Evidence-informed Policymaking (EIPM) Measures Inventory – User Guide

Contents

Figures

Tables

3ie project team

Shannon Shisler, Ashiqun Nabi, Tomasz Kozakiewicz, Etienne Lwamba, Sanghwa Lee, Megha Bhattacharyya, Pierre Marion, Suvarna Pande and Christine Kelly

Advisory group

Name	Role	Affiliation
Marie Gaarder	Co-Academic Director	3ie
Thomas Kelly	Co-Academic Director	3ie
Andy Hinsley	FCDO lead	FCDO
Yaw Ansu	Advisory group	Development Bank of Ghana
Michael Donovan	Advisory group	Penn State University
James Georgalakis	Advisory group	IDS
Rob Lloyd	Advisory group	ITAD
Laurenz Mahlanza-Langer	Advisory group	Pan-African Collective for Evidence (PACE)
Kathryn Oliver	Advisory group	London School of Hygiene and Tropical Medicine
Pieter Serneels	Advisory group	University of East Anglia
Louise Shaxson	Advisory group	INASP
Birte Snilstveit	Advisory group	3ie
Maroof Syed	Advisory group	Centre for Economic Research in Pakistan (CERP)

About this guide

Measuring the use of evidence in policymaking involves examining the characteristics and capacities of those who use evidence, the pathways and processes through which evidence is translated, and the outcomes linked to evidence-informed policymaking (EIPM). While various measurement tools have been created for this purpose, finding the right tool can often prove difficult due to the lack of organised databases.

This *EIPM Measures Inventory* systematically gathers measurement tools relevant for all aspects of EIPM. Measures are mapped onto the high-level conceptual framework of evidence use in policy developed by a team from the FCDO Research Commissioning Centre (RCC) consortium, 2025. The framework conceptualises the links among evidence actors (producers, users, intermediaries and other key stakeholders), pathways of and conditions that moderate evidence translation, and outcomes of evidence use within policy contexts.

The inventory organises the measurement tools by constructs associated with each component of the conceptual framework (see the separate user guide and narrative report for the conceptual framework for information on its components here. This allows users to choose the right tools to measure the elements or pathways of evidence use that they are interested in. The inventory can also be used as a standalone methodological resource to explore the properties and characteristics of EIPM measures.

This guide offers an overview of the inventory and how it can be used, and information on the selection, characterisation, and organisation of the included measures.

Key definitions

- **Measure**: A tool or instrument used to collect data on a particular outcome area or construct (such as the capability to use evidence) for the purpose of assessing the effect of an intervention, policy, or programme, or to measure the conditions that support EIPM (e.g., characteristics of policymakers or organisations, relationships between policymakers and knowledge brokers, perceptions about the quality of evidence, etc.)
- **Construct**: A concept of interest to be measured.
- **Indicator:** A way to measure and monitor a given milestone, outcome, or construct.
- **Reliability**: The extent to which a measure consistently measures the same thing each time. For detailed definitions of specific types of reliability, see Appendix 1.
- **Validity**: The extent to which a measure adequately captures the construct of interest. For detailed definitions of specific types of validity, see Appendix 1.
- **Evidence**: Reflecting the role of the FCDO and RCC in commissioning research, our focus for the conceptual framework was primarily on evidence generated from research, by which we mean a systematic investigative process employed to increase or revise current knowledge. The measures inventory

adopted a corresponding focus on research evidence, but included tools referring more broadly to 'evidence' without further definition.

- **Evidence-informed policymaking**: Refers to the integration of evidence into policymaking at various levels or branches of governance and in various aspects of the policy process, from agenda setting and policy formulation to policy decision-making and implementation. We conceive of policy broadly to include governmental strategy or national planning, public sector investment decisions, programme implementation, legislation, reform, regulation, or similar.
- **Policymakers:** For the purpose of this inventory, policymakers refer to any individual working in a government department at any level of government (i.e., national and sub-national), including elected officials or civil servants, that either could or should contribute to a policy process. We also include individuals working in multilateral organisations, such as agencies and funds in the United Nations system, the World Bank and Inter-American Development Bank, among others.
- **Barriers to evidence use:** Factors that hinder the integration of evidence into policy processes, such as time constraints, lack of research skills, or political pressures.
- **Enablers of evidence use:** Factors that facilitate the use of evidence in policymaking, such as strong researcher-policymaker relationships, clear communication of findings, or supportive organisational cultures.
- **Evidence intermediaries:** Following Breckon and Boaz (2023), these are individuals or organisations for whom evidence is central to their everyday mission. They operate with closeness to government and facilitate the transfer, exchange, and translation of knowledge between researchers and policymakers.
- **Evidence ecosystem:** The interconnected network of actors (e.g., researchers, policymakers, practitioners), institutions (e.g., universities, think tanks, government agencies), and processes (e.g., research production, synthesis, dissemination) involved in generating and using evidence for policy decisions.

About the EIPM Measures Inventory

What's in the inventory

The ***EIPM Measures Inventory*** presents a rich database of measures and related guidance to help explore and choose the right tools for tracking and assessing the use of research evidence in policymaking.

The inventory provides detailed information on each measure, including validity, reliability, and example use cases. This helps users choose the most suitable tool for their needs.

How we developed the inventory

We searched for studies on evidence use from selected sources (see details in the section below on search strategy and Table 1). We identified relevant studies based on specific eligibility criteria (see Table 2). From each eligible study, we identified the measure(s) of evidence use and extracted key characteristics related to what behaviour or construct is measured, where the measure has been used, whether there is evidence that the measure is valid, among other information (see Table 3).

A panel of internal and external experts provided guidance on identifying sources for finding studies, developing eligibility criteria and deciding which characteristics to extract.

Where and how to access the inventory

The inventory is available directly through this link: https://eipm-measures.softr.app/

In the future, the inventory will also be accessible through an interactive conceptual framework, allowing users to click on a component or construct in the conceptual framework and be directed to available measures of that domain.

How to navigate the inventory

Measures Database - landing page: As shown in Figure 1, at the top of the landing page the user can find an overview of the three ways the measures can be explored: by conceptual framework element, by measure, by underlying study. The resources page also refers users to additional guidance and documentation.

The page also includes frequently asked questions whose answers will become visible to the user after clicking on a question. At the bottom of the page is a Google form link that allows users to submit feedback about the inventory but also suggestions for measures, papers or resources that are currently missing.

