



Impact Data and Evidence Aggregation Library

About IDEAL: Introduction & Use Cases

World Bank Team

June 9, 2025

1

Introduction to IDEAL

Alaka Holla



How much evidence is there on using pay-for-performance for general practitioners?

Is childcare a good idea for both women and children?

Is preschool more effective when parents also receive counselling?

Which policies improve foundational learning the most?

Status quo: How do people answer these questions right now?



1. They do a quick search on Google or Google Scholar and focus on the studies that have clearly stated, often positive, results.
2. They check if JPAL, IPA, or the World Bank has a nice brief or they check the introduction to a paper they like on the topic.
3. Maybe they look for existing meta-analyses.
4. They ask Alaka, who then takes a deep breath.

What would we need to answer these questions?



How much evidence is there on using pay-for-performance for general practitioners?

Is childcare a good idea for both women and children?

Is preschool more effective when parents also receive counselling?

Which policies improve foundational learning the most?

What would we need to answer these questions?



1. A compilation of the **universe** of **estimated treatment effects, across disciplines**, that is **up-to-date**.
2. Some way of **standardizing effect sizes** across studies.
3. Some detail on the evaluated **interventions**.
4. Some detail on what is **measured**.
5. A distinction across **contexts**.
6. Some indication of **quality** - i.e. the correct information has been extracted.



There are some existing databases...

POLICY IMPACTS

The Policy Impacts Library MVPF Explained Our Research Events About SUBMIT RESEARCH

The Policy Impacts Library

This library provides a standardized database of MVPF estimates derived from rigorous empirical research. The details page highlights policy providers in-depth information on the way in which the MVPF is calculated.

The Generalizer Begin Analysis



Assessing the generalizability of a completed evaluation.

For researchers evaluating the findings from a study, The Generalizer compares the final sample of schools that took part in the evaluation with the relevant population of schools in the United States as well as each of the 50 states. Summary information is provided indicating the similarity between the sample and population, as well as guidelines regarding how this similarity affects the generalizability of findings.

IES WWC What Works Clearinghouse Search Go

Keyword Enter keyword(s) Search

26 Results filtered by:

Product Type	Grade Level	Highest Evidence Tier	Name (Release Date)
Intervention Report	K-2	TIER 2 MODERATE	Early Risers (Children Identified With or at Risk for an Emotional Disturbance) (June 2019)
Intervention Report	K-12	TIER 3 PROMISING	Early Days is a multi-year prevention program for elementary school children demonstrating early aggressive and disruptive behavior. The intervention model includes two child-focused components and two parent/family components. The Child Skills Component is designed to teach skills that reduce aggression and increase self-control.
Functional Behavioral Assessment-based Interventions (December 2016)			Functional behavioral assessment (FBA) is an individualized problem-solving process for addressing student problem behavior. An assessment is conducted to identify the underlying cause of a student's problem behavior. This assessment process involves collecting information about the environmental conditions that precede the problem behavior and the...
Check & Connect (Dropout Prevention) (May 2015)			Check & Connect is a dropout prevention strategy that relies on close monitoring of student performance, academic support, social support, and other supports. The program has two main components: "Check" and "Connect." The Check component is designed to continually assess student engagement through close monitoring of student performance and...
Fast Track: Elementary School (Children Identified With or at Risk for an Emotional Disturbance) (January 2014)			Fast Track: Elementary School (Children Identified With or at Risk for an Emotional Disturbance) (January 2014)

Filters

Publication Date Since 2018 (last 5 years) Since 2015 (last 10 years) Since 2003 (last 20 years)

Intervention Report K-2 TIER 2 MODERATE

Topic Literacy STEM Social Emotional Learning and Behavior High School Completion

Intervention Report K-12 TIER 3 PROMISING

Population 9-12 TIER 3 PROMISING

Find a study Search

Status Recruiting and not yet recruiting studies All studies

Condition or disease (For example: breast cancer)

U.S. National Library of Medicine ClinicalTrials.gov

Find Studies About

ClinicalTrials.gov is a database of privately and publicly funded research studies conducted around the world.

Cochrane Library Import evidence. Improve decisions. Better health.

