A Metadata Schema for Data from Experiments in the Social Sciences [*]

Jack Cavanagh[†]   Jasmin Fliegner[‡]   Sarah Kopper[†]   Anja Sautmann[§]

December 28, 2022

Abstract

The use of randomized controlled trials (RCTs) in the social sciences has greatly expanded, resulting in newly abundant, high-quality data that can be reused to perform methods research in program evaluation, to systematize evidence for policymakers, and for replication and training purposes. However, potential users of RCT data often face significant barriers to discovery and reuse. We propose a metadata schema that standardizes RCT data documentation and can serve as the basis for one – or many, interoperable – data catalogs that make such data easily findable, searchable, and comparable, and thus more readily reusable for secondary research. The schema is designed to document the unique properties of RCT data. Its set of fields and associated encoding schemes (acceptable formats and values) can be used to describe any dataset associated with a social science RCT. We also make recommendations for implementing a catalog or database based on this metadata schema.

1 Introduction

The use of randomized control trials (RCTs) in the social sciences has greatly expanded over the past two decades, from economics and political science to public health and education.[1] In parallel, journals, funders, policy organizations, and organizations promoting open science have emphasized that original research data be publicly accessible for others to analyze and use. As a result of this concerted push, hundreds of original RCT datasets from a wide variety of contexts and populations are already published and in principle accessible to researchers, with the potential to benefit numerous areas of research.

However, there remain to date significant barriers to the discovery and use of existing RCT data. Published datasets are scattered across data repositories, journal and university websites, and researcher homepages. In many data repositories, filter options and documentation fields are broad, often consisting of free form text fields, and data quality and provenance are hard to assess. RCT data are often published with a focus on replicating the analysis in a specific paper, and the properties specific to RCTs and features of the data have to be pieced together from inspecting datasets, data appendices, and readme files. Combining datasets across studies is hindered by a lack of harmonization and documentation. Any work in this regard by individual researchers is lost to the next person who wants to conduct a similar study.

In this paper, we propose a metadata schema that can serve as the basis for a catalog of RCT data (or, more broadly, any social science experimental data). The metadata schema defines a set of fields with encoding schemes that can be used to describe datasets associated with social science experiments. The encoding scheme defines the acceptable formats and values to complete the field, such as free text entry, dates, numeric values, or controlled vocabularies (multiple choice options).[2]

The proposal organizes fields into thematic modules and includes whether a field should be optional or mandatory. A core objective of this proposal is standardization. Standardizing the documentation of specific types of data allows harmonization, aggregation, and cross-referencing. Data repositories that follow a standardized schema can be easily made searchable by internet search engines such as Google Dataset Search. As much as possible, our schema therefore conforms with common data description standards (such as those of the Data Documentation Initiative (DDI)), and with existing schemata for experiments as well as survey data from the social sciences (such as the World Bank's Microdata Catalog, the AEA RCT Registry, and ClinicalTrials.gov).

In addition, we added fields that we considered particularly useful for secondary research. Datasets documented with the schema can be easily searched or filtered by many different criteria, from the type

---

[1]We use the term RCT to describe an experiment that uses randomization to assign some intervention or treatment to participants but is conducted outside of a controlled environment (e.g. a laboratory.)

[2]We follow the definitions in ISO (2021).

and time period of intervention, to features of the randomized research design (such as stratification or clustering) to contents of the data (such as how take-up/treatment compliance was measured). These options do not currently exist in commonly used data repositories in the social sciences (e.g. Harvard Dataverse, openICPSR, or the World Bank Microdata Catalog). We formulated a set of principles to guide decision making in selecting the final set of fields. These aim to balance the effort to contribute metadata, the usefulness of the metadata for different research purposes, and the complexity of the information collected. For example, we wanted to make the entry of new metadata straightforward for the majority of RCT datasets, while providing enough flexibility to describe unusual research designs or unique data properties.[3] In addition to our proposal, we also formulated a set of recommendations for implementing a catalog or database based on this metadata schema.

With this schema, we hope to provide a public good to the research community that facilitates the creation of inter-operable catalogs, with the goal of speeding up scientific discovery, reducing duplication of effort, and creating a systematic overview of existing RCT research. Research facilitated by RCT data catalogs on external validity, impact heterogeneity, and generalizability of policy impacts beyond the immediate study populations has the potential to help policymakers, funders, or impact investors make decisions about policy programs. Reuse of RCT data can in turn bolster investments in primary data collection and spur methods improvements that will make future experiments faster and more robust. Better access to RCT data can also offer new research opportunities for scholars who do not have the resources to undertake costly primary data collection themselves. Finally, original data is a citable contribution to science independent of an associated paper or report. Enhancing the visibility of the data and the data citation helps ensure that those responsible for creating the data receive credit (some of whom may not be co-authors of the academic study).

The next section gives a brief overview of areas of research that make use of secondary RCT data. Section 3 describes the process undertaken to create the metadata schema. Section 4 walks through the schema in detail, and section 5 describes considerations for creating a metadata catalog based on the schema.

2   Uses of Secondary Experimental Data in Research

RCT data have many properties that make them useful for testing hypotheses and garnering insights that were not necessarily the focus of the original study. Randomization provides a credible exogenous source of variation in the data. Many RCTs collect representative data on large populations, often of groups that are underrepresented (e.g. because they are not part of the formal economy) or of particular interest for policy research (e.g. eligible for certain benefits). RCT datasets often contain indicators and variables of high policy relevance and may use innovative measurement methods such as lab-in-the-field preference measures.

---

[3]We also aimed to make the schema usable to describe laboratory experiments, although this is not the main focus.

In addition, as RCTs as a method mature, researchers have turned their attention to consolidating and systematizing the RCT-based evidence, as well as expanding and improving experimental methodology and econometric analysis methods. In short, many emerging areas of secondary research could benefit from improved access, systematic cataloging, and harmonized documentation of RCT data. Here we provide a brief overview of research that reuses RCT data and informed the development of the schema.

Combining evidence. Meta-analysis techniques such as Bayesian hierarchical models (BHM) can increase external validity and generalizability of experimental results. Recent examples include Bandiera et al. (2021), who combine 16 laboratory and field experiments to estimate the impact of performance pay on women, and Meager (2019), who estimates the impact of micro credit on income-generating activities and consumption. When available data is not catalogued, meta-studies run the risk of overlooking less prominent studies.[4]

Assessing estimation and prediction methods. In seminal work, LaLonde (1986) used experiments as a benchmark to assess the bias in non-experimental estimation methods. The literature that followed (Fraker and Maynard, 1987; Dehejia and Wahba, 1999, 2002; Glazerman et al., 2003, and many more) has matured to the point of being able to draw conclusions about the full distribution of bias (e.g Chaplin et al., 2018). Researchers have also used experimental data for making out-of-sample predictions and evaluating external validity, e.g. by comparing different prediction methods or quantifying site selection bias (e.g. Hotz et al., 2005; Allcott, 2015; Gechter et al., 2019).

Measurement. Researchers use existing data to validate methods of measurement for important concept and indicators. Recent work has for example examined sources of measurement error in agricultural data (e.g. Beegle et al., 2012; Rosenzweig and Udry, 2019) and measurement methods for women's agency (e.g. Donald et al., 2020; Jayachandran et al., 2021).

Estimating structural models. Structural models can exploit the experimental variation for identification and enrich experimental data, for example by evaluating the role of underlying preferences and behavioral factors for take-up decisions as a means to conduct welfare analysis or make predictions for the effects of new policies (e.g. Todd and Wolpin, 2006, 2010; Meghir et al., 2019; Guiteras et al., 2019).

Statistical learning and machine learning. A rapidly growing literature applies machine learning methods to RCT data. Examples include regularization methods to discipline covariate selection or the identification of treatment effect heterogeneity (e.g. Chernozhukov et al., 2018; Chernozhukov et al., 2018), and new sampling methods, e.g. multi-arm bandits and related adaptive experimental algorithms (e.g. Dimakopoulou et al. (2018); Caria et al. (2021); Kasy and Sautmann (2021)). Existing RCT datasets can be used to optimize algorithms, check large- or small-sample behavior, and simulate thousands of trials to improve speed and evaluate performance without incurring costs or burdening subjects.

---

[4]An issue potentially exacerbated by biases inherent to the publication process (Andrews and Kasy, 2019).

Epistemology, RCT methodology, and research transparency. Researchers have begun to examine large sets of studies to understand the "political economy" of designing, conducting, and publishing experiments (e.g., Andrews and Oster (2019) on external validity bias; Gechter and Meager (2022) on the role of pre-existing infrastructure for site selection; Höffler (2017) on how data publication influences citation rates; Anderson and Magruder (2017) on how to reduce the number of false discoveries; or Christensen and Miguel (2018) on the adherence to transparent research practices).

Summary statistics. RCT data can help both researchers and the broader public to better understand underrepresented populations or get an overview of the body of experimental evidence. Especially in low-income contexts, RCT data delivers a detailed picture of populations that are typically not well-represented in available data – neither government data, such as (formal) labor market statistics, nor private data, such as bank records. RCT data can be used to extract stylized facts, conduct exploratory research or power calculations, and more. For example, the non-profit AidGrade (2019) compiled a database of standardized effect sizes and standard errors to facilitate simple forms of comparative analysis (e.g. Vivalt, 2015, 2019).

The metadata schema is designed to benefit all these different applications by

- enabling filtering datasets on criteria such as unit of randomization or intervention assignment strategy;
- collecting information specific to RCTs such as interventions, treatment arms, or treatment compliance;
- facilitating the combination of multiple RCT datasets by documenting features such as available covariates, time period covered in the data, or inclusion/exclusion criteria;
- recording external resources such as registry entries, ethics review protocols, and academic publications, and documenting information such as whether a pre-analysis plan exists or who funded or partnered in the implementation of the RCT.

3   Creating the Metadata Schema

We followed the process outlined in ISO (2021) in creating the metadata schema. Throughout, we were advised by a group of data scientists and data librarians from Harvard Dataverse, Harvard Medical School, Harvard Business School, MIT, ISPS/Yale, DDI, Vivli and the World Bank Microdata Catalog.

We started by reviewing the types of research that use RCT and experimental data (as summarized in section 2), conducting a survey with researchers who have re-used RCT data or expressed interest in methods research,[5] and researching existing metadata schemata for social science and experimental data. The four main sources of metadata fields we ended up using host some of the largest collections of information on existing RCTs and RCT data (at the time of writing).

---

[5]Expressions of interest were collected through the Research Methods Initiative of Innovations for Poverty Action (IPA) and the Global Poverty Research Lab, or through surveys of affiliated researchers of the Abdul Latif Jameel Poverty Action Lab (J-PAL).

For information related to survey data, we focused on schemata based on the **Data Documentation Initiative (DDI)**, an international standard for documenting survey data (DDI, 2021), in particular the fields used in the Harvard Dataverse and the International Household Survey Network (IHSN) template of the World Bank Microdata Catalog. The **Harvard Dataverse** is an accredited, cross-disciplinary data repository that enables any researcher or institution to publish and archive data and code. J-PAL and IPA maintain a Dataverse data collection called the "Datahub for Field Experiments in Economics and Public Policy" that currently hosts over 200 RCT datasets. The repository builds on a suite of tools for the publication of scholarly data (King, 2007), and its metadata schema is mapped to the DDI Codebook. The schema is organized in blocks that can be used to customize the metadata documentation in individual data collections (Harvard Dataverse, 2021). We considered all blocks used by the J-PAL/IPA Datahub.

The **World Bank Microdata Library** hosts survey and other data from multiple institutions, including the World Bank's own research departments, which publish their data under the World Bank's Open Data Policy. It aggregates multiple named collections, including those of the Development Economics Research Group (DECRG), the Development Impact Evaluation Unit (DIME), and the Strategic Impact Evaluation Fund (SIEF),[6] and all metadata can be accessed through an Automated Programming Interface (API) (The World Bank, 2022). The World Bank's IHSN microdata template has four sections compatible with DDI – Document Description, Study Description, Datasets, and Variables Groups – and an External Resources section compatible with the Dublin Core metadata standard (IHSN, 2022). We considered all of these for the schema and adopted many usage recommendations from Dupriez et al. (2021).

