# CARE-Emory Evaluation Capacity Strengthening Series

# **Session 5**: Sampling and design of evaluation

## Professor Matt Freeman, Emory University

## August 20, 2020

[**Zoom recording**](https://careorg.zoom.us/rec/share/2fRZdZ-q9k9JBZXRwXzjcfcBLJ65T6a8hyNMrPIOmRoeTVXrHqtIFwxKRo1v98sc) Password: H%3Vh^0B

**Slide deck**:

## 

## Introduction

* Why do we do M&E
  + There are a lot of reasons and a lot of stakeholders to whom we are accountable
  + That means that we must do a lot of thinking about the data need to collect, be very clear what those data are for, and then plan for data.
  + Many times, we design an M&E system without thinking how the system feeds into all the reasons/purposes for M&E and for all the stakeholders.
  + Same thing for evaluations
  + This is a very intentional process
  + Cannot design the system until you clearly define purpose (system needs to be right-fit)
* What is monitoring?
  + Ongoing, internally facing, often used for course correcting
  + Should be systematic
* How is evaluation different from monitoring?
  + Complementary, mutually reinforcing components of a single system

|  |  |
| --- | --- |
| **Monitoring**   * Links activities and resources to objectives * Routinely collects data * Usually more focused on what the organization is (or is supposed to be) doing * Feed this information back into operation to mangers to make real-time decisions | **Evaluation**   * Assess causal contribution of activities to outcomes and impacts * Collected at discrete time points * Usually more focused on drawing conclusions at one point in time * Used to improve future operations, inform policy, provide recommendations for future projects |

## Evaluation design

* Basic **evaluation design types** 🡪
  + **Formative**: informs development of program, ToC, M&E approach
    - Exploratory; can be mixed methods (definitely include qualitative methods)
    - Needs assessment to understand context (barriers and facilitators of change)
    - Piloting to shape interventions
  + **Process**: informs if program conducted as envisioned (per protocol)
    - Need a protocol – e.g., program needs to articulate clearly what is intended so that can be assessed
  + **Summative**: informs contribution of program to change; can look change in the following ways 🡪
    - **Equity**: populations not reached or impacted?
    - **Sustainability**: long-term change?
    - **Efficacy**: impact under controlled settings
    - **Efficiency** (Cost-effectiveness): understand whether program gaining benefit under resource constraints
    - **Effectiveness**: impact in real-word setting
* Types of **summative evaluation** 🡪
  + **Descriptive**: determines what is taking place and describes process, conditions, and organizations (formative)
  + **Normative**: compares what is taking place to what should be taking place (adequacy and process) (e.g., did the intervention happen and how; did change occur—non-experimental assessment of change; before and after-design, for example)
  + **Causal**: assess difference intervention makes (impact – probability and plausibility designs [experimental vs quasi-experimental)
* How do we identify counterfactuals / controls – for experimental designs
* What are the challenges to identifying controls – sometimes means that you cannot do an experimental design to generate probability evidence (the ‘gold’-standard)
* **Random** (experimental-probability) vs **non-random** (quasi-experimental-plausibility) control groups
* Provision (can we provide it, in the real world), utilization (will people access it), coverage (can you provide at scale), impact (at population-level)
* Considerations: capacity, need, cost, ethics, etc.

## Sampling for evaluation

* Why sample: can’t do (don’t need to do) a census, e.g., talk to everyone
* Sampling can be random and non-random.
  + Do non-random when: Care about variability, don’t need prevalence, want to explore sub-groups (or can’t do random)
  + Do random when: want representative sampling, have sampling frame, sample size determines how representativeness. (Think about sampling and error.)
* **Non-random sampling approaches**
* Expert sample
* Purposive sample
* Snowball sample
* Convenience sample
* Deviant cases
* **Random sampling approaches**
* Simple == every unit has equal chance of inclusion
* Systematic == every unit equal chance of inclusion, dependent on first unit (Select every fifth person)
* Stratified == different sub-groups have pre-assigned chance of inclusion (randomly select units within strata – like a district, sex, or age – and then randomly select within those strata; strata can also be randomly selected)
* Cluster == each cluster has equal chance of inclusion; within cluster all included; select sub-set of units where you are intervening.
  + More cost-effective and logistically easier than straight random sample; but need to have the right number of units and may miss divergent clusters
* **Sampling and error**
* sample gives an **estimate** of the **true** value of a measure
* random err – difference from true value that results from natural variation in sample cannot avoid)
* system err – error you introduce in how you do study, do everything you can to avoid it (want to understand stunting in kids; sample from schools; systematically excluding kids who don’t go to school and those kids might be more likely to be stunted)

(want to understanding birthweight; look at list of all kids born (vital registry) but not all kids born are on that list)

* Three factors in **determining sample size** (need 2 of 3 to calculate):
  + Power—minimal detectable effect—sample size
* **Internal validity**: error free results; valid for population in which you conducted study
  + Reduce systematic error as far as possible; this is why we randomize (both for participant selection and for treatment-control allocation)
  + If not random, do matching – plausibility evidence; not probabilistic!
* **External validity**: generalizable result; valid for other populations; always need to discuss external validity in your write-up