1 The Controlled Vocabularies: 0.9.0

Below is version 0.9.0 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 from the original source 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.

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)	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	

	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		

	F. Study was designed to analyze (Source: New CV)	**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)		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.	Panel	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 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 (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	