Figure 1: Measures database landing page

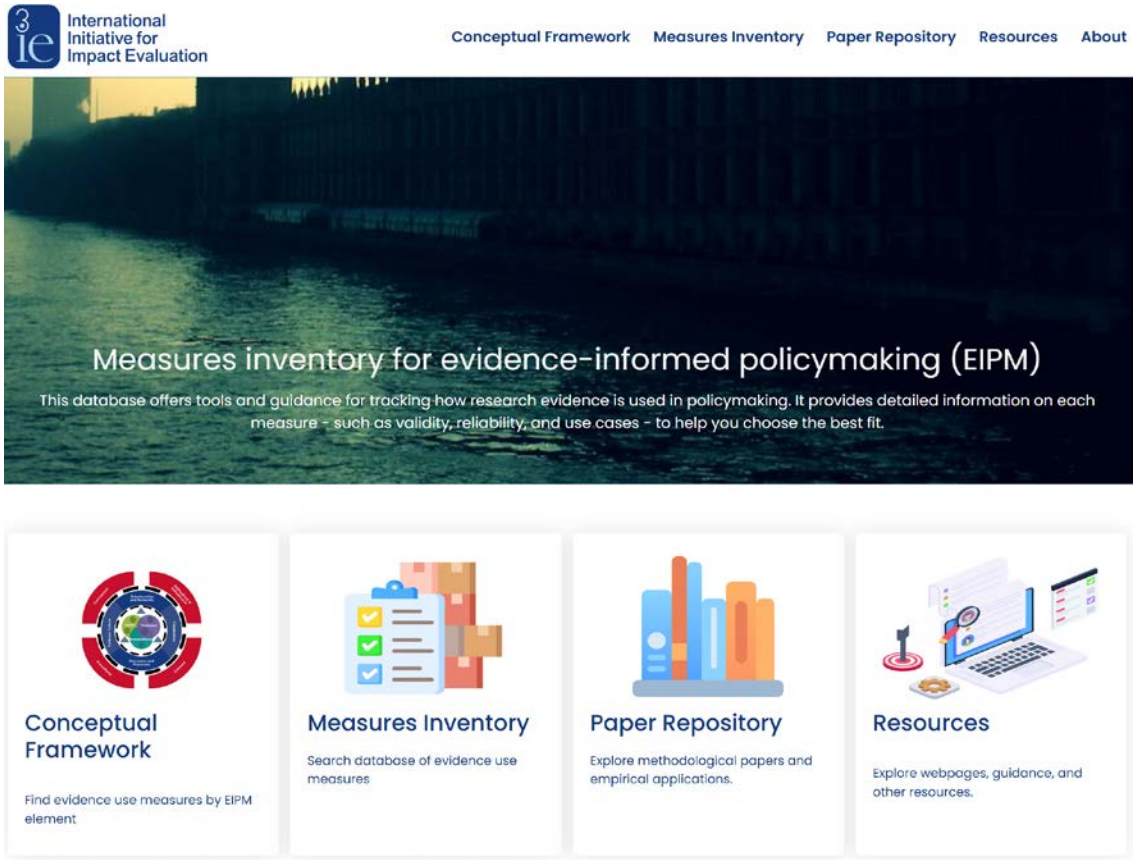
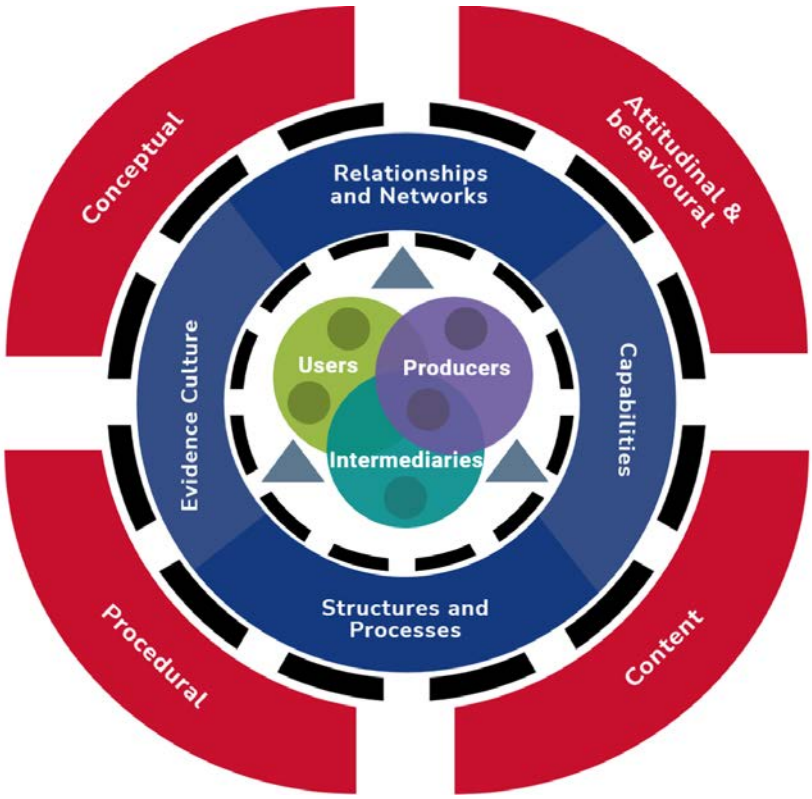


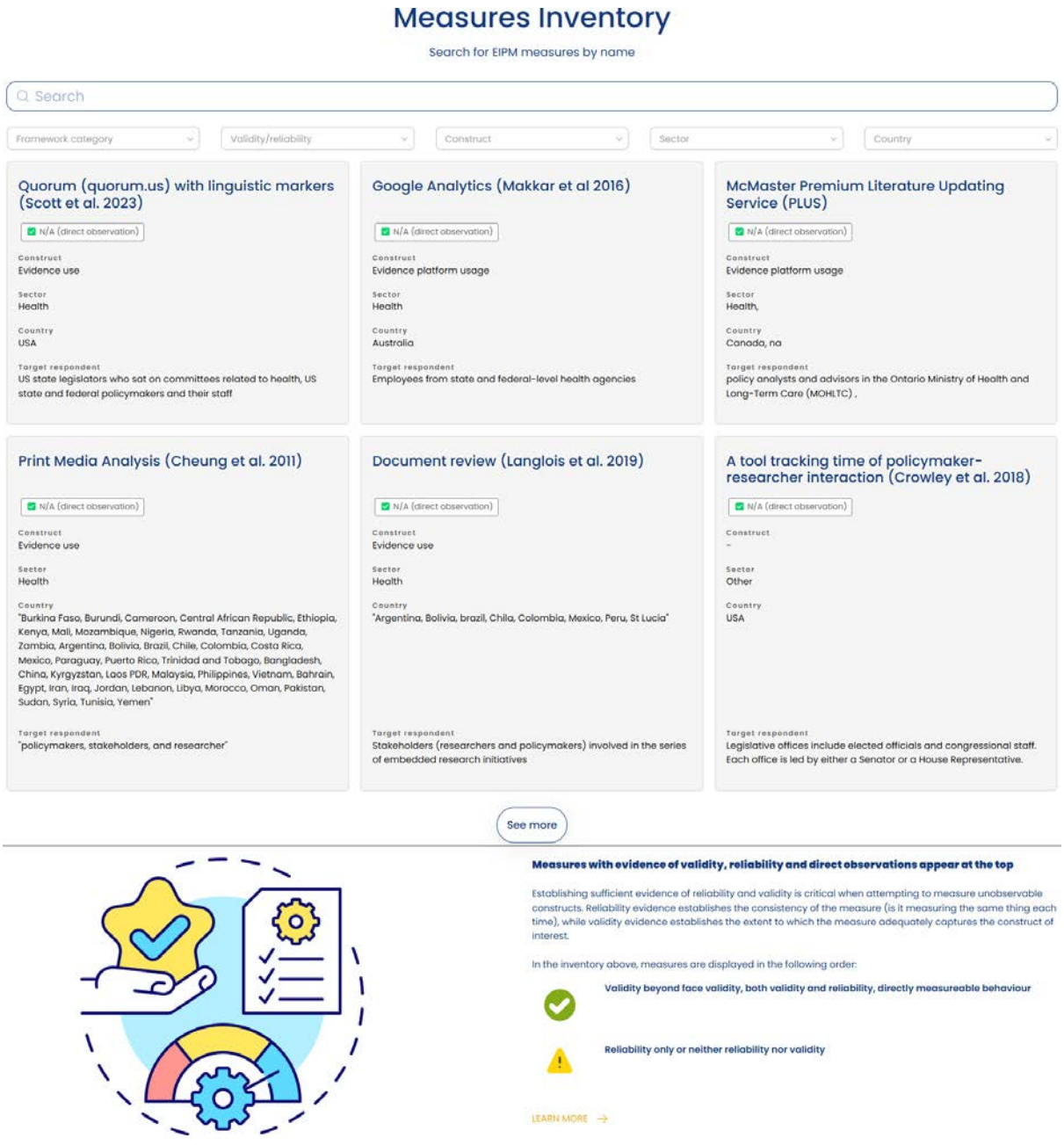
Figure 2: EIPM conceptual framework



Measures Database – 'Conceptual Framework' page: This option directs users to the EIPM conceptual framework developed by a team from the FCDO RCC consortium (see Figure 2). We recommend referring to the conceptual framework to help identify constructs of interest for measurement and associated measurement tools for each.

Measures Database – 'Measures Inventory' page: As shown in Figure 3, the inventory lists all the EIPM measures as tiles that include basic information about the measure and previous use cases. See Table 3 below for the full list of measure characteristics. For measures appearing in multiple papers / use cases, all the respective options are shown.

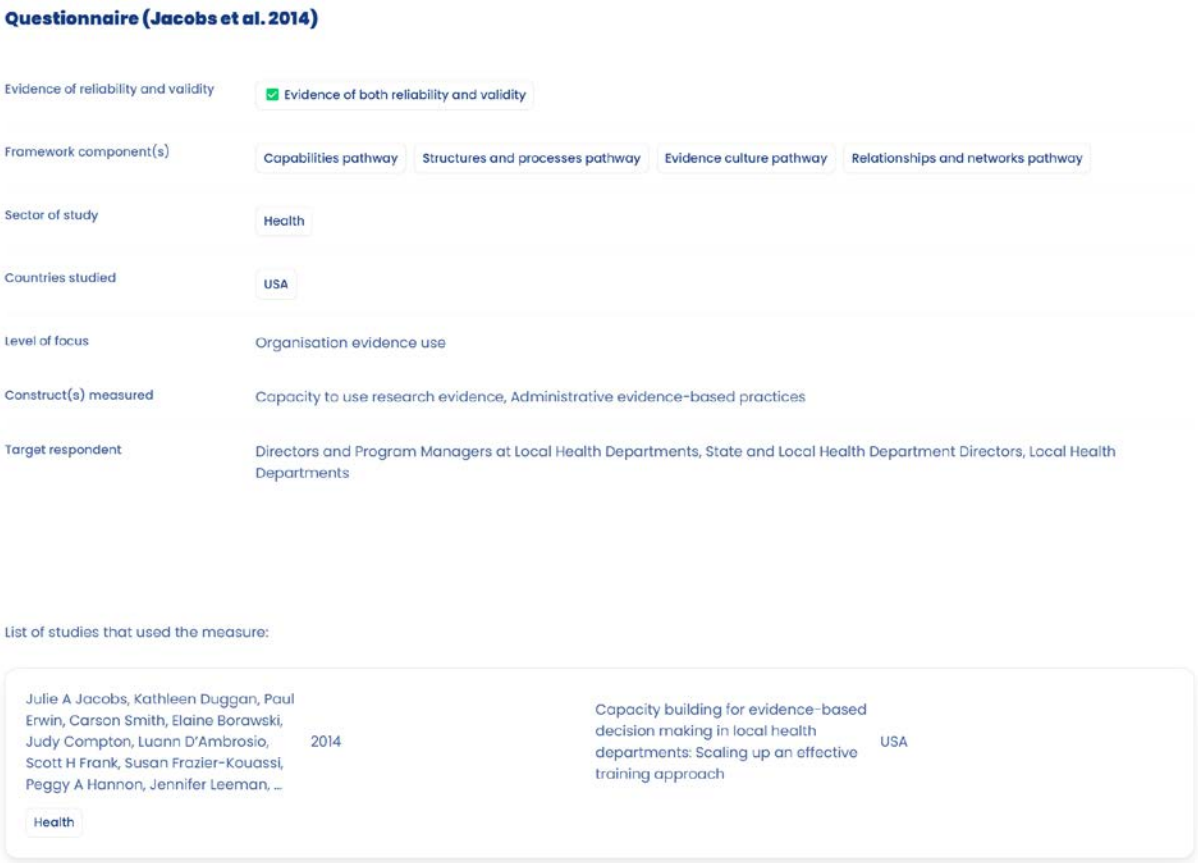
Figure 3: Measures inventory page



The inventory allows users to filter the measures displayed by the conceptual framework category, construct name, sector, country, and evidence of validity/reliability. Users can filter on multiple characteristics simultaneously, or multiple items within a characteristic, by selecting all the items of interest. By default, measures are sorted by evidence of validity/reliability, prioritising those with sufficient psychometric evidence or measures of directly observable behaviours (where psychometric evidence is not needed) – shown with a tick mark icon. See the section below on 'Evidence of reliability and validity' for more information on how this was assessed.

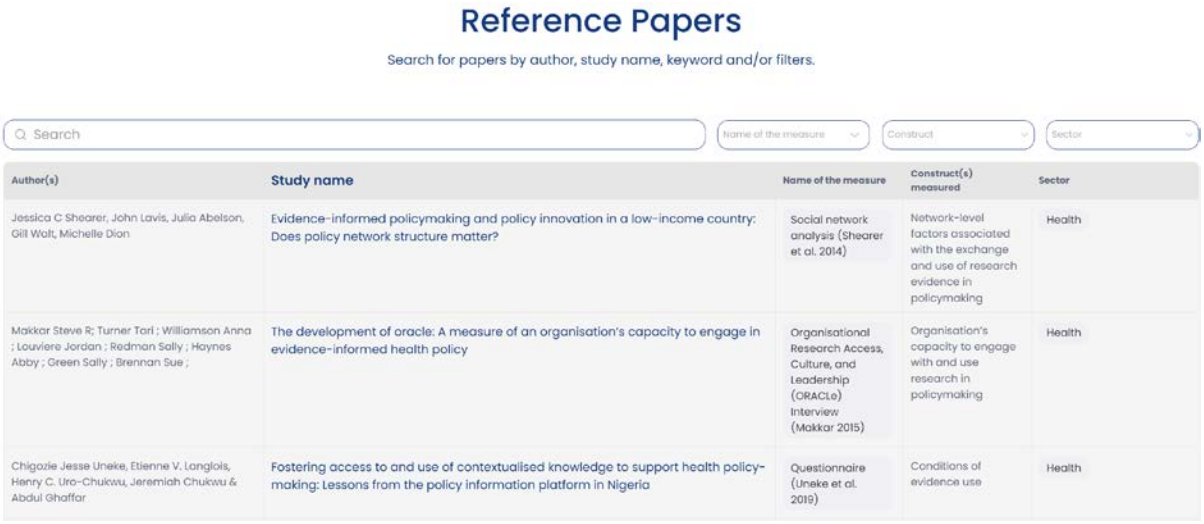
Measures Database – 'Individual measure' page: Clicking on a measure tile in the 'Measures inventory' page (Figure 3) directs the user to the individual measure page (Figure 4). This page provides more detailed information about the measure, including the construct(s) it measures and validity and reliability assessments. At the bottom of the page is a list of all identified studies that introduced or re-used the measure.

Figure 4: Individual measure page



Measures Database – 'Paper Repository': As shown in Figure 5, the Reference Paper repository lists all the studies that contained at least one of the EIPM measures listed in the inventory. It also allows users to filter by measure name, construct and sector.

Figure 5: The Reference Paper repository



Where and how did we search for the measures?

We relied on the following sources to identify potentially relevant measures of EIPM:

Existing reviews on evidence use

- The inventory of EIDM measurement indicators by Pan-African Collective for Evidence (PACE) (PACE 2024)
- PACE Evidence Gap Map of EIDM studies (Nduku et al. 2024a)
- PACE-3ie SRs of EIDM interventions (Nduku et al. 2024b)
- Reviews identified during scoping[1]
- Use of Research Evidence: A Methods Repository (https://uremethods.org)

Grey literature sources

- Relevant organisations and repositories (see Table 1 for full list)

Other sources

- Citation tracking
- Suggestions by experts

Table 1: List of grey literature sources

Organisation	**Link**
African Institute for Development Policy (AFIDEP)	https://www.afidep.org

[1] Christine et al. 2011, Combaz, Connor, and Georgalakis 2013, Gitomer and Crouse 2019, Hayter and Morales H. 2023, Jabali et al. 2024, Langer, Tripney, and Gough 2016, Lawlor et al. 2019, Milat and Li 2017, Oliver et al. 2014, Ouimet et al. 2024