Most existing data catalogs do not cover fields related specifically to the design of RCTs. For these, we turned to two important trial registries. The **AEA RCT Registry** of the American Economic Association is likely the most complete record of past and ongoing RCTs in economics, including unpublished studies (AEA RCT Registry, 2022). The registry metadata contains information specific to RCTs that is not typically included in survey schemata, such as the intervention, outcome measures, and the reviewing IRB. The US National Library of Medicine and the National Institutes of Health maintain the trial registry and results database **ClinicalTrials.gov** for clinical trials in the United States (McCray and Ide, 2000). All clinical studies of drugs and devices controlled by the Food and Drug Administration (FDA) must be registered here. The registry has detailed metadata field definitions and maintains an API feed (ClinicalTrials.gov, 2022). Many repositories draw from or link to this registry (e.g. Vivli, 2022; ISRCTN, 2022). The fields on interventions, study arms, and outcome measures provided a model for our schema.

Roughly 270 metadata fields were under consideration for inclusion in the final schema. Some metadata

---

[6]Impact evaluation often means RCT in this context but can also mean other rigorous analysis methods aimed at estimating causal impacts.

schemata provided valuable insights even if their fields were ultimately not adopted; the full list of schemata considered can be found in the GitHub repository for this project. The following **principles for the RCT metadata schema** guided decision-making on (i) which metadata fields to include, (ii) the encoding scheme for each field including whether to create a controlled vocabulary (multiple choice options), and (iii) the cardinality of each field (i.e., whether the field is optional vs. mandatory and if it can be repeated):

1. The primary purpose of the schema is to provide information that helps identify RCT datasets for use in secondary research.

2. The schema primarily describes the design of the data collection and content of the data, not the academic study or analysis results.

3. Preference is given to DDI-compliant over existing and to existing over newly created fields.

4. Field definitions should make information comparable across studies.

5. The level of detail collected must balance usefulness for the purpose above with the effort required to create an RCT metadata record for contributors.

6. The schema must balance ease of use with completeness.

Note that principle 1 and 2 set the schema apart from data repositories that focus on making the analysis in the original study replicable (and thus primarily record aspects of the data that were already exploited for research). Item 2 also excludes information such as estimated treatment effect sizes, which depend on the analysis method applied. Items 3 and 4 aim to make the schema interoperable with existing catalogs. Item 4 led us to sometimes amend the original field definitions and provide controlled vocabularies wherever possible (see below). Item 5 ruled out information that could be difficult to obtain or verify, especially ex-post (such as information on intervention cost or budget). Item 6 means that the defaults of the schema focus on the most common data structures, while optional free text fields allow for supplying additional information.

Iterations of the metadata schema were extensively reviewed and tested by the authors, supported by a group of J-PAL staff, an external group of experts in RCT data reuse or metadata schemata, and the advisory group mentioned above. Part of the testing consisted of completing the metadata fields for a set of 29 RCTs. Most of these test datasets are hosted on the J-PAL Dataverse and were chosen based on aspects of their design, with the aim of testing both "typical" RCTs and "edge cases" (for a full list, see the supplementary material on the GitHub repository for this project). The purpose was both to test the metadata schema in its entirety, including the sequencing and structure of the fields, and to select, develop and test associated controlled vocabularies. We also tested whether the field definitions were easy to understand, and whether there was any difficulty obtaining the requested information for a given RCT. Tester feedback led us for example to provide a definition of what constitutes a "dataset"; see Section V. Data below.

A Note on the Controlled Vocabularies

Though not formally part of the schema, the controlled vocabularies (CVs) are an important part of our proposal. CVs are primarily used for filtering, and as with multiple-choice survey questions, they need to "partition" the space of possible options, meaning the set of individual entries must cover the entire universe of options without overlap (i.e. options that could fit two or more entries). Moreover, entries need to be balanced, in the sense that they need to be specific enough to help users narrow down the set of studies of interest, but broad enough so that each entry applies to more than a small number of studies.

The catalog testing process[7] led to numerous adaptations to existing CVs, both removal/aggregation of entries and addition of new entries. Two examples are the fields *Kind of Data* and *Mode of data collection* (CVs G and I in Appendix B). These fields are part of the DDI and appear in many schemata including the IHSN, which uses a truncated version of the DDI CV (Dupriez et al., 2021). However, testers found the IHSN CV not well-suited to describe administrative data or data from laboratory or "lab-in-the-field" experiments. This led us to add back relevant items from the DDI CV as well as create new entries.

In some cases we developed entirely new CVs, such as CV E in Appendix B, which describes available types of covariates at the cluster or group level. We see some of the newly proposed CVs as under development. In these cases, we provide a version 0.9 in the GitHub repository of this project, with the aim of updating to version 1.0 based on a larger body of RCT datasets.

4   The Metadata Schema in Detail

In what follows, we give an overview of the metadata schema, divided into modules I to VII. Appendix A contains a corresponding table of all metadata fields, with the CVs we suggest listed in Appendix B. In this section, we provide supplementary information about the schema for back-end users or *contributors* who create metadata schema entries, *data users* who are perusing a catalog based on this schema or reading individual metadata entries, and data stewards and application programmers integrating the schema into an existing application or building a new catalog, also called *catalog owners*. We included illustrative examples that may be helpful for establishing a standardized way to describe an RCT's research design. Some details serve users who are less familiar with the conventions and practices for social science RCTs; researchers who routinely work with RCT data may find the information in Appendices A and B sufficient.

For each metadata field, the table in Appendix A contains a name, a short description, an encoding scheme, and the "cardinality" of the field, that is, whether the field is optional or mandatory, and whether

---

[7]Testers were asked to comment on the overall suitability of each CV for documenting RCTs, as well as the individual options within each CV. Testers could also suggest new CV entries. For test fills, testers both provided free-form text responses and selected all applicable choices from each existing CV (if any) in order to evaluate the CV's coverage as well as potential ambiguities.

it is unique or "repeatable". We also included suggestions for implementation for an adopting organization (see section 5 for more). A cross-walk with other metadata schemata is posted in the Github repository. Note that in many cases we adopted existing fields but modified the wording of definitions for clarity in the context of social science experiments.

What Constitutes an RCT Metadata Record?   An RCT is an experiment combined with data collection on the subjects or experimental units. It is principally defined by the study population and unit of randomization, the intervention, and the randomization procedure used to create comparable treatment arms. The schema is designed assuming that each top-level metadata record corresponds to exactly one *study* or RCT, in which a set of interventions were randomized in a sample representative of some population described in section II. We describe how to delineate the one or more individual *datasets* that are part of an RCT below in section V on "Data".

I   Basic Information

The **basic information** fields in the schema provide summary information on the RCT.

The section contains all the information needed to cite RCT data, including the *study title* as well as *author/data owner names* and *affiliations*. Original data is a contribution to science separate from publications based on the data. Contributors may consider crediting a larger or different set of individuals from the academic article. If the data is stored in a repository or other citable location, the metadata record should include the same citation information that is also provided with the data.

Similarly, the *abstract* describes the purpose, nature, and scope of the RCT and data, and may contain more or different information from the paper abstract. The *topic classification* uses a CV to describe the area of research. While we tested CVs such as CESSDA, IHSN, the World Bank themes, and the J-PAL/IPA sectors, ultimately an organization adopting the metadata schema may choose a topic CV based on its own requirements and use case. For example, an economic journal may choose to use the JEL codes (AEA, 2022). The *version* and *version date* fields provide version control for the metadata record.

II   Study Population

The **study population** section provides information related to the study as a whole and who was included in it, covering location, study population, and study sample.

Contributors can choose the *country of intervention* from a CV (ISO country code), with the option to add free-text detail on *geographical coverage*, as well as any *inclusion and exclusion criteria* for the intervention studied. In a policy context, these might be formal eligibility criteria for a social program or benefit; the researchers might also apply other research-related conditions for inclusion into the treatment and control

groups. Jointly, the geographical information and inclusion/exclusion criteria describe the sampling frame from which the randomization units in the treatment and control groups are (randomly) drawn (see also section IV on sampling method).

The second set of fields concerns the unit of randomization as the primary unit of statistical analysis (as opposed to the unit of observation; see below). A *randomization unit* can be an individual experimental unit or a group (a cluster). Contributors are asked to choose the randomization unit from a controlled vocabulary. The CV we recommend expands the DDI CV considerably to account for frequently occurring units of randomization in social science experiments, such as households, businesses, or schools. The CV allows for separate description of physical units (e.g., a production line, a class room) and administrative or legal units (e.g., all employees of a firm, all students at the same grade level). This distinction can matter for example with interventions that exhibit physical spillovers, such as health interventions. Health interventions are for example often assigned at the school, classroom, or grade level (see e.g. Parker et al., 2021).

With individual-level randomization, the unit of randomization is typically the same as the unit of observation. With cluster-level randomization, each randomization unit may contain several observation units or even different types of observation units.[8] A unit counts as a *targeted* randomization unit or cluster if it was intended for inclusion in the study (either to receive an intervention or in the control group), even if the intervention was ultimately not offered or received as intended, or if no outcomes were measured. A unit counts as an *actual* randomization unit if at least one outcome was measured for one observation unit within the cluster post-intervention. There may be a variety of reasons for a discrepancy between actual and targeted sample sizes. This could be random variation (e.g. patients or job seekers visiting a facility on a given date), but also implementation errors. Note that even if a targeted unit is assigned the experimental intervention as planned, there may be non-compliance, i.e. subjects may not take up the intervention or circumvent or counteract it. Compliance is covered in section IV; *actual study sample size* should include non-compliers. The fields in this section pool observations across waves and treatment arms to provide high-level information on the size of the study. For a breakdown by arm see section III.

Cluster-randomized studies may randomly assign different units of observation, say, buyers and sellers. Note, however, that the unit of randomization is "one level higher", i.e. (for example) the market in which buyers and sellers interact. Even though individual buyers and sellers are randomized into treatment, the random variation used in the analysis comes from the different treated shares on both sides of the market (in the field "Study was designed to analyze" in section IV, contributors would in this case report "general equilibrium effects"). Similarly, in the cross-over design in Lopez et al. (2022) (see below), patients arriving

---

[8]The unit – or units – of observation are recorded in section V. Data below, how randomization units are sampled is in section IV. Study Design, and how randomization units are assigned to treatment arms is detailed in section III. Outcomes and Interventions.

at clinics at different days received different treatment arms. However, the level of randomization is the clinic. In the edge case where an academic study uses two separate randomization procedures and presents the treatment effects from each separately – i.e. uses the randomization at different levels for identification – contributors may choose to create two metadata records, which can be linked in section VII. Otherwise we recommend reporting the higher level unit of randomization (e.g. the market vs. the buyer or seller).

Example: The study "Targeting the Poor: Evidence from a Field Experiment in Indonesia" by Alatas et al. (2012) compared different methods for targeting aid to poor households. To create their sample, the authors chose three provinces in Indonesia, then randomly selected 640 villages from those provinces (stratified by geographic location, see below). The choice of the three provinces should be described as part of the inclusion/exclusion criteria. The description should also include that larger villages with more than 100 households per sub-village on average were excluded from sampling in one district. The *randomization unit* in this study was the village, while the *unit of observation* was the household. Both the *targeted* and *actual* sample sizes were 640 villages and sub-villages and 5756 households.

III   Outcomes and Interventions

This section concerns the tested interventions and outcomes. The first set of fields describes the outcome measures collected. The fields in this section are repeated for each outcome variable. Contributors provide a short free text *name*, a *category* chosen from a controlled vocabulary, an optional free-text *description*, and a yes/no answer as to whether the outcome was measured at least once prior to the intervention ("at baseline"). Baseline outcome measures are relevant for treatment effect estimates but can also serve a range of secondary research purposes, such as summary statistics of the study population. We tested different CVs for the outcome categories but ultimately decided against making a recommendation before carrying out further testing with more (and more varied) RCTs. The free text description field can be used to provide additional information such as the unit of measurement, data format and type, or any transformations applied. This is especially useful when comparing data from different studies.