Cochrane Reviews Trials Clinical Answers About Help About Cochrane

Filter your results

Date Publication date Last month Last 6 months Last 9 months Last year Custom Range

Results matching * in All Text Select all (8961) Export selected citations Show all previous Order by Date Title Author Last updated Results per page 25 50

8961 Cochrane Reviews matching * in All Text

Cochrane Database of Systematic Reviews issue 1 of 12, January 2023

Results matching * in All Text Select all (8961) Export selected citations Show all previous Order by Date Title Author Last updated Results per page 25 50

Explore 439,057 research studies in all 50 states and in 221 countries.

See listed clinical studies related to the coronavirus disease (COVID-19)

ClinicalTrials.gov is provided by the U.S. National Library of Medicine.



The ideal database doesn't exist. Yet.

1. A compilation of the **universe** of **estimated treatment effects**, **across disciplines**, that is **up-to-date**.
2. Some way of **standardizing effect sizes** across studies.
3. Some detail on the evaluated **interventions**.
4. Some detail on what is **measured**.
5. A distinction across **contexts**.
6. Some indication of **quality** - i.e. the correct information has been extracted.



This has led to a situation where:

1. Evidence aggregation is slow, manual, and limited to researchers with a lot of resources; biased by the questions they pose.
2. There are evidence gaps in terms of outcomes, interventions, and country contexts.
3. Visibility and publication biases and shortfalls in research transparency distort the evidence base.

What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews



David K. Evans and Anna Popova

Over the course of just two years, at least six reviews have examined interventions that seek to improve learning outcomes in developing countries. Although the reviews ostensibly have the same objective, they reach sometimes starkly different conclusions. The first objective of this paper is to identify why reviews diverge in their conclusions and how future reviews can be more effective. The second objective is to identify areas of overlap in the recommendations of existing reviews of what works to improve learning. This paper demonstrates that divergence in the recommendations of learning reviews is largely driven by differences in the samples of research incorporated in each review. Of 229 studies with student learning results, the most inclusive review incorporates less than half of the total studies. Across the reviews, two classes of programs are recommended with some consistency. Pedagogical interventions that tailor teaching to student learning levels—either teacher-led or facilitated by adaptive learning software—are effective at improving student test scores, as are individualized, repeated teacher training interventions often associated with a specific task or tool. Future reviews will be most useful if they combine narrative review with meta-analysis, conduct more exhaustive searches, and maintain low aggregation of intervention categories. Education, Impact Evaluation, Human Capital. JEL codes: O15, I21, I28, J13

David K. Evans, Anna Popova, What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews, *The World Bank Research Observer*, Volume 31, Issue 2, August 2016, Pages 242–270

The ideal database doesn't exist. Yet. Enter IDEAL!



1. A compilation of the **universe** of **estimated treatment effects, across disciplines**, that is **up-to-date**.
2. Some way of **standardizing effect sizes** across studies.
3. Some detail on the evaluated **interventions**.
4. Some detail on what is **measured**.
5. A distinction across **contexts**.
6. Some indication of **quality** - i.e. the correct information has been extracted.



IDEAL

A **joint library** of RCTs in low & middle-income countries with *accurate, reliable, and consistent* information on:

- Estimated impacts and their precision
- Study design
- Intervention details
- Research guidance for users

A **platform** for making data available to inform program and research designs and steer investments.

Collaboration



Northwestern



What Works Clearinghouse

2 Use Cases for IDEAL

Anja Sautmann



Use cases in both policy and research

Aggregating
Evidence for
Policy

Evaluating
the Body of
Evidence

Econometric
Methods



1. Aggregating Evidence for Policy

Most important purpose: evidence aggregation

- Compare (determinants of) effects of the same intervention type on a range of outcomes
- Compare different interventions for effectiveness for the same outcome
- From simple stock-taking to formal meta-analysis