Organisation	Link
Africa Centre for Systematic Reviews and Knowledge Translation (ACSRKT)	https://chs.mak.ac.ug/afcen
Alliance for Health Policy and Systems Research	http://www.who.int/alliance-hpsr/en
Asian Development Bank	https://www.adb.org
Centre for Evidence and Implementation	https://www.ceiglobal.org
Centre for Global Development	https://www.cgdev.org
Centre for Science and Policy	https://www.csap.cam.ac.uk
Coalition for Evidence-Based Policy (US)	http://coalition4evidence.org
eBase Africa	https://www.ebaseafrica.org
EPPI Centre, UCL Institute of Education, University College London (UK)	https://eppi.ioe.ac.uk/cms
Hewlett Foundation	https://hewlett.org
Innovations for Poverty Action (IPA)	https://www.poverty-action.org
Inter-American Development Bank	https://www.iadb.org
International Initiative for Impact Evaluation (3ie)	https://www.3ieimpact.org/evidence-hub
International Network for Government Science Advice (INGSA)	https://www.ingsa.org
itad	https://www.itad.com
National Bureau of Economic Research, USA	https://www.nber.org
Observatory for Public Sector Innovation	https://oecd-opsi.org/bi-projects
OECD	https://www.oecd.org/southafrica
On Think Tanks	https://onthinktanks.org
Overseas Development Institute (ODI)	https://odi.org/en
Results for Development	https://r4d.org
Strengthening Evidence use for Development Impact (SEDI), Oxford Policy Management	https://www.opml.co.uk/projects/strengthening-the-use-of-evidence-for-development-impact

Organisation	Link
The Evidence-informed Policy Network (EVIPNet)	httpshttps://www.who.int/initiatives/evidence-informed-policy-network
USAID Evaluations Clearinghouse	http://dec.usaid.gov
West Africa Capacity-building and Impact Evaluation (WACIE), 3ie	https://www.3ieimpact.org/research/west-africa-capacity-building-and-impact-evaluation
William T Grant Foundation	https://wtgrantfoundation.org
World Bank Evaluation office	https://ieg.worldbankgroup.org
World Health Organization (WHO)	https://www.who.int

Note: We searched the following keywords in these sources: evidence-based policy, evidence-informed policy, knowledge translation, evidence use, barriers (to evidence use), and facilitators (of evidence use). We were unable to search an additional list of identified grey literature sources due to time constraints.

What measures are eligible?

We used the following eligibility criteria to decide whether to include a measure in the inventory:

Table 2: Eligibility criteria for measures

Criteria	Description
Population	• Developed for and/or used in any country. • Focuses on measuring evidence used by public policymakers. • Can include other evidence actors only if policymakers' evidence use is clearly distinguishable/ reported separately • Excludes measures assessing donors', funders', or knowledge brokers' performance in producing or promoting evidence. • Excludes evidence used by practitioners (e.g., doctors, nurses, teachers, etc.) • Covers evidence use at individual, group, or organisational levels.
Aspects of evidence use	• Measures of evidence use and / or conditions that facilitate (or hinder) evidence use, including capability, opportunity and motivation to use evidence / facilitating conditions • Designed for diagnosis of capacity for evidence use, tracking evidence use over time, or evaluating interventions that promote evidence use by policymakers.

Criteria	Description
	• Identified measures were mapped onto one or more of the following components from the EIPM conceptual framework: o Outcomes of evidence use ▪ Conceptual use (evidence reshapes intellectual frameworks without immediate policy changes) ▪ Changes in policy content (evidence informs new or revised policies, plans, and strategies) ▪ Changes in policy procedures (evidence enhances policy development, debate, decision-making, and implementation) ▪ Changes in attitudes/behaviours (evidence shifts attitudes and behaviours of decision-makers and stakeholders). o Pathways of change ▪ Establishing supportive structures (institutional mechanisms like policies, processes, and infrastructures) ▪ Creating evidence culture (values, beliefs, norms shaping evidence use) ▪ Building relationships (connections among evidence producers, intermediaries, users, and stakeholders), ▪ Strengthening capacity (enhancing abilities to generate, interpret, and apply evidence). o Actors' attributes / baseline characteristics (features of evidence ecosystem actors) and; o Moderating/contextual factors (barriers and facilitators influencing evidence generation, communication, and use).
Topic / policy focus area	• Measures developed for use across policy areas or a specific policy area.
Language	• Primary focus on measures available in English. • Measures in French, Spanish, Italian, Portuguese, Arabic, and other languages are included when in-house multilingual resources are available.
Timeframe	• Any, until November 2024.

We did not include measures or methods for research funders or researchers to track the use, impact or quality of individual pieces or programmes of research evidence, such as impact evaluations or systematic reviews. For example,

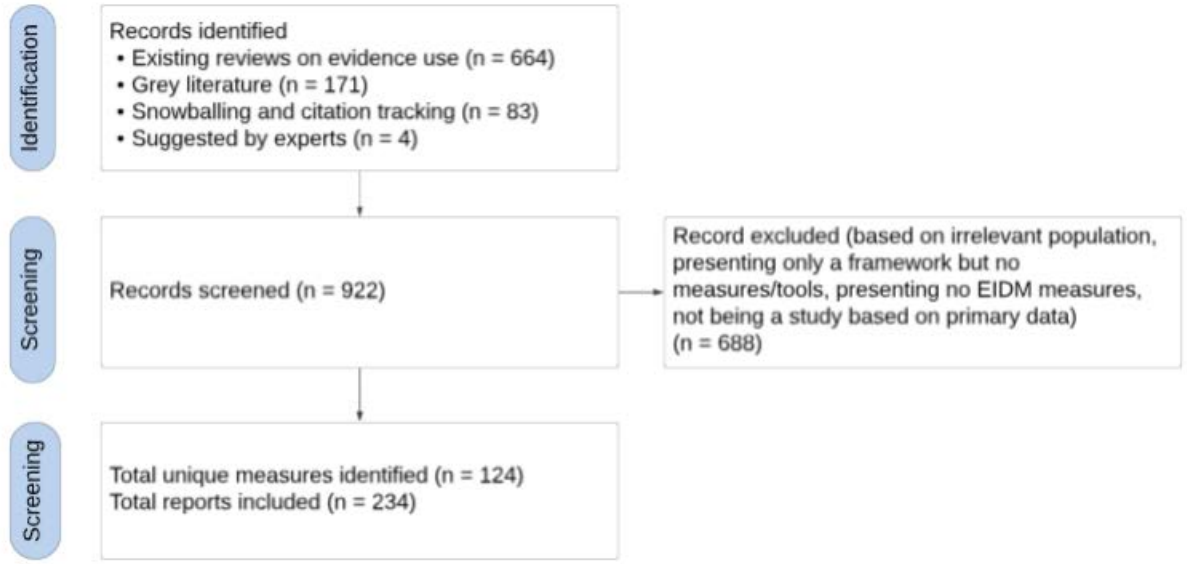
- The Research Excellence Framework (REF), the UK system for assessing the quality of research in UK higher education institutions, and associated outcomes of such assessment.
- Studies using contribution analysis as a method of tracking the impact and use of a piece of research.

Where identified papers reported use of an existing EIPM measure, we used citation tracking to identify the original source and characteristics of the measure. Similarly, if the process for developing and validating a measure was reported separately to its practical application, we sourced the related methodological information.

What did we find?

We screened over 900 records for potential inclusion (Figure 6) and ultimately included 124 measures of EIPM.

Figure 6: Number of studies identified, selected and included in the database



What characteristics of the measures did we document?

We extracted the following characteristics of each measure (Table 3)

Table 3: Extracted characteristics of measures

Characteristics of the measure	Description
Name of the measure	The official name or the name created by the author, if available. In other cases, a descriptive name entailing the measure type and the authors' names/year.
Constructs measured	We use this term broadly to represent the concept, phenomenon, outcome, or result being measured (either through direct observation or inferred from responses or indicators).

