IDEAL

Impact Data and Evidence Aggregation Library

IDEAL Data Extraction

Fei Yuan & Sergio Puerto

June 9, 2025

INTRODUCTION

IDEAL aims to produce a high-quality multidisciplinary data library for evidence aggregation.

Understanding the objective, design and process of data extraction.

Agenda

1. Objective of data extraction
2. Design of data extraction instrument
3. Quality assurance measures
4. Data entry mask
5. Data extraction workflow
6. Training program

1

Objective of data extraction

How data extraction fits in the IDEAL project

Core component of IDEAL outputs

Data extraction is a key input for multiple IDEAL outputs.

- The IDEAL library will be an open-source data product, so data extraction is the backbone of the project.
    - Pilot - extracted data from 1,000 RCT papers.

- Data extraction complements other IDEAL outputs, such as RCT classification tools, survey fields, supervision protocol and data entry mask, by testing and providing feedback.

Multiple data extraction methods

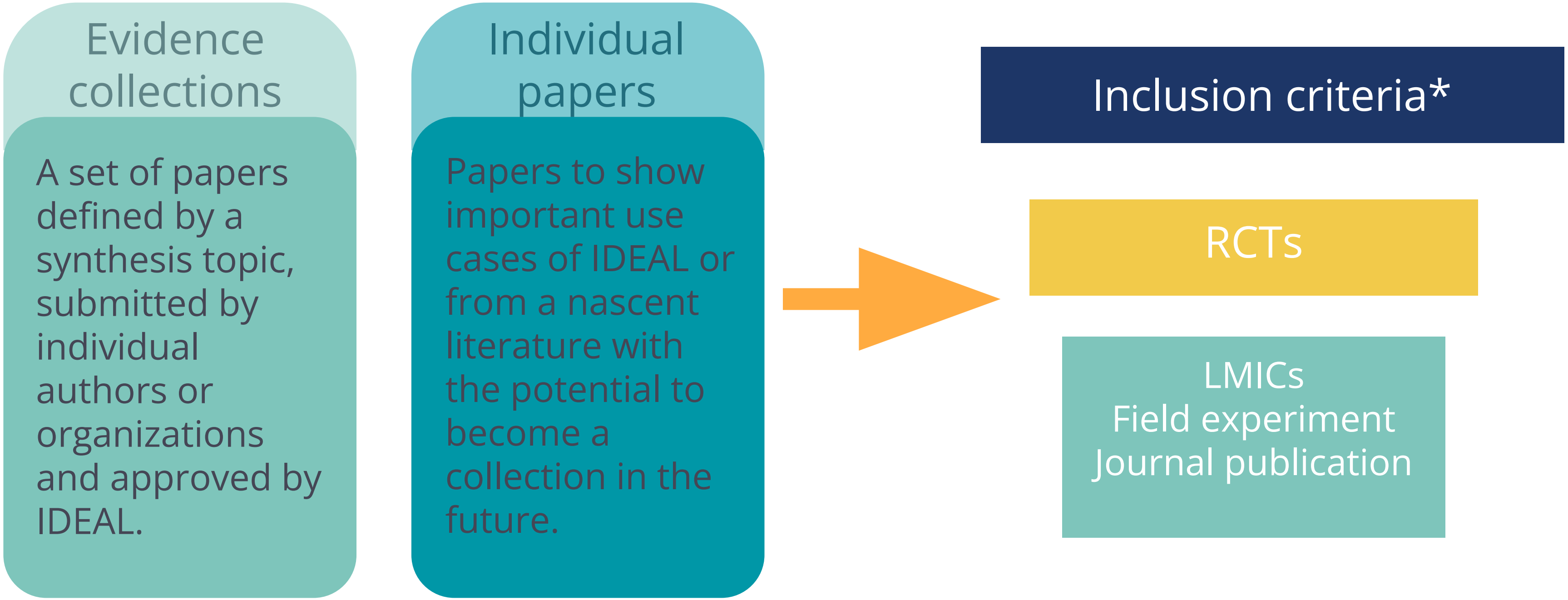
- Manual coding by **human coders**
- Automated data extraction using LLM-based or machine learning-based tools
- Hybrid: **Manual coding** with assistance from automation tools

These methods will be used at distinct phases of the project and during various stages of data extraction.

Manual coding

Producing an accurate and reliable initial dataset covering multiple disciplines is essential for developing automation tools.