The next part of the module records the actual interventions and their assignment to study arms. A social science RCT may test the effects of a policy change, encouragement, information, or other process or action. In laboratory experiments, experimental treatments can include complex variation of the incentives, game forms, or information provided that govern the interactions of participants. An *intervention* is any experimental manipulation of the participants' environment. An *arm* is a randomly selected subgroup of participants that receives none, one, or multiple interventions as part of the study. A canonical RCT consists of one treatment arm receiving the intervention and a control group arm that does not receive the intervention. In practice, RCT designs are often more complex, with multiple study arms and different

interventions or intervention levels. The schema proceeds by listing all interventions and arms, and then matching none, one, or multiple interventions to each arm. This structure closely follows *ClinicalTrials.gov*; to our knowledge, no metadata schema used in the social sciences provides this level of detail.[9]

Contributors are first asked to list the study *interventions* by assigning a short *name*, classifying the *intervention type*, and then providing an optional free text *description*. The CV for the intervention type is still in development, as the CVs tested did not provide the right balance between broad classifications and detailed options for the kinds of interventions that occur frequently in social science RCTs. Note that the intervention type is one of three key fields that either alone or in combination help users narrow down the content of the study; the other two are the topic area (under I. Basic Information), and the outcome measures (see above). All three may be different: for example, Oster and Thornton (2012) randomize the distribution of menstrual cups in Nepal; they measure take-up by direct recipients as well as individuals in their social networks as the outcome; and the aim of the research is to understand peer effects in technology adoption. Finalizing the CVs for these fields will require testing with a larger body of studies. The new CVs will be posted on the GitHub repository for this project.

Next, contributors can select the *intervention assignment strategy* from a CV and then optionally provide a more complete description. The CV options are adapted from *ClinicalTrials.gov* and include parallel, factorial, and cross-over assignment, as well as the option "other".[10] Contributors can also provide additional information, such as whether the random assignment was carried out using stratification. We encourage listing out all stratification variables. Stratification and other procedures aimed at improving balance such as re-randomization typically reduce variance but also need to be accounted for in the analysis.

Finally, the section records the *study arms*. Contributors can give an identifying *name* to each arm, list the *targeted* and *actual number of randomization units* in the arm, and then indicate which intervention(s) this group received (if any: a control group may receive no interventions). This allows users to back out, for example, how many subjects received the same intervention across treatment arms, and (for factorial designs) which intervention combinations are observed in the data.

Some experimental designs are common in the social sciences but less so in clinical trials. We make some recommendations on how to describe these designs using the options provided. One example is randomized phase-in, meaning that the intervention is randomly assigned to start at different times for different experimental arms, and the comparison between the arms (already) receiving the intervention and the arms not (yet) receiving the intervention in each period is used to estimate the treatment effect. We recommend

---

[9]For example, the AEA RCT Registry and YARD only have free text fields to describe the intervention(s) and do not differentiate between *interventions* and *arms*.

[10]Note that we dropped "single group assignment" and "sequential assignment" as these assignment strategies do not create a comparison group and therefore do not constitute an RCT by the common understanding of this term in the social sciences.

using the "crossover design" option, defining one arm for each study group that starts the intervention at a different time, and explaining the timing of the phase-in and duration for which each arm receives the intervention in the free-form text field.

Example: In Barrera-Osorio et al. (2020), 101 private secondary schools in Uganda were randomly assigned to receive per-student vouchers from the government, 51 starting in the 2011 school year, and 50 starting in the 2012 school year. The authors use the difference in intervention start date to estimate the short-term impact of the public-private partnership program on student enrollment and performance. Other examples of studies using phase-in designs are given in Bouguen et al. (2020).

In more standard cross-over designs, each arm receives different (possibly all) interventions sequentially, and only their order is randomly assigned. Such designs are frequently used in laboratory experiments. A cross-over experiment with two treatment conditions A and B might then be defined as having two arms and two interventions, where both arms receive both A and B (but one arm receives intervention A first and the other receives intervention B first).

Example: In Lopez et al. (2022), the two clinic-based interventions consisted of discount vouchers for a free course of malaria treatment, given either to physicians to pass on to patients at their discretion ("doctor voucher"), or directly to patients before the consultation with the physician ("patient voucher"). The days on which each clinic received either no intervention, the doctor voucher intervention, or the patient voucher intervention were selected based on a randomized schedule. This design could be described as a cross-over design with 60 arms (each clinic) that each received both interventions. The rotational calendar that was used can be described in the free-text field. This cross-over design is unusual in that every clinic was randomized into a different schedule.

Factorial (cross-randomization) designs combine two (or more) types of interventions with different levels or intensities, leading to arms that each receive different combinations of (levels of) interventions. The simplest factorial designs have two interventions, and the "levels" might simply consist of either receiving the intervention, or not receiving it. The four treatment arms then receive A and B, only A, only B, or neither. More complex factorial designs might involve multiple levels of each type of intervention.

Example: Cohen et al. (2015) cross-randomized several subsidy levels (from 0% to 92%) for malaria medication and malaria tests to understand how the availability of affordable testing affected demand for malaria treatment. This design may be recorded by defining one intervention for each subsidy level and each subsidized good that appears in the study. Factorial assignment encompasses "fractional" factorial designs which may drop some of the cells that would be created by a full cross-combination of all intervention levels.

To some degree, the definition of arms and interventions is up to the metadata contributor. For example, some policies or programs consist of a bundle of different types of interventions, such as a health consultation

combined with a discount on a health product. An RCT may only test some combinations of these components; e.g., a family planning consultation with or without a discount on birth control, but not a discount without a consultation. Such designs are technically closer to a parallel design than a factorial design, since the effect of the discount cannot be independently assessed from the effect of the consultation. That said, contributors may still choose to define the two interventions "discount" and "consultation" (as in a factorial design) rather than "consultation plus discount" and "consultation." This can make sense especially if the two intervention types differ substantively and data users might be looking for interventions of only one type. Contributors should still select the intervention assignment strategy that best fits (in this case, parallel).

Recently, adaptive experimental designs have received increased attention in the social sciences. One type of experiment uses the information learned during early observations or waves in the experiment to alter the assignment shares of the different treatment arms (e.g., Kasy and Sautmann, 2021; Caria et al., 2021). Even though the arm size is ex-ante unknown, we recommend using the "factorial" or "parallel" CV entries, using the arm sizes targeted in each stage of the adaptive design, and the free-text field to describe the adaptive assignment strategy. Other experiments study the optimal treatment of a given experimental unit over time, see e.g. Almirall et al. (2014). This may include randomized changes to the treatment over time. Here, the contributor might choose the option "other" and describe the assignment strategy in the free text field.

The last two fields in this section record the overall time period of all interventions. The timing of the individual intervention is not recorded separately in order to reduce the burden of completing the intervention fields, but if the timing of an intervention is important to the design, contributors can either use the free-text field or choose to define separate interventions that distinguish time of treatment receipt. This may be appropriate in phase-in or cross-over designs as above; if the intervention de facto changes over time (e.g. a remedial tutoring program with a changing curriculum); or if treatment effects are expected to differ significantly based on timing or length of the treatment.

IV Study Design

This module provides further information on the research design of the study. The first field asks whether the study builds on or extends a prior RCT. This could be the case if a new outcome is measured, or an intervention is added, but the randomization and sample of at least some of the original study are retained. This information can be important for assessing statistical power or identifying related data sources.

The next set of fields records how the randomization units in the study were sampled from the sampling frame described by the fields in section II. Contributors define the sampling *type* using the associated CV detailed in Appendix B and can then (optionally) *describe the sampling method*. When the sampling strategy is simple, the "type" field may be sufficient. For example, an RCT may include the entire population in a

location – all students in a district's schools, etc. – in which case the sampling method is "1. Total universe (population)"; no further explanation is required. When sampling was carried out in multiple stages that involve different forms of probability selection (Option 2.5: Probability - Multistage), or using a mixed strategy (Option 4: Mix of probability and non-probability sampling), providing a description of the process is very helpful to data users. Note that *sampling type* refers to the method for sampling randomization units. If different, information on the sampling of the *observational* units may be provided in section V. The *Study sampling method: Description* field also allows contributors to provide more detail on how the sample size was chosen (e.g., a sample determined by ex ante power calculations vs. an RCT at scale), and why targeted and actual numbers of randomization units may not be the same (see above).

Example: The Indonesia study by Alatas et al. (2012) sampled randomization units using a stratified multistage design. The authors randomly selected 640 villages from the included three provinces based on a 30/70 urban/rural split, and then randomly selected one sub-village (neighborhood) from each village. On their own, the two sampling stages could be described as "2.3.1 Probability - Stratified: Disproportional stratified" and "2.1 Probability - Simple random," respectively. In the CV, the contributor should select "2.5 Probability - Multistage" and describe the two selection stages for the village and sub-village in the free text field. Additional helpful context for the *Study sampling method: Description* field could include that, even though the targeted and actual sample sizes are equal, five of the originally selected villages were replaced prior to the randomization for various reasons.

The other elements of the Study Design section provide additional information relevant for the variability and external validity of the treatment effect estimates. First, contributors can select the types of *covariates* available in the data from a CV, with the decision to indicate types of available covariates rather than variable-level information aiming to not overly burden contributors. After testing different CVs, we propose a CV adapted from GESIS (Hoffmeyer-Zlotnik, 2016) to describe individual-level covariates, and a new vocabulary to describe covariates at the cluster level, such as household or other group-level characteristics. Information on covariates is needed for meta-analyses and can also be useful for methodological research. For instance, Tabord-Meehan (2018) uses data from an experiment on increasing charitable donations by Karlan and Wood (2017) to demonstrate a new method of adaptive stratification that uses information from a first experimental wave to select "stratification trees" in the second wave.

The next field of the schema includes a new optional field that describes which forms of treatment effects *the study was designed to analyze*. This includes the "intent to treat" effect, average treatment effect, and local average treatment effect or average treatment effect on the treated. Note that the latter two imply that compliance with the treatment assignment is known (see below). A study that is designed to analyze "4. Heterogeneous treatment effects or effects by subgroup" is powered to detect treatment effects in

each population subgroup. The researchers may conduct disproportional stratified sampling and oversample subgroups that constitute a small share of the population in order to estimate treatment effects in this subgroup. A study designed to measure "5. General equilibrium effects" might measure outcomes for groups other than the directly affected group and randomize the interventions at the market level, rather than the individual level. An example might be to measure the effects of providing the unemployed with job search assistance on salaries and firms. A study that captures "6. Spillovers or externalities" measures effects of treating one unit on other units in the vicinity. This is often done by varying the share of treated units within a cluster and requires collecting outcome data on untreated units.

Example: Crépon et al. (2013) randomly varied the share of unemployed job seekers in a city receiving a job placement assistance program in order to study displacement effects on those not receiving the program.

These design features are specific to social science RCTs, and to our knowledge there exists to date no CV for them. We consider this CV under development.

The last field of this section describes compliance with the randomized intervention assignment. In some situations, treatment assignment is not identical with treatment receipt or take-up. Those assigned to the intervention may not actually receive it, and conversely those not assigned to it may nonetheless gain access.

Example: Imperfect compliance is particularly common in so-called encouragement designs. In an experiment with around 1500 small firms in Tajikistan, Okunogbe and Pouliquen (2022) vary whether firms are trained and provided assistance for filing their taxes electronically in order to estimate the effect of e-filing on tax payments and other outcomes. About 60% of the control, but 93% of the treatment group adopt e-filing. This is an example of "two-sided non-compliance":[11] some firms in the treatment group do not use e-filing while many in the control group do.

The degree of treatment compliance is important for external validity and potential selection effects and has also been used in methodological research; for example, Bernard et al. (2022) use imperfect compliance RCTs to estimate the bias of observational methods in practice. Partial compliance is more common in the social sciences than in laboratory or medical trials, which are typically closely controlled. At the same time, it is often difficult to unambiguously define and measure. We ask contributors to use the free-form *compliance description* field to explain what forms of non-compliance are in principle possible, how compliance was measured, and what the actual rates of non-compliance are.

V   Data

The **data** section of the metadata schema describes the actual data available to users, arranged in one or more datasets. For data with restricted access, access modalities can be described in the External Resources

---

[11]See Angrist and Pischke (2009) for the terminology of one-sided and two-sided compliance.

section. Data that are not accessible to anyone but the original researchers should not be described.

More than one *dataset* may be associated with a given RCT. For example, the contributors may have collected census or administrative data for a larger sample than the ultimate study population, or specific information on two different populations affected by the RCT (e.g. buyers and sellers of a good). Datasets are distinct from data files; a dataset may be broken up into several files and even stored in multiple locations, for instance restricted-access GPS data vs. publicly accessible de-identified survey responses.[12] In general, a set of records may constitute its own dataset if it contains information central to the study, such as a separate outcome measure or the sampling frame, and (i) consists of observational units from a distinct study population (or sample from the study population) or (ii) is based on an independent data source (e.g. with a specific mode of data collection).

What delineates a dataset is ultimately up to the metadata contributor, but we recommend keeping the number of datasets to the minimum needed to describe the data well for users; often the data associated with a given RCT can be characterized as a single dataset. For example, the same dataset may contain several rounds of data collection, and the metadata schema allows multiple "cycles" (including repeated cross-sections). Even data from different sources can often be treated as part of the same dataset.

In some cases it can be useful to define a separate dataset to describe the data in sufficient detail. For example, two measures of the same outcome, such as measures of crop productivity obtained through in-person audits and satellite imagery, may have large discrepancies in coverage and diverging numbers of observations. In cluster-randomized studies, data may be available both at the cluster and individual levels, and these datasets should be described separately if each contains primary outcome measures. Similarly, a full census in the study area followed by sampling a subset of the population for the interventions and endline data collection warrants defining separate datasets for each data collection round.

For each dataset, contributors are asked to define a short description or *name*, then provide information on the types and number of *observational units*, first in total and then, further below, per arm. This information is collected at the dataset level in order to allow for multiple units of observation, for example outcomes measured at the teacher and the student levels. The per-arm fields are optional to allow description of data collected prior to treatment assignment (e.g. a study population census). A unit counts as a *targeted* observation unit if it was selected or intended for data collection. This may be an estimated number. A unit counts as an *actual* observation unit if at least one outcome was measured for it. With clustered randomization designs, the targeted and actual numbers of randomization and observation units may all be different. Targeted and actual number of observations may differ in particular if there is survey attrition, which gives an indication of data quality and can help answer methodological questions (e.g. when researching

---

[12]Conversely, a data file may contain several datasets (e.g. an excel file with several sheets).

study designs intended to limit attrition).

Contributors are also asked to define the time method (i.e. whether the data contains one or more cross-sectional samples or panel data) and the number of *cycles* (i.e. waves or rounds of data collection), list the *modes of data collection* in detail, describe the sampling method, and specify whether there are sampling weights. Note that researchers may technically carry out separate power calculations and devise sampling procedures for different units of observation within the same study. In order to collect this information in one place, any information related to power calculations (including the determination of the targeted number of units of observation) should be included in the field IV.3 "Study sampling method: Description". The method of sampling for the randomization unit is included in IV.2-3, but the method of sampling the unit of observation within the unit of randomization and any discrepancy between targeted and actual units of observation is recorded here in field V.1.I.

A free text field allows contributors to provide *data collection notes*, which may describe for example how the observational units in the dataset were sampled, when during the experiment data was collected (e.g. at baseline, midline, or endline), and what quality controls were in place for the data collection. Finally, contributors are asked to provide information on the timing of data collection by specifying the time period covered in each data cycle, and, if different (e.g. in retrospective surveys) the dates of data collection.

VI   Ethics and Research Transparency

This section allows contributors to provide information that may be relevant to the legitimacy, external validity, credibility, and robustness of the study and its data. This includes information on *ethics review* conducted, specifically the *reviewing institution(s)* and *protocol number(s)*, as well as information on funding or supporting bodies and implementation partners. Here, survey firms, collaborating government agencies, and other parties can be named.

Contributors can also indicate whether any documentation is available to users on the ethics of the study, such as consent forms or a structured ethics appendix, as proposed by Asiedu et al. (2021), and to what degree records related to registration or pre-specification exist. The CV includes options such as "pre-results acceptance" for studies accepted into a journal based on the pre-specified research design, and "populated pre-analysis plan" for cases where the researchers produced a document containing the analysis exactly as pre-specified, typically separate from the research paper (see Banerjee et al. (2020)). The CVs for both fields were newly developed.

VII   External Resources

This section provides information on external resources available for the study, including the location of the described data. For each resource, contributors can provide a *type* and free-form *description*, *citation*, *link (DOI/URL)*, and information on the *access policy* for this resource.

If available, contributors should not only point to the data itself, but also to resources describing the data and study such as academic publications, reports, codebooks, ethics protocols, etc. The controlled vocabulary for the resource *type* is adapted from the World Bank Microdata Catalog but allows as additional types "database or data repository entry" (to describe the locations of the study's datasets), "trial registration", "pre-analysis plan", "populated pre-analysis plan", and "research ethics documentation". The last four entries complement the information in section VI.

Data from a single RCT may be stored in different locations (e.g. replication data in a repository like Harvard Dataverse, access-restricted identifying data with the researcher, administrative data with a separate data provider); each resource should be linked. Conversely, different external resources, such as the data itself along with replication code, are sometimes stored in the same location or with the same citation information. The CV allows contributors to enter the information once and tick all options that apply under *type* and provide additional detail in the *description*.

5   Creating a Metadata Catalog Based on the Metadata Schema

We close with a few considerations for implementing a catalog based on the proposed schema, with a focus on maximizing the quality, consistency, and usability of the metadata and catalog.

Catalog: An RCT metadata catalog consists principally of a back-end data entry interface; front-end search, filtering, and display functions; and an underlying database of metadata fields. Catalog owners may consider modifying the schema. For users interested in adopting only a subset of the metadata fields, we made an effort to minimize dependencies across modules (with the exception of core properties of the RCT such as the study arms, which appear in several modules). Catalog owners may also be interested in expanding the catalog beyond recording the properties of the data, for example by adding fields for analysis results, such as standardized treatment effect sizes, or contextual information, such as intervention costs.[13]

To fully support interoperability and bulk uses of the metadata, we recommend an *API interface* to access and download the catalog contents. This functionality also allows the catalog to be read by data search engines like Google Dataset Search, increasing reach and accessibility. Since the CVs for this schema are

---

[13]These are not part of our schema because they often require a separate effort: verification (e.g. by replicating the analysis), modification (e.g. by standardizing effect sizes or harmonizing analysis approaches), and possibly additional research/estimation of data not in the public domain (such as the fixed and variable intervention costs).

under development, we recommend that early adopters use free text fields after "other" options to collect information for potential expansions or modifications to the CVs.[14] Finally, *user accounts* allow contributors to track and edit catalog entries after they have been posted, data owners to claim entries, and front-end users to save searches or individual records, while a *data review system* contributes to ensuring data quality control and avoiding version control issues, duplication of entries, or so-called "ghost entries". For this purpose, incomplete entries could be regularly flagged for purging.

Back-end: The data entry interface for contributors should minimize the time and effort (and hence potential for errors) required to complete an entry to help maintain data quality. This includes making use of existing information where available. Especially for bulk entry of existing RCT data, a sophisticated implementation could draw on existing public records where available, such as the API interfaces of Clinical-Trials.gov and the World Bank Microdata Catalog,[15], or the downloadable AEA RCT Registry data. New catalog entries could also be partially pre-filled by scraping web data from external resources such as registry entries and academic articles. This could best be achieved by beginning the data entry with the links to (some) external resources (which would also permit a duplication check within the catalog).

Within a metadata entry, information filled early on can provide inputs to later fields, such as pre-filling the treatment arms that were recorded in the "outcomes and interventions" section (III) in the "data" section (V). The catalog can also prompt contributors for information based on earlier inputs, such as suggesting entries for ethics documentation (section VI) or prior studies (section IV) in the "external resources" section VII. Automated cross-checks and data validation can additionally be applied to numeric and multiple-choice entries. We make specific suggestions in the programming notes in Appendix A. These features minimize errors and duplication of effort and improve consistency. Incomplete entries should be regularly automatically saved to prevent data loss.

An intuitive interface and extensive user support can also help with data entry. For example, CVs should be implemented as multiple choice radio buttons or "select all that apply" tick boxes; at the end of an entry in a loop, the navigation should allow users to choose between adding another entry or moving to the next field. In addition to the field and CV definitions in the Appendix, the interface could provide help "bubbles" containing longform instructions and detailed examples that could be drawn from this article. This could be especially useful for describing the intervention assignment and study design in sections III and IV.

Front-end: Careful interface design for front-end users can facilitate a quick overview over multiple datasets as well as individual RCTs. An individual study view could pull selected information from various modules and arrange it for better reading. An example is the cross-reference between which arms and

---

[14]The authors encourage submitting suggestions for such changes through the GitHub for this project.

[15]This is facilitated by the crosswalk we created to link individual fields in our schema to those in other schemas; see this project's GitHub repository.

interventions, which would be most intuitively displayed as a table. Ideally, users can also save searches or selected studies (see above) and visualize various metadata fields for this set (e.g. show the share of entries with a specific value for CV fields). Front-end users should also have access to the same longform/help information as back-end users to facilitate interpretation of the data.

The most important usability features for a catalog are the search and filter options available. Boolean operators AND, OR, and NOT improve filtering within CVs or across fields, e.g. they could allow users to find all studies outside of a specific country, or studies that cover specifically early childhood and primary education. Mathematical operators ($>$, $\leq$, etc.) on dates and numerical entries can for example help find studies within certain time periods, or of a certain sample size. A WYSIWYG mask for constructing a search within and across fields helps first-time users, while a free-text entry field supporting advanced search functions makes the exact criteria applied replicable for other users, e.g. for meta-analysis purposes.

Just as with the metadata schema itself, standardization and free access are paramount for fostering the reuse of RCT data and promoting equitable access. We therefore encourage catalog implementers to make user access free and to use open-source programming to allow other catalog owners to adopt useful features.

References

AEA (2022). JEL Classification System/EconLit Subject Descriptors. URL: https://www.aeaweb.org/econlit/jelCodes.php?view=jel (07/22/2022).

AEA RCT Registry (2022). The American Economic Association's registry for randomized controlled trials. URL: https://www.socialscienceregistry.org (12/09/2022).

AidGrade (2019). Aidgrade. URL: http://www.aidgrade.org/ (08/25/2019).

Alatas, V., A. Banerjee, R. Hanna, B. A. Olken, and J. Tobias (2012). Targeting the poor: Evidence from a field experiment in Indonesia. *American Economic Review 102*(4), 1206–40.

Allcott, H. (2015). Site selection bias in program evaluation. *Quarterly Journal of Economics 130*, 1117–1165.

Almirall, D., I. Nahum-Shani, N. E. Sherwood, and S. A. Murphy (2014). Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. *Translational Behavioral Medicine 4*(3), 260–274.

Anderson, M. L. and J. Magruder (2017). Split-sample strategies for avoiding false discoveries. Technical Report 23544.

Andrews, I. and M. Kasy (2019). Identification of and correction for publication bias. *American Economic Review 109*(8), 2766–2794.

Andrews, I. and E. Oster (2019). A simple approximation for evaluating external validity bias. *Economics Letters 178*, 58–62.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Asiedu, E., D. Karlan, M. Lambon-Quayefio, and C. Udry (2021). A call for structured ethics appendices in social science papers. *Proceedings of the National Academy of Sciences 118*(29), e2024570118.

Bandiera, O., G. Fischer, A. Prat, and E. Ytsma (2021). Do women respond less to performance pay? building evidence from multiple experiments. *American Economic Review: Insights 3*(4), 435–54.

Banerjee, A., E. Duflo, A. Finkelstein, L. F. Katz, B. A. Olken, and A. Sautmann (2020). In praise of moderation: Suggestions for the scope and use of pre-analysis plans for rcts in economics. Technical report, National Bureau of Economic Research.

Barrera-Osorio, F., P. de Galbert, J. Habyarimana, and S. Sabarwal (2020). The impact of public-private partnerships on private school performance: Evidence from a randomized controlled trial in uganda. *Economic Development and Cultural Change 68*(2).

Beegle, K., C. Carletto, and K. Himelein (2012). Reliability of recall in agricultural data. *Journal of Development Economics 98*(1), 34–41. Symposium on Measurement and Survey Design.

Bernard, D., G. Bryan, S. Chabé-Ferret, J. de Quidt, J. Fliegner, and R. Rathelot (2022). How biased are observational methods in practice? accumulating evidence using randomised controlled trials with imperfect compliance. *Ongoing work*.

Bouguen, A., Y. Huang, M. Kremer, and E. Miguel (2020). Using randomized controlled trials to estimate long-run impacts in development economics. *Annual Review of Economics 68*(2).

Caria, S., G. Gordon, M. Kasy, S. Quinn, S. Shami, and A. Teytelboym (2021). An adaptive targeted field experiment: Job search assistance for refugees in Jordan. *Working paper*.

Chaplin, D. D., T. D. Cook, J. Zurovac, J. S. Coopersmith, M. M. Finucane, L. N. Vollmer, and R. E. Morris (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study-comparisons. *Journal of Policy Analysis and Management 2*(37), 403–429.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernandez-Val (2018). Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments. *arXiv e-prints*, arXiv:1712.04802v3.

Christensen, G. and E. Miguel (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature 56*(3), 920–80.

ClinicalTrials.gov (2022). ClinicalTrials.gov is a database of privately and publicly funded clinical studies conducted around the world. URL: https://clinicaltrials.gov (12/09/2022).

Cohen, J., P. Dupas, and S. Schaner (2015). Price subsidies, diagnostic tests, and targeting of malaria treatment: evidence from a randomized controlled trial. *American Economic Review 105*(2), 609–45.

Crépon, B., E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora (2013). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *The Quarterly Journal of Economics 128*(2), 531–580.

DDI (2021). Document, Discover and Interoperate. URL: https://ddialliance.org/ (08/27/2019).

Dehejia, R. H. and S. Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association 94*(448), 1053–1062.

Dehejia, R. H. and S. Wahba (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics 84*(1), 151–161.

Dimakopoulou, M., Z. Zhou, S. Athey, and G. Imbens (2018). Estimation Considerations in Contextual Bandits. *arXiv e-prints*, arXiv:1711.07077v4.

Donald, A., G. Koolwal, J. Annan, K. Falb, and M. Goldstein (2020). Measuring women's agency. *Feminist Economics 26*(3), 200–226.

Dupriez, O., D. M. Sanchez Castro, and M. Welch (2021). Quick reference guide for data archivists. URL: https://guide-for-data-archivists.readthedocs.io (02/22/2021).

Fraker, T. and R. Maynard (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *The Journal of Human Resources 22*(2), 194–227.

Gechter, M. and R. Meager (2022). Combining experimental and observational studies in meta-analysis: A mutual debiasing approach. *Working paper*.

Gechter, M., C. Samii, R. Dehejia, and C. Pop-Eleches (2019). Evaluating ex ante counterfactual predictions using ex post causal inference. *arXiv e-prints*, arXiv:1806.07016v2.

Glazerman, S., D. M. Levy, and D. Myers (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science 589*, 63–93.

Guiteras, R., J. Levinsohn, and A. M. Mobarak (2019). Demand estimation with strategic complementarities: Sanitation in Bangladesh. CEPR Discussion Paper No. DP13498.

Harvard Dataverse (2021). Dataverse documentation v. 5.3. URL: https://guides.dataverse.org/en/5.3/ (02/22/2021).

Höffler, J. (2017). Replication and economics journal policies. *American Economic Review 107*(5), 52–55.

Hoffmeyer-Zlotnik, J. H. P. (2016). Standardisation and harmonisation of socio-demographic variables.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics 125*, 241–270.

IHSN (2022). DDI Metadata Editor (Nesstar Publisher 4.0.10). URL: https://ihsn.org/software/ddi-metadata-editor (07/22/2022).

ISO (2021). ISO/TC46/SC11N800R1 Building a metadata schema – where to start. URL: https://committee.iso.org/files/live/sites/tc46sc11/files/documents/N800R1%20Where%20to%20start-advice%20on%20creating%20a%20metadata%20schema.pdf (02/15/2021).

ISRCTN (2022). ISRCTN Registry. URL: https://www.isrctn.com/ (09/12/2022).

Jayachandran, S., M. Biradavolu, and J. Cooper (2021). Using machine learning and qualitative interviews to design a five-question women's agency index.

Karlan, D. and D. H. Wood (2017). The effect of effectiveness: Donor response to aid effectiveness in a direct mail fundraising experiment. *Journal of Behavioral and Experimental Economics 66*, 1–8.

Kasy, M. and A. Sautmann (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica 89*(1), 113–132.

King, G. (2007). An introduction to the Dataverse network as an infrastructure for data sharing. *Sociological Methods & Research 36*(2), 173–199.

LaLonde, R. J. (1986). Evaluating the econometric evaluation of training programs with experimental data.

American Economic Review 76, 604–620.

Lopez, C., A. Sautmann, and S. Schaner (2022). Does patient demand contribute to the overuse of prescription drugs? *American Economic Journal: Applied Economics 14*(1), 225–60.

McCray, A. T. and N. C. Ide (2000). Design and implementation of a national clinical trials registry. *Journal of the American Medical Informatics Association 7*(3), 313–323.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *AEJ: Applied Eocnomics 11*, 57–91.

Meghir, C., A. M. Mobarak, C. D. Mommaerts, and M. Morten (2019, July). Migration and informal insurance: Evidence from a randomized controlled trial and a structural model. Working Paper 26082, National Bureau of Economic Research.

Okunogbe, O. and V. Pouliquen (2022). Technology, taxation, and corruption: Evidence from the introduction of electronic tax filing. *American Economic Journal: Economic Policy 14*(1), 341–72.

Oster, E. and R. Thornton (2012). Determinants of technology adoption: Peer effects in menstrual cup take-up. *Journal of the European Economic Association 10*(6), 1263–1293.

Parker, K., M. Nunns, Z. Xiao, T. Ford, and O. C. Ukoumunne (2021). Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the United Kingdom: a methodological systematic review. *BMC Medical Research Methodology 21*(1), 152.

Rosenzweig, M. R. and C. Udry (2019). External validity in a stochastic world: Evidence from low-income countries. *The Review of Economic Studies*.

Tabord-Meehan, M. (2018). Stratification trees for adaptive randomization in randomized controlled trials. *arXiv e-prints*, arXiv:1806.05127.

The World Bank (2022). Microdata library. URL: https://microdata.worldbank.org (12/09/2022).

Todd, P. E. and K. I. Wolpin (2006). Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *American Economic Review 96*(5), 1384–1417.

Todd, P. E. and K. I. Wolpin (2010). Structural estimation and policy evaluation in developing countries. *Annual Review of Economics 2*(1), 21–50.

Vivalt, E. (2015). Heterogeneous treatment effects in impact evaluation. *American Economic Review 105*(5), 467–70.

Vivalt, E. (2019). Specification searching and significance inflation across time, methods and disciplines. *Oxford Bulletin of Economics and Statistics 81*(4), 797–816.

Vivli (2022). A global clinical research data sharing platform. URL: https://vivli.org/ (09/12/2022).

Appendices

A   The Metadata Schema

Below is the proposed schema in full. The first two columns name the field and provide a short description. Column 3 defines the encoding scheme which describes whether the field permits only controlled entries, free text, numeric values, etc. Column 4 specifies the "cardinality" of the field, or the minimum and maximum number of entries. A minimum of 0 means the field is optional. A maximum of n means the field can be repeated multiple times. For example, a field with cardinality (0..1) is optional and unique (such as the abstract of the study), whereas a field with cardinality (1..n) must contain at least one entry and can contain multiple (such as the list of authors). Fields may be designated as optional if we consider them useful for potential secondary uses of the data, but they may be not applicable in all cases or not known to the contributor. The last column is not part of the schema but provides some notes on the suggested implementation of an RCT data catalog at the back and front end, such as data entry support and display options.

As described in the text, many metadata fields below are adapted or taken from existing schemata, and where possible we kept field definitions similar to their sources to facilitate the mapping of fields. Sections I (Basic Information), II (Study Population), V (Data) and VII (External Resources) are in large parts similar or identical to World Bank IHSN fields or the underlying DDI fields. Section III (Outcomes and Interventions) borrows extensively from ClinicalTrials.gov. Some AEA RCT Registry fields are referenced in section VI (Ethics and Research Transparency). All sections except I contain at least some new fields, and section IV (Study Design) is new in many parts. The GitHub repository contains a full crosswalk between schemata.

Table Legend

Bolded text	Denotes a set of repeatable questions (a "loop")	
Encoding scheme	CV	Controlled vocabulary
Cardinality	0..1	Optional and non-repeatable
	0..n	Optional and repeatable
	1..1	Mandatory and non-repeatable
	1..n	Mandatory and repeatable

The Metadata Schema
I. Basic Information

Field		Description	Encoding	Cardin-ality	Programming Notes
I.1.	Title	The name of the study.	Free text	1..1	
I.2.	**Authors/owners:** The person(s), corporate body, or agency responsible for the substantive and intellectual content of the data. This list may differ from the authors named on an associated paper or grant.				
I.2.A.	Authors/owners: Name	Use "surname, first name" format.	Free text	1..n	Loop: assign unique ID to reference each author.
I.2.B.	Authors/owners: Affiliation	Author's affiliated institution at the time of data creation. Can be the same as above if the owner is an agency.	Free text	1..n	
I.3.	Abstract	A summary describing the purpose, nature, and scope of the RCT and data collection, special characteristics of its contents, and major subject areas covered.	Free text	0..1	
I.4.	Topic classification	The broad substantive topic(s) covered by the data.	CV	1..n	Backend: format as "Select all that apply."
I.5.	Version	Version number of the study entry at the appropriate level.	Numeric	N/A	Automatically generated.
I.6.	Version date	Version date of the study entry at the appropriate level.	Format: YYYY-MM-DD	N/A	Automatically generated.

II. Study Population

Field		Description	Encoding	Cardin-ality	Programming Notes
II.1.	Country of intervention	The country or countries in which the intervention was implemented, even if the study did not cover the entire country.	CV: ISO country codes.	1..n	Backend: format as "Select all that apply."
II.2.	Geographical coverage	The geographic level at which the data is representative, within the country of intervention and conditional on the inclusion / exclusion criteria. Provides the total geographic scope of the data and, if needed, additional geographic selection criteria. Entries may be region or state names, along with qualifiers such as "urban areas only", etc. Note that a study can for example have national coverage even when some districts are not included, as long as all districts were eligible for sampling as part of the sampling strategy.	Free text	1..1	Reference IV.2-3"Study Sampling Method".
II.3.	Inclusion/exclusion criteria	The criteria to determine eligibility for inclusion in the study and randomized assignment. In general, it should be possible to tell from the country, geographical coverage, unit of randomization, and inclusion / exclusion criteria whether a given individual or unit (hypothetical or real) is a member of the population that is the object of the research and from which the sample was drawn.	Free text	1..1	Reference IV.2-3"Study Sampling Method".
II.4.	Unit of randomization	The level of treatment assignment: individuals, locations, facilities, groups, etc. Also referred to as the level of clustering. The level of treatment assignment/unit of randomization may be the same as the the unit of observation.	CV A	1..1	
II.5.	Unit of randomization: Targeted study sample size	The targeted number of randomization units pooled over all study arms and periods or phases of random assignment (waves). Include if an approximate target was used.	Numeric	0..1	
II.6.	Unit of randomization: Actual study sample size	The actual number of randomization units pooled over all study arms and periods or phases of random assignment (waves). Count only randomization units for which an outcome of one observational unit was measured at least once across all post-intervention data collection cycles.	Numeric	1..1	

III. Outcomes and Interventions

Field	Description	Encoding	Cardin-ality	Programming Notes
III.1. Outcomes: Measurements used to determine the effect of an intervention/treatment/program on experimental subjects or units. Please repeat the information for each main outcome measure.				
III.1.A. Outcome: Name	A brief descriptive name used to refer to the outcome measure.	Free text	1..n	Loop: assign unique ID to reference each outcome.
III.1.B. Outcome: Category	The broad category of the specific outcome measure.	CV	1..n	CV under development. Collect responses for updating the CV.
III.1.C. Outcome: Description	Additional information about the outcome measure, such as importance to the analysis (e.g., primary vs. secondary outcome), unit of measurement (e.g., meters), format/data type (e.g. categorical), distribution class (for numeric outcomes, e.g. count, binary, real numbers), range of possible values (e.g. 0-100), as well as a description of how the outcome was constructed (if relevant).	Free text	0..n	
III.1.D. Outcome: Collected pre-treatment?	Is a measurement of this outcome available before any treatment or notification of treatment took place ("at baseline")?	CV: Yes / No	1..n	
III.2. Interventions: An intervention is defined as a process or action that is the focus of an RCT or experiment. The intervention may be a policy change (such as the right to buy an amount of subsidized food), an experimental condition (such as a high or low cost of contributing to a public good in a lab experiment), an encouragement, nudge, or information treatment (such as text messages or TV ads), etc. Different variants of a process or action are a distinct intervention if they are separately randomly assigned. Receiving no treatment is not an intervention. Please repeat the information for each intervention tested in the study.				
III.2.A. Intervention: Name	A brief descriptive name used to refer to the intervention.	Free text	1..n	Loop: assign unique ID to reference each intervention.
III.2.B. Intervention: Type	The category or type of intervention.	CV	1..n	CV under development. Collect "other" responses for updating the CV.
III.2.C. Intervention: Description	Free text description of the details of the intervention.	Free text	0..n	
III.3. Intervention assignment strategy	The strategy used for assigning interventions to study arms.	CV B	1..1	
III.4. Assignment strategy description	A description of the intervention assignment strategy. If relevant, provide details such as the timing of the different interventions in a given study arm in more complex designs such as phase-in and crossover. If the treatment assignment was carried out using stratified randomization, please explain here how the strata were formed, and if possible, name the stratification variables.	Free text	0..1	
III.5. Number of arms	The number of subgroups of participants in the randomized trial that receive none, one, or several specific interventions (i.e., arms) according to the trial's protocol. For a trial with multiple periods or phases of random assignment (waves) that have different numbers of arms, the maximum number of arms from all periods or phases.	Numeric	1..1	Backend: Restrict to integers > 1.

III. Outcomes and Interventions (continued)

Field	Description	Encoding	Cardinality	Programming Notes
III.6. **Arms:** Subgroups of participants that receive none, one, or several specific interventions according to the trial's protocol. Please repeat the information for each study arm.				
III.6.A. Arm: Name	A brief descriptive name used to refer to the study arm.	Free Text	1..n	Loop: assign unique ID to reference each arm. Use III.5 to generate required number of arms. Pre-populate with generic names, e.g. Arm 1, Arm 2.
III.6.B. Arm: Targeted sample size.	The targeted number of randomization units assigned to this study arm across all periods or phases of random assignment (waves).	Numeric	0..n	Backend: Restrict to integers > 0. Cross check sum with II.5.
III.6.C. Arm: Actual sample size.	The actual number of randomization units assigned to this study arm across all periods or phases of random assignment (waves). Count only randomization units for which at least one outcome of one observation unit was measured post intervention.	Numeric	1..n	Backend: Restrict to integers ≥ 0. Cross check sum with II.6.
III.6.D. Arm: interventional cross-reference	Indicate which interventions are provided in this arm of the study.	Free text	1..n	Implement as checkboxes using III.2.A via unique intervention IDs generated.
III.7. Intervention start date	The first date when the administration of any of the interventions (after random assignment) began. Please enter the earliest start date of all interventions. If any element of the date is unspecified, use "X" as input.	Format: YYYY-MM-DD	1..1	Backend: give examples, e.g. 2016-05-XX or 202X-XX-XX.
III.8. Intervention end date	The last date when the administration of any of the interventions ended. Please enter the last end date of all interventions. any element of the date is unspecified, use "X" as input.	Format: YYYY-MM-DD	1..1	

IV. Study Design

Field	Description	Encoding	Cardinality	Programming Notes
IV.1. Prior work	Does this study extend or rely on any prior study? Examples are collecting additional outcomes for interventions randomly assigned in a previous study, expanding the sample, or adding a treatment arm.	CV: Yes / No / Unknown	1..1	If "yes" is selected, reference VII. External Resources for information on the prior study.
IV.2. Study sampling method: Type	The type of sampling method used to select the randomization units to be included in the experiment. If sampling is performed in several stages, please select "Probability – multistage," or "Mix of probability and non-probability sampling" and provide additional details in the description field.	CV C	1..1	
IV.3. Study sampling method: Description	An overall description of the procedure for sampling the randomization units included in the study; if the sampling was performed in several stages, consider listing them out with an explanation. Include a description of the method used to obtain the targeted number of randomization and observation units (e.g., power calculations), along with any information related to the sampling that is relevant to users comparing targeted and actual units of randomization.	Free text	0..1	Backend: reference unit of observation information in V.1.1.
IV.4. Covariates: Individual	Please select all individual-level covariate categories included in this study.	CV D	0..n	Backend: format as "Select all that apply." If no option is selected, ask user to confirm.

29

IV. Study Design (continued)

Field	Description	Encoding	Cardin-ality	Programming Notes
IV.5. Covariates: Group	Please select all cluster- or group-level covariate categories included in this study.	CV E	0..n	Backend: format as "Select all that apply." If no option is selected, ask user to confirm.
IV.6. Study was designed to analyze	Please select all types of treatment effects the study was designed to measure or analyze (i.e., the randomization was designed accordingly and the data includes the necessary information, such as intervention take-up).	CV F	0..n	Backend: format as "Select all that apply." If no option is selected, ask user to confirm. Collect "other" responses for updating the CV.
IV.7. Compliance	Please describe what forms of noncompliance with any of the interventions are possible or observed, and, if available, how treatment compliance is measured in the data and what the take-up rates are. Noncompliance occurs when not all units take up or receive the assigned intervention, or when at least some units receive an intervention they were not assigned.	Free text	0..1	Reference the implications of selected options "LATE or TOT" and "ATE" for compliance in IV.6.

V. Data

Field	Description	Encoding	Cardin-ality	Programming Notes
V.1. **Datasets:** Information about the datasets included in this study and the methodology employed in data collection. Datasets are distinct from data files. A set of records may constitute a separate dataset if it contains information central to the analysis, such as an outcome measure, and (i) consists of observational units from a distinct study population or (ii) comes from an independent data source or mode of data collection. Please repeat the following elements for each dataset.				
V.1.A. Dataset: Name	A brief descriptive name used to refer to the dataset.	Free text	1..n	Loop: assign unique ID to reference each dataset.
V.1.B. Dataset: Unit of observation	The basic unit of analysis or observation that the dataset describes. The unit of observation may be the same as the unit of randomization.	CV A	1..n	Backend: multiple choice (1 option per dataset).
V.1.C. Dataset: Unit of observation: Targeted sample size	The targeted number of observation units pooled over all study arms and periods or phases of random assignment (waves). Include if an approximate target was used.	Numeric	0..n	Backend: Restrict to integers > 0.
V.1.D. Dataset: Unit of observation: Actual sample size	The actual number of observation units included in the dataset.	Numeric	1..n	Backend: Restrict to integers > 0.
V.1.E. Dataset: Kind of data	Please select all types of data included in the dataset.	CV G	1..n	Backend: format as "Select all that apply."
V.1.F. Dataset: Time method	The time method or time dimension of the dataset.	CV H	1..n	Backend: multiple choice (1 option per dataset).
V.1.G. Dataset: Number of cycles	How many cycles (data collection or measurement rounds) are in the dataset?	Numeric	1..n	Backend: Restrict to integers > 0. Cross-validate with V.1.F. (e.g. panel data vs. only 1 included cycle).
V.1.H. Dataset: Mode of data collection	The manner(s) in which the interview was conducted or information was gathered.	CV I	0..n	Backend: format as "Select all that apply."
V.1.I. Dataset: Observational unit sampling method: Description	A description of the procedure used to select the observational units within the randomization units if the unit of observation is different from the unit of randomization. If sampling was performed in several stages, consider listing them out with an explanation. Include any information related to the sampling that is relevant to users comparing targeted and actual units of observation.	Free text	0..n	Backend: reference power calculations used to determine targeted number of observations in IV.3.

V. Data (continued)

Field		Description	Encoding	Cardin-ality	Programming Notes
V.1.J.	Dataset: weights	The sampling procedures used may make it necessary to apply weights to produce accurate statistical results. Are sampling weights included in this dataset?	CV: Yes / No	1..n	
V.1.K.	Dataset: Notes on data collection	Brief description of the data collection or compilation. Include any relevant information such as which of the dataset's cycles was collected pre-treatment, during treatment, or post-treatment; reasons for differences between time period covered by the data and dates of data collection; quality assurance protocols such as number of call-backs; etc.	Free text	0..n	
V.1.L. Dataset: Cycles: Information on the time period covered by the data and, if different, period of data collection in each cycle. These are often identical but may differ in retrospective surveys or administrative data. Please repeat the information for each cycle (wave or round) included in this dataset.					
V.1.L.i.	Dataset: Cycle: Cycle name	A brief descriptive name used to refer to the cycle (data collection or measurement rounds), such as study population census, baseline, endline, etc.	Free text	1..n	Loop: use unique dataset ID and assign unique ID to reference each cycle within the dataset.
V.1.L.ii.	Dataset: Cycle: Start of time period covered	Start date of the time period covered by the data in this data collection cycle. If any element of the date is unspecified, use "X" as input.	Format: YYYY-MM-DD	1..n	
V.1.L.iii.	Dataset: Cycle: End of time period covered	End date of the time period covered by the data in this data collection cycle. If any element of the date is unspecified, use "X" as input.	Format: YYYY-MM-DD	1..n	
V.1.L.iv.	Dataset: Cycle: Start of data collection	Start date of the data collection, if different from the start date of the time period covered by this data collection cycle. If any element of the date is unspecified, use "X" as input.	Format: YYYY-MM-DD	0..n	
V.1.L.v.	Dataset: Cycle: End of data collection	End date of the data collection, if different from the end date of the time period covered by this data collection cycle. If any element of the date is unspecified, use "X" as input.	Format: YYYY-MM-DD	0..n	
V.1.M.Dataset: Arms: Please repeat this information for each treatment arm of this study.					
V.1.M.i.	Dataset: Arm: Name	A brief descriptive name used to refer to the study arm.	Free text	1..n	Loop: use unique arm ID and unique dataset ID to reference each arm within the dataset. Cross check/pre-fill with arm names in III.6.A.
V.1.M.ii.	Dataset: Arm: Targeted number of observational units	The targeted number of observational units in this arm. Include if an approximate target was used.	Numeric	0..n	Cross check sum with V.1.C.
V.1.M.iii.	Dataset: Arm: Actual number of observational units	The actual number of observational units in this arm included in the dataset. Leave empty if this data set does not have experimental arms (e.g. study population census prior to randomization).	Numeric	0..n	Cross check sum with V.1.D.

31

VI. Ethics and Research Transparency

Field	Description	Encoding	Cardinality	Programming Notes
VI.1. Ethics Review: Include information on any ethics review conducted.				
VI.1.A. Ethics Review: Reviewing institution	The name or hosting institution of the ethics review body.	Free text	0..n	
VI.1.B. Ethics Review: Review protocol number	IRB protocol number or case reference.	Free text	0..n	
VI.2. Research ethics documentation	Select all documentation available discussing the ethics of the research or documenting the consent process.	CV J	0..n	Back end: format as "select all that apply". Collect "other" responses for updating the CV.
VI.3. Registration/pre-specification	Was the experiment registered or pre-specified? Select all documentation available with time-stamped/version-controlled records.	CV K	0..n	Back end: format as "select all that apply". Collect "other" responses for updating the CV.
VI.4. Funding agency/sponsor	The source(s) of funds for production of the work. Please list all organizations (local, national, or international) that have materially contributed, in cash or in kind, to the data collection or compilation.	Free text	0..n	
VI.5. Implementation partner	Other parties or persons that have played a significant role in implementing the interventions or collecting the data. Please name individuals' affiliations and roles in their organization at the time of implementation.	Free text	0..n	

VII. External Resources

Field	Description	Encoding	Cardinality	Programming Notes
VII.1.Resources: Information on any related materials. Include the location(s) of the data, separating locations with different access conditions, as well as other information helpful to data users, such as related publications, information on prior studies the work extends or builds on, questionnaires or codebooks, and any ethics documentation or research-transparency related records.				Cross-reference/prefill information in IV.1 (prior work/related studies), V.1 (listed datasets), VI.2 (research ethics documentation), and VI.3 (registration/pre-specification)
VII.1.A.External resource: Type	Please select all external resource types included in this location or citation.	CV L	1..n	Check boxes to select all that apply.
VII.1.B.External resource: Description	A brief description or name of the resource(s).	Free text	0..n	
VII.1.C.External resource: Citation	Complete bibliographic reference containing all of the elements of a citation that can be used to cite the work following a standard format such as APA, MLA, Chicago, etc.	Free text	1..n	
VII.1.D.External resource: Link (DOI/URL)	The DOI or, if DOI is not available, URL of the resource. Leave blank if neither is available.	Free text	0..n	
VII.1.E.External resource: Access policy	Is access to the resource restricted in any way? If known, provide a description of the restrictions and/or the process for accessing the resource.	Free text	0..n	

B   The Controlled Vocabularies

Below is a list of the controlled vocabularies for text fields in the metadata schema, labeled alphabetically for referencing. The first two columns contain parent categories and detailed child categories. In some controlled vocabularies, the parent category can be selected, whereas in others, the user has to select one of the child categories (following the conventions of the source CV); this is indicated by the use of italics for the parent. The third column contains "Notes" on the CV options.

