for d) are not reported. Score "Probably No" if b) is addressed, but authors did not present results for all outcomes announced in the

			for exogeneity (p≤0.05 is required to reject the null hypothesis of exogeneity); (b) the coefficient of the selectivity correction term (Rho) is significantly different from zero (P<0.05) (Heckman approach). - For studies using multivariate regression analysis, authors conduct appropriate specification tests (e.g. testing robustness of results to the inclusion of additional variables, or (very rare) reporting results of multicollinearity test etc.).	method section OR did not meet requirement d) although reported. Score “No” if authors use uncommon or less rigorous estimation methods such as failure to conduct multivariate analysis for outcomes equations OR if some important outcomes are subsequently omitted from the results or the significance and magnitude of important outcomes was not assessed.
6: Reporting bias - Justification	Analysis reporting bias - Justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
7: Other bias - Assessment	7 - Other risks of bias: Is the study free from other sources of bias?	1= Yes, 4 = No	Score “Yes” if the reported results do not suggest any other sources of bias. Score “No” if other potential threats to validity are present, and note these here (e.g. coherence of results, survey instruments used are not reported)	
7: Other bias - Justification	Other risks of bias - Justification	Open answer	Justification for coding decision (Include a brief summary of justification for rating, mentioning your response to all sub questions, cite relevant pages).	
8: External validity	8 - External validity	Open answer	Open answer- what do authors say about external validity, if anything?	

Appendix 3: Search strategy

The search for the impact evaluations included in this systematic review was implemented as part of the Evidence Gap Map on addressing root causes and drivers of irregular migrations (Berretta et al., forthcoming). This appendix summarizes that search strategy for the EGM. More details can be found in the EGM report along with an example of a search string.

Authors of the EGM followed two different approaches depending on whether the intervention domain had been explored recently by other evidence mapping efforts or not. For the former, the authors leveraged pre-existing search strategies, while for the later they devised a search strategy comprising key words and Boolean operators.

Updated searches

For the domain on strengthening resilience against shocks and stressors, numerous categories were taken from the Mapping evidence of what works to strengthen resilience to shocks and stressors (Berretta et al., 2022). To update the search, the following databases were used:

- CAB Abstracts (EBSCO)
- CAB Global Health (OVID)
- Africa-Wide (EBSCO)
- Academic Search Complete (EBSCO)
- APA PsycInfo (OVID)
- Web of Science (SSCI)
- Econlit (EBSCO)
- Social Science Research Network (SSRN)
- World Bank (EBSCO Discovery)
- Agris (EBSCO Discovery)
- RePEc (EBSCO Discovery)
- Campbell library

For the domain on violence prevention, numerous categories were taken from The effects of rule of law interventions on justice outcomes: an evidence gap map (Sonnenfeld *et al.*, 2023). To update the search, the following databases were used:

- Scopus
- Social Science Citations Index
- International Political Science Abstracts
- Communication & Mass Media Complete
- Research Papers in Economics (RePEc)

New search strategies

Two domains in the EGM by Berretta and colleagues (Forthcoming) had not been covered by previous EGMs: Economic opportunities and Orderly and safe migration management. Given the nature of interventions within those domains, reported changes in outcomes are expected to occur in a number of development sectors. As such, the strategy considered sector specific databases where appropriate. The following databases were searched using :

- Scopus
- Social Science Citations Index
- International Political Science Abstracts
- Research Papers in Economics (RePEc)
- CAB Abstracts
- Africa-Wide
- Academic Search Complete
- Web of Science
- Econlit
- Social Science Research Network (SSRN)
- World Bank
- Campbell library

Grey literature searches

Berretta and Colleagues (Forthcoming) searched for grey literature on the websites of 102 organizations. These organizations were selected on the basis of their action and work in migration related matters such as the International Organization of Migration (IOM), the Center for Migrant Studies, the Global Forum on Migration and Development, and IZA World of Labor, among others. Other website from referential international development and research organizations were also searched including Abdul Latif Jameel Poverty Action Lab (J-Pal), the United Nations Evaluation Group, the United States – Development Experience Clearing House, the AEA RCT Registry protocols, and others. A complete list of organizations and websites are presented in the appendix of the EGM report.

Other searches

Berretta and Colleagues (Forthcoming) also implemented forward and backward citation tracking of included papers. The authors used the software Publish and Perish and Citation Tracer to facilitate this search. In backward citation tracking, they reviewed eligible studies from the bibliographies of included studies. Finally, a public call for relevant papers was published via blog.