Characteristics of the measure	Description
Type of measure	The method or instrument used to collect data, such as questionnaires, interviews, observational tools, and others.
Level of focus of the measure	• Individual level: focusing exclusively on an individual's view, attitude or action. • Group level: focusing on a group/ team/ unit/ department within an organisation. • Organisational level: focusing exclusively on organisation's view, attitude or action (even if the respondent is an individual representing the organisation).
Reported evidence of validity and reliability	Where relevant (e.g., the behaviour is not directly observed and thus psychometric evidence is crucial to ensure that the construct is accurately captured), we indicate whether the measure has been through a process to assess its validity (e.g., content validity, construct validity, criterion-related validity, face validity, ecological validity, internal validity, external validity, incremental validity) and reliability (e.g., test-retest reliability, inter-rater reliability, parallel-forms (or alternate-forms) reliability, internal consistency, intra-rater reliability, generalizability reliability). See Appendix 1 for detailed descriptions of each type of validity and reliability.
Evidence of reliability and validity (psychometric evidence)	We code for psychometric evidence based on the reported evidence of validity and reliability. 0 – No evidence of validity or reliability 1 – Evidence of reliability 2 – Evidence of validity (beyond face validity) 3 – Evidence of both reliability and validity N/A – the measure cannot be tested for psychometric properties (e.g., the behaviour it measures is directly observable)
Sectors where the measure is used	Sectors in which the measure was used in the study (e.g., agriculture, fishing and forestry, education, energy and extractives, financial sector, health, social protection, industry, trade and services, information and communication, public administration, transportation, water, sanitation and hygiene, and others)
Countries where the measure is used	Country where the measure is used in the study
Target participants of the studies using the measure	Population/participants of the study that employed the measure

Key features of the measures at a glance

124 measures included in the database

238 studies analysed

Figure 7: Measures by conceptual framework components

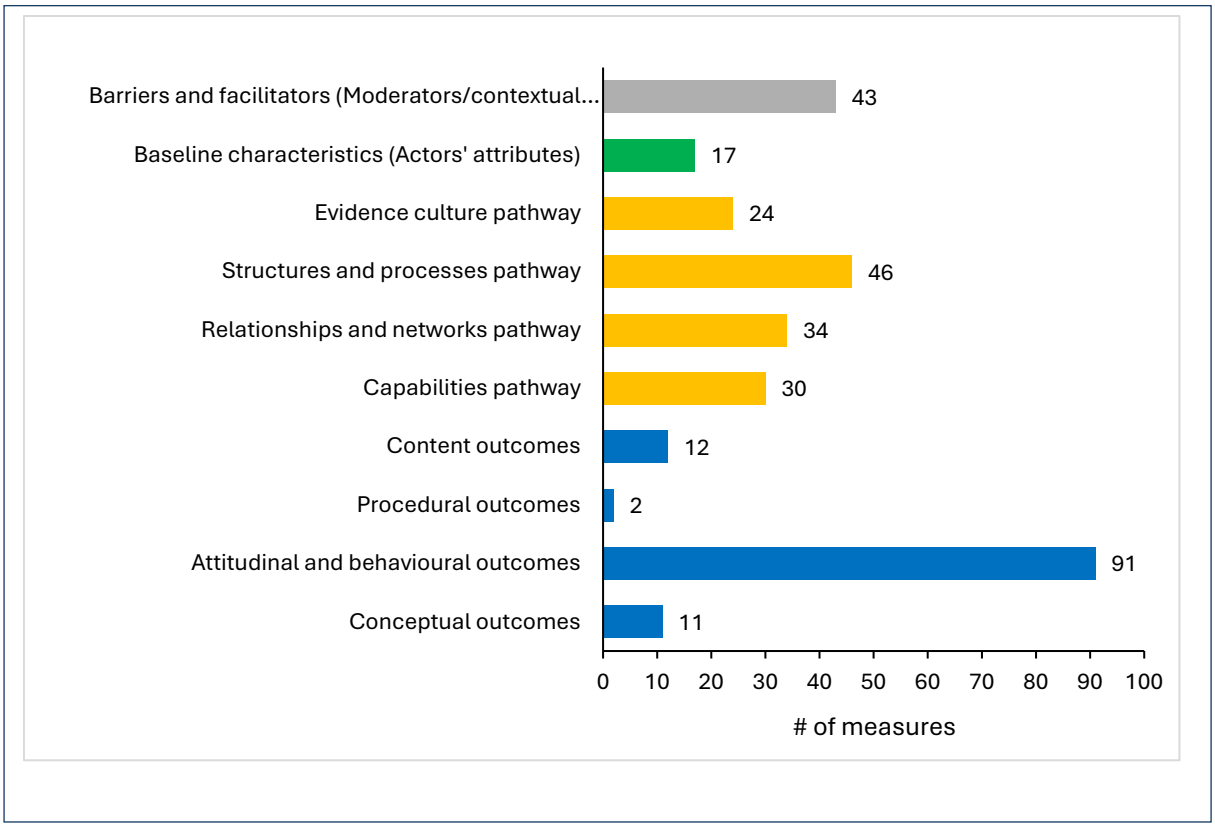


Figure 8: Types of measures

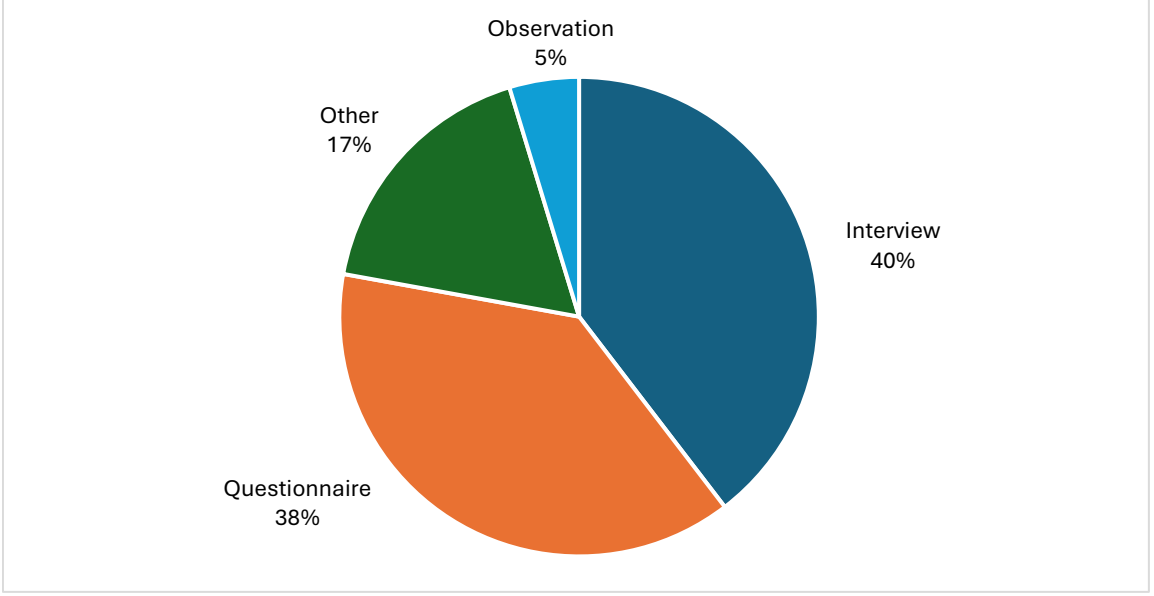


Figure 9: Sectors of measures use

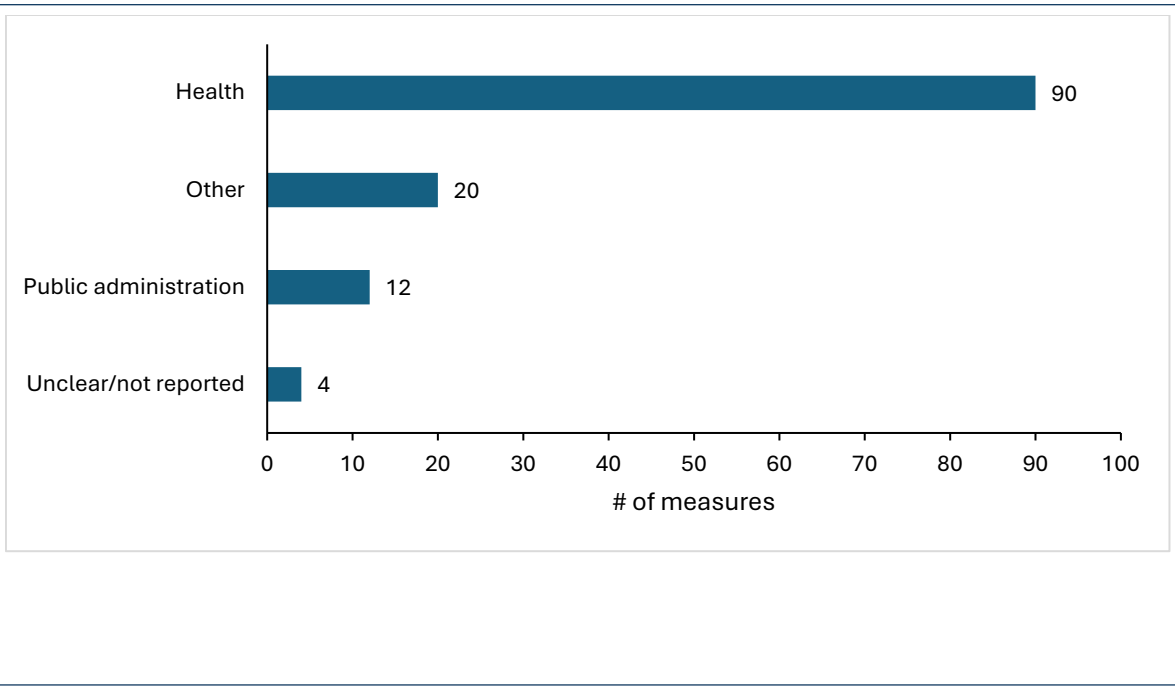


Figure 10: Levels of evidence use targeted by measures

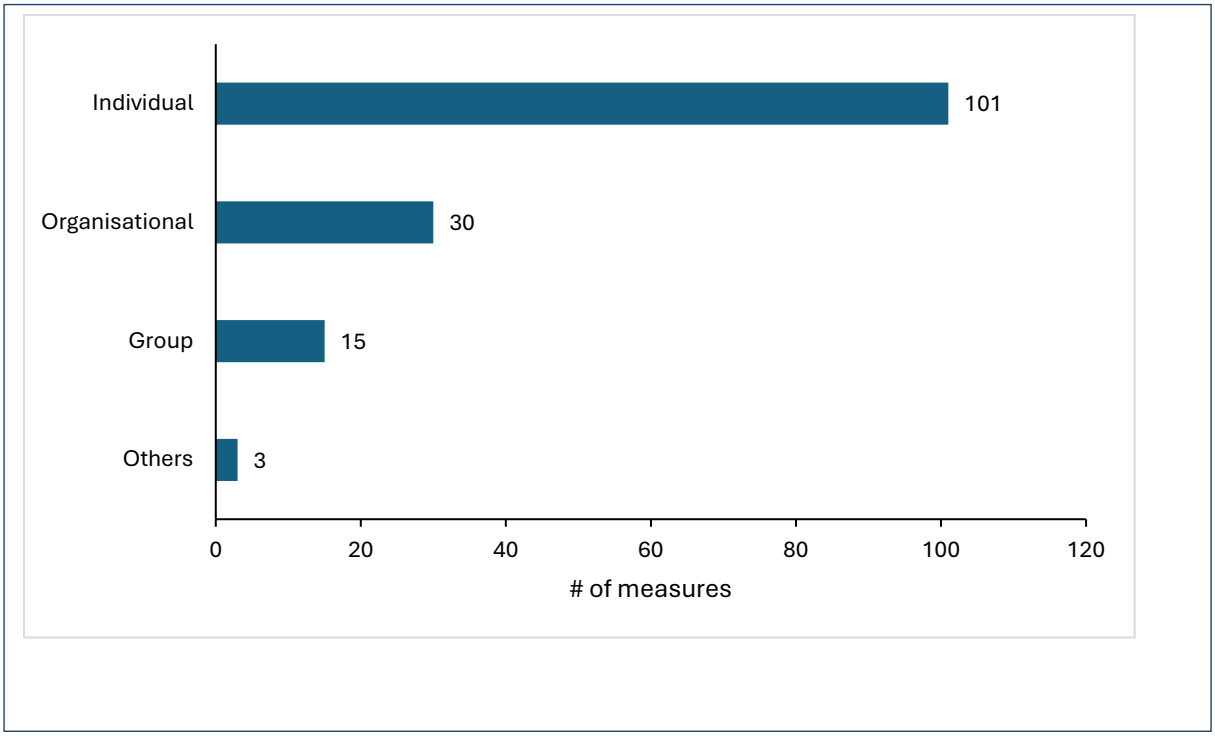


Figure 11: Countries where measures were used

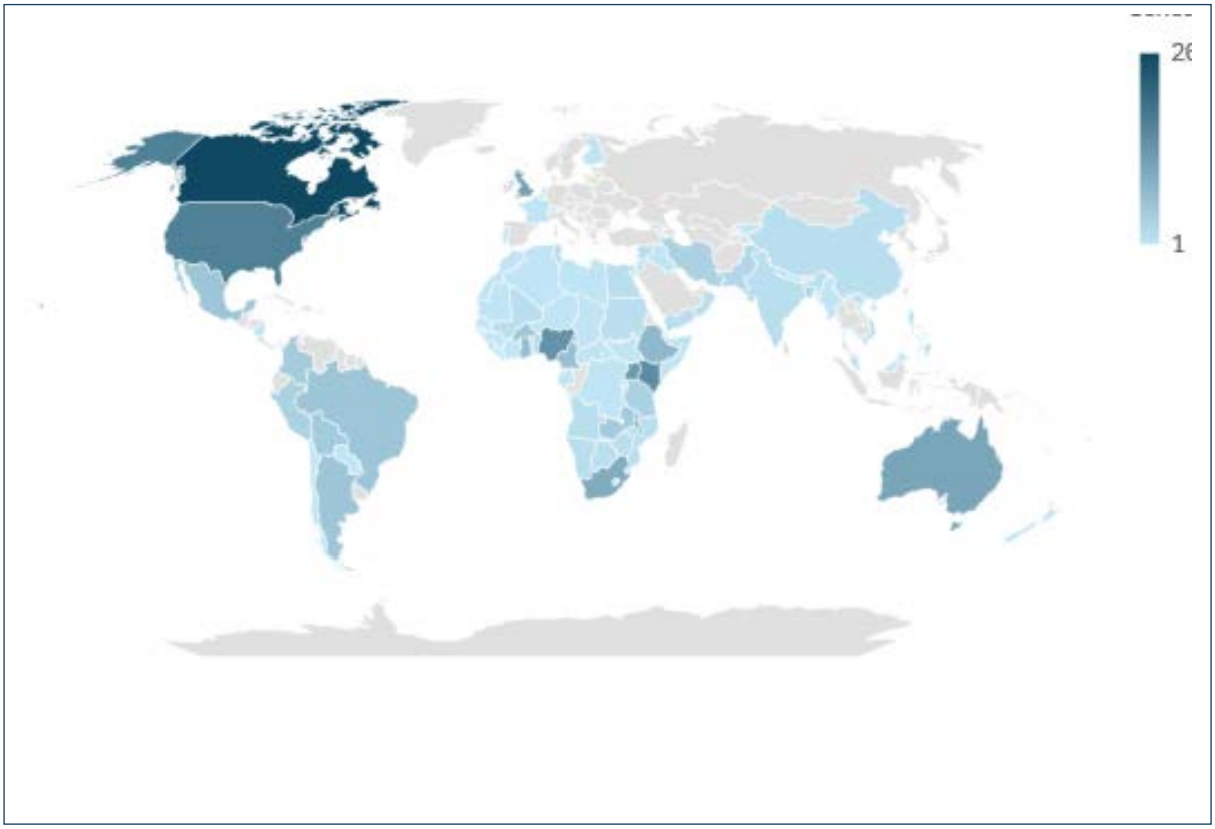


Table 4: Top 10 countries of measures use

Countries	No. of measures
Canada	26
USA	17
Global/ unspecified	16
Kenya	15
Nigeria	14
Uganda	13
UK	12
Australia	11
South Africa	10
Malawi	10

Evidence of reliability and validity

Establishing sufficient evidence of reliability and validity is critical when attempting to measure unobservable constructs. Reliability evidence establishes the consistency of the measure (is it measuring the same thing each time), while validity evidence establishes the extent to which the measure adequately captures the construct of interest. A measure can be reliable without being valid (e.g., it is measuring the same thing each time, but it's measuring the *wrong* thing). However, a measure cannot be valid without also being reliable. Thus, we have established evidence of validity as the minimum requirement needed to recommend a measure for use in future studies. The exception to this rule is studies with *only* evidence of face validity, which is insufficient to confirm validity[2]. Measures with only validity evidence (beyond face validity), or with evidence of both validity and reliability, both meet the minimum criteria and are displayed with a green tick icon in the measures inventory. For measures with no evidence of either validity or reliability, or for measures with only reliability evidence, we recommend that further psychometric validation take place prior to using the measure in future research. This is signalled to users by a yellow warning sign on the relevant measure tile.