Data sources



Evidence collections

A set of papers defined by a synthesis topic, submitted by individual authors or organizations and approved by IDEAL.

Individual papers

Papers to show important use cases of IDEAL or from a nascent literature with the potential to become a collection in the future.

Inclusion criteria*

RCTs

LMICs
Field experiment
Journal publication

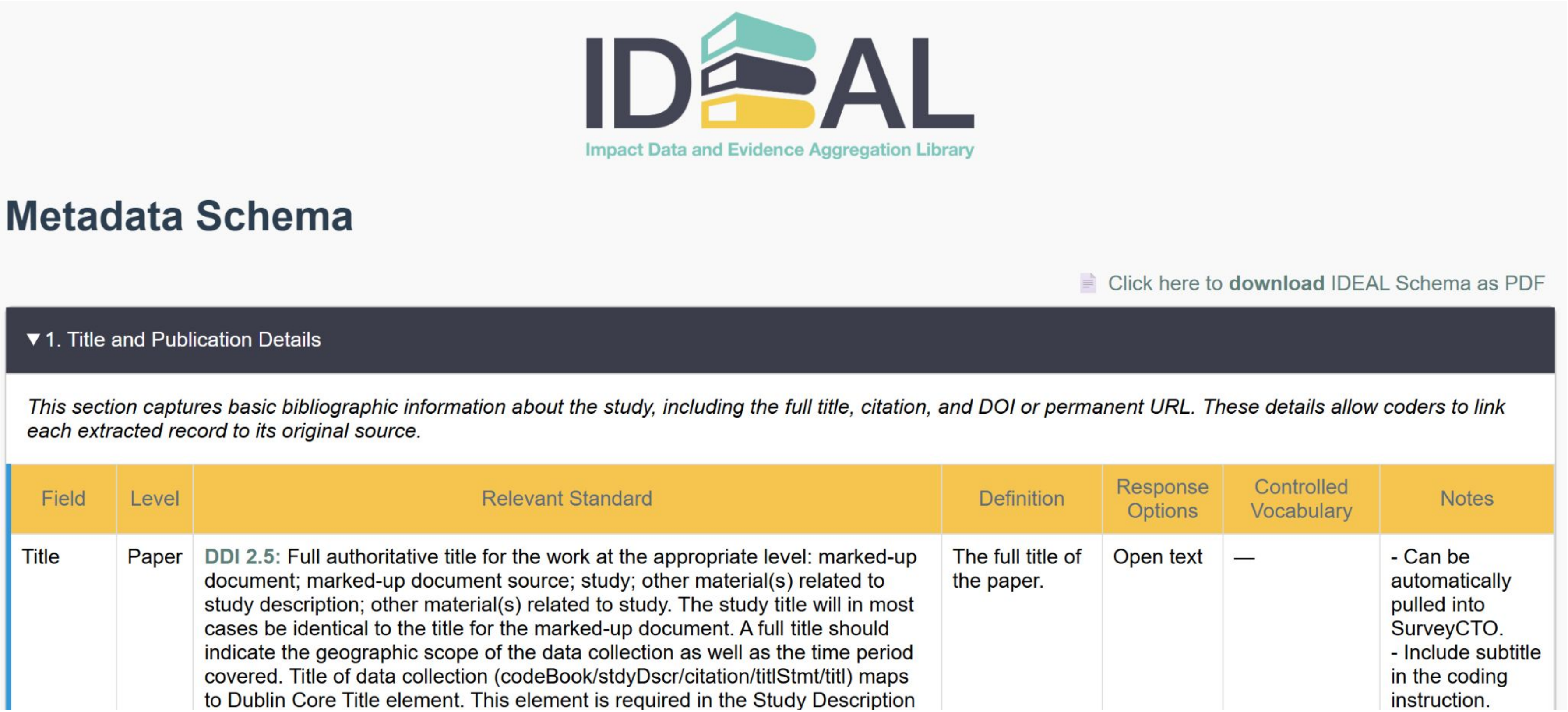
2

Design of data extraction instrument

Metadata schema and survey field

What information to extract from papers?

- IDEAL metadata schema
  - A minimum set of fields needed to standardize, aggregate, and analyze average effects across studies, applicable to all RCTs in the social sciences in low- and middle-income countries.
  - Followed international standards (e.g. DDI, clinicaltrials, AEA registry), based on a crosswalk of multiple coding instruments, and consulted experts.



IDEAL

Impact Data and Evidence Aggregation Library

Metadata Schema

Click here to **download** IDEAL Schema as PDF

▼1. Title and Publication Details

This section captures basic bibliographic information about the study, including the full title, citation, and DOI or permanent URL. These details allow coders to link each extracted record to its original source.

Field	Level	Relevant Standard	Definition	Response Options	Controlled Vocabulary	Notes
Title	Paper	DDI 2.5: Full authoritative title for the work at the appropriate level: marked-up document; marked-up document source; study; other material(s) related to study description; other material(s) related to study. The study title will in most cases be identical to the title for the marked-up document. A full title should indicate the geographic scope of the data collection as well as the time period covered. Title of data collection (codeBook/stdyDscr/citation/titlStmt/titl) maps to Dublin Core Title element. This element is required in the Study Description	The full title of the paper.	Open text	—	- Can be automatically pulled into SurveyCTO. - Include subtitle in the coding instruction.

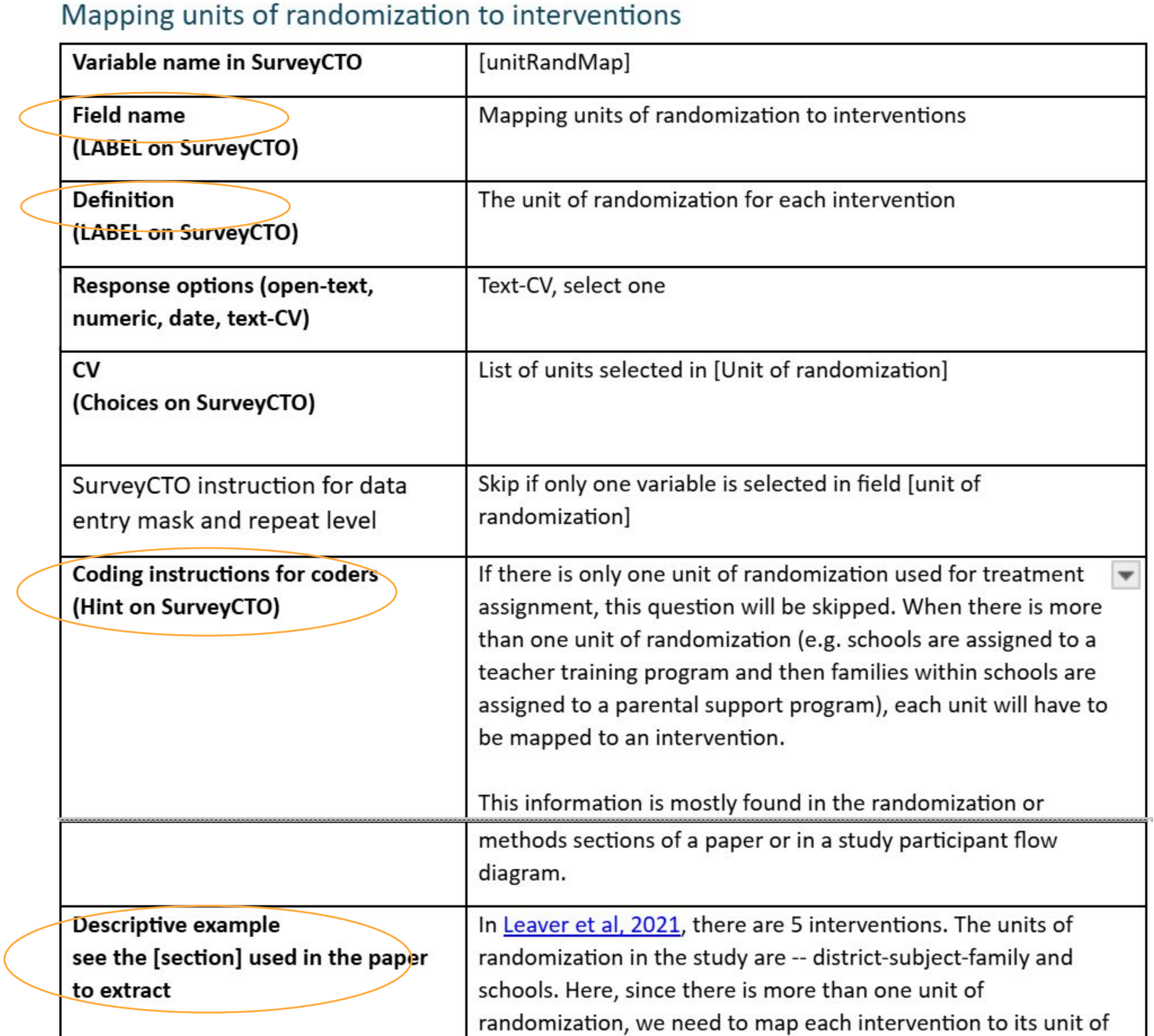
Metadata Overview

1. Title and publication details	6. Interventions
2. Resources	7. Outcomes
3. Partners and funders	8. Estimates
4. Topics and objectives	9. Quality and robustness
5. Sampling	10. Coding version tool

IDEAL survey fields and coding protocol

- Guided by the metadata schema, the survey fields were designed to capture all relevant information from each study.
  - One metadata concept may need multiple survey fields to capture, e.g. intervention details.

- Each survey field includes detailed coding instructions and descriptive examples to facilitate data extraction.
  - We will go over every field in this training!

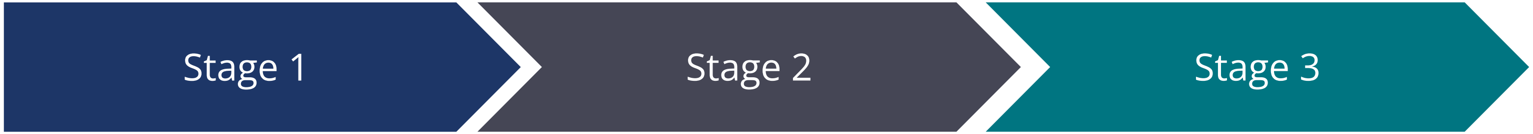


Mapping units of randomization to interventions

Variable name in SurveyCTO	[unitRandMap]
Field name **(LABEL on SurveyCTO)**	Mapping units of randomization to interventions
Definition **(LABEL on SurveyCTO)**	The unit of randomization for each intervention
Response options (open-text, numeric, date, text-CV)	Text-CV, select one
CV **(Choices on SurveyCTO)**	List of units selected in [Unit of randomization]
SurveyCTO instruction for data entry mask and repeat level	Skip if only one variable is selected in field [unit of randomization]
Coding instructions for coders **(Hint on SurveyCTO)**	If there is only one unit of randomization used for treatment assignment, this question will be skipped. When there is more than one unit of randomization (e.g. schools are assigned to a teacher training program and then families within schools are assigned to a parental support program), each unit will have to be mapped to an intervention. This information is mostly found in the randomization or methods sections of a paper or in a study participant flow diagram.
Descriptive example **see the [section] used in the paper to extract**	In Leaver et al, 2021, there are 5 interventions. The units of randomization in the study are -- district-subject-family and schools. Here, since there is more than one unit of randomization, we need to map each intervention to its unit of

Three-stage survey instrument

Survey fields are split into three stages for coding to ensure data quality and streamline the workflow



Stage 1

Structural fields about experimental design, outcome measures and exhibits.

Stage 2

Identification of eligible treatment effects for data extraction.

Stage 3

Study details & treatment effect estimates and precision.

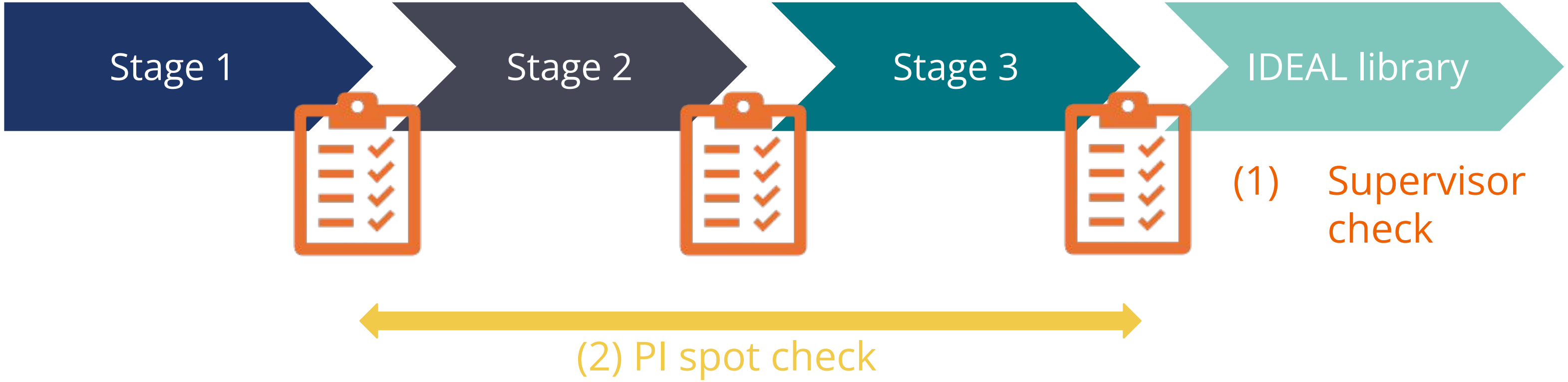
3 Quality assurance

With supervision protocol

Quality assurance measures

- All papers in the pilot will be double-coded for all three stages.

- All data will be checked by a supervisor ("supercoder") and spot checked by PIs.
  - Across survey stages, only supervisor-checked data can go to the next stage for coding. Timely feedback provided to coders.



Calculation of quality metrics

Accuracy

- A given entry is **accurate** if it represents what is described in the manuscript and verified by a supervisor and a PI.
- By coder, by paper, and by field

[illegible]

$$\frac{\#\ fields\ verified\ accurate}{Total\ number\ of\ fields\ in\ the\ paper}$$

Reliability

- **Reliability = 100%** if two coders extract the same information from the paper in the same way.
- Inter-coder reliability by field and by paper.

$$Reliability\ of\ a\ field =$$

$$\frac{\#\ double\ coded\ extractions\ scored\ as\ reliable}{Total\ number\ of\ unique\ papers\ coded}$$

4

Data entry mask

How does coding work in practice?

Data entry mask

- One data-entry survey form per stage.
    - Each survey form follows the coding protocol with build-in quality check features, repeated groups, and skip logics to facilitate data entry.

- 3 data-entry + 3 survey checks + a separate set of PI check surveys.
    - SurveyCTO platform (for training and pilot phase)
    - Custom data-entry portal compatible with automation tools

Data entry mask

Stage 1:

- Study-level information about experimental design
- Exhibit-level information on “structural” fields that characterize treatment effects

Stage 2:

- Mapping of structural factors to individual treatment effects in exhibit

Stage 3: Two separate survey forms

1) Study Details: all other relevant information about intervention, outcomes, etc.
2) Estimates: point estimates, precision values, etc.

SurveyCTO data-entry form

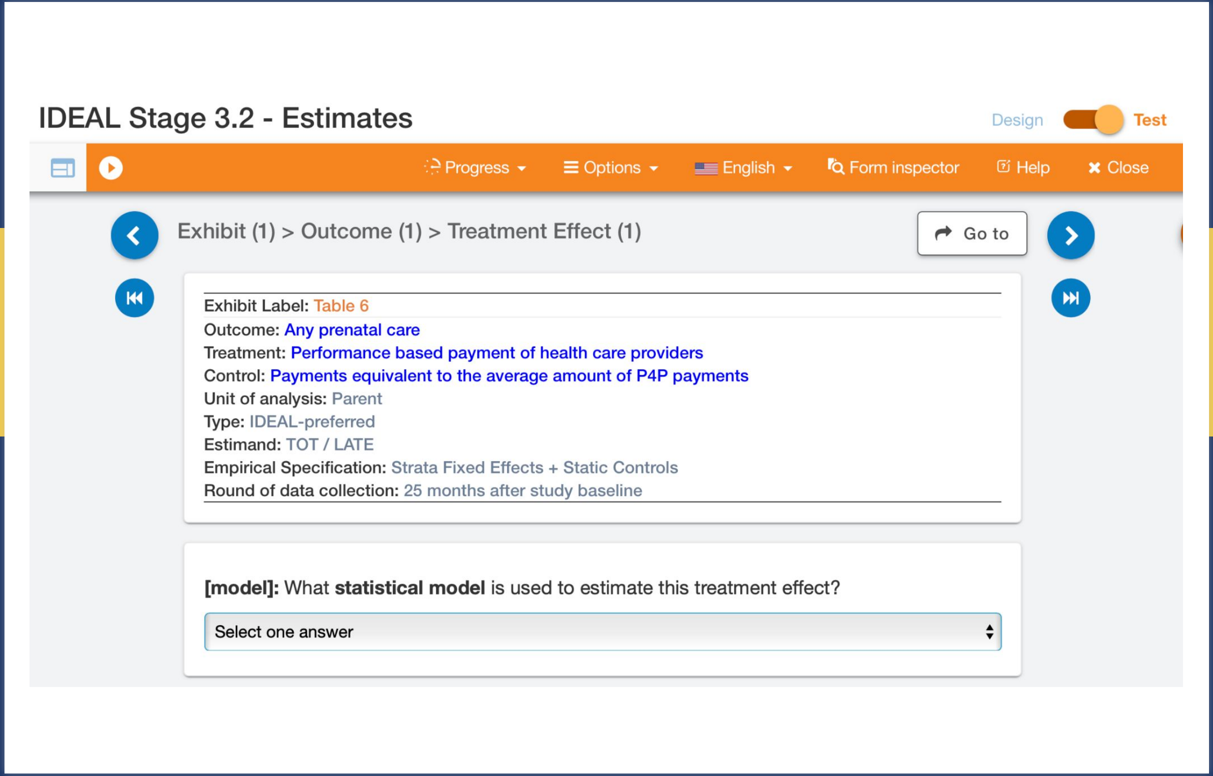


Exhibit (1) > Outcome (1) > Treatment Effect (1)

Exhibit Label: Table 6

Outcome: Any prenatal care

Treatment: Performance based payment of health care providers

Control: Payments equivalent to the average amount of P4P payments

Unit of analysis: Parent

Type: IDEAL-preferred

Estimand: TOT / LATE

Empirical Specification: Strata Fixed Effects + Static Controls

Round of data collection: 25 months after study baseline

[model]: What **statistical model** is used to estimate this treatment effect?

Select one answer

General coding principles

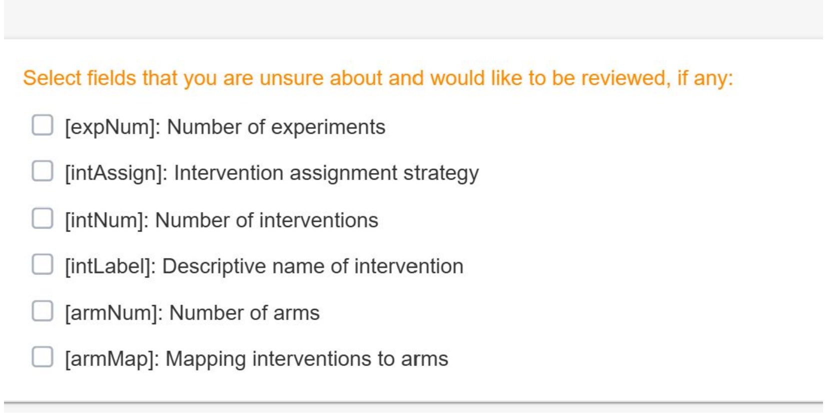
- **The goal is to accurately describe the evidence contained in the paper.**
  - Stick to the paper and report what’s in the paper.
  - Avoid unnecessary inference, speculation or calculation.
  - Be concise and look for facts.

- **Adhere closely to the coding protocols. Reach out for help.**
  - It helps the pilot to know the limitations of the protocol and survey forms.
  - Similar issues will appear for several papers, many of them already resolved.

Coding support

Field-specific questions: for things you already coded but you are unsure about

- Use SurveyCTO request-for-review sections.
- Supervisors will address in the review.
- Coder report with the submitted entry and selected entry for each stage.



Select fields that you are unsure about and would like to be reviewed, if any:

□ [expNum]: Number of experiments

□ [intAssign]: Intervention assignment strategy

□ [intNum]: Number of interventions

□ [intLabel]: Descriptive name of intervention

□ [armNum]: Number of arms

□ [armMap]: Mapping interventions to arms

Coding support

Paper-specific question: for issues that prevents you from coding

- Reach out to the supervisor of the paper. The supervisor’s name will be in the paper folder.
- Supervisor can respond to your question or escalate it to PIs.
- In some cases, the paper can be triaged.

General coding support:

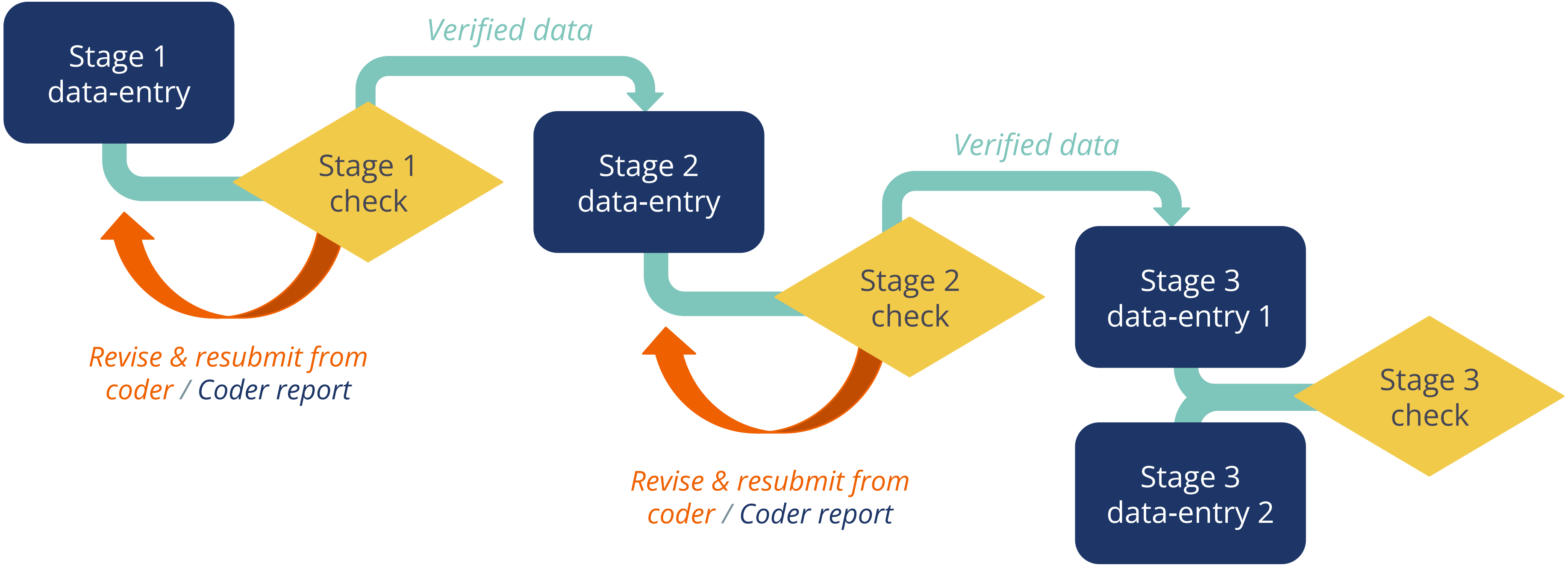
- Weekly drop-in office hours with supervisors
- Office hours by appointment with supervisors
- Quarterly coder meetings with PIs

5

Data extraction workflow

How does coding work in practice?

Data extraction workflow



Papers

Paper assignment

- Each coder receives a randomized set of papers from existing IDEAL collections.
- On SurveyCTO, each coder has their own “paper list” to code.
- Each paper has an unique ID (please use ID for support).

Access to papers

- You will have access to a Google Drive folder for each paper (main paper + supplementary materials).

Progress tracking

SurveyCTO entries will be tracked.

- Completed entries reported by SurveyCTO.
- Status tracking on GitHub by paper and by stage - done by supervisors.
- Regular notifications to coders and supervisors. Check your coder folder for reports and feedback.

Weekly schedule

- A weekly email with the number of papers to code and new assignment notification from the management team. [~ 20 hours per week]
- Reach out to us if you need a different schedule so we can plan ahead.

Performance assessments

Coder performance monitoring

- Progress on assignments
- Frequency of resubmission requests from supervisors
- Accuracy rate and inter-coder reliability

From coder to supervisor

- High-performing coders will be recommended as supervisors
- Supervisor has a slight different work program and is required to work **5-6** more hours per week

6 Training Program

Overview of this training

Different types of training sessions

This week:

- Overview of IDEAL
- Concept reviews of IDEAL fields
- Walkthroughs of IDEAL fields
- SurveyCTO demos

Next week:

- Practice sessions (asynchronous)
- Office hours

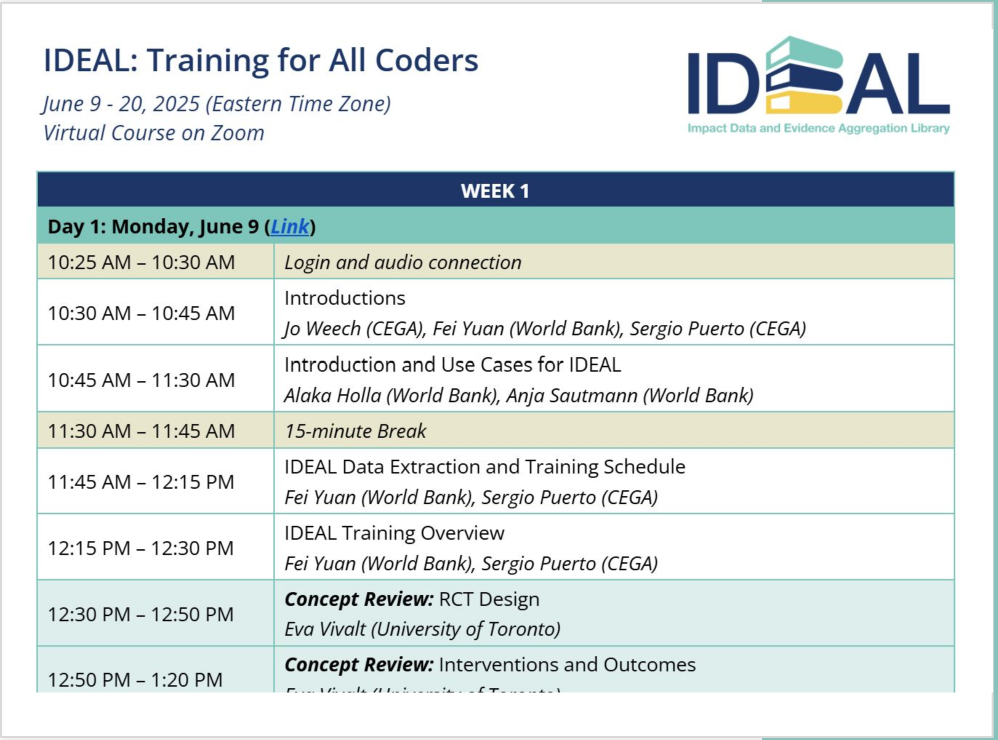
Later in the summer:

- Lectures on metascience and research transparency

Training schedule

Week 1 -
Live sessions
(June 9 - 12)

Four hours per day
10:25 - 14:30 ET



IDEAL: Training for All Coders

June 9 - 20, 2025 (Eastern Time Zone)
Virtual Course on Zoom

IDEAL
Impact Data and Evidence Aggregation Library

WEEK 1	
Day 1: Monday, June 9 (*Link*)	
10:25 AM – 10:30 AM	*Login and audio connection*
10:30 AM – 10:45 AM	Introductions *Jo Weech (CEGA), Fei Yuan (World Bank), Sergio Puerto (CEGA)*
10:45 AM – 11:30 AM	Introduction and Use Cases for IDEAL *Alaka Holla (World Bank), Anja Sautmann (World Bank)*
11:30 AM – 11:45 AM	*15-minute Break*
11:45 AM – 12:15 PM	IDEAL Data Extraction and Training Schedule *Fei Yuan (World Bank), Sergio Puerto (CEGA)*
12:15 PM – 12:30 PM	IDEAL Training Overview *Fei Yuan (World Bank), Sergio Puerto (CEGA)*
12:30 PM – 12:50 PM	***Concept Review:*** RCT Design *Eva Vivalt (University of Toronto)*
12:50 PM – 1:20 PM	***Concept Review:*** Interventions and Outcomes *Eva Vivalt (University of Toronto)*

Training schedule

Week 2 -
Practice sessions
(June 16-19)

Offline practice assignments
Office Hours

Lecture series

June 18:
Future of foreign aid - Dean Karlan

July 2:
Effect size and generalizability - Eva Vivalt

July 9:
Research transparency - Edward Miguel

Management team

-Fei Yuan, World Bank, fyuan@worldbank.org
-Sergio Puerto, CEGA, sergiopuerto@berkeley.edu
-Jo Weech, CEGA, jweech@berkeley.edu

Supervisors

-Jennie Barker, CEGA, jlbarker@berkeley.edu
-Mike Gibson, World Bank, mtgibson@umd.edu
-Winnie Mughogho, World Bank, mughoghowinnie@gmail.com
-Eric Dee, AidGrade, eric.dee@mail.utoronto.ca
Salim Benhachmi, AidGrade, salim.benhachmi@mail.utoronto.ca

Stay tuned and stay connected!

Thank you for listening

Fei Yuan
fyuan@worldbank.org

Sergio Puerto
sergiopuerto@berkeley.edu

IDEAL

Impact Data and Evidence Aggregation Library