Options added to existing CVs are indicated by underlined text. If a CV is labeled as modified, but no entries are underlined (as in Controlled vocabulary "D. Covariates: Individual"), this indicates that some categories were dropped or consolidated or that the "Notes" were edited or added.

Table Legend

italics	Parent categories in italics cannot be selected and are displayed for organizational purposes only. Selecting one of the child categories is required.
<u>underline</u>	Underlined fields were added or modified from the original source.

Table 2: Controlled Vocabularies

A. Unit of Obs./Randomization (Source: Adapted from DDI)

		Notes
1.	Individual	Any individual person, irrespective of demographic characteristics, professional, social or legal status, or affiliation.
	1.1 Political/social leader	
	1.2 Health provider	e.g. Doctors, nurses, midwives, etc.
	1.3 Patient	
	1.4 Education provider	e.g. Teachers, principals, etc.
	1.5 Student	
	1.6 Farmer	
	1.7 Employee	
	1.8 Business owner	
	1.9 Voter	
	1.10 Public servant	
	1.11 Parent	
	1.12 Other	
2.	Organization or legal entity	Any kind of formal administrative and functional structure - includes associations, institutions, agencies, businesses, political parties, schools, etc.
	2.1 Firm or business	
	2.2 Legal or administrative division of a firm or business	e.g. Department
	2.3 Farm or agricultural business	
	2.4 School	
	2.5 Legal or administrative division of a school	e.g. subjects, cohorts, grades
	2.6 University/college	
	2.7 Legal or administrative division of a university/college	e.g. majors, cohorts
	2.8 Hospital, health clinic or doctor's office	
	2.9 Other organization or legal entity	
3.	Family	Two or more people related by blood, marriage (including step-relations), or adoption / fostering, or who identify as a couple, and who may or may not live together.
	3.1 Nuclear family	
	3.2 Extended family	
	3.3 Parent(s) with dependent children	
	3.4 Couples	
	3.5 Other	
4.	Household	A person or group of people who share common living arrangements or certain amenities, resources, or facilities. This may include pooling some or all of their income and wealth and collectively consuming certain types of goods and services, mainly housing and food.
5.	Housing Unit	A house, apartment, mobile home, group of rooms, or single room that is occupied (or intended for occupancy) as separate living quarters in which the occupants live and eat separately from other building occupants.

A. Unit of Observation/Randomization (Cont.)

		Notes
6.	Other group	Two or more individuals assembled together or having some unifying relationship.
7.	Event/process	Any type of incident, occurrence, or activity. Events are usually one-time, individual occurrences, with a limited, or short duration. Examples: criminal offenses, riots, meetings, elections, sports competitions, terrorist attacks, natural disasters like floods, etc. Processes typically take place over time, and may include multiple "events" or gradual changes that ultimately lead, or are projected to lead, to a particular result. Examples: court trials, criminal investigations, political campaigns, medical treatments, education, athletes' training, etc.
8.	Geographic unit	Any entity that can be spatially defined as a geographic area, with either natural (physical) or administrative boundaries.
	8.1 Physical division of a firm or business	e.g, plants, production lines
	8.2 Physical division of a school or university/college	e.g. classrooms, buildings
	8.3 Agricultural plot or physical unit	e.g., stable, greenhouse
	8.4 Census tract, zip code, or other neighborhood-level administrative unit based on geographic division	
	8.5 Village, community, or other town-level geographic division	
	8.6 District, province, or other upper-level geographic division	
9.	Time unit	Any period of time: year, week, month, day, or bimonthly or quarterly periods, etc.
10.	Text unit	Books, articles, any written piece/entity.
11.	Other	

B. Intervention Assign. Strategy (Source: Adapted from CT.gov)

		Notes
1.	Parallel	Arms are assigned to one (or no) intervention in parallel for the duration of the intervention(s).
2.	Factorial	Two or more interventions are partially or fully cross-randomized to arms and evaluated in parallel.
3.	Crossover	Arms are assigned to different interventions or combinations of interventions (including no intervention) during different phases of the study.
4.	Other	

C. Study Sampling Method (Source: DDI)

	Study Sampling Method (Source: DDI)	Notes
1.	Total universe (population)	All units (individuals, households, organizations, etc.) of a target population are included in the randomization. For example, if the target population is defined as the members of a trade union, all union members are invited to participate in the study. Also called "census" if the entire population of a regional unit (e.g. a country) is selected.
2.	Probability	All units (individuals, households, organizations, etc.) of a target population have a non-zero probability of being included in the randomization sample and this probability can be accurately determined. Use this broader term if a more specific type of probability sampling is not known or is difficult to identify.
	2.1 Simple random	All units of a target population have an equal probability of being included in the randomization sample. Typically, the entire population is listed in a "sample frame", and units are then chosen from this frame using a random selection method.
	2.2 Systematic random	A fixed selection interval is determined by dividing the population size by the desired sample size. A starting point is then randomly drawn from the sample frame, which normally covers the entire target population. From this starting point, units for the randomization sample are chosen based on the selection interval. Also known as interval sampling.
	2.3 Stratified	The target population is subdivided into separate and mutually exclusive segments (strata) that cover the entire population. Independent random samples are then drawn from each segment. For example, in a national public opinion survey the entire population is divided into two regional strata: East and West. After this, randomization units are drawn from within each region using simple or systematic random sampling. Use this broader term if the specific type of stratified sampling is not known or difficult to identify.
	2.3.1 Stratified: Proportional stratified	The target population is subdivided into separate and mutually exclusive segments (strata) that cover the entire population. Independent random samples are then drawn from each segment. Use this broader term if the specific type of stratified sampling is not known or difficult to identify.
	2.3.2 Stratified: Disproportional stratified	The target population is subdivided into separate and mutually exclusive segments (strata) that cover the entire population. In disproportional sampling the number of units chosen from each stratum is not proportional to the population size of the stratum when viewed against the entire population. The number of sampled randomization units from each stratum can be equal, optimal, or can reflect the purpose of the study, like oversampling of different subgroups of the population.
	2.4 Cluster	The target population is divided into naturally occurring segments (clusters) and a probability sample of the clusters is selected. Data are then collected from all units within each selected cluster. Sampling is often clustered by geography, or time period. Use this broader term if a more specific type of cluster sampling is not known or is difficult to identify.
	2.4.1 Cluster: Simple random	The target population is divided into naturally occurring segments (clusters) and a simple random sample of the clusters is selected for randomization. Data are then collected from all units within each selected cluster. For example, for a sample of students in a city, a number of schools would be chosen using the random selection method, and then all of the students from every sampled school would be included.
	2.4.2 Cluster: Stratified random	The target population is divided into naturally occurring segments (clusters); next, these are divided into mutually exclusive strata and a random sample of clusters is selected from each stratum. Data are then collected from all units within each selected cluster. For example, for a sample of students

C. Study Sampling Method (Cont.)

	Notes
	in a city, schools would be divided into two strata by school type (private vs. public); schools would be then randomly selected from each stratum, and all of the students from every sampled school would be included.
2.5 Multistage	Sampling is carried out in stages using smaller and smaller units at each stage, and all stages involve a probability selection. The type of probability sampling procedure may be different at each stage. For example, for a sample of students in a city, schools are randomly selected in the first stage. A random sample of classes within each selected school is drawn in the second stage. Students are then randomly selected from each of these classes in the third stage.
3. Non-probability	The selection of randomization units (individuals, households, organizations, etc.) from the target population is not based on random selection. It is not possible to determine the probability of each element to be sampled. Use this broader term if the specific type of non-probability is not known, difficult to identify, or if multiple non-probability methods are being employed.
3.1 Availability	The sample selection is based on the units' accessibility/relative ease of access. They may be easy to approach, or may themselves choose to participate in the study (self-selection). Researchers may have particular target groups in mind but they do not control the sample selection mechanism. Also called "convenience" or "opportunity" sampling.
3.2 Purposive	Randomization units are specifically identified, selected and contacted for the information they can provide on the researched topic. Selection is based on different characteristics of the independent and/or dependent variables under study, and relies on the researchers' judgement. The study authors, or persons authorized by them have control over the sample selection mechanism and the universe is defined in terms of the selection criteria. Also called "judgement" sampling. Some types of purposive sampling are typical/deviant case, homogeneous/maximum variation, expert, or critical case sampling.
3.3 Quota	The target population is subdivided into separate and mutually exclusive segments according to some predefined quotation criteria. The distribution of the quotation criteria (gender/age/ethnicity ratio, or other characteristics, like religion, education, etc.) is intended to reflect the real structure of the target population or the structure of the desired study population. Non-probability samples are then drawn from each segment until a specific number of randomization units has been reached.
3.4 Respondent assisted	Randomization units are identified from a target population with the assistance of units already selected (adapted from "Public Health Research Methods", ed. Greg Guest, Emily E. Namey, 2014). A typical case is snowball sampling, in which the researcher identifies a group of units that matches a particular criterion of eligibility. The latter are asked to recruit other members of the same population that fulfill the same criterion of eligibility (sampling of specific populations like migrants, etc.).
4. Mix of probability and non-probability sampling	Sample design that combines probability and non-probability sampling within the same sampling process. Different types of sampling may be used at different stages of creating the randomization sample. For example, for a sample of minority students in a city, schools are randomly selected in the first stage. Then, a quota sample of students is selected within each school in the second stage. If separate samples are drawn from the same target population using different sampling methods, the type of sampling procedure used for each sample should be classified separately.
5. Other	

37

D. Covariates: Individual (Source: Adapted from GESIS) | E. Covariates: Higher (Source: New CV)

#	D. Covariates: Individual	#	E. Covariates: Higher
1.	Sex	1.	Housing/property characteristics or amenities
2.	Age	2.	Demographics of household members or household structure
3.	Race/ethnicity	3.	Household assets - ownership or debt
4.	Religion	4.	Household income
5.	Citizenship	5.	Farm characteristics
6.	Marital status/registered partnership	6.	Demographic characteristics of town, village or other governmental unit
7.	Education	7.	Geographic characteristics of town, village or other governmental unit
8.	Labor status	8.	Ethno-political characteristics of town, village, or other governmental unit
	8.1 Description of employment	9.	Crime, violence, or legal enforcement indicators
	8.2 Description of professional activity	10.	Firm-level characteristics
	8.3 Professional status	11.	School characteristics
	8.4 Attachment to the labor force	12.	Hospital or clinic characteristics
	8.5 Previous employment	13.	Other
9.	Income		
10.	Other		

F. Study was designed to analyze (Source: New CV) | Notes

#	Study was designed to analyze	Notes
1.	ITT	The data allows estimation of the effect of being assigned to treatment, also called intent to treat effect or ITT (i.e., treatment assignment is recorded in the data; the default).
2.	LATE or TOT	The data allows estimation of the effect of receiving treatment, also called local average treatment effect (LATE) or effect of treatment on the treated (TOT) (i.e., treatment compliance or take-up is recorded in the data).
3.	ATE	The study allows identification of the average effect of treatment in the study population, also called average treatment effect or ATE (i.e., treatment compliance is automatic/perfect; this may be the case for e.g. laboratory experiments).
4.	Heterogeneous treatment effects or effects by subgroup	The study was designed to allow for the identification of heterogeneous treatment effects or effects by subgroup for one or more covariates.
5.	General equilibrium effects	The randomization was designed to be able to identify general equilibrium effects (e.g., cluster randomization to measure cluster-level effects on prices, labor market outcomes, etc.).
6.	Spillovers or externalities	The study was designed to measure spillover effects or externalities caused by the intervention (e.g., cluster randomization with varying saturation and data collected on everyone in the cluster).
7.	Interaction effect of different interventions	The study's interventions were assigned to arms in a way that allows the analysis of interaction effects (e.g., factorial designs).
8.	Effect of varying treatment intensity	The study was designed such that distinct arms were assigned different intensities of a broader intervention (e.g., a cash transfer that has $20, $40, and $60 arms).
9.	Other	Any other design features that permit estimating the effect of an intervention on units in the study population in a specific way.