This requirement does not apply to measures of directly observable behaviour (e.g., counts of web clicks, amount of time spent interacting with an evidence platform, etc.), and thus will show an N/A in this field.

The reliability of qualitative interviews was coded based only on the methodology used to create the interview questions/interview guides. While some papers present information on the reliability and validity of their qualitative *analysis*, this would not be relevant to the reliability and validity of the interview itself. Most typically, for qualitative

[2] Face validity refers to the degree to which a test appears to measure what it is supposed to measure, based on a superficial assessment. Though not a rigorous scientific form of validity, it can affect the test-taker's perceptions and motivation and has utility as an initial check of items and is appropriate in the earliest stages of test development (but is not sufficient evidence to establish the psychometric properties of the test).

interviews, the process for establishing the validity of the questions in the interview involves a thorough literature review and consultations with experts to devise potential questions, then a testing of those questions, followed by revisions, again in consultation with experts.

Table 5: Measures by conceptual framework components and evidence of validity and reliability

Conceptual framework themes	ToC components	Total # of measures identified	# of measures with sufficient psychometric evidence*	# of directly observable measures ‡
EIPM outcomes	Conceptual	11	5 (45%)	0
	Attitudinal and behavioural	91	34 (37%)	7
	Content	2	0 (0%)	1
	Procedural	12	3 (25%)	5
EIPM pathways of change	Structures and processes	30	12 (40%)	2
	Evidence culture	34	11 (32%)	2
	Relationships and networks	46	16 (35%)	3
	Capabilities	24	12 (50%)	2
Actors' attributes	Baseline characteristics	17	8 (47%)	0
Moderating/ contextual factors	Barriers and facilitators	43	18 (42%)	1

*Notes: * Only applies to measures that do not measure directly observable behaviours. We consider sufficient psychometric evidence to be (i) evidence of both reliability and validity or (ii) evidence of validity beyond face validity. ‡ These directly observable measures (such as, counts of website visits, time tacking, or document review) cannot be tested for psychometric properties.*

Feedback and omissions

If there are any measures that meet the eligibility criteria for the inventory that are not included here, or any published psychometric evidence for existing measures that has not been identified, please let us know via this Google form.

References

Christine, Clar, Campbell Susan, Davidson Lisa, and Graham Wendy. 2011. “What Are the Effects of Interventions to Improve the Uptake of Evidence from Health Research into Policy in Low and Middle-Income Countries?” UK Foreign, Commonwealth & Development Office (FCDO).

Combaz, Emilie, Melanie Connor, and James Georgalakis. 2013. “Knowledge Translation in the Global South: An Exploratory Mapping of the Literature.” IDS Working Paper 592. Brighton: Institute of Development Studies. https://doi.org/10.19088/IDS.2023.033.

Gitomer, Drew, and Kevin Crouse. 2019. “Studying the Use of Research Evidence: A Review of Methods.” William T. Grant Foundation Monograph. New York, NY: William T. Grant Foundation.

Hayter, Emily, and Marcela Morales H. 2023. “Review Policy Labs and Evidence Use in Education.” Zurich: Jacobs Foundation.

Jabali, Seyyed Hadi, Shahram Yazdani, Hamid Pourasghari, and Mohammadreza Maleki. 2024. “From Bench to Policy: A Critical Analysis of Models for Evidence-Informed Policymaking in Healthcare.” *Frontiers in Public Health* 12 (March). https://doi.org/10.3389/fpubh.2024.1264315.

Langer, Laurenz, Janice Tripney, and David Gough. 2016. “The Science of Using Science Researching the Use of Research Evidence in Decision-Making.” London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London. https://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/Science%202016%20Langer%20report.pdf?ver=2016-04-18-142701-867.

Lawlor, Jennifer, Kristen Mills, Zachary Neal, Jennifer Watling Neal, Camren Wilson, and Kathryn McAlindon. 2019. “Approaches to Measuring Use of Research Evidence in K-12 Settings: A Systematic Review.” *Educational Research Review* 27 (June):218–28. https://doi.org/10.1016/j.edurev.2019.04.002.

Milat, Andrew J., and Ben Li. 2017. “Narrative Review of Frameworks for Translating Research Evidence into Policy and Practice.” *Public Health Research & Practice* 27 (1): 2711704. https://doi.org/10.17061/phrp2711704.

Nduku, P., Ategeka, J., Madonsela, A., Mdlalose, T., Galada, T., Mutanha, T., Kuchenmüller, T., et al. 2024a. The Art and Science of Using Evidence: An Evidence Map. Technical Report. in press.

Nduku, P., Stevenson, J., Ategaka, J., Mdlalose, T., Mutanha, T., Shisler, S., Pande, S. & Mahlanza-Langer, L. 2024b. What works to increase the use of evidence for policy decision-making: A systematic review. New Delhi: International Initiative for Impact Evaluation.

Oliver, Kathryn, Simon Innvar, Theo Lorenc, Jenny Woodman, and James Thomas. 2014. “A Systematic Review of Barriers to and Facilitators of the Use of Evidence by

Policymakers.” *BMC Health Services Research* 14 (1): 2. https://doi.org/10.1186/1472-6963-14-2.

Ouimet, Mathieu, Morgane Beaumier, Adrien Cloutier, Alexandre Côté, Éric Montigny, François Gélineau, Steve Jacob, and Stéphane Ratté. 2024. “Use of Research Evidence in Legislatures: A Systematic Review.” *Evidence & Policy* 20 (2): 226–43. https://doi.org/10.1332/174426421X16656568731041.

PACE. 2024. “EIDM Measurement: An Interactive Inventory of Indicators.” Pan-African Collective for Evidence NPC. 2024. https://prezi.com/view/ybFXNH4XZHtPcK1mb7R8/.

Appendix 1

Glossary of definitions we used for coding the various types of validity and reliability, with examples.

Types of validity evidence: We looked for evidence of the following types of validity:

- **Content Validity**: Ensures that the test fully represents the domain or content it aims to measure.
    - For example, if a test is measuring mathematical ability, it should cover all relevant mathematical skills, not just a subset (e.g., cognitive tests).
    - Typically established by (1) defining the construct, (2) thorough literature review, (3) generate an item pool, (4) gather an expert panel/focus group, (5) experts provide feedback on the items, suggest revisions or removal.
- **Construct Validity**: Reflects how well a test measures the theoretical construct or concept it claims to measure. This type includes:
    - **Convergent Validity**: When a test correlates highly with other measures of the same construct that have already established validity (unlike concurrent validity, the measures *do not* have to occur at the same time).
        - For example, researchers might compare the results of a new measure of capability to use evidence with an existing validated measure of capability. They would then statistically analyze the results to ensure that the two measures of capability would have a high correlation with one another. This would confirm that these measures are indeed assessing the same construct.
        - Researchers may also mention conducting a factor analysis to see if the items load onto the expected factors that are supposed to represent the construct.
    - **Discriminant Validity**: When a test does not correlate with measures of unrelated constructs.
        - For example, researchers might look at the results of a measure of capability to use evidence and a separate measure of motivation to use evidence. They would then statistically analyze the results to ensure that the measure of capability would have a low correlation with the measure of motivation. This would confirm that these measures are indeed assessing unique constructs.
    - **Criterion-Related Validity**: Evaluates how well one measure predicts an outcome based on another, external criterion. It has two subtypes:

    - **Predictive Validity**: The extent to which the test predicts future performance or behaviour.
        - For example, in our example paper, "*the average individual and/or consensus scores were higher for higher-end than lower-end research use organizations on every questionnaire item"* - in other words, the scores would be predictive of the level of research use in an organization
    - **Concurrent Validity**: The extent to which the test correlates with outcomes measured at the same time.
        - For example, if a new depression scale is being validated, concurrent validity would be established by administering the new scale and an established depression scale simultaneously to the same group of participants. A high correlation between the two scales would indicate strong concurrent validity.