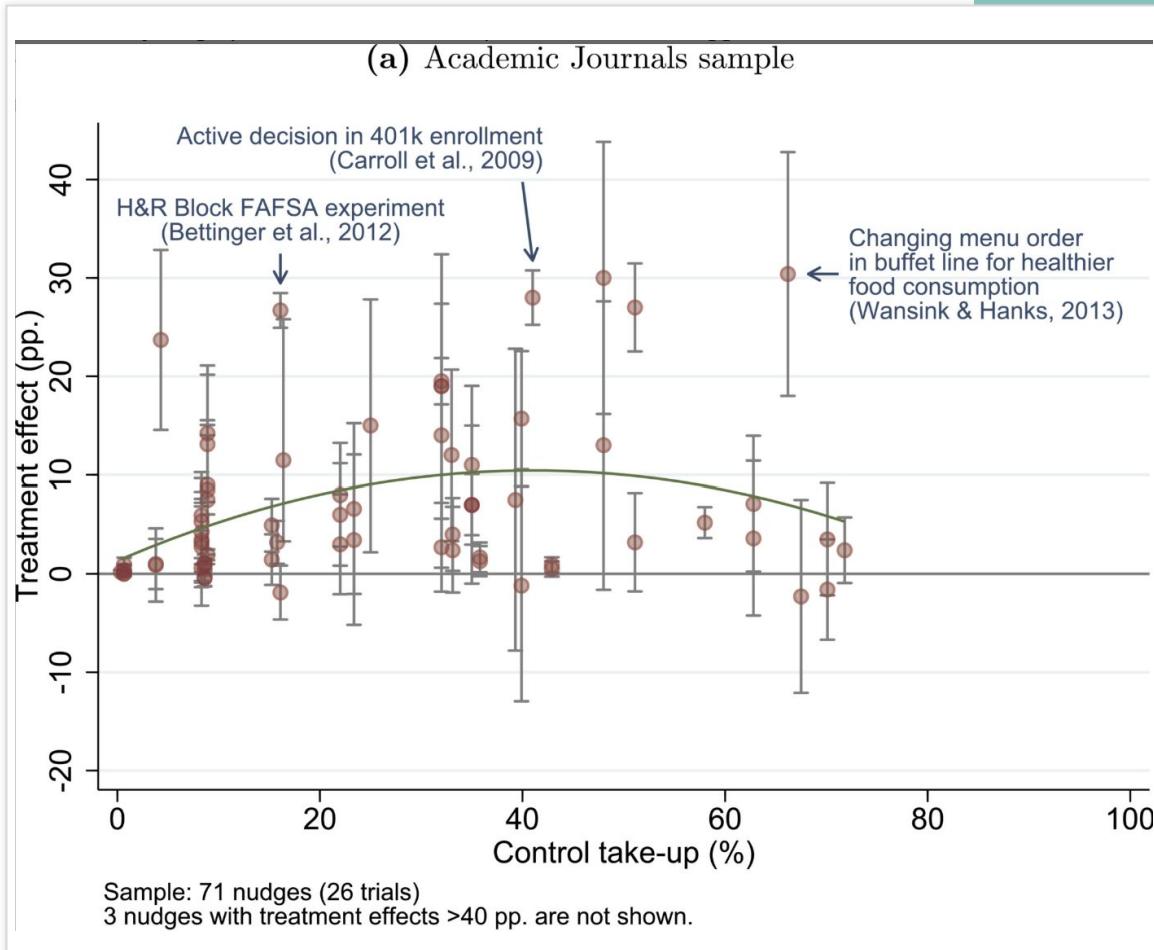


Aggregating Evidence for Policy: Example 1

- Comparing the same intervention across different settings/implementations
 - “**RCTs to Scale: Comprehensive Evidence from Two Nudge Units**” (Della Vigna & Linos, 2022, *Econometrica*)
- Lots of buzz around nudges: very low cost and effects seemingly large.

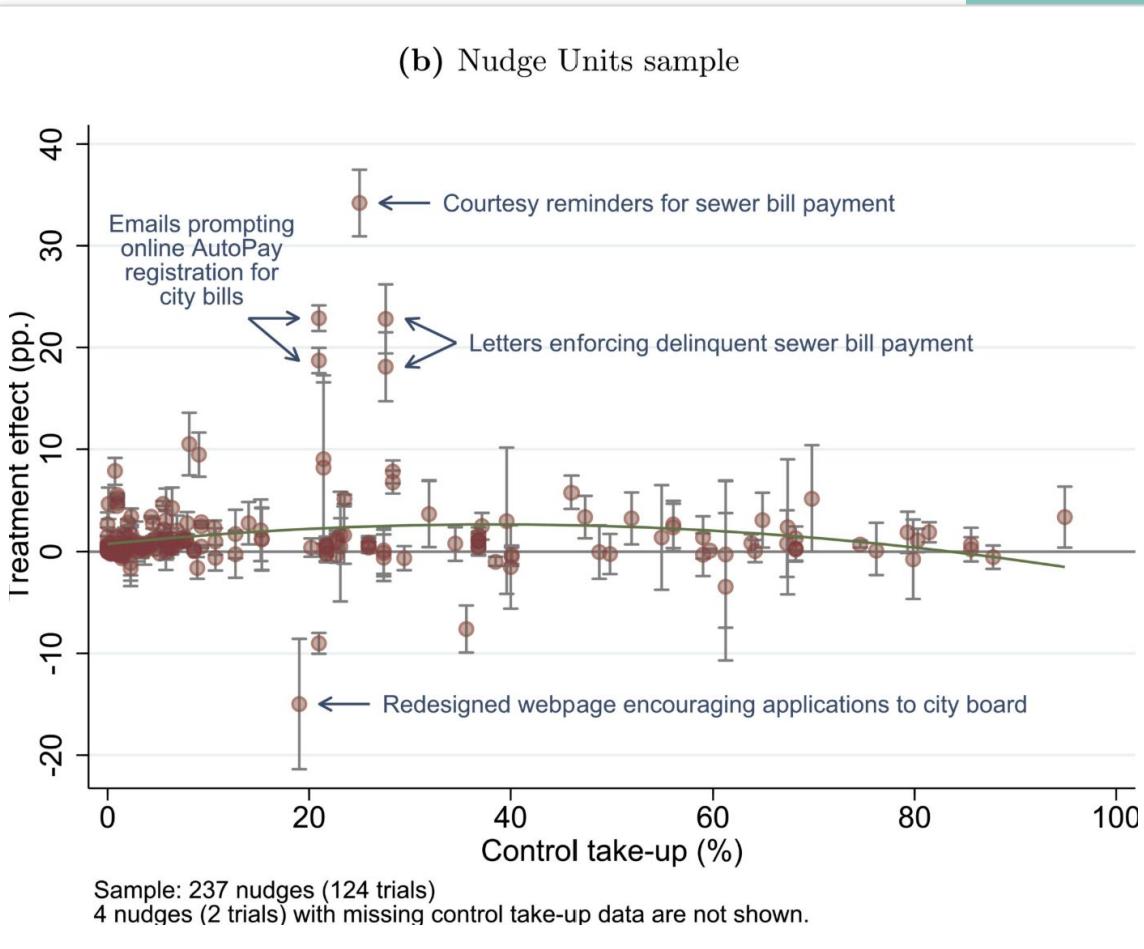
“Academic” nudges

Academic papers on
“nudge” interventions:
average increase in
take up from 26% to
nearly 35%.



Nudge unit nudges

In data from 126 trials,
23 million observations:
Average increase in
take up from 17.3% to
18.7%.





Aggregating Evidence for Policy: Example 2

- Comparing interventions that target the same or related outcomes
 - **Paper:** “Cost-effectiveness of 14 Global Fund recommended interventions for HIV/AIDS, malaria, syphilis, and tuberculosis in 128 countries: a meta-regression analysis” (Silke et al., Lancet Global Health, 2024)
- Incremental cost effectiveness ratio (ICER) estimates from Tufts CEA and GH CEA registries (cost/utility ratios, cost-per-DALY-averted)
- Fill data gaps to provide countries with a “league table” of interventions



2. Evaluating the Body of Evidence

Another important purpose: understand what kind of evidence we currently have (vs. what we would like to have?)

- Simplest case: evidence gaps – on specific outcomes, regions/countries, interventions...
- But: can also ask how studies are done, and how that drives (published) estimates
- Example: Understanding what drives heterogeneity in effect sizes
“How much can we generalize from impact evaluations?”, Eva Vivalt, JEEA 2020



Evaluating the Body of Evidence: Example

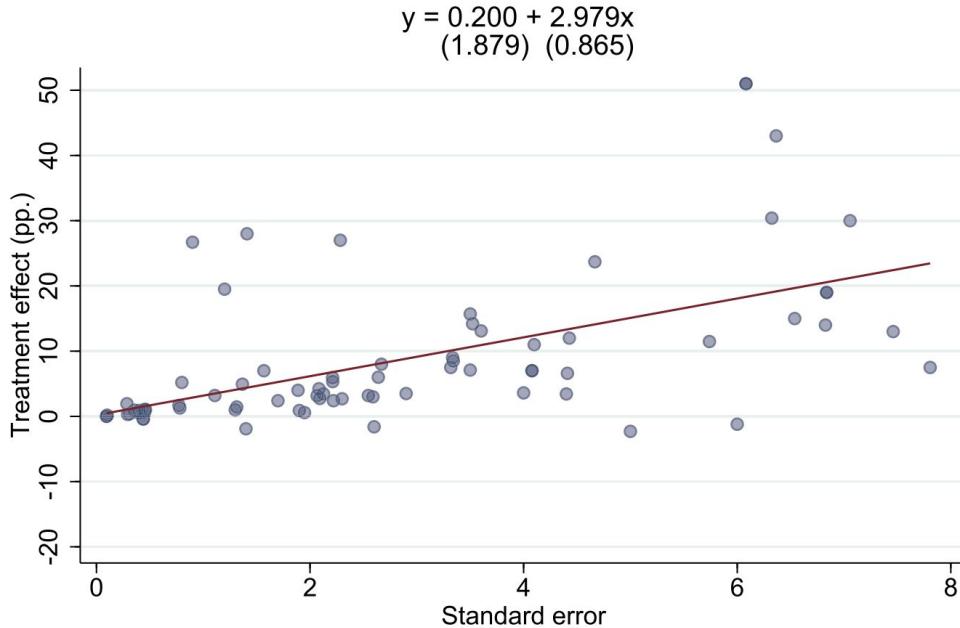
“Identification of and Correction for Publication Bias”, Andrews & Kasy, AER 2019

- Conditional publication probability (as a function of study results): bias if < 1 for effect estimates that are not significant
 - [Note: an estimate can be not significant even if the true effect is not zero]
- With a large enough set of estimates:
 - **Relationship between SE and effect size is a measure of bias**

SE vs. TE size

With publication bias,
published effect sizes from
smaller studies tend to be
larger.

(a) Academic Journals



Estimated TE as a function of standard error
from **academic nudge trials**.



Econometric or other methods

- Many practical and theoretical open questions, and constant innovation, in designing randomized studies and analyzing the resulting data
- Data from existing studies with suitable designs can help **simulate new sampling or estimation methods**, Ex.: “Adaptive treatment assignment for policy choice”, Kasy & Sautmann 2021
- **Body of evidence can...**
 - show how, and how often, specific methods are used in practice
 - **show whether methodological choices affect precision or size of TEs**



Econometric or other methods: Example

Muralidharan et al. “Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments”, ReStat 2025

- From the paper:
 - 27 out of 124 experiments in top five journals over 10 years use factorial designs (cross-randomization)
 - 19 of them do not include interaction effects between cross-randomized treatments in the estimation
 - Inference is only correct if the true interaction effect is zero
- After re-estimating the models including interaction terms:
 - Median absolute change in point estimates is 96%
 - 26% of estimates change sign, 53% (29 out of 55) of TE estimates are not significant at 5% level anymore

3 The IDEAL Pilot 2025

Alaka Holla



Goals of the pilot



Data extraction from 1,000 RCTs using IDEAL survey fields



Test the minimum set and build “controlled vocabularies”



Speed up extraction with machine learning methods



Present the prototype to key audiences

IDEAL outputs



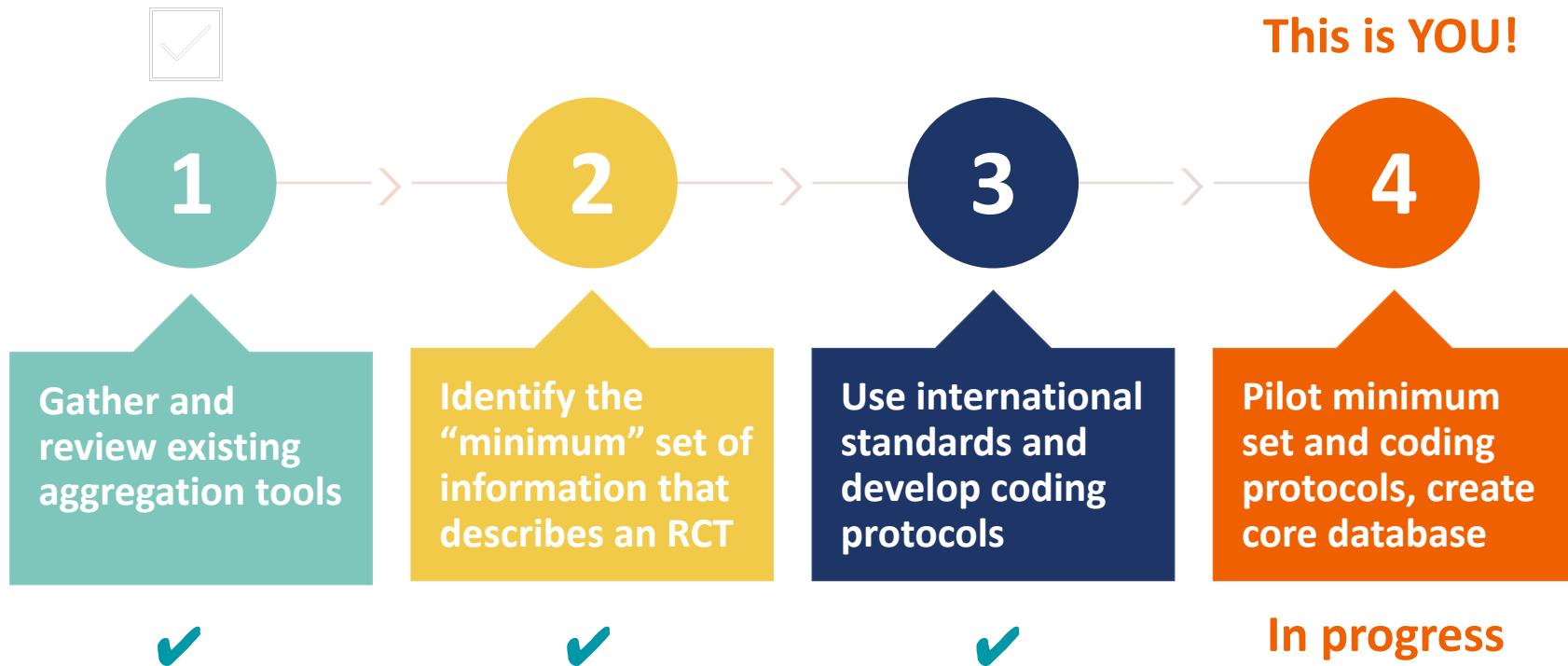
1	RCT classifier package	Open-source tools to automate search for published studies and classification by empirical method and thematic focus.
2	Metadata schema for minimum fields	A minimum set of fields needed to standardize, aggregate, and analyze average effects across all RCTs in the social sciences
3	Survey fields to capture schema fields	3-stage survey instrument with detailed coding protocol to consistently capture schema fields.
4	Data entry masks	Open-source survey forms to extract information on survey fields from published papers
5	Quality assurance and supervisor protocol	Procedures, survey forms, and code for quality checks of extracted data that yield calculations of accuracy and inter-rater reliability for the data in library.



IDEAL outputs

6	Training resources	Training packages with manuals, slides, videos, and practice papers to guide data extraction following the IDEAL survey instruments.
7	Papers and extracted data	Data extracted from 1,000 pilot papers
8	Data, data documentation and code	Data and metadata published in MicroData Library and statistical software package to calculate average effect sizes from downloaded IDEAL data.
9	Partnerships	Partnerships with other development agencies, research institutes, and universities to test and use IDEAL schema and data.
10	Library	A user interface and a public repository with all IDEAL outputs including process documents.

Progress so far





Where we are going

This is YOU!

4

Create highest quality core database of 1,000 RCTs

5

Refine protocols and train ML algorithms for rapid expansion

6

Build a prototype of the library for policy and research audience

7

A “go-to” open access library of RCTs that grows with the evidence

In progress

Thank you
for listening

Alaka Holla
aholla@worldbank.org

Anja Sautmann
asautmann@worldbank.org

