Appendices

A The Metadata Schema

Below is the proposed schema in full. The first four 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 two columns are not part of the schema but describe the origin of the field and notes on the suggested implementation for an adopting organization. Note that a number of fields with a named "origin" have been modified from their original source to accommodate the needs of social science experiments. We mark cases where the field is new but inspired by another schema, with "similar to."

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
Field origins	WB	World Bank/IHSN
	DDI	Data Documentation Initiative
	AEA	AEA RCT Registry
	CT	ClinicalTrials.gov
	New	New field
Note on field origins	The "origins" column denotes the original location of the field; however, we have modified the wording (but not substance) of many definitions for clarity in the context of social science RCTs.	

The Metadata Schema

I. Basic Information

Field	Description	Encoding	Cardinality	Origin	Programming Notes
I.1. Title	The name of the study	Free text	1..1	WB	
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	WB	
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	WB	
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. If relevant, please feel free to insert the paper abstract and add any additional information relevant for the data not included elsewhere. Please note that information on available accompanying material can be provided in the *Data* section and the *External Resources* sections.	Free text	0..1	WB / DDI	
I.4 Topic classification	The broad substantive topic(s) covered by the data.	CV	1..n	WB / DDI	Checkboxes to select all that apply.
I.5. Version	Version number of the study entry at the appropriate level.	Numeric	N/A	DDI	Automatically generated.
I.6. Version date	Version date of the study entry at the appropriate level.	Format: YYYY-MM-DD	N/A	DDI	Automatically generated.

II. Study Population

Field	Description	Encoding	Cardinality	Origin	Programming Notes
II.1. Country of intervention	The country or countries in which the intervention was implemented. Please repeat the element for each country, even if the study did not cover the entire country.	CV	1..n	WB	Use ISO country codes.
II.2. Geographical coverage	The geographic level at which the data is representative, conditional on the inclusion/exclusion criteria below. Includes the total geographic scope of the data and, if needed, additional geographic selection criteria. Entries may be country 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 in the sample, as long as all districts were eligible for sampling as part of the sampling strategy.	Free text	0..1	WB / DDI	
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 description of the 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. In some cases the chosen sample from this population is only representative after applying sampling weights; this is indicated by the yes/no sampling weight field for each dataset below. Any additional information about the criteria used in the construction of sampling weights should be included here (see also "Study sampling method" below).	Free text	1..1	DDI	
II.4. Unit of randomization and sample size: Kind and number of randomization units, in integer format. May be the same as the *unit of observation.* In a clustered design this is the kind and number of clusters, pooled across all waves and treatment arms.					
II.4.A. 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 may be the same as the the unit of observation.	CV A	1..n	New (similar to AEA / WB)	
II.4.B. Unit of randomization: Targeted study sample size	The targeted number of randomization units pooled over all units and waves.	Numeric	1..n	New (similar to AEA / WB)	
II.4.C. Unit of randomization: Actual study sample size	The actual number of randomization units pooled over all arms and waves. Count only randomization units for which at least one outcome of one observation unit was measured post intervention.	Numeric	1..n	New (similar to AEA / WB)	

III. Outcomes and Interventions

Field	Description	Encoding	Cardinality	Origin	Programming Notes
III.1. Outcomes: Measurements used to determine the effect of an intervention/treatment/program on participants or experimental units. Please repeat the information for each main outcome measure.					
III.1.A. Outcomes: Name	A brief descriptive name to refer to the outcome measure.	Free text	1..n	CT	Each outcome should be assigned a number by the system for easy reference.
III.1.B. Outcomes: Category	The broad category of the specific outcome measure.	CV	1..n	New	CV under development
III.1.C. Outcomes: 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	CT	
III.1.D. Outcomes: Collected pre-treatment?	Was this outcome measured before any treatment or notification of treatment took place ("at baseline")?	Yes / No	0..n	New	Radio buttons to select one response.
III.2. Number of arms	The number of subgroups of participants in the randomized trial that receives 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 that have different numbers of arms, the maximum number of arms from all periods or phases.	Numeric	1..1	CT	
III.3. **Arms:** Subgroups of participants that receive none, one, or several specific interventions according to the trial's protocol					
III.3.A. Arm: Name	A short name used to identify the arm.	Free Text	0..n	CT	Each arm should be assigned a number by the system for easy reference.
III.3.B. Arm: Targeted sample size.	The targeted number of randomization units assigned to this study arm.	Numeric	1..n	New (similar to AEA / WB)	
III.3.C. Arm: Actual sample size.	The actual number of randomization units assigned to this study arm, defined as the number of randomization units for which at least one outcome was measured.	Numeric	1..n	New (similar to AEA / WB)	

III. Outcomes and Interventions (continued)

Field	Description	Encoding	Cardinality	Origin	Programming Notes
III.4. **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 rice), 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 letters advertising a loan product), etc. Different variants of a process or action are a distinct intervention if they are randomly assigned. Receiving no treatment is not an intervention.					
III.4.A. Intervention: Name	A brief descriptive name used to refer to the intervention.	Free text	1..n	CT	Each intervention should be assigned a number by the system for easy reference.
III.4.B. Intervention: Type	The general type of intervention. Please select "other" if none of the multiple choice options is a good fit and enter a free text type.	CV	1..n	CT	CV under development
III.4.C. Intervention: Description	Free text description of the details of the intervention.	Free text	0..n	CT	
III.5. Intervention assignment strategy	The strategy for assigning interventions to the randomization units in a trial.	CV B	1..1	CT	
III.6. Assignment strategy description	A description of the intervention assignment strategy. Stratification variables (if any) can be listed under "Stratified randomization" below. Any other information about the treatment assignment should be provided here.	Free text	0..1	CT	
III.7. Stratified randomization	If the treatment assignment was carried out using stratified randomization, please explain here how the strata were formed. If possible, name the stratification variables. Write N/A if no stratification took place.	Free text	0..1	WB	
III.8. Arm or group / interventional cross-reference	Indicate which interventions are provided in each arm of the study, using the cross-reference check boxes.	Check-boxes	1..1	CT	Checkboxes using the intervention numbers generated in "Intervention: Name"
III.9. Intervention start	The date when the administration of the intervention (after random assignment) began. This refers to all interventions. If the study involved multiple interventions or e.g. several cohorts with different start dates, please enter the earliest start date of all interventions. If the exact day in the month and / or the month is unknown, please enter "99" for the day and / or the month.	Format: YYYY-MM-DD	1..1	AEA	
III.10. Intervention end	The date when the administration of the intervention ended. This refers to all interventions. If the study involved several interventions or e.g. several cohorts with different end dates, please enter the end date of the intervention which ended last. If the exact day in the month and / or the month is unknown, please enter "99" for the day and / or the month.	Format: YYYY-MM-DD	1..1	AEA	

IV. Study Design

Field	Description	Encoding	Cardinality	Origin	Programming Notes
IV.1. Prior work	Does this study extend or rely on any prior RCTs? Examples are collecting additional outcomes for interventions randomly assigned in a previous study, expanding the sample, or adding a treatment arm. If yes, please cite the study as an external resource.	Yes / No / Unknown	1..1	New (similar to AEA)	Radio buttons to select one response.
IV.2. Study sampling method: Type	The type of sampling method used to select the units to be included in the randomized experiment. If sampling is performed in several stages, please select "Probability – multistage," "Non-probability," or "Mix of probability and non-probability sampling" and provide additional details in the description field.	CV C	1..1	New (similar to WB)	
IV.3. Study sampling method: Description	An overall description of the sampling procedure; 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 method that is relevant to users comparing targeted and actual units of randomization and units of observation.	Free text	0..1	New (similar to WB)	
IV.4. Covariates: Individual	Please select all individual-level covariate categories included in this study.	CV D	0..n	New	Checkboxes to select all that apply.
IV.5. Covariates: Group	Please select all group-level covariate categories included in this study.	CV E	0..n	New	Checkboxes to select all that apply.
IV.6. Does a measure of treatment receipt exist?	Does the data contain a measure of take-up or treatment receipt? If yes, please expand under *Compliance*. Choose "not relevant" if treatment receipt is automatic upon assignment.	Yes / No / Not Relevant	1..1	New	Radio buttons to select one response.
IV.7. Compliance description	Please describe whether noncompliance with any of the interventions is possible and, if available, how it is measured 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. An example is "one-sided imperfect compliance," where arms that were not assigned the intervention are prevented from receiving it, but some units in the treated arm(s) do not take up the intervention.	Free text	1..1	New	
IV.8. The study was designed to analyze	Please select all of the following that the project was designed to measure or analyze.	CV F	0..n	New	Checkboxes to select all that apply.

V. Data

Field	Description	Encoding	Cardinality	Origin	Programming Notes
V.1. Dataset: Information about the datasets included in this study and the methodology employed in data collection. Datasets are distinct from data files. A set of observations may constitute a separate dataset if it contains information central to the analysis, 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 descriptive name for the dataset.	Free text	1..n	New (similar to WB)	Assigned a number by the system for easy reference.
V.1.B. Dataset: Unit of observation	The basic unit(s) of analysis or observation that the dataset describes: individuals, families/households, groups, physical locations, events such as court cases, facilities, institutions/organizations, administrative units, etc.	CV A	1..n	New (similar to WB)	Checkboxes to select all that apply.
V.1.C. Dataset: Sample size - Unit of observation: Targeted	The targeted number of observation units pooled over all treatment arms and waves.	Numeric	1..n	New (similar to WB)	
V.1.D. Dataset: Sample size - Unit of observation: Actual	The actual number of observation units pooled over all treatment arms and waves. Count all units for which an outcome was measured at least once across all post-intervention data collection cycles.	Numeric	1..n	New (similar to WB)	
V.1.E. Dataset: Kind of data	Types of data included. Please select all categories that apply.	CV G	1..n	WB	Checkboxes to select all that apply.
V.1.F. Dataset: Time method	The time method or time dimension of the dataset.	CV H	1..n	WB	Checkboxes to select all that apply.
V.1.G. Dataset: Number of cycles	How many cycles (waves or rounds) are in the dataset?	Numeric	1..n	New	
V.1.H. Dataset: Mode of data collection	The manner in which the interview was conducted or information was gathered. Leave empty if mode unknown.	CV I	0..n	WB	Checkboxes to select all that apply.
V.1.I. Observational unit sampling method: Description	A description of the procedure used to select observational units within randomization units. 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 method that is relevant to users comparing targeted and actual units of randomization and units of observation.	Free text	0..n	New (similar to WB)	
V.1.J. Dataset: Are there sampling weights?	The sampling procedures used may make it necessary to apply weights to produce accurate statistical results. Are sampling weights included in this dataset?	Yes / No	1..n	WB	Radio buttons to select one response.

V. Data (continued)

Field		Description	Encoding	Cardinality	Origin	Programming Notes
V.1.K.	Dataset: Notes on data collection	Brief description of the data collection or compilation. Include 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; how the sample was selected from the study population; quality assurance protocols such as number of call-backs; etc.	Free text	0..n	WB	
V.1.L.	**Dataset: Timing of Cycles:** Information on the time period covered by the data and 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	Brief name or description of the cycle (e.g. baseline, endline).	Free text	1..n	WB	
V.1.L.ii.	Time period covered: Start	Start date of the time period covered in this data collection cycle. If the exact day in the month and / or the month is unknown, please enter "99" for the day and / or the month and explain in the data collection notes.	Format: YYYY-MM-DD	1..n	WB	
V.1.L.iii.	Time period covered: End	End date of the time period covered in this data collection cycle. If the exact day in the month and / or the month is unknown, please enter "99" for the day and / or the month and explain in the data collection notes.	Format: YYYY-MM-DD	1..n	WB	
V.1.L.iv.	Dates of data collection: Start	Start date of the data collection, if different from the start date of the time period covered. If the exact day in the month and / or the month is unknown, please enter "99" for the day and / or the month and explain in the data collection notes.	Format: YYYY-MM-DD	0..n	WB	
V.1.L.v.	Dates of data collection: End	End date of the data collection, if different from the end date of the time period covered. If the exact day in the month and / or the month is unknown, please enter "99" for the day and / or the month and explain in the data collection notes.	Format: YYYY-MM-DD	0..n	WB	
V.1.M.	**Dataset: Arms:** Please repeat this information for each treatment arm of this study.					
V.1.M.i.	Dataset: Arm: Name	Name used to identify the arm in the study.	Free text	1..n	New (similar to CT)	Should match arm names provided in the study design.
V.1.M.ii.	Arms: Number of observational units: Targeted	The targeted number of observational units in this arm.	Numeric	1..n	New (similar to AEA)	
V.1.M.iii.	Arms: Number of observational units: Actual	The actual number of observational units in this arm included in the dataset. Count all units for which an outcome was measured at least once across all post-intervention data collection cycles.	Numeric	1..n	New (similar to AEA)	

VI. Ethics and Research Transparency

Field	Description	Encoding	Cardinality	Origin	Programming Notes
VI.1. Ethics Review: Include information on any ethics review conducted.					
VI.1.A. Reviewing institution	The name or hosting institution of the ethics review body.	Free text	0..n	AEA	
VI.1.B. Review protocol number	IRB protocol number or case reference.	Free text	0..n	AEA	
VI.2. Research ethics documentation	Are any documents available discussing the ethics of the research or documenting the consent process? Please tick all that apply and provide references in the "external resources" section.	CV J	0..n	New	Check boxes to select all that apply.
VI.3. Registration/pre-specification	Was the experiment registered or pre-specified? Please tick all that apply and provide references in the "external resources" section. Only include time-stamped/version-controlled records.	CV K	0..n	New	Check boxes to select all that apply.
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	WB	
VI.5. Implementation partner	Please list any other parties or persons that have played a significant role in implementing the study or collecting the data. Please name individuals' affiliations and roles in their organization at the time of implementation.	Free text	0..n	AEA	

VII. External Resources

Field	Description	Encoding	Cardinality	Origin	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.					
VII.1.A. External resource: Type	For locations or citations that contain multiple resource types, please tick all included.	CV L	1..n	WB	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	WB	
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	WB	
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	WB	
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	New	

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 ("Value") contains the verbatim text that is filled into the metadata schema field. These text entries are typically coded as one word, with periods separating parent and child categories and capitalization indicating separate words. For example, the subcategory "Morbidity and mortality" of the parent "Demography" (Controlled vocabulary "A. Topic classification/outcome category", adopted from CESSDA) is coded as "Demography.MorbidityAndMortality" when entered into the schema field. We also adopted this coding in newly proposed CVs. The last 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 "I. 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.
underline	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 providers	e.g. Doctors, nurses, midwives, etc.
1.3 Patient	
1.4 Education providers	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.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, household, or housing unit	
3.1 All individuals sharing certain amenities, resources, or facilities	
3.2 Nuclear family	
3.3 Extended family	
3.4 Parent(s) with dependent children	
3.5 Couples	
3.6 Shared accommodation or building	
3.7 Other	

A. Unit of Observation/Randomization (Cont.)		Notes
4.	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.
5.	Geographic unit	Any entity that can be spatially defined as a geographic area, with either natural (physical) or administrative boundaries.
	5.1 Physical division of a firm or business	e.g. plants, production lines
	5.2 Physical division of a school or university/college	e.g. classrooms, buildings
	5.3 Agricultural plot or physical unit	e.g., stable, greenhouse)
	5.4 Census tract, zip code, or other neighborhood-level administrative unit based on geographic division	
	5.5 Village, community, or other town-level geographic division	
	5.6 District, province, or other upper-level geographic division	
6.	Time unit	Any period of time: year, week, month, day, or bimonthly or quarterly periods, etc.
7.	Text unit	Books, articles, any written piece/entity.
8.	Other group	Two or more individuals assembled together or having some unifying relationship.
9.	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	Use free-text entry to describe the intervention assignment strategy used in the study.

C. 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 Method", 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	

D. Covariates: Individual (Source: Adapted from GESIS)	E. Covariates: Higher (Source: New CV)
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	
11. N/A: No individual-level analysis	

F. Study was designed to analyze (Source: New CV)	Notes
1. 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.
2. General equilibrium effects	The randomization was designed to be able to identify general equilibrium effects and the data contains the necessary information (e.g., cluster randomization to measure cluster-level effects on prices, labor market outcomes, etc.).
3. 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).
4. Interaction of different treatments	The study's interventions were assigned to arms in a way that allows the analysis of interaction effects (e.g., crossover designs).
5. 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).
6. Other	Any other noteworthy design or data features (e.g., data collected through new or innovative methods or data collected about a novel topic).

G. Kind of Data (Source: Adapted from DDI)		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 budgetdiaries	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.	Panel	Datasets that contain baseline and endline surveys that track the same participants included here.
4.	N/A (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 Telephone: PATI	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		K. Registration/pre-specification		Notes
1.	IRB protocol	1.	Trial registration	Entry in any trial registry
2.	Consent process description	2.	Trial pre-registration	Pre-registration in any trial registry
3.	Other	3.	Registration in WHO-accredited clinical trial registry	
		4.	Registered/time-stamped pre-analysis plan	

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	