- **Face Validity**: Refers to the degree to which a test appears to measure what it is supposed to measure, based on a superficial assessment. Though not a rigorous scientific form of validity, it can affect the test-taker's perceptions and motivation and has utility as an initial check of items and is appropriate in the earliest stages of test development (but is not sufficient evidence to establish the psychometric properties of the test).
    - Typically, researchers would share the test with a group of laypeople, such as potential test takers or colleagues, and ask for feedback on whether the items seem relevant and appropriate for measuring the construct.

- **Ecological Validity**: Assesses whether the results of a test can be generalized to real-world settings or everyday life (e.g., findings are applicable outside of the research environment)
    - One way to establish this type of validity is to ensure that the study conditions match the practical realities of the policy-making environment. For example, conduct the study in the government offices or policy making environments, ensure that participants capture the diversity in policy making contexts, structure tasks that mimic the type of decision policymakers actually make and align them with the actual process of EIPM, evaluate how real world constraints impact EIPM (e.g., time constraints, resource limitations, political pressures, etc.).

- **Internal Validity**: Concerns whether the design and methods of a study allow for accurate conclusions about the relationships between variables within the context of the study.

    - Elements assessed during risk of bias process (e.g., assignment mechanism, controlling for confounding, establishing a control group, implement blinding if possible, ensure similar timing of measurement of treatment and control groups, use valid and reliable outcome measures, address attrition, use appropriate statistical techniques).

- **External Validity**: Relates to the generalizability of the test results beyond the specific study or sample.

    - To establish external validity, demonstrate that participants are representative of the target population, conduct the study in multiple settings and with multiple populations.

- **Incremental Validity**: Examines whether the test adds meaningful predictive power or insight above and beyond other existing measures.

    - This would involve comparing a new measure to an existing measure and analyzing the extent to which the new measure adds unique explanatory power (e.g. the addition of the new measure significantly improves the model's ability to predict the outcome – this is typically done through a hierarchical regression analysis and is indicated by the change in the R^2 value).

Types of reliability evidence: We looked for evidence of the following types of reliability:

- **Test-Retest Reliability**: Assesses the consistency of a test over time. The same test is administered to the same group of people at two different points in time, and the scores are compared. High correlation between the two sets of scores indicates good test-retest reliability.

    - Potential tests include:

        - **Pearson's correlation coefficient (*r*):** the most commonly used statistic for assessing test-retest reliability, especially when the data is continuous.

        - **Intraclass Correlation Coefficient (ICC):** accounts for both agreement and consistency, considering the variability within and between subjects.

        - **Spearman's Rank-Order Correlation (ρ):** Used when data are ordinal or not normally distributed. Spearman's correlation assesses the consistency of rankings between two time points.

        - **Cohen's Kappa (κ):** Commonly used for test-retest reliability when data is categorical (e.g., diagnostic classifications).

        - **Percent agreement:** A simple measure of the proportion of identical responses between two administrations of a test.

- **Inter-Rater Reliability**: Evaluates the level of agreement between two or more independent observers or raters. This is critical when subjective judgments are involved (e.g., scoring essays or behavioural observations). High inter-rater reliability means that different raters give similar ratings for the same subject.
    - Potential tests include:
        - **Cohen's Kappa (κ):** suitable when there are only two raters and the data are categorical (e.g., yes/no, diagnostic categories).
        - **Fleiss' Kappa:** extends Cohen's Kappa to situations where there are more than two raters.
        - **Intraclass Correlation Coefficient (ICC):** commonly used for interval or ratio data, particularly when there are more than two raters. ICC is preferred when the focus is on assessing consistency across multiple ratings.
        - **Krippendorff's Alpha (α):** Measures reliability across any number of raters and for any level of measurement (nominal, ordinal, interval, or ratio). Krippendorff's Alpha is versatile and can handle missing data, making it suitable for complex or real-world data where not all raters rate all items.
        - **Percent Agreement:** Often used as a basic measure of agreement, particularly in categorical data, where it's easy to interpret and calculate.
        - **Spearman's Rank-Order Correlation (ρ):** suitable when the data is ordinal and there are two raters who rank items.
- **Parallel-Forms (or Alternate-Forms) Reliability**: Assesses the consistency of the results of two equivalent versions of a tool designed to measure the same construct. The two forms are administered to the same group (preferably in a counterbalanced way that alternates the order of questions to minimize any order effects). Statistics to measure parallel forms reliability include:
    - **Pearson Correlation Coefficient (r):** most common statistic to assess parallel forms reliability.
    - **Intraclass Correlation Coefficient (ICC):** ICC is particularly useful when the two forms should produce nearly identical scores. This method can be especially relevant if the forms are intended to be interchangeable.

- **Mean Difference Analysis (Paired Samples t-Test):** This test is used alongside correlation measures to confirm that the two forms yield similar average scores. A non-significant result ($p > 0.05$) suggests that there is no substantial difference in scores between the two forms.
- **Bland-Altman Plot:** This graphical method assesses agreement between two measurement forms by plotting the differences between scores against their averages, identifying any systematic bias or limits of agreement.

- **Internal Consistency Reliability**: Examines the consistency of results across items within a measure. If all items on a test measure the same underlying construct, they should produce similar results. There are different ways to assess internal consistency:
  - **Cronbach's Alpha**: The most common measure, which reflects the average correlation among all items in a test.
  - **Split-Half Reliability**: Involves dividing the test into two equal halves (e.g., odd vs. even items) and correlating the scores of the two halves.
  - **Kuder-Richardson Formula (KR-20 or KR-21)**: Used for dichotomous (yes/no or true/false) items to measure internal consistency.
  - **Average Inter-Item Correlation:** This method calculates the average of all correlations between pairs of items, providing an overall estimate of internal consistency. This approach is especially useful for shorter tests or when it is preferable to see individual item correlations.
  - **McDonald's Omega (ω):** McDonald's Omega is an alternative to Cronbach's Alpha that does not assume equal contribution of all items, making it more accurate for tests with varying item loadings.
- **Intra-Rater Reliability**: Similar to inter-rater reliability but focuses on the consistency of the ratings made by a single rater over multiple occasions. High intra-rater reliability indicates that the same individual provides consistent assessments over time. Typical statistics include:
  - **Intraclass Correlation Coefficient (ICC):** widely used for measuring intra-rater reliability, especially with continuous or ordinal data, as it quantifies the consistency of ratings by a single rater across multiple instances.
  - **Cohen's Kappa (κ):** Cohen's Kappa measures the agreement between ratings made by the same rater on two different occasions when the data is categorical. It accounts for chance agreement.

- **Weighted Kappa:** Particularly useful when ratings are on an ordinal scale (e.g., a 5-point scale), where differences between adjacent categories are not as severe as differences between distant categories.

- **Pearson Correlation Coefficient (r):** Used for continuous data when the focus is on the consistency of the relationship, rather than exact agreement.

- **Spearman's Rank-Order Correlation (ρ):** measures the consistency of the rank order of ratings made by the same rater on two occasions, particularly useful for ordinal data.

- **Generalizability Reliability:** assesses how well observed scores generalize across different facets, which are aspects of measurement conditions that can vary and impact scores. Facets might include raters, test items, testing occasions, or settings, depending on the measurement context. While traditional reliability focuses on a single error source (e.g., test-retest or inter-rater reliability), generalizability theory examines multiple sources of error simultaneously, providing a more comprehensive view of reliability.

  - **Generalizability Coefficient (*g* or *G*):** quantifies the reliability of a measure across specified facets. A higher *g* value indicates greater reliability or consistency across the different facets, meaning the observed scores are more likely to generalize well across these conditions.