Impact Data and Evidence Aggregation Library (IDEAL)
Data Extraction Guide

1. Overview of IDEAL

The Impact Data and Evidence Aggregation Library (IDEAL) is a collaborative initiative to create a comprehensive, open-source and open-access database of randomized controlled trials (RCTs) conducted in low- and middle-income countries across disciplines. IDEAL systematically documents impact evaluation studies to support evidence synthesis, meta-analysis, and evidence-informed policy and practice. This guide describes IDEAL's metadata schema, the survey, data extraction process, and quality assurance measures, providing researchers, policymakers, reviewers, and potential partners with a comprehensive understanding of how IDEAL operates.

Objectives and Principles. IDEAL is guided by four core principles that shape all aspects of the library's design and implementation:

- **Transparency:** IDEAL operates as an open-source, open-access global public good. All methodological decisions, coding protocols, and quality assurance measures are documented and publicly available. The library provides clear information about what data is collected, how it is extracted, and what standards are applied, enabling users to understand, use, and appropriately interpret IDEAL's data.
- **Reproducibility:** IDEAL's processes and outputs are designed to be replicated and used by third parties. Detailed protocols document each step of the review and coding process, along with an open-source data entry mask. The open-source approach ensures that partner organizations can adopt and adapt IDEAL's methods fully or even partially, and that the broader research community can verify and build upon IDEAL's work.
- **Minimum set of fields:** Rather than attempting to capture every possible detail about each study, IDEAL focuses on a carefully selected minimum set of fields essential for evidence synthesis and quality assessment. This approach reflects the library's multi-disciplinary scope, recognizing that the relevance and importance of specific information can vary across disciplines and research domains. In addition, it balances comprehensiveness with feasibility, enabling systematic coding at scale while capturing the most critical information for users to conduct meta-analyses, systematic reviews, and evidence-based decision-making.
- **Scalability and sustainability:** IDEAL's design prioritizes methods that can be implemented efficiently across large numbers of studies. The staged workflow, quality control procedures, and technology platform are all structured to support systematic data extraction from thousands of RCTs by various users while maintaining high data quality.

IDEAL Collaboration. IDEAL is a collaborative initiative launched in June 2023, bringing together member and partner organizations committed to building open-access evidence infrastructure for development research. The initiative operates through a technical working group consisting of Principal Investigators from IDEAL's member and partner organizations, who meet regularly to develop the registry's methodological outputs. IDEAL is an ongoing initiative that welcomes new collaborators constantly. The leading organization is the Strategic Impact Evaluation Fund at the World Bank with collaborators including the Development Research Group and the Development

Impact Group at the World Bank, the Center for Effective Global Action at the University of California, Berkeley, AidGrade, Northwestern University, Innovations for Poverty Action, and the International Initiative for Impact Evaluation (3ie).

This collaborative model reflects IDEAL's commitment to building sustainable, distributed capacity for evidence synthesis rather than centralizing all activities within a single institution. By training coders and supervisors across multiple organizations and establishing shared standards and protocols, IDEAL aims to create an infrastructure that can scale efficiently while maintaining quality.

IDEAL Outputs. IDEAL's infrastructure is being developed across four main outputs:

1. **Schema and extraction protocol:** IDEAL's metadata schema defines the conceptual framework for what information is extracted from each study, developed through consultation with global standards and a crosswalk of 15 evidence aggregation instruments. This conceptual schema is operationalized through structured data extraction instruments built in SurveyCTO, which translate the schema into specific fields, response options, validation rules, and skip logic. Together, these components enable consistent and systematic data extraction at scale.
2. **Library of RCTs**: IDEAL is building a comprehensive library of randomized controlled trials conducted in low- and middle-income countries. Each study in the library is coded using the standardized schema, creating consistent, comparable information across all studies. The library is designed to be interoperable, able to pull information from other platforms that use the same schema.
3. **Access platform**: IDEAL is developing a platform that allows researchers to download extracted data for their own analyses and presents information in accessible and interactive ways for non-technical audiences including policymakers and practitioners.
4. **Pedagogical resources:** To support broader adoption of evidence synthesis methods and transparency about IDEAL's approach, IDEAL is developing comprehensive pedagogical resources. These include training courses and technical materials for coders and external users, documentation of IDEAL's methodology and standards (including this review guide), and open-source data files that explain the assumptions underlying standard packages for meta-analysis in statistical software. These resources aim to mainstream evidence aggregation methods across the research community and enable other organizations to adopt or adapt IDEAL's approaches.

Meta-data Schema

The Impact Data and Evidence Aggregation Library (IDEAL) aims to present consistent information on treatment effects from randomized controlled trials (RCTs) conducted in the social sciences in low- and middle-income countries. While the goal is not to eliminate the need for meta-analytic researchers to read original studies, IDEAL will always present effect sizes that have been standardized across studies and information that will permit library users to restrict their comparisons by study attributes, such as context, experimental design, measurement, and intervention design and to understand what information relevant for evidence aggregation may be contained in the study, such as the estimation of heterogeneous treatment effects or the presence of a publicly available dataset.

To construct a set of minimum fields for this, the IDEAL team first conducted a desk-review, consulting the data collection tools used by researchers who had conducted meta-analysis, the fields used by existing evidence repositories (such as AidGrade and ClinicalTrials.gov), and guidelines used in evidence aggregation, such as PRISMA and GRADE. Next the team consulted external experts including more than 20 specialists in social sciences, evidence aggregation, and meta-analysis who participated in external working group meetings in September and October 2023. The meta-data schema was ultimately approved by the IDEAL Steering Committee in November 2024.

The Survey

To operationalize the data extraction for the minimum set of fields in the meta-data schema, the IDEAL team has developed a set of survey fields to capture relevant information from each individual paper through a series of working group meetings. Data extraction is currently conducted by human coders on a survey mask developed using Open Data Kit (ODK) tools in SurveyCTO. The initial dataset coded and checked by humans will serve as the ground truth data for future automated data extraction tools that would be integrated into IDEAL.

Staged Data Extraction Workflow. IDEAL employs a three-stage data extraction workflow designed to manage complexity, ensure quality, and create logical dependencies between different types of information:

- **Stage 1** extracts the structural characteristics of experiments, identifying how many experiments a paper reports, what interventions and study arms exist, which outcomes have treatment effects, what empirical specifications are used, when data was collected, and which exhibits (tables, figures, text) contain results. This structural mapping determines what questions appear in later stages.
- **Stage 2** systematically locates treatment effects by matching outcomes, arm comparisons, and specifications identified in Stage 1 with the actual results reported in each exhibit. IDEAL's prioritization hierarchy guides this process, ensuring that the most analytically appropriate estimates are identified for each outcome while also documenting author-preferred specifications when they differ.
- **Stage 3** collects comprehensive details about experiments, interventions, outcomes, samples, and the identified treatment effects. This includes intervention descriptions, outcome definitions, sample characteristics, implementation context, and the numerical values and precision measures for treatment effects.

Quality checkpoints between stages ensure that only validated information flows forward, with supervisor review after Stages 1 and 2, and double-coding employed in Stage 3 where the bulk of detailed information is extracted.

Roles and Responsibilities. IDEAL's quality assurance system depends on clearly defined roles:

- **Coders** extract information from published papers following detailed protocols and training. Coders complete all three stages for assigned papers, working systematically through the

SurveyCTO instruments and incorporating supervisor feedback to continuously improve their work.

- **Supervisors** review coding submissions after Stages 1 and 2, selecting or correcting answers to create validated datasets for subsequent stages, and providing detailed feedback to coders. Supervisors also reconcile discrepancies between the two independent coders in Stage 3 to produce final consensus data. Supervisors also host regular office hours to answer questions from coders.
- **Principal Investigators (PIs)** conduct spot checks on a subset of supervisor-reviewed papers (i.e. 20% initially and then 10%), provide additional quality oversight, establish ground truth for training papers, and make final decisions on methodological questions or complex coding issues that arise during implementation.
- **Data management team** monitors the data extraction process by tracking progress, maintaining the data entry mask and processing submitted data. The management team organizes regular check-ins with coders, supervisors and PIs to provide updates on technical and process issues in data extraction.

Using This Guide. This guide is organized into six main sections following this overview:

- **Section 2** describes the screening criteria that determine which studies are eligible for IDEAL review
- **Section 3** defines what constitutes an "experiment" in IDEAL's framework and how to identify single versus multiple experiments
- **Section 4** details the core components IDEAL extracts from each experiment and why this information matters
- **Section 5** explains the three-stage data extraction workflow and how information flows through the system
- **Section 6** describes IDEAL's quality assurance procedures, training requirements, and reliability metrics

Together, these sections provide a complete picture of how IDEAL transforms published experimental research into a systematic, searchable database of evidence.

2. Screening & Scope

The IDEAL Registry applies systematic screening criteria to identify eligible studies for inclusion. This section describes the specific requirements studies must meet to be considered for coding, including manuscript characteristics, topic and geographic scope, design characteristics and reporting of treatment effects. These criteria ensure that the registry maintains methodological consistency and includes only those studies that meet minimum quality and relevance thresholds.

2.1 Study Eligibility Requirements

Eligible research report. To be eligible for IDEAL coding, studies must meet the following requirements.

- **Peer-Reviewed.** Studies must be published in a peer-reviewed journal to be eligible for IDEAL inclusion. Peer review provides an independent assessment of methodological quality and ensures that studies have undergone systematic evaluation by subject matter experts prior to publication. This requirement ensures a baseline level of methodological scrutiny and accessibility for the initial registry development. Future phases may expand to include other publication venues and formats.
- **Availability.** Studies must have a DOI to guarantee permanent access. This requirement ensures that IDEAL registry users can access and review the original studies and facilitates transparency in the coding process.
- **Timeframe** Studies published from [YEAR] onward are eligible.
- **Language.** Studies must be available in English.

Eligible countries and topics. Studies must be implemented in **low- and middle-income countries**. IDEAL defines low- and middle-income countries according to World Bank income group classifications at the time of the study's implementation. This geographic restriction focuses the registry on contexts most relevant to international development policy and practice, while recognizing the distinct challenges and opportunities present in resource-constrained settings. Studies conducted entirely in high-income countries are not eligible, regardless of their methodological rigor or policy relevance.

IDEAL adopts a **cross-disciplinary approach** to topical coverage, encompassing field experiments across diverse substantive areas including but not limited to education, health, governance, agriculture, labor markets, finance, environment, and social protection. This broad scope reflects the registry's goal of providing comprehensive documentation of experimental evidence relevant to development policy and practice.

Eligible study design. The study design must be based on **randomized controlled trials (RCTs)** that evaluate field or policy interventions. Random assignment to treatment and control conditions provides the most rigorous approach to causal inference by ensuring that treatment and comparison groups are statistically equivalent in expectation, thereby mitigating selection bias and enabling credible estimation of intervention impacts.

Exclusions. Some experimental studies are excluded from the IDEAL review even if they employ random assignment.

- Lab-in-the-field experiments are not eligible for inclusion. While these studies may use randomization and occur in field settings, they typically involve artificial decision-making tasks or controlled experimental environments that differ substantively from policy-relevant field interventions.
- Design interventions are excluded from the registry. These studies randomize the order, framing, or wording of survey questions to examine measurement effects rather than testing substantive policy or programmatic interventions. These experiments fall outside the scope of policy-relevant field trials that IDEAL aims to catalog.
- Hypothetical experiments are not eligible. Studies must report results from interventions that were actually implemented. Proposed experiments, pilot protocols without results, or

simulations of potential interventions do not meet the registry's implementation requirement.

Eligible treatment effects reporting. To be eligible for inclusion, studies must report **at least one quantitative treatment effect** with available data on magnitude of the effect and statistical precision. This information may appear in tables, figures, or text within the manuscript. The estimated effect can be reported using any appropriate metric (e.g., mean differences, proportions, odds ratios, hazard ratios). Acceptable forms of precision information include standard errors, confidence intervals, p-values, or test statistics from which precision can be calculated. Studies that report only descriptive statistics, qualitative findings, or treatment effect estimates without data do not meet this minimum reporting threshold. This requirement ensures that included studies contribute quantitative evidence on intervention impacts that can inform future meta-analyses and evidence synthesis efforts. Studies may report effects on any outcome domain relevant to the intervention.

3. Experiments in IDEAL: Definitions and Scope

IDEAL reviews published manuscripts that report results from randomized controlled trials. Within each manuscript review, IDEAL identifies and codes one or more experiments. This section defines what IDEAL considers an "experiment," provides guidance for determining whether a manuscript contains a single experiment or multiple experiments, and describes the intervention assignment strategies such as parallel, factorial, crossover, and adaptive designs that IDEAL documents for each experiment. Because published papers vary widely in their complexity, some reporting a single straightforward trial while others presenting multiple related or independent experiments with sophisticated randomization schemes, clear definitions are essential. These distinctions are critical for users conducting meta-analyses, as they affect sample independence and appropriate statistical methods for evidence synthesis.

3.1 Defining an Experiment

For the purposes of IDEAL review, an **experiment** is defined as a **unified research design** that uses randomization to create a valid counterfactual to evaluate the causal effects of one or more interventions on a **defined sample.**

The critical question in determining whether a paper contains one or multiple experiments is: Were the randomization procedures designed and implemented as parts of an integrated research design, or do they represent independent studies?

When Does a Manuscript Contain Multiple Experiments?

A manuscript contains multiple experiments when it reports results from **fundamentally separate research designs** that operate independently. Key indicators of multiple experiments include:

- **Independent samples:** Different populations or geographic areas were sampled and randomized separately. For example, a manuscript reporting an education RCT conducted in

100 Kenyan schools and a separate health RCT conducted in 50 Tanzanian villages contains two experiments. A paper studying an RCT in one region of a country and another RCT in another region in the same country also has two experiments.

- **Unrelated interventions tested separately:** Different interventions were evaluated through separate, non-integrated randomization schemes. For example, if a manuscript reports that 80 schools were randomized to teacher training versus control in 2015, and separately reports that 60 different schools were randomized to student tutoring versus control in 2017, these are two distinct experiments.
- **Sequential studies without integration:** The manuscript reports results from studies conducted at different times with different samples, even if testing similar interventions. While results may be presented together for comparison, the underlying experiments were independent.

Exclusions When a manuscript includes both eligible field experiments and ineligible designs (such as lab-in-the-field experiments or a hypothetical intervention), only the eligible field experiment(s) are coded.

Complex Designs Within One Experiment

Many experiments employ sophisticated designs involving multiple stages or dimensions of randomization. These design features do not create separate experiments when they are part of an integrated research plan. Examples of complex single experiments include:

- A study that randomizes schools to teacher training and independently randomizes schools to instructional materials, creating four experimental conditions (Training+Materials, Training only, Materials only, Neither). This is one experiment testing two interventions and their potential interaction, even though it involves two dimensions of randomization.
- A study might first randomize villages to receive a cash transfer program or not, then randomize individuals within program villages to receive cash transfers or not, while all individuals in control villages receive no transfers. Despite having different units of randomization at each stage (villages, then individuals), this is one experiment with an integrated design that tests both direct effects of transfers and spillover effects within treated villages.
- A study randomizes districts to an intervention, then within treated districts randomizes schools to different implementation approaches. If designed as a unified study, this represents one experiment with nested randomization stages that work together to answer related research questions.
- A study randomizes schools to receive an intervention in Year 1, Year 2, or Year 3, using earlier cohorts as controls for later cohorts. This is one experiment with a phased-in design, even though treatment timing varies across units.
- A study that randomizes units to Intervention A, Intervention B, Intervention C, or Control is one experiment comparing multiple treatment variants against a common control group.

The defining characteristic of these complex designs is that all randomization stages or dimensions were planned and executed as components of a single research design to answer related research questions.

3.2 Intervention Assignment Strategies

Within each experiment, IDEAL documents the overall design structure that determines how interventions are allocated to study arms. Understanding these assignment strategies is important for interpreting results and conducting appropriate statistical analyses. IDEAL classifies experimental designs into the following categories:

- **Parallel Design:** The most common strategy in randomized controlled trials. Each intervention is assigned to only one arm, creating distinct groups that can be compared (e.g., treatment vs. control, or multiple treatment arms with different interventions). Units are randomly assigned to separate groups and maintain their assignment throughout the study.
    - Example: Schools are randomized into three distinct groups, recognition program, in-kind performance rewards, and control group.
- **Factorial Design.** Used when evaluating two or more interventions both alone and in combination. At least one intervention is assigned to more than one study arm, allowing researchers to test individual effects and interaction effects between interventions. This design enables examination of whether interventions work better in combination than separately.
    - Example: Four groups are created, psychosocial stimulation only, micronutrient supplementation only, both interventions combined, and control (neither intervention).
- **Crossover Design.** Each study arm receives different interventions (including no intervention) at different phases of the study. This category also includes phase-in or stepped-wedge designs where intervention rollout is randomized and every unit ultimately receives the program. **Important note:** If the study's endline data collection occurs before units receive interventions beyond their initial assignment, this is not classified as crossover but rather as parallel design with staggered implementation.
    - Example: Doctors and patients switch between control and treatment conditions on different days, creating within-unit comparisons.
- **Adaptive Design.** The randomization rule can change during the trial based on experimental data collected during the study. Assignment probabilities may be modified based on interim results from previous waves or outcomes observed under current interventions. These designs are particularly useful when researchers want to maximize the number of units receiving effective treatments while still maintaining experimental rigor. **Important note: The IDEAL survey cannot accommodate studies using an adaptive experimental design.**
    - Example: In multi-wave trials, the proportion of units assigned to each treatment arm changes across waves based on results from prior waves. Or the timing of switching interventions depends on outcomes observed under the current intervention, with more effective treatments receiving higher assignment probabilities in later waves.

- **Other.** For assignment strategies that don't fit the above categories. When this classification is used, IDEAL coders specify the nature of the design in detail.

3.3 Multiple Manuscripts from One Experiment

Multiple manuscripts may present results from a single experiment. IDEAL codes these manuscripts separately but identifies them as **related studies** ex post to inform users that the samples are not independent. Common scenarios include:

- **Different follow-up periods:** One manuscript reports short-term effects at 12 months while another reports long-term effects at 36 months of the same participants receiving the same intervention. These manuscripts present different temporal analyses of one experiment.
- **Different outcomes:** One manuscript reports education outcomes while another reports health outcomes from the same experimental sample and intervention. These are different manuscripts analyzing different outcomes from one experiment.
- **Different analytic samples:** One manuscript analyzes the full experimental sample while another focuses on a pre-specified subgroup. These represent different analyses of one experiment.
- **Sequential interventions to the same sample:** If a study first implements Intervention A, collects endline data, publishes results in one manuscript, then later implements Intervention B to the same experimental sample and publishes results in a second manuscript, IDEAL codes these manuscripts separately but labels them as **related studies** because they share the same participants. Users should be aware that results from these manuscripts are not independent, a critical consideration for meta-analysis.

3.4 Related Studies from Overlapping Samples

To the extent possible, given the scale of the IDEAL library, IDEAL identifies and flags related studies when different experiments share substantial portions of their samples, even if the interventions or randomization approaches differ. "Substantial overlap" typically means that a majority of participants in one study also participated in the other, though reviewers use judgment based on the specific context.

For instance, if a new experiment is overlaid on an existing RCT's sample, both studies would be coded and flagged as related when this relationship can be determined from the published literature. This transparency about sample overlap, where detectable, enables appropriate statistical methods when synthesizing evidence and helps users avoid treating non-independent samples as independent observations in meta-analyses.

4. IDEAL Coding Components

IDEAL organizes information into five interconnected components that together provide a comprehensive picture of the experimental research: the design structure, the population studied, the interventions tested, the outcomes measured, and the effects estimated. These components were identified through consultation with global standards (including the Data Documentation Initiative

and clinical trial registries) and a detailed crosswalk of 15 evidence aggregation instruments used by leading research organizations.

This standardized approach enables diverse users to locate studies relevant to their questions and extract the information needed for their purposes. Because IDEAL documents what study authors report without imposing quality ratings, users can apply their own frameworks and standards to the data.

4.1 Core Components

1. Experimental Design Structure: IDEAL collects the basic architecture of each experiment, including how many experiments a paper evaluates, what study arms exist within each experiment, and how units were assigned to these arms. Understanding experimental structure is fundamental to interpreting any findings. It clarifies what comparisons are being made and how treatment assignment was determined. This information allows evidence synthesis across studies with similar designs and enables appropriate statistical methods that account for design features like clustering or stratification.

This component includes:

- Number of distinct experiments in the paper
- Study arm labels and descriptions
- Unit(s) of randomization—the level at which random assignment occurred (individuals, households, classrooms, schools, communities, or other clusters)
- Randomization method (simple, blocked, stratified, or other approaches)
- Number of units assigned to each arm and overall

2. Study Sample: IDEAL collects characteristics that define who participated in the study, how they were identified and recruited, and at what scale the intervention operated. Population characteristics determine the scope of inference for any study's findings and are essential for assessing external validity. Documenting eligibility criteria, target populations, and sampling approaches enables evidence synthesis across similar populations and supports assessments of whether findings might generalize to other contexts. Recording sample sizes at randomization and analysis provides the foundation for calculating attrition, a key quality indicator for RCTs.

This component includes:

- Eligibility criteria (who could participate) and exclusion criteria (who could not)
- Target population description
- Geographic and institutional scale of implementation
- Sampling and recruitment approach
- Sample sizes at randomization and at analysis

3. Interventions: IDEAL collects detailed descriptions of what participants in each study arm experienced, including both treatment interventions and comparison conditions. Comprehensive intervention documentation is necessary for multiple purposes: assessing the feasibility of

replication, understanding what aspects of interventions may have driven effects, determining whether interventions across studies are similar enough for evidence synthesis, and evaluating whether findings might transfer to different implementation contexts. Complete descriptions of comparison conditions are equally important, as the contrast between treatment and comparison determines what the effect estimate actually represents.

This component includes:

- Intervention content and components
- Intensity: duration, frequency, dosage, or amount delivered
- Delivery mode and mechanisms (in-person, remote, who delivered it)
- Implementation context: geographic location, institutional settings
- Implementation fidelity and adherence when reported
- Complete comparison condition descriptions (not just "control" but what comparison participants actually experienced)

4. Outcomes and Measurement: IDEAL collects information about outcomes for which treatment effect estimates are reported, including what constructs they measure and how measurement occurred. Detailed outcome documentation enables identification of studies measuring relevant constructs and assessment of construct validity. Understanding measurement approaches is essential for evidence synthesis, as different instruments purporting to measure the same construct may not be directly comparable. Documentation of units of analysis and timing clarifies what effects represent.

This component includes:

- Outcome variable definitions
- Measurement instruments and approaches used
- Unit of analysis (which may differ from the unit of randomization)
- Reference periods (outcomes measured "in the past week" versus "past month")
- Timing of data collection rounds

5. Treatment Effect Estimates: IDEAL collects treatment effect estimates that include both effect magnitude (point estimates) and precision (standard errors, confidence intervals, or p-values), drawn from exhibits in the main text of papers. Treatment effects are the primary output of experiments and the foundation for evidence synthesis. However, studies often report multiple estimates for the same outcome using different samples, analytical approaches, or model specifications. Systematic documentation of which estimates are reported and how they were produced is essential for reliable meta-analysis and evidence synthesis. IDEAL's prioritization approach ensures comparable estimates across studies while also documenting author preferences, allowing users to examine how analytical choices affect conclusions.

4.2 IDEAL's approach to selecting treatment effect estimates

To select treatment effect estimates, IDEAL follows a systematic prioritization approach that balances standardization with documentation of study-specific decisions.

Estimand: IDEAL prioritizes intention-to-treat (ITT) effects. ITT estimates include all participants as originally assigned regardless of whether they received the intervention. These effects answer the policy-relevant question: "What is the impact of offering this intervention?" IDEAL collects Local Average Treatment Effect (LATE), Treatment-on-the-Treated (TOT), or other estimands when ITT estimates are unavailable or when LATE/TOT/other is the authors' preferred estimand. LATE and TOT measure effects for those who comply with the assignment or those who actually participated, respectively.

Sample: IDEAL prioritizes effects for the full experimental sample. Subgroup effects (for example, effects separately for boys and girls) are collected when full-sample results are unavailable or when subgroup analysis is the study's primary research question.

Analytical specification: Studies often report multiple specifications—different combinations of statistical controls—for the same outcome. IDEAL selects specifications according to a hierarchy that prioritizes models adjusting for design features and baseline outcomes: first, models with stratification controls (when stratification was used) plus baseline outcome measures; second, models with stratification controls only; third, models with stratification controls plus other baseline covariates, choosing the simplest model when multiple options exist.

Author preferences: When study authors explicitly indicate a preferred specification that differs from IDEAL's hierarchy, both the IDEAL-preferred and author-preferred estimates are collected. This allows examination of how analytical choices affect conclusions while maintaining standardized estimates across studies.

Completeness and location: Eligible treatment effects must appear in the main text (in tables, figures, or narrative) and include both a point estimate and at least one precision measure. Supplementary materials may be consulted to find precision information for effects presented in the main text, but effects appearing only in appendices without main-text presentation are not coded.

5. The IDEAL Data Extraction Process

IDEAL's data extraction process translates the conceptual components described in Section 4 into a systematic, staged workflow with integrated quality controls. This section provides an overview of how coding is implemented, the rationale for IDEAL's three-stage approach with supervisor checkpoints, and how information flows from paper to database. The process employs strategic double-coding at the final stage where the bulk of detailed information is collected, balancing data quality with coding efficiency. Detailed field-by-field coding protocols and instructions are maintained separately; this section focuses on the overall architecture and logic of the extraction process to help partners and users understand how IDEAL transforms published manuscripts into structured, reusable data.

5.1 Three Stages of Data Extraction

IDEAL uses a three-stage architecture for data extraction rather than a single comprehensive survey. This design serves multiple purposes: it manages the complexity of extracting detailed information

from diverse experimental papers, ensures quality through built-in checkpoints before information flows forward, and creates logical dependencies where earlier stages define the scope and structure of later stages. Each stage builds systematically on the previous one, with supervisor review between stages ensuring that only accurate structural information propagates through the workflow. The result is an efficient process that maintains high data quality while avoiding the need for extensive ex-post data cleaning or reconciliation.

Stage 1: Mapping the Experimental Structure

Stage 1 identifies the architecture of what is being tested and where results appear in the manuscript. Coders document the number of experiments in the paper, the interventions and study arms within each experiment, the outcomes for which treatment effects are reported, the empirical specifications used in analyses, the rounds of data collection, and critically, the exhibits (tables, figures, or text sections) that contain treatment effects.

This stage comes first because these structural elements determine what questions appear in later stages. Most importantly, identifying exhibits in Stage 1 organizes all subsequent coding—Stages 2 and 3 present questions exhibit-by-exhibit, allowing coders to work systematically through the paper's results table-by-table or figure-by-figure rather than jumping around the manuscript.

The data collection platform includes automated validation checks that prevent logical inconsistencies and data entry errors during all stages. For example, if a coder indicates there are three interventions but only provides labels for two, the system flags this discrepancy. Once the coder completes Stage 1, a supervisor reviews and approves the outputs before the coder can proceed to Stage 2. This early checkpoint is critical because errors in identifying the experimental structure would cascade through all subsequent stages.

Stage 2: Locating Treatment Effects

Stage 2 uses the structural information from Stage 1 to identify which specific treatment effects exist for each outcome and to determine which estimates are eligible for extraction using IDEAL's preferred specifications. The survey automatically generates questions for all possible combinations of outcomes, arm comparisons, and empirical specifications identified in Stage 1, then presents these questions organized by exhibit.

For example, when a coder reaches Table 3 (identified in Stage 1 as containing treatment effects), the survey asks about each potential treatment effect that could appear in that table: "Does Table 3 contain a treatment effect for math scores comparing Treatment A to Control using Specification 1 for measurement period X?" This exhibit-based organization allows coders to focus on one table or figure at a time, systematically confirming which treatment effects are present.

Survey logic incorporates IDEAL's prioritization hierarchy for specifications, stopping the search once preferred estimates are found for each outcome. This efficiency feature means coders do not answer unnecessary questions about lower-priority specifications when higher-priority ones are available. As in Stage 1, the data collection platform includes automated validation checks that catch errors and inconsistencies as coders work.

After completing Stage 2, a supervisor reviews and approves the entered data before the coder proceeds to Stage 3. This checkpoint ensures that only the eligible treatment effects—those actually present in the paper and matching IDEAL's selection criteria—are flagged for detailed extraction in the next stage.

Stage 3: Extracting Detailed Information

Stage 3 collects comprehensive details about the experiment, interventions, outcomes, and treatment effect estimates. This is the most extensive stage, covering the bulk of IDEAL's metadata including intervention descriptions (content, intensity, delivery mechanisms, implementation context), outcome definitions (measurement instruments, units of analysis, time periods), sample characteristics (eligibility criteria, sample sizes at randomization and analysis), and treatment effect values (point estimates, standard errors, confidence intervals, p-values, sample sizes used in estimation).

When extracting treatment effect estimates, questions are organized by exhibit, mirroring the structure from Stage 2. For each treatment effect identified in Stage 2, coders enter the numerical values and precision measures found in that specific table or figure. All collected data is automatically associated with the correct experiments, arms, outcomes, and specifications identified in previous stages, eliminating the need for ex-post linking or matching.

Because Stage 3 captures the most detailed information and involves numerical data entry where errors could occur, this stage undergoes double-coding. Two independent coders complete Stage 3 for each paper, with discrepancies between the two coded entries resolved through supervisor review (described in Section 6).

Data Flow and Integration

Information flows seamlessly across the three stages through the data collection platform, SurveyCTO, preloading functionality. Data collected in Stage 1—such as intervention labels, outcome names, and exhibit identifiers—automatically populate the relevant questions in Stage 2. Similarly, the treatment effects confirmed in Stage 2 determine which detailed extraction questions appear in Stage 3. By the end of Stage 3, all data resides in one integrated relational dataset with clear linkages between structural elements, treatment effect locations, and detailed estimates. No ex-post linking across stages is required.

5.2 The SurveyCTO Platform

IDEAL uses SurveyCTO as its data collection platform in the pilot stage, chosen for its accommodation of ODK tools that are widely used in development research and fieldwork. The platform's automated validation catches common errors in real-time during Stages 1 and 2—for example, flagging if a coder indicates three interventions but only labels two, or if responses create logical inconsistencies. Dynamic survey logic ensures that questions in later stages automatically populate based on earlier responses, so Stage 2 generates questions for each outcome identified in Stage 1, and Stage 3 presents extraction prompts for each treatment effect confirmed in Stage 2.

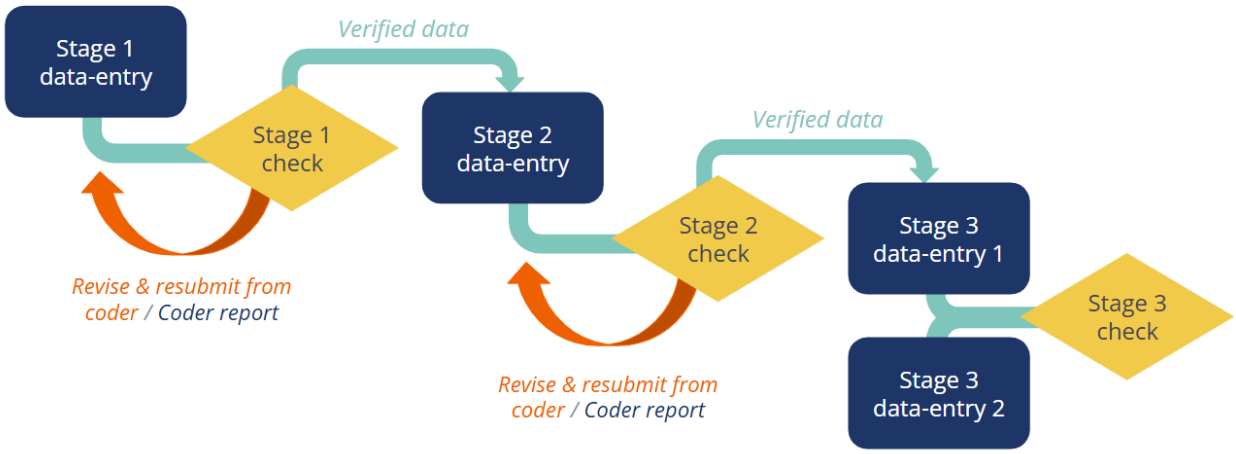
The platform also supports supervisor review and approval workflows between stages, preventing coders from proceeding until their work has been validated. Detailed survey forms translate IDEAL's conceptual framework into structured questions with appropriate skip patterns, validation rules, and consistency checks, all designed to facilitate coding, minimize errors and ensure data quality.

5.3 Coding Workflow in Practice

The complete workflow for each paper proceeds as follows. A coder completes Stage 1, mapping the experimental structure and identifying exhibits. A supervisor reviews this work and either approves it (allowing the coder to proceed) or flags issues for correction. Once Stage 1 is approved, the same coder proceeds to Stage 2, systematically working through each exhibit to locate treatment effects. After Stage 2 completion, a supervisor again reviews and approves before the coder moves forward.

Stage 3 follows a different approach. Two independent coders each complete Stage 3 for the paper, extracting detailed information without seeing each other's work. Supervisors then compare the two codings, identify discrepancies, and work with coders to resolve differences and reach consensus on the final values. This double-coding at Stage 3 ensures reliability of the detailed data—intervention descriptions, outcome definitions, and treatment effect estimates—while maintaining efficiency in the earlier structural stages where supervisor review provides sufficient quality control.

The time required varies by paper complexity. A straightforward two-arm trial with a few outcomes presented in simple tables requires less time than a complex factorial design with 20 outcomes reported across multiple tables with various specifications. Training requirements and quality assurance procedures are detailed in Section 6.



6. Quality Assurance and Reliability

IDEAL employs a multi-layered quality assurance system to ensure accurate and consistent data extraction across all coded papers. The approach combines comprehensive training, built-in

validation checks during coding, structured supervisor review between coding stages, double-coding of detailed information in Stage 3, and systematic calculation of quality metrics. This section describes each component of IDEAL's quality assurance framework and the metrics used to monitor and improve coding quality during the pilot phase.

6.1 Training and Preparation

Coders are hired through a competitive process with a screening coding exercise. Before coding live papers, all coders complete a comprehensive training program. The training package includes detailed coding protocols, training slides and videos for concepts and fields prepared by Principal Investigators, and training videos demonstrating how to complete surveys on SurveyCTO.

During training, coders complete 10 practice papers to develop proficiency with the coding system. For these training papers, supervisors extract the ground truth—the correct answers for each field—which is then reviewed by PIs to ensure accuracy. This ground truth serves two purposes: providing guidance to coders as they learn the system, and calculating accuracy metrics to assess coder readiness.

Throughout the training period and during coding periods, office hours sessions are hosted to answer coders' questions and provide additional support. [PLACEHOLDER: Information needed on certification/readiness criteria before coders can begin coding live papers, if formal assessment exists beyond practice papers.]

6.2 Built-in Quality Controls During Coding

The SurveyCTO platform incorporates automated validation checks and skip logics in all stages of data extraction to prevent errors in real-time. These checks catch logical inconsistencies, such as when a coder indicates there are three interventions but only provides labels for two, or when responses create contradictions across related fields. They also streamline coding by skipping irrelevant questions based on the previous answers to related fields. For example, the mapping of interventions to study arms will be skipped if the assignment strategy is parallel as a study arm will be automatically generated for each intervention. These built-in controls reduce coding errors at the point of data entry and improve efficiency by alerting coders to problems immediately rather than discovering them during later review stages.

6.3 Supervisor Review Process

IDEAL implements two rounds of quality checks during the coding process for each paper: after Stage 1 and after Stage 2. These supervisor reviews serve dual purposes—building a dataset with correct answers for the next coding stage, and providing feedback to coders for continuous improvement.

In Stages 1 and 2, the supervisor reviews the information in each field. The supervisor can correct minor errors or provide feedback to the coder and request resubmission of Stage 1 or 2 if major errors are made in the coding. When revisions are submitted, the review process starts over

Supervisor reviews are built into the coding workflow through SurveyCTO. Coders can only progress to the next stage after the supervisor completes and approves the review, ensuring that only validated information flows forward.

While Stages 1 and 2 rely on supervisor review for quality control, Stage 3 employs double-coding. Two independent coders complete Stage 3 for each paper, extracting detailed information without seeing each other's work. Because Stage 3 captures the bulk of IDEAL's metadata—intervention descriptions, outcome definitions, sample characteristics, and treatment effect values—double-coding at this stage ensures reliability of the final data while maintaining efficiency in earlier stages.

After both coders complete Stage 3, supervisors compare the two sets of coding, identify discrepancies, and select or enter the preferred values that enter the dataset. [PLACEHOLDER: Specific information needed on the reconciliation process—how discrepancies are resolved and criteria for final decisions.]

6.4 PI Spot Check

An additional quality assurance layer occurs through PI spot checks. PIs from each organization conduct spot checks on a random subset of fields and papers after supervisor review. Initially, PIs check 20% of fields for 20% of randomly selected papers; this may be reduced to 10% of fields later in the pilot as quality stabilizes.

If a PI flags an error or has comments during the spot check, the supervisor responsible for that paper corrects the corresponding data in the surveys. The PI spot check is conducted through a standalone survey on SurveyCTO and serves as an additional quality assurance measure independent of the main coding workflow.

6.5 Quality Metrics

IDEAL calculates two primary quality metrics during the pilot phase: accuracy and reliability. These metrics use data from supervisor checks and PI spot checks and are calculated at multiple levels to provide detailed insights into coding quality.

The accuracy rates will be calculated using the comparison between the supervisor checked data and the PI cleared data. Therefore, the statistics will be based on the set of papers that go through a PI check.

Accuracy. A given entry is considered accurate if it represents what is described in the manuscript and is verified by a PI for information that requires inference. Accuracy is calculated at several levels:

Accuracy of a paper = (Number of fields verified accurate) / (Total number of fields in the paper)

Accuracy of a section per paper = (Number of fields verified accurate in the section) / (Total number of fields in the section)

Average accuracy of a supervisor = Sum of accuracy per paper across all n papers coded / (Total number of papers coded by that coder)

Reliability. A given entry is considered reliable if two coders extract the same information from the paper in the same way. Importantly, two coders could extract information in different ways and still both be accurate. For example, defining interventions as "training" and "cash" versus "training" and "training plus cash" could both accurately describe a trial but would not be coded identically. IDEAL aims for coders to be both accurate and consistent with each other.

Given the IDEAL multi-stage coding workflow, reliability for Stage 1 and Stage 2 fields is calculated by comparing the entries between a coder and a supervisor. For Stage 3, when double coding is performed by two coders, the reliability is calculated based on the entries by two coders.

Reliability is calculated as:

Reliability of a field = (Number of double-coded extractions scored as reliable) / (Total number of unique papers coded)

Average reliability of a section = (Average reliability of fields in the section) / (Total number of unique papers coded)

[PLACEHOLDER: Information needed on target reliability thresholds, if any exist, and how reliability data is used to identify training needs or problematic fields.]

[PLACEHOLDER: Appendix table to list the fields to be included in the calculation of accuracy and reliability.]

6.7 Ongoing Quality Monitoring

- PI-cleared data processing to check and remove outliers.
- Feedback request form available on the IDEAL data platform.

[PLACEHOLDER: Information needed on processes for ongoing quality monitoring beyond the formal metrics—spot checks of completed papers, tracking patterns in discrepancies, identifying fields needing clearer guidance, iterative improvements to training and protocols based on pilot experience.]