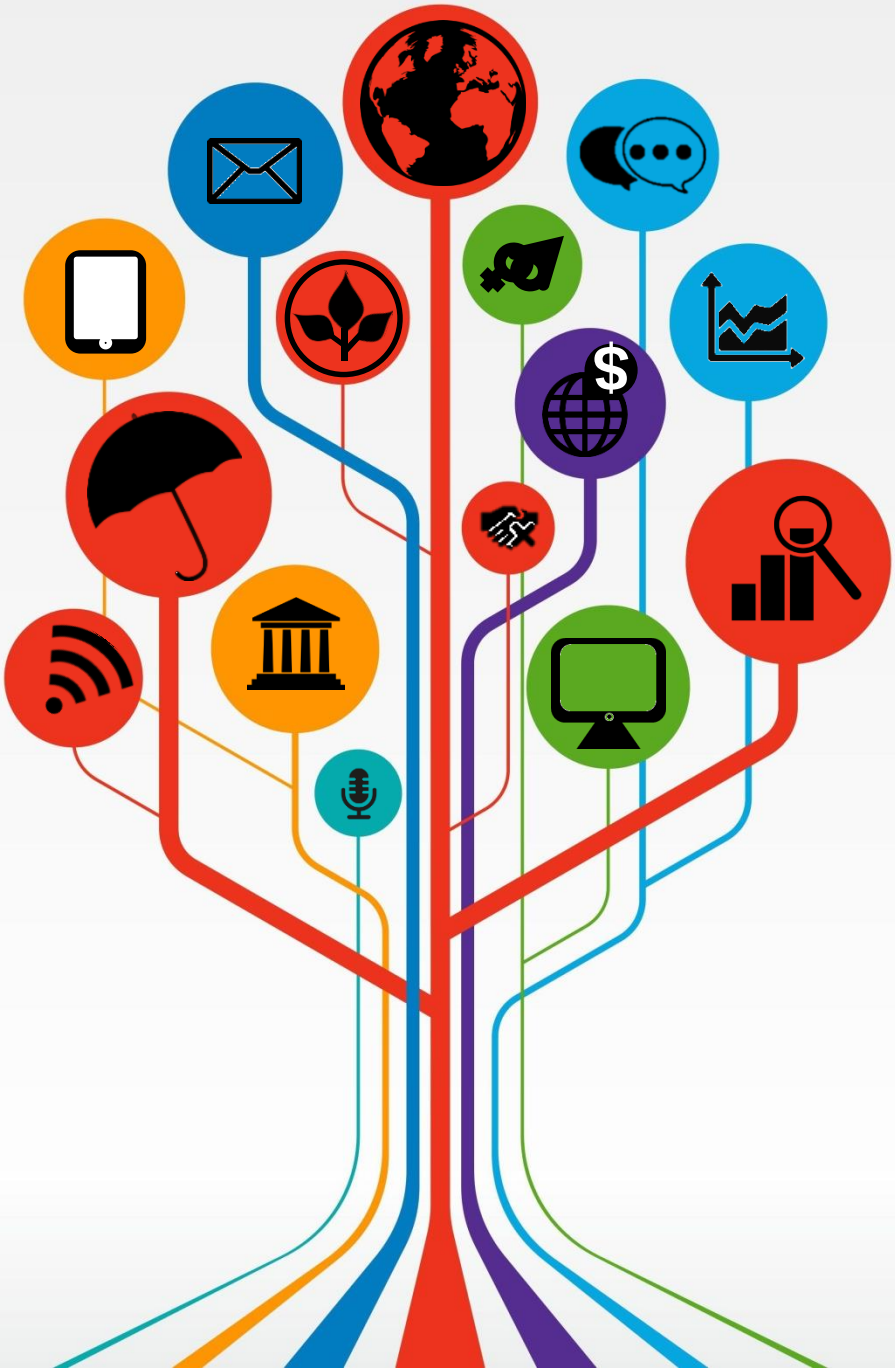
FIELD COORDINATOR WORKSHOP

Manage Successful Impact Evaluations

10 - 14 JUNE 2019
WASHINGTON, DC



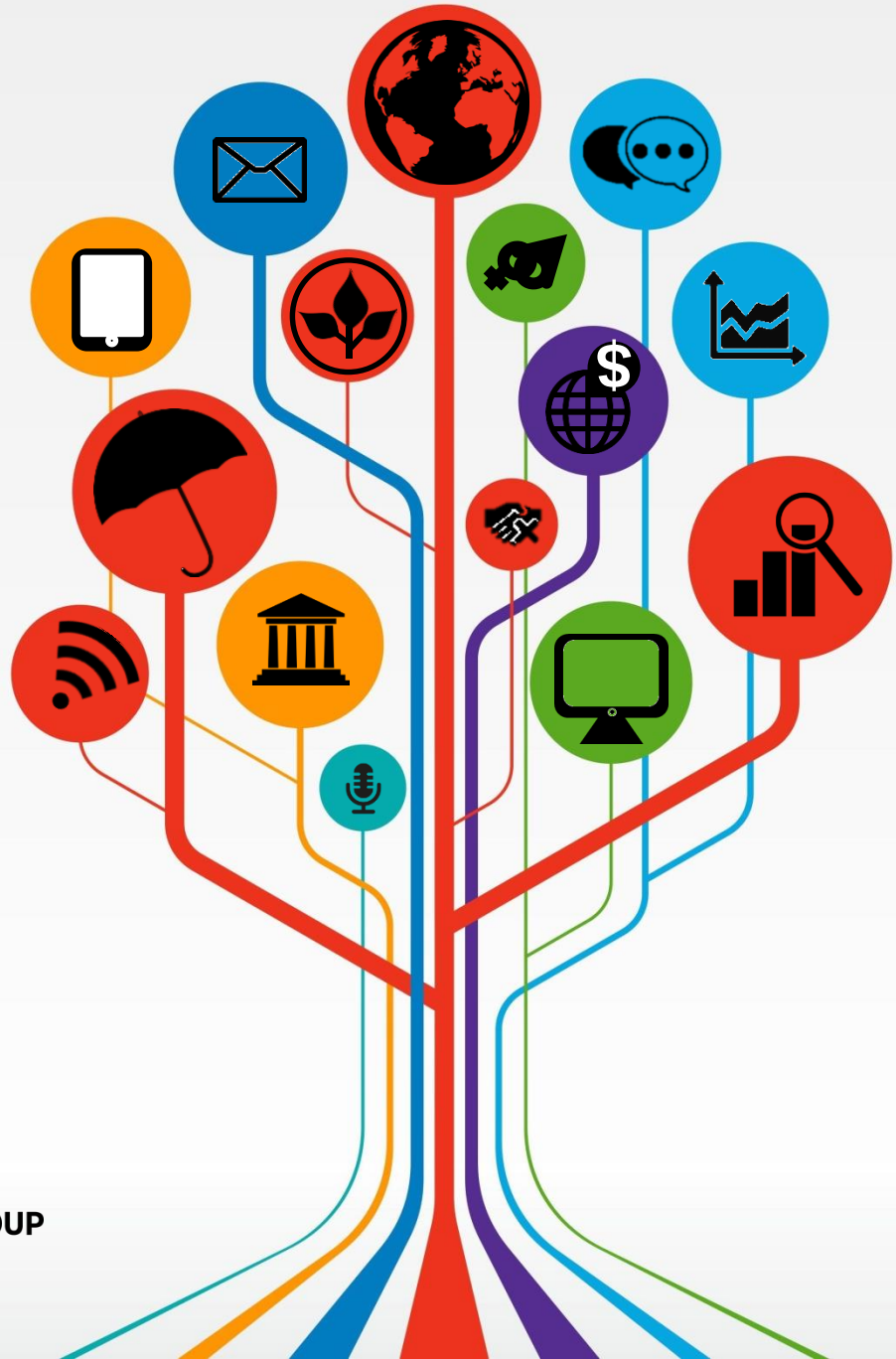
i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GR

Sampling for IE
Track 2

Maria Jones

13 June 2019



i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

outline

Main objective:

