Session 3:
How to understand the results of evidence syntheses & meta-analysis

C4ED – EUTF

October 2024

Center for Evaluation and Development

European Commission

Objectives of Session 3

Recap of statistical fundamentals: Effect size, Confidence interval, Statistical significance

Interpretation of results: Reading forest plots, Reading funnel plots, Heterogeneity, Contextualizing findings, Evaluating robustness

Next steps: implications, dissemination and further research

Recap on statistical fundamentals

For more information, please refer to the slides from the previous workshops

Effect size

- Quantitative measure of the impact an intervention has on an outcome.
- Measured as regression coefficient, odds ratio, mean difference or risk ratio
- Shows whether the intervention increased or decreased the outcome variable: increase (+) or decrease (-)

This is not the same as whether the change is good or bad!!

Variable	Income coefficient	95% CI	p-value
TVET participation	0.05	[-0.18, 0.28]	0.67
TVET participation	0.25	[0.05, 0.45]	0.02

For more information, please refer to the slides from the previous workshops

Confidence interval (CI)

- Range with upper and lower bound within which the true effect lies with a probability of 95% (or other level if indicated as such)
- Effectively a measure of precision:
  - Narrow CI shows a precise estimate of effect size
  - Wide CI shows less certainty about the effect size
- If the CI includes zero, then the effect is statistically insignificant

Variable	Income coefficient	95% CI	p-value
TVET participation	0.05	[-0.18, 0.28]	0.67
TVET participation	0.25	[0.05, 0.45]	0.02

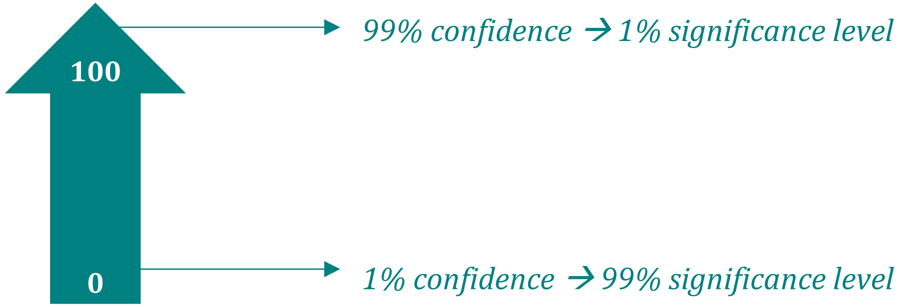
→ *More uncertainty about results*

→ *Less uncertainty about results*

For more information, please refer to the slides from the previous workshops

Statistical significance

- Statistical significance levels determine the point at which observed effects are unlikely to be due to chance alone
  - Typically, this likelihood/probability (significance level) is set at 5%
  - **Lower** significance levels imply **higher** confidence in results



99% confidence → 1% significance level

1% confidence → 99% significance level

  - Significance/Confidence levels reflected by asterisk (no *, no significance)

    Usually *, ** or *** in papers and reports for 10, 5 and 1% significance levels

Statistical significance contd.

- P-values reflect the probability that the observed effect is due to random chance
    - Example: A p-value of 0.09 means there is a 9% chance that the observed effect is due to random variation if the ‘true’ effect is zero.
- P-values are used to determine statistical significance
- Statistical significance of results is reflected by a p-value < significance level

Variable	Income coefficient	95% CI	p-value
TVET participation	0.05	[-0.18, 0.28]	0.67
TVET participation	0.25	[0.05, 0.45]	0.02

→ *Insignificant at 5% significance level*

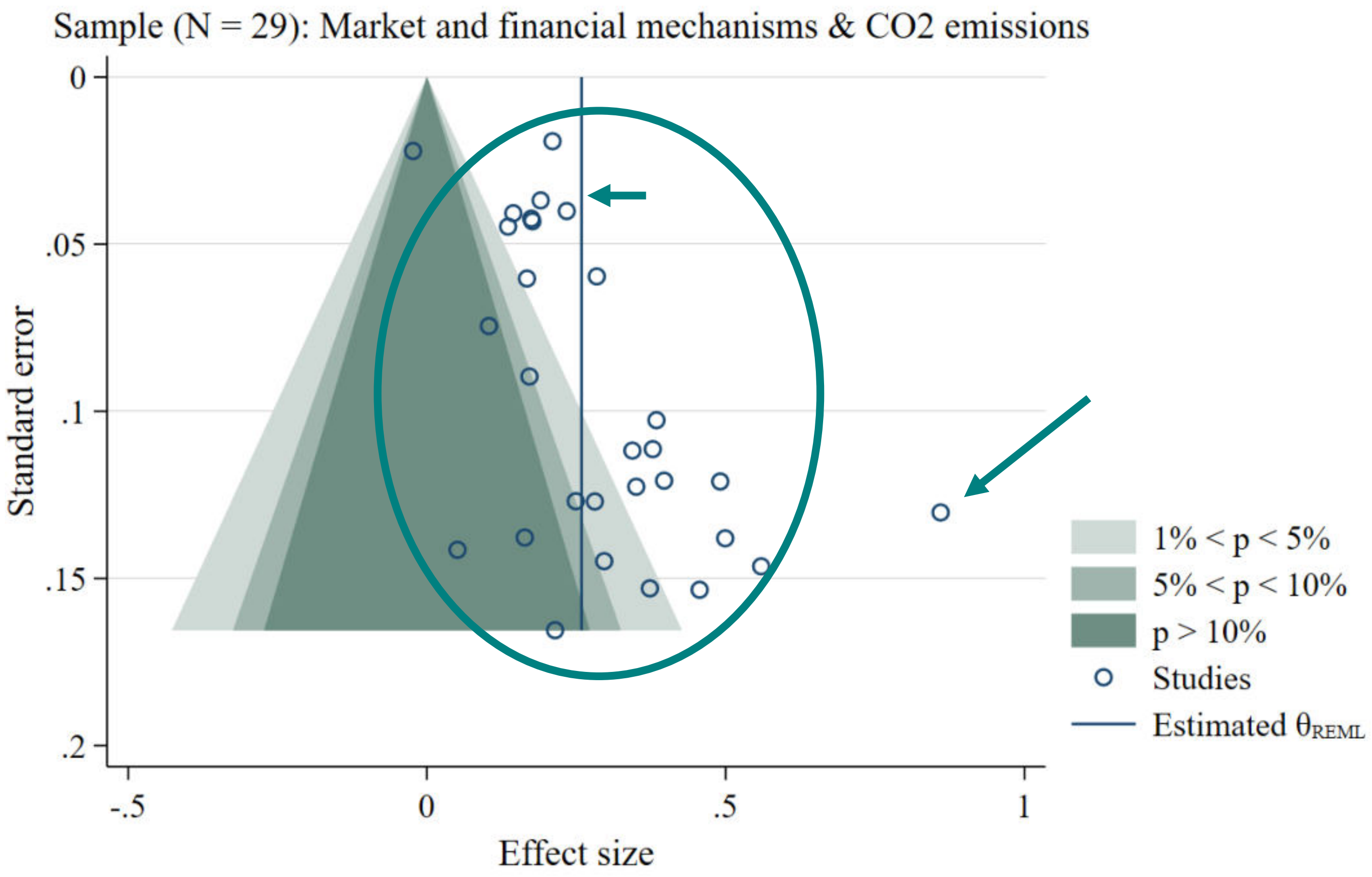
→ *Significant at 5% significance level*

For more information, please refer to the slides from the previous workshops

Interpretation of results

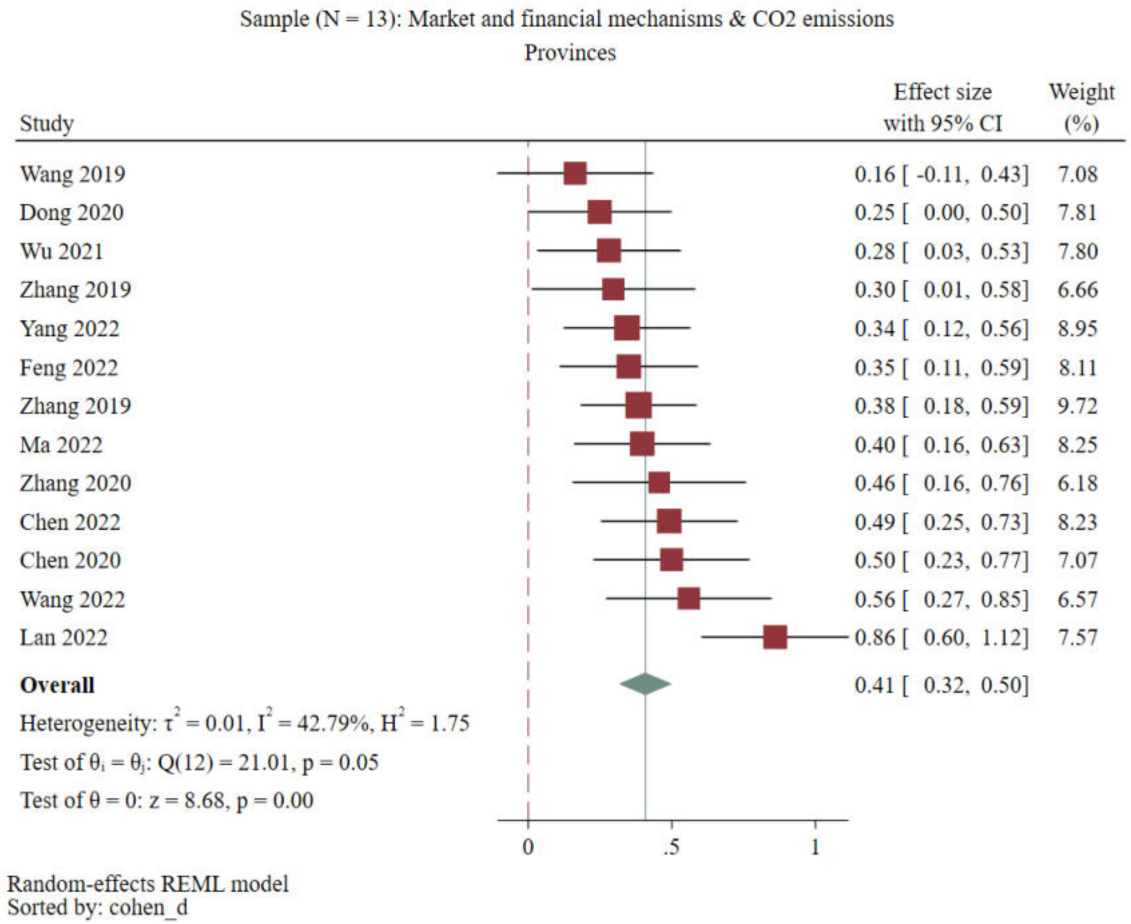
Reading funnel plots

- Effectively a scatter plot of effect size and precision
  - Effect size on the x-axis
  - Precision on the y-axis, here standard error
- Vertical line depicts overall estimated effect
- Sometimes outliers can be identified
- Asymmetry around effect size may be indication of publication bias or small-study effects



Reading forest plots

- Effect size of each study is a red square. Size reflects the weight.
- The study's confidence interval is the horizontal line
- The overall effect size is the green diamond, where the edges represent its confidence interval
- Unit matters → helps for interpretation
- Effect size and confidence interval noted on the right
- Effect line (green) and line of no effect (red dashed) are vertical
  - → All but two studies don't include 0 in their CI - estimate a statistically significant effect
  - → All but one study have estimated overall effect size within their CI



Sample (N = 13): Market and financial mechanisms & CO2 emissions

Provinces

Study	Effect size with 95% CI	Weight (%)
Wang 2019	0.16 [-0.11, 0.43]	7.08
Dong 2020	0.25 [0.00, 0.50]	7.81
Wu 2021	0.28 [0.03, 0.53]	7.80
Zhang 2019	0.30 [0.01, 0.58]	6.66
Yang 2022	0.34 [0.12, 0.56]	8.95
Feng 2022	0.35 [0.11, 0.59]	8.11
Zhang 2019	0.38 [0.18, 0.59]	9.72
Ma 2022	0.40 [0.16, 0.63]	8.25
Zhang 2020	0.46 [0.16, 0.76]	6.18
Chen 2022	0.49 [0.25, 0.73]	8.23
Chen 2020	0.50 [0.23, 0.77]	7.07
Wang 2022	0.56 [0.27, 0.85]	6.57
Lan 2022	0.86 [0.60, 1.12]	7.57
Overall	0.41 [0.32, 0.50]	

Heterogeneity: $\tau^2 = 0.01$, $I^2 = 42.79\%$, $H^2 = 1.75$

Test of $\theta_i = \theta_j$: $Q(12) = 21.01$, $p = 0.05$

Test of $\theta = 0$: $z = 8.68$, $p = 0.00$

Random-effects REML model
Sorted by: cohen_d

Heterogeneity

- Heterogeneity describes how different effect sizes are across studies
- If the studies are rather different, the question is: Why?
- Most common measure of heterogeneity: I^2
    - Quantifies the % of total variation across studies due to heterogeneity rather than chance
- Typical thresholds are
    - Low heterogeneity: 0%-25%
    - Moderate heterogeneity: 25%-50%
    - High heterogeneity: 50%-75%
    - Very high heterogeneity: >75%

Contextualizing findings

Important to think about practical significance

- Importance of the effect size in real-world terms
- Statistical significance doesn’t matter if effect size negligible in practical terms
  - → Statistically significant effect of TVET that increases yearly income by 3USD is not practically significant. If it increases yearly income by 5000USD, this is highly practically significant

Subgroup analysis

- Particularly important if studies are heterogeneous
- Examine how effect sizes differ by different subgroups
  - E.g. by gender, age group, educational background, geographic context, etc.
- Differences may be highly relevant to policy makers

Evaluating robustness

Sensitivity analysis

- Leave-one-out analysis to check robustness (Would results change if one study were missing?)
- Compare methods of handling missing data

Heterogeneity

- Check heterogeneity for patterns in driving factors of effect sizes

Overall quality of evidence (see last slide set)

- Investigate whether excluding low-quality studies change results significantly
- Are there systematic differences in effect sizes by study quality?

Publication bias assessment

- Investigate whether publication bias is suspected
- Compare multiple methods for correcting for publication bias

Assess applicability and generalizability

Next steps

Implications for decision making

Informing best practice

- E.g. Programme design

Policy formulation

- Evidence-based policies
- Regulatory decisions

Resource allocation

- Prioritize more effective interventions
- Inform cost-effectiveness analysis

Dissemination strategies

Reporting

- Clear and concise
- Adapt content to target audience
- Use of visual aids

Use multiple dissemination channels to reach a wide audience

- Academic journals, policy briefs, blogs, social media, conferences, etc.

Engage diverse stakeholders through

- Collaborative efforts
- Feedback mechanisms to improve communication strategies
- Educational workshops
- Stakeholder-specific messages

END OF SESSION 3

Session 4:
Using Evidence in Project and Policy Planning

C4ED – EUTF

October 2024

Center for Evaluation and Development

European Commission

Objectives of Session 4

Identify and understand role of evidence synthesis in project decision-making process as well as opportunities and challenges of its integration

Explore strategies for translating evidence into actionable recommendation and intervention

Understand the integration of evidence into actionable recommendation through a case study

Role of evidence synthesis in project planning and policy formulation

Roles of evidence synthesis in policy and project planning

- Minimize bias and enhance objectivity
- Support transparent and justifiable decisions
- Facilitating consensus among stakeholders
- Identifying knowledge gaps and research priorities

Opportunities

Integrate the best available evidence to design future projects

- *Identify trades in a TVET training that result in positive employment and income outcomes*

Inform on the different approaches used in the past to reach a certain goal

- *Identify if classroom training or apprenticeships or a combination of both results in likelihood of starting a business*

Opportunities

Identify Dos and Don’ts

- *Best practices*
- *The most impactful*
- *The cheapest approaches*
- *The most cost effective*

Risks and pitfalls to avoid

- *Potential undesired impacts,*
- *Unexpected factors that can hinder the project*

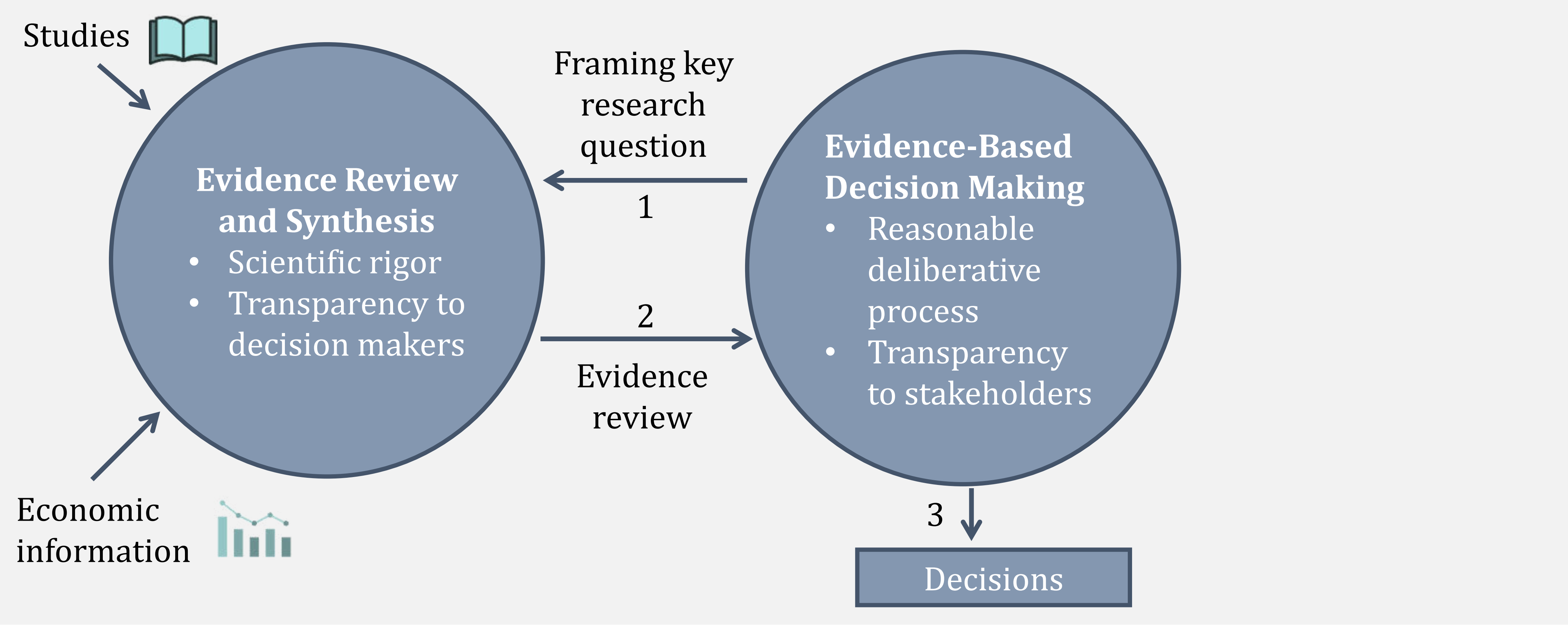
Challenges

- Mismatch between research outcomes and organizational goals, limited stakeholder engagement, and data accessibility issues can hinder integration efforts
- Data availability, quality and applicability issues
- Integrate diverse sources and methods can be complex
- Overcoming resistance from conventional practices
- Gap in capacity for conducting evidence synthesis and interpretation
- Keeping evidence relevant in rapidly changing fields

Strategies for translating evidence into action

- Utilize a variety of evidence sources
- Translate complex evidence into actionable recommendations
- Align recommendations with organizational goals
- Engage policy makers early to ensure recommendation feasibility
- Communicate evidence effectively to diverse audiences
- Explore dynamic relationship between evidence synthesis & evidence-based decision making

Evidence synthesis and evidence-based decisions



Studies

Economic information

Evidence Review and Synthesis

- Scientific rigor
- Transparency to decision makers

Framing key research question

1

2

Evidence review

Evidence-Based Decision Making

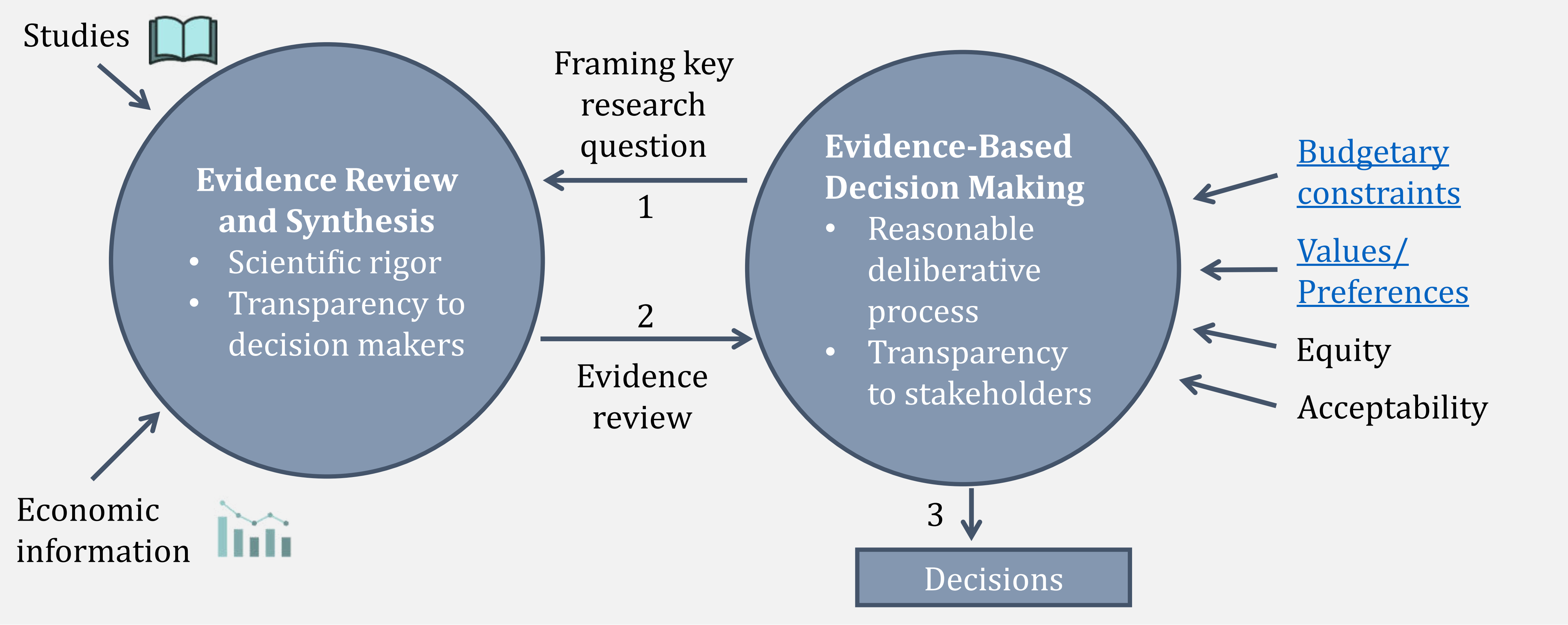
- Reasonable deliberative process
- Transparency to stakeholders

3

Decisions

Source: Teutsch & Berger (2005)

Evidence synthesis and evidence-based decisions


Studies
Framing key research question
1
Evidence Review and Synthesis
- Scientific rigor
- Transparency to decision makers
2
Evidence review
Evidence-Based Decision Making
- Reasonable deliberative process
- Transparency to stakeholders
Budgetary constraints
Values/ Preferences
Equity
Acceptability
Economic information
3
Decisions

Source: Teutsch & Berger (2005)

Budget constraints

- Scope and scale determination of interventions
- Optimize resource allocation
- Factor in cost effectiveness analysis is critical
- Financial planning and fund utilization
- Balance costs and outcomes to reflect financial realities
- Factor-in project sustainability
- Consider strategic budget at question formulation stage

Go back

Values and preferences in decision making

- Stakeholder values and preferences shape program outcomes
- Integrating values early in the process can help align outcomes
- Diversity of stakeholder values and preferences require balancing
- Local contextual values could influence design and implementation
- Managing multiple diverse values is challenging but essential for program design and implementation
- Communicating value driven decisions enhances stakeholder buy-in

Go back

Equity

- Equitable consideration is critical in avoiding unjust exclusivity
- Evidence generated should be looked at in line with the goal of reducing disparities between groups (economic status, race, gender or geographic location)
- Evidence synthesis help identify existing disparities
- Recommendations or learnings from evidence synthesis which inform decision making should not perpetuate existing disparities or create new ones
- Equity is critical for the design of programs that are transformational and sustainable
- Stakeholder involvement is critical in addressing equity in program planning

Go back

Acceptability

- Acceptability is not a one-time assessment but continuous process
- Evidence-based decision making should consider acceptability which depends highly on community engagement and satisfaction
- Acceptability highly dependent on cultural and social compatibility
- Addressing acceptability has to be strategic to overcome barriers
- Clear communication in project planning in the face of evidence obtained from evidence synthesis is critical
    - Ensures stakeholders remain informed

Stakeholder consultations

- Adapt synthesized evidence to local contexts
- Use stakeholder feedback to refine interventions
- Continuous evaluation and adaptation based on new data
- Critical role of stakeholders in the synthesis process for relevance and applicability
- Enhances project ownership and sustainability through engagements
- Resolves conflicts and aligns interests via evidence-based dialogue

Case study
(Tripney and Hombrados, 2013)

Case study: Background

Study Title: *Technical and vocational education and training (TVET) for young people in low- and middle-income countries: a systematic review and meta-analysis (Tripney and Hombrados, 2013)*

Overview of the systematic review’s scope: 26 studies across predominantly Latin America, analyzing TVET's impact on youth employment

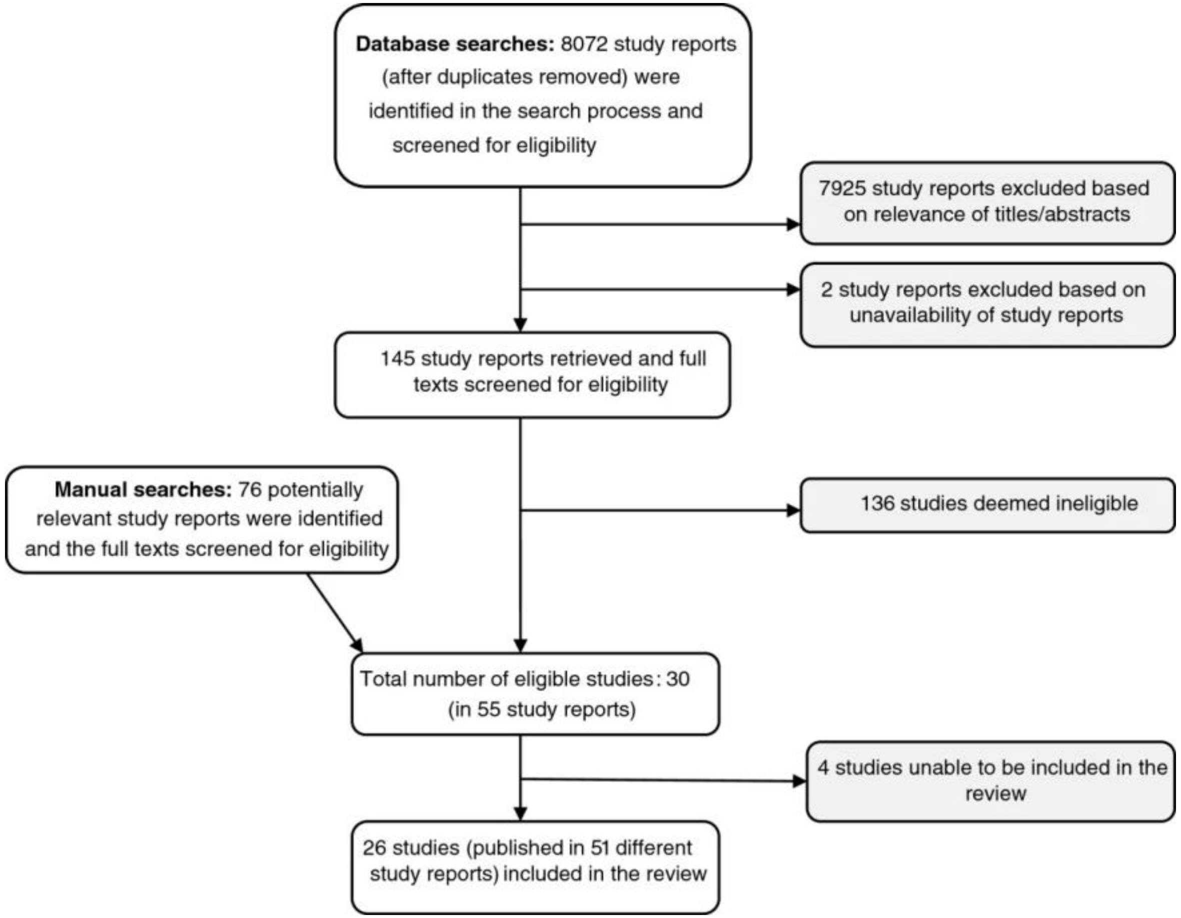
Case study: Methodology

Study methodology

Significant challenges

Limited geographic scope, underrepresentation of TVET types (Apprenticeships), small number of RCTs

(Tripney and Hombrados, 2013)



Database searches: 8072 study reports (after duplicates removed) were identified in the search process and screened for eligibility

7925 study reports excluded based on relevance of titles/abstracts

2 study reports excluded based on unavailability of study reports

145 study reports retrieved and full texts screened for eligibility

Manual searches: 76 potentially relevant study reports were identified and the full texts screened for eligibility

136 studies deemed ineligible

Total number of eligible studies: 30 (in 55 study reports)

4 studies unable to be included in the review

26 studies (published in 51 different study reports) included in the review

Case study: Key findings

Small but statistically significant positive effects on:

- Paid employment (13.4%)
- Formal employment (19.9%)
- Monthly earnings (12.7%)
- Treatment effect on self-employment earnings and weekly hours worked not significant

(Tripney and Hombrados, 2013)

Case study: Limitations

- Not all eligible studies could be included into the evidence synthesis
- Several methodological issues were identified
- The methods for comparing effective sizes are complex and need further research
- Not enough RCTS which look at causal estimates of the impacts of TVET
- Study limited to Latin American and Caribbean countries

Thus, conclusions could be under or over estimation of the impact of TVET

Case study: Recommendations for research knowledge gap

⚠ **Caution** when interpreting or generalizing findings due to methodological issues

Study recommendations:

- Enhance rigor
- Broaden research scope
- Assess intervention components

- *(Tripney and Hombrados, 2013)*

Case study: Recommendations for decision-making

- Engage a wide range of **stakeholders in program planning**
- Emphasize the need for **stakeholders to actively participate in commissioning robust research designs**, such as RCTs and QEDs, to generate reliable evidence supporting TVET's effectiveness
- Highlight **the importance of providing budgetary allocation** among stakeholders for rigorous outcome research across broader range of TVET programs & geographical settings
- **Prioritise cost effectiveness analyses** in future impact evaluations to get more data on costs of TVET interventions
- Implement TVET programs that respect and incorporate local cultural norms and values

END OF SESSION 4

Session 5:
Evidence from EUTF interventions for future programming

C4ED – EUTF
October 2024

Center for Evaluation and Development

European Commission

Objectives of Session 5

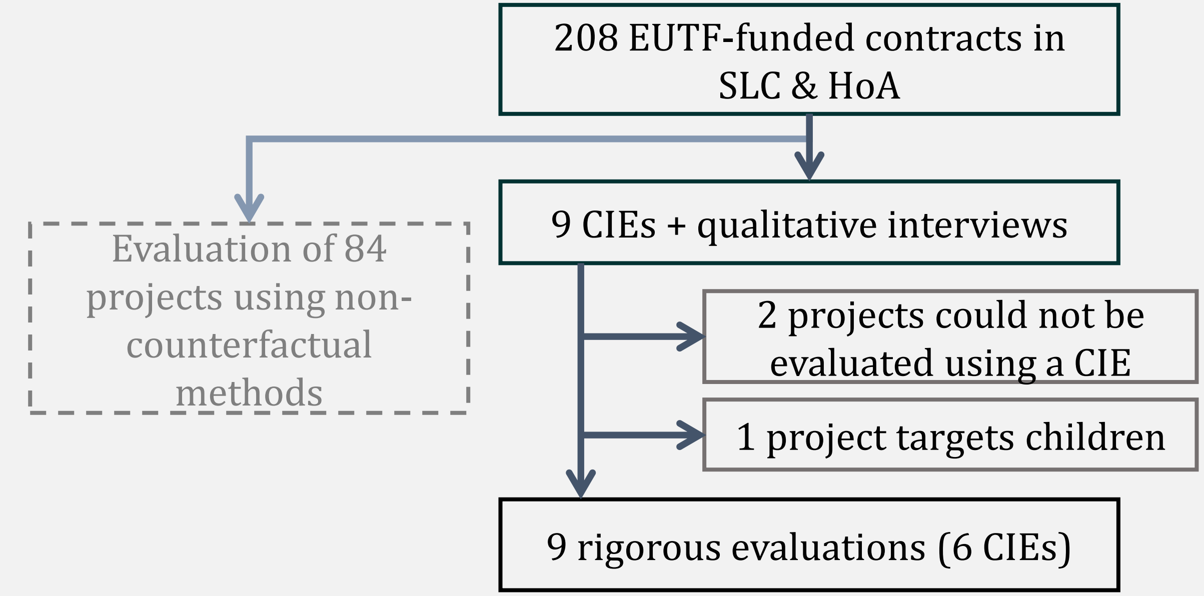
Share conclusions from the Counterfactual Impact Evaluations (CIEs) of the EUTF-funded projects.

Use synthesized conclusions and brainstorm on concrete solutions to overcome the challenges faced by EUTF-funded projects

Identify potential recommendations

Background

Overarching goal of the portfolio evaluation: Measure and understand the impacts of strengthening skills and improve employment of vulnerable groups to reduce <u>irregular</u> migration



208 EUTF-funded contracts in SLC & HoA

Evaluation of 84 projects using non-counterfactual methods

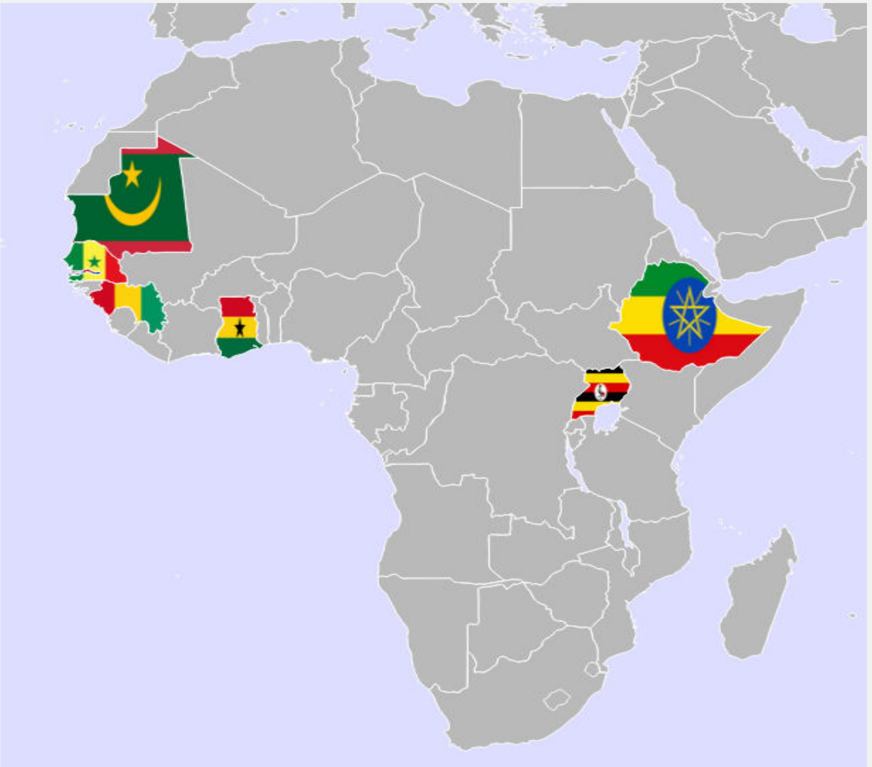
9 CIEs + qualitative interviews

2 projects could not be evaluated using a CIE

1 project targets children

9 rigorous evaluations (6 CIEs)

Figure 1: Location of projects with CIEs



Source: C4ED

3 sets of key findings & lessons learnt

1. **Dropouts and no-shows** + breakout session
2. **Impacts on employment** + breakout session
3. **(Impacts on) migration** + breakout session

Key findings #1: Dropouts and no-shows

Training projects often face issues with no-shows and dropouts

→ Efficiency issue: projects are not running at full capacity

→ Ethical issue: individuals that would like and can participate are not selected

Main causes:

- Costs
- Other NGOs and development agencies
- Quality and relevance of training
- Social constructs and gender roles

To limit no-shows and dropouts…

Invest in selection process to identify suitable candidates:

- Ensure that interests and expectactions are aligned with the curricula
- Ensure that candidates are willing to attend the training

Adapt the training to targeted population and context:

- Timing & duration must allow the participant to comply with his/her obligations/aspirations
- Provide services to facilitate attendance
- Promote inclusion of marginalised groups
- Coordinate with interventions provided by other entities

+ Build a waiting list of eligible candidates to deal with no-shows and dropouts.

Key findings #2: impacts on employment

1. **It takes time to find a stable job after benefiting from a project (>2 years)**
2. **If possible, beneficiaries tend to open their own business**
3. **Technical trainings are usually useful but deemed insufficient to open a business**
4. **Impacts are limited on vulnerable profiles** such as females and refugees because they face specific challenges such as:
    - Household obligations
    - Lack of foundational knowledge
    - Lower access to capital
    - Limited social network

To promote (inclusive and decent) employment…

To support entrepreneurial initiatives, it is key to:

- Promote entrepreneurial skills
- Provide access to capital (start-up kits, access to loans…)

Support must consider different needs…

- Trades:
  - Type and amount of capital needed to start a firm?
  - Links and experience needed to find wage-employment in the existing private sector?
- Vulnerable profiles (females, refugees, returning migrants)

Findings #3: (Impacts on) migration

1. In Sahel Lake Chad (SLC), few projects managed to enroll the targeted population: returning migrants

Costs

Long term project strategies *versus* short term constraints

Other NGOs and development agencies

Psychological & emotional challenges

Findings #3: (Impacts on) migration

1. In Sahel Lake Chad (SLC), few projects managed to enroll the targeted population: returning migrants
2. Beneficiaries do no show clear willingness to migrate outside the country.
3. Not all projects intend explicitly to reduce the intention to migrate.
4. Employment & income-related outcomes seem disconnected to the intentions to migrate

To reduce irregular migration…

1. **Actively collaborate with institutions dealing with (potential) migrants …**

   … to accurately target (potential) migrants/refugees.

   … to understand needs in the specific context.

2. **Promotion of employment is not the key to reduce intention to migrate:**

   - Need to further explore other aspects such as macroeconomic and political stability?
   - Other factors?

END OF SESSION 5

Session 6:
Integrating AI into Evidence Synthesis and Evaluation

C4ED – EUTF

October 2024

Center for Evaluation and Development

European Commission

Objectives of Session 6

Understand current state if AI in evidence synthesis

Explore some AI tools in and platforms used in evidence synthesis and other tasks

Explore the challenges and future directions in using AI for evidence synthesis

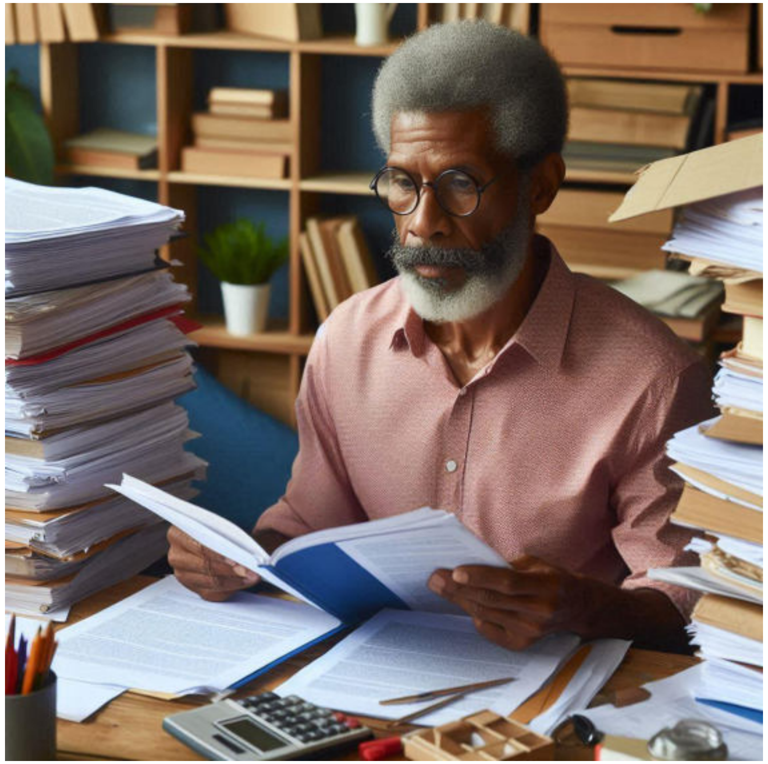
European Commission

Current state of AI in evidence synthesis

What is AI?

> *“Artificial intelligence can be defined as the ability of the software systems to carry out tasks that usually require human intelligence: vision, speech, language, knowledge and search.”*
>
> — World bank (2024)

Why AI?

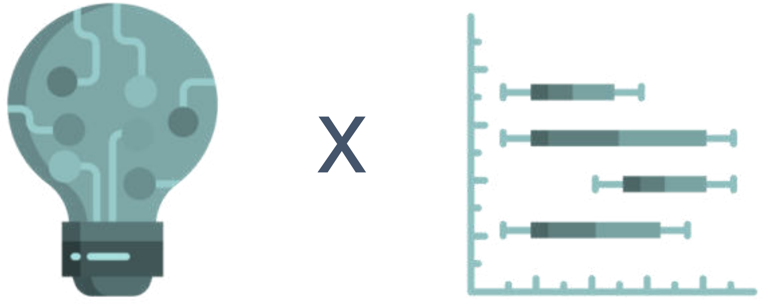


- Evidence synthesis can be expensive if done properly
- Inability to keep up with pace of production of research and studies being published
- Over 5.14 million academic articles alone are published every year

Why AI?

- Conventional manual evidence synthesis processes can be time consuming
- Time constraints do not help to inform decision-making at critical times
- Due to continued change and update of information, bodies of research and evidence quickly become outdated
- Need for efficiency and reliability in generating evidence without incurring heavy human and time cost
- Individual classification decisions can introduce inconsistencies in how insights of the same type are classified

How is AI being used in evidence synthesis?



- Automatically clustering and visualizing results
- Study classification
- Screening studies for eligibility
- Automatically finding studies
- Information and data extraction

AI tools and platforms

AI tools - ChatGPT

- Uses deep learning to understand and generate human-like text making summarization of complex evidence easy
- Interactive Q&A: Ask any question and get answers right away like speaking with a human
- Language helper: Can translate texts from many languages, making more evidence available for synthesis
- Current model GPT-4 knowledge base includes information up to September 2023

AI tools - perplexity

- AI powered search engine and chatbot utilizes advanced technology to provide accurate and comprehensive answers to user queries
- Transparency: Shows sources of its answers and provides citations
- Personalization: Ability to personalize answers based on past history and users' interests

AI tools - SciSpace

SCISPACE

- AI powered assistant to assist research tasks literature review, data extraction
- AI powered interactive PDF analysis simplifies and summarizes complex studies
- Facilitates meta-analysis by combining multiple studies
- Able to include filters like research gaps, future research, methods used, problem statement, variables
- Tools includes citation generator with various formats

AI tools – (Other)

- Consensus
- Heuristics
- OpenRead
- Explainpaper

- World Bank’s Development Team (DIME)
- Evidence for policy makers
  - Impacts of interventions
  - Avoids hallucinations and generic responses

ImpactAI

Opportunities, challenges and future directions of AI in evidence synthesis and evaluation

Opportunities of using AI in evidence synthesis and evaluations

- Increased productivity
- Higher quality results
- Improved performance
- AI beneficial for all users
- Levelling and enhancing of abilities
- Limitations on tasks beyond capacity of AI

Implications for the evaluation field

- Evaluators must engage with emerging AI technologies or risk becoming less relevant in the field.
- Data scientists, lacking awareness of key evaluation issues, might take over more evaluation tasks
- Active engagement with AI is crucial to ensure evaluators remain central and relevant in evaluation practice
- Research will be critical for helping the evaluation sector keep up with emerging AI approaches

Evidence synthesis using AI

- IEG Experiment: ***Setting up Experiments to Test GPT for Evaluation***
- A study by Independent Evaluation Group (IEG) of the World Bank
- Study assesses the integration of generative AI models into evaluation
- More common use of AI has been more discriminative. i.e. decisions about boundaries and classes of texts.
- These do not generate anything new
- Aims of using GPT are to improve ***speed***, ***enhanced capabilities***, ***new insight***, and ***improved quality***

Evidence synthesis using AI

- Total of nine experiments spanning various stages of the evaluation process (pre-analysis, analysis, post-analysis)
- Experiment span across user profile needs (data scientists, analysts, and team leaders)
- Experiment included output types like text, images and programming codes
- Experiment results could be compared with output already generated by an evaluation team

What worked?

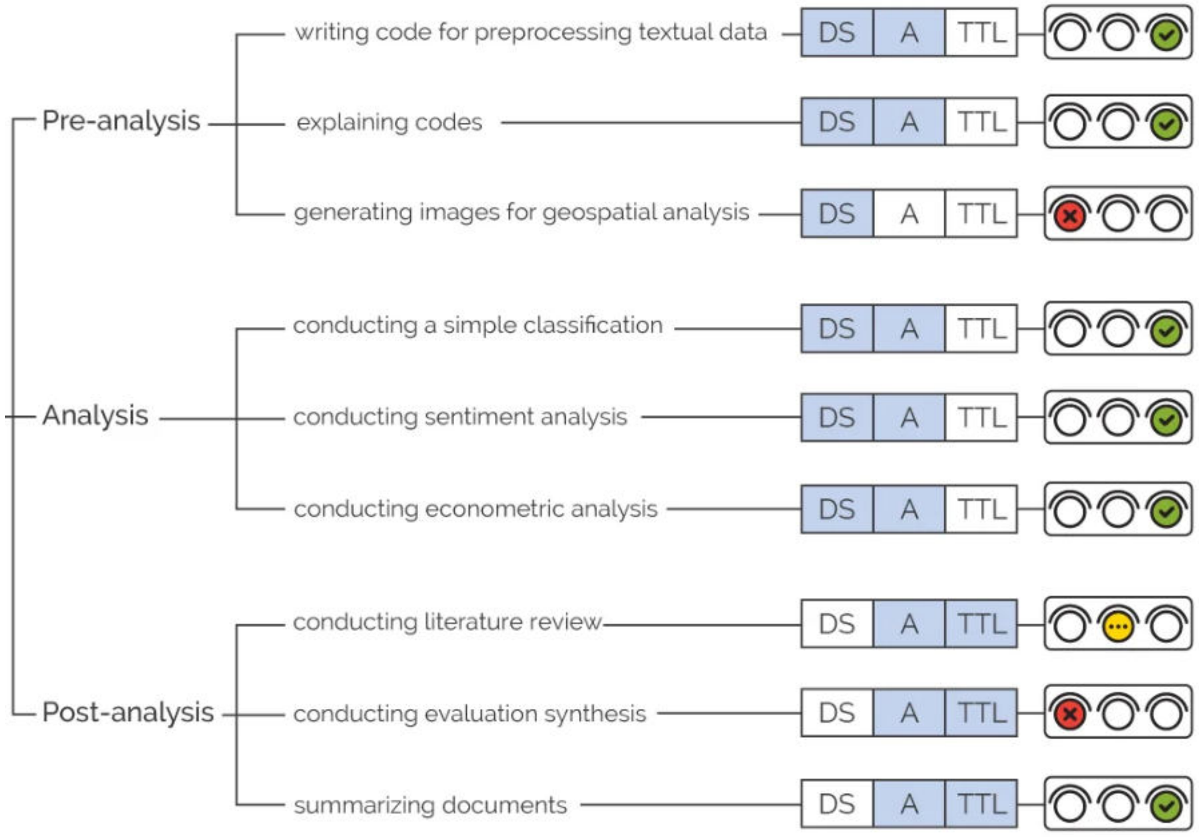
- Writing code for preprocessing textual data
- Explaining Programming code
- Conducting simple classification
- Conducting sentimental analysis
- Conducting econometric analysis
- Summarizing individual documents

Unfulfilled promises

- Generating synthetic images for data augmentation in the use of spatial analysis
- Conducting a literature review (fabricated evidence and hallucination)
- Conducting an evaluation synthesis (generic responses)

Conclusion

Figure 1. Recommended Uses of GPT for Evaluation Practice


Pre-analysis
writing code for preprocessing textual data — DS A TTL — ○ ○ ✓
explaining codes — DS A TTL — ○ ○ ✓
generating images for geospatial analysis — DS A TTL — ✗ ○ ○
Analysis
conducting a simple classification — DS A TTL — ○ ○ ✓
conducting sentiment analysis — DS A TTL — ○ ○ ✓
conducting econometric analysis — DS A TTL — ○ ○ ✓
Post-analysis
conducting literature review — DS A TTL — ○ … ○
conducting evaluation synthesis — DS A TTL — ✗ ○ ○
summarizing documents — DS A TTL — ○ ○ ✓

Source: Independent Evaluation Group.
Note: A = analyst; DS = data scientist; TTL = task team leader.

Way forward

European Commission

- Enhance Efficiency and Quality
- Support High-Level Analysis
- Human-AI Collaboration
- Improve AI Models for monitoring, evaluation and learning
- Ethical and Inclusive Practices

Discussion question

Considering the rapid advancements in AI and its integration into evidence synthesis, to what extent do you believe AI will reshape our roles as monitors and evaluators or as project planners and implementers?

END OF SESSION 6