G. Kind of Data (Source: Adapted from DDI Definition)

	Kind of Data	Notes
1.	Sample survey data	Survey data collected from a sample of an underlying population.
2.	Census/enumeration data	Data that covers a complete population.
3.	Administrative records data	Information collected, used, and stored primarily for administrative (i.e., operational) rather than research purposes.
4.	Aggregate data	Data at a level of aggregation higher than the units represented in the study, such as country or state-level average household income.
5.	Clinical data	Data either collected during the course of ongoing patient care or as part of a formal clinical trial program.
6.	Event/transaction data	Data that describes an event or transaction, such as data recording sales/business transactions.
7.	Observation data/ratings	Data collected as they occur (for example, observing behaviors, events, etc.), without attempting to manipulate any of the independent variables.
8.	Process-produced data	Paradata or process metadata: Information about data cleaning and transformation processes.
9.	Time budget diaries	Data collected from respondent-produced diaries that contain information on their time use.
10.	Choice experiments for preference eliciation	
	10.1 Incentivized	Data produced from choice experiments with real-world incentives.
	10.2 Hypothetical	Data produced from hypothetical choice experiments (i.e., those that do not have any real-world implications for the respondents.)
11.	Economic games with participant interaction	Laboratory or "lab-in-the-field." Data collected from laboratory or lab-in-the-field games played by the respondents, such as dictator or trust games, with real-world incentives.
12.	Measurement and tests	
	12.1 Educational	Assessment of knowledge, skills, aptitude, or educational achievement by means of specialized measures or tests. Includes standardized testing.
	12.2 Physical	Assessment of physical properties of living beings, objects, materials, or natural phenomena. For example, blood pressure, heart rate, body weight and height, as well as time, distance, mass, temperature, force, power, speed, GPS data on physical movement and other physical parameters or variables, like geospatial data.
	12.3 Psychological	Assessment of personality traits or psychological/behavioral responses by means of specialized measures or tests. For example, objective tests like self-report measures with a restricted response format, or projective methods allowing free responses, including word association, sentence or story completion, vignettes, cartoon test, thematic apperception tests, role play, drawing tests, inkblot tests, choice ordering exercises, etc.
13.	Textual data	Data taken or coded from texts, including but not limited to documents, reports, or speeches.
14.	Other	

H. Time Method (Source: Adapted from ADA)

		Notes
1.	One-time cross-sectional data	
2.	Repeated cross-sectional data	
3.	<u>Panel</u>	Datasets that contain baseline and endline surveys that track the same participants included here.
4.	Does not apply (admin or similar)	
5.	Other	

I. Mode of Data Collection (Source: Adapted from DDI)

		Notes
1.	Interview	A pre-planned communication between two (or more) people - the interviewer(s) and the interviewee(s) - in which information is obtained by the interviewer(s) from the interviewee(s). If group interaction is part of the method, use 'Focus group'.
	1.1 Face-to-face interview	Data collection method in which a live interviewer conducts a personal interview, presenting questions and entering the responses. Use this broader term if not CAPI or PAPI, or if not known whether CAPI/PAPI or not.
	1.1.1 Face-to-face: CAPI/CAMI	Computer-assisted personal interviewing. Data collection method in which the interviewer reads questions to the respondents from the screen of a computer, laptop, or a mobile device like tablet or smartphone, and enters the answers in the same device. The administration of the interview is managed by a specifically designed program/application.
	1.1.2 Face-to-face: PAPI	Paper-and-pencil interviewing. The interviewer uses a traditional paper questionnaire to read the questions and enter the answers.
	1.2 Telephone interview	Interview administered on the telephone. Use this broader term if not CATI, or if not known whether CATI or not.
	1.2.1 Telephone: CATI	Computer-assisted telephone interviewing. The interviewer asks questions as directed by a computer, responses are keyed directly into the computer and the administration of the interview is managed by a specifically designed program.
	1.2.2 <u>Telephone: PATI</u>	The interviewer uses a traditional paper questionnaire to read the questions and enter the answers; the survey is conducted through a telephone.
	1.3 Email	Interviews conducted via e-mail, usually consisting of several e-mail messages that allow the discussion to continue beyond the first set of questions and answers, or the first e-mail exchange.
	1.4 Web-based	An interview conducted via the Internet. Examples include interviews conducted within online forums or using web-based audio-visual technology enabling the interviewer(s) and interviewee(s) to communicate in real time.
2.	Self-administered questionnaire	Self-administered questionnaire includes knowledge tests and preference elicitation.
	2.1 Paper	Self-administered survey using a traditional paper questionnaire delivered and/or collected by mail (postal services), by fax, or in person by either interviewer, or respondent.
	2.2 Email	Self-administered survey in which questions are presented to the respondent in the text body of an e-mail or as an attachment to an e-mail, but not as a link to a web-based questionnaire. Responses are also sent back via e-mail, in the e-mail body or as an attachment.

I. Mode of Data Collection (Cont.)		Notes
	2.3 SMS/MMS	Self-administered survey in which the respondents receive the questions incorporated in SMS (text messages) or MMS (messages including multimedia content) and send their replies in the same format.
	2.4 Web-based	Computer-assisted web interviewing (CAWI). Data are collected using a web questionnaire, produced with a program for creating web surveys. The program can customize the flow of the questionnaire based on the answers provided, and can allow for the questionnaire to contain pictures, audio and video clips, links to different web pages etc. (adapted from Wikipedia).
	2.5 CASI	Computer-assisted self-interview (CASI). Respondents enter the responses into a computer (desktop, laptop, Palm/PDA, tablet, etc.) by themselves. The administration of the questionnaire is managed by a specifically designed program/application but there is no real-time data transfer as in CAWI, the answers are stored on the device used for the interview. The questionnaire may be fixed form or interactive. Includes VCASI (Video computer-assisted self-interviewing), ACASI (Audio computer-assisted self-interviewing) and TACASI (Telephone audio computer-assisted self-interviewing).
3.	Self-administered writings and/or diaries	Narratives, stories, diaries, and written texts created by the research subject.
	3.1 Email	Narratives, stories, diaries, and written texts submitted via e-mail messages.
	3.2 Paper	Narratives, stories, diaries, and written texts created and collected in paper form.
	3.3 Web-based	Narratives, stories, diaries, and written texts gathered from Internet sources, e.g. websites, blogs, discussion forums.
4.	Observation	Research method that involves collecting data as they occur (for example, observing behaviors, events, etc.), without attempting to manipulate any of the independent variables.
	4.1 Field observation	Observation that is conducted in a natural environment. Field observation is defined as interactions, not designed by the researcher.
	4.1.1 Participant field observation	Type of field observation in which the researcher interacts with the subjects and often plays a role in the social situation under observation. Note: "Field observation" is defined as interactions not designed by the researcher.
	4.1.2 Non-participant field observation	Observation that is conducted in a natural, non-controlled setting without any interaction between the researcher and his/her subjects.
	4.2 Laboratory observation	Observation that is conducted in a controlled, artificially created setting. Note: "Laboratory observation" is defined as researcher-designed economic games between participants
	4.2.1 Computer interactions: Participant	Computer-based economic games in which the researcher interacts with the subjects and often plays a role in the situation under observation.
	4.2.2 Computer interactions: Non-participant	Computer-based economic games that are conducted without any interaction between the researcher and his/her subjects.
	4.2.3 Computer interactions: Bot participant	Computer-based economic games in which a bot interacts with the subjects and often plays a role in the situation under observation
	4.3.1 In-person interactions: Participant	Type of laboratory observation in which the researcher interacts with the subjects and often plays a role in the social situation under observation. Example: Observation of children's play in a laboratory playroom with the researcher taking part in the play.
	4.3.2 In-Person interactions: Non-participant	Type of laboratory observation that is conducted without any interaction between the researcher and his/her subjects.

I. Mode of Data Collection (Cont.)

		Notes
5.	Recording	Registering by mechanical or electronic means, in a form that allows the information to be retrieved and/or reproduced. For example, images or sounds on disc or magnetic tape.
6.	Content coding	As a mode of secondary data collection, content coding applies coding techniques to transform qualitative data (textual, video, audio or still-image) originally produced for other purposes into quantitative data (expressed in unit-by-variable matrices) in accordance with pre-defined categorization schemes.
7.	Aggregation	Statistics that relate to broad classes, groups, or categories. The data are averaged, totaled, or otherwise derived from individual-level data, and it is no longer possible to distinguish the characteristics of individuals within those classes, groups, or categories. For example, the number and age group of the unemployed in specific geographic regions, or national level statistics on the occurrence of specific offences, originally derived from the statistics of individual police districts.
8.	Other	Use if the mode of data collection is known, but not found in the list.

J. Research ethics documentation (Source: New CV)

		Notes
1.	IRB protocol	
2.	Description of consent process	
3.	Consent forms text or dialogue	
4.	Record of consent in the data	
5.	Structured ethics appendix	See Asiedu et al. (2021)
6.	Other	

K. Registration/pre-specification (Source: New CV)

		Notes
1.	Trial registration	Entry in any trial registry
2.	Trial pre-registration	Pre-registration in any trial registry
3.	WHO-accredited clinical trial registry	Any entry (pre- or post-registration) in a WHO-accredited clinical trial registry
4.	Pre-analysis plan	Registered/time-stamped pre-analysis plan
5.	Pre-results acceptance	Pre-results acceptance in an academic journal
6.	Public pre-results document	Other public pre-results proposal or document
7.	Populated pre-analysis plan	Populated pre-analysis plan separate from the research paper
8	Other	

L. External Resources Types (Source: IHSN)

		Notes
1.	Database or data repository entry	Location of data included in this study.
2.	*Document*	
	2.1 Administrative	This includes materials such as the survey budget; grant agreement with sponsors; list of staff and interviewers, etc.
	2.2 Analytical	This includes documents that present analytical output (academic papers, etc.). This does not include the descriptive survey report.
	2.3 Questionnaire	This includes the actual questionnaire(s) used in the field.
	2.4 Reference	Any reference documents that are not directly related to the specific dataset, but that provide background information regarding methodology, etc. For international standard surveys, this may for example include the generic guidelines provided by the survey sponsor.
	2.5 Report	Survey reports, studies and other reports that use the data as the basis for their findings.
	2.6 Technical	Methodological documents related to survey design, interviewer's and supervisor's manuals, editing specifications, data entry operator's manual, tabulation and analysis plan, etc.
	2.7 Other	Miscellaneous items.
3.	Pre-analysis plan	pre-analysis plan, if separate from the trial registration.
4.	Populated pre-analysis plan	Populated pre-analysis plan, if separate from the trial registration/pre-analysis plan.
5.	Research ethics documentation	Any documentation related to research ethics, such as IRB or other ethics review protocols, consent process, consent forms, structured ethics appendix, etc.
6.	Program	Programs generated during data entry and analysis (data entry, editing, tabulation and analysis). Include replication files here.
7.	Table	Tabulations such as confidence intervals that may not be included in a general report.
8.	Audio	Audio type files.
9.	Map	Any cartographic information.
10.	Photo	
11.	Video	Video type files provided as additional visual information.
12.	Website	Link to related website(s).
13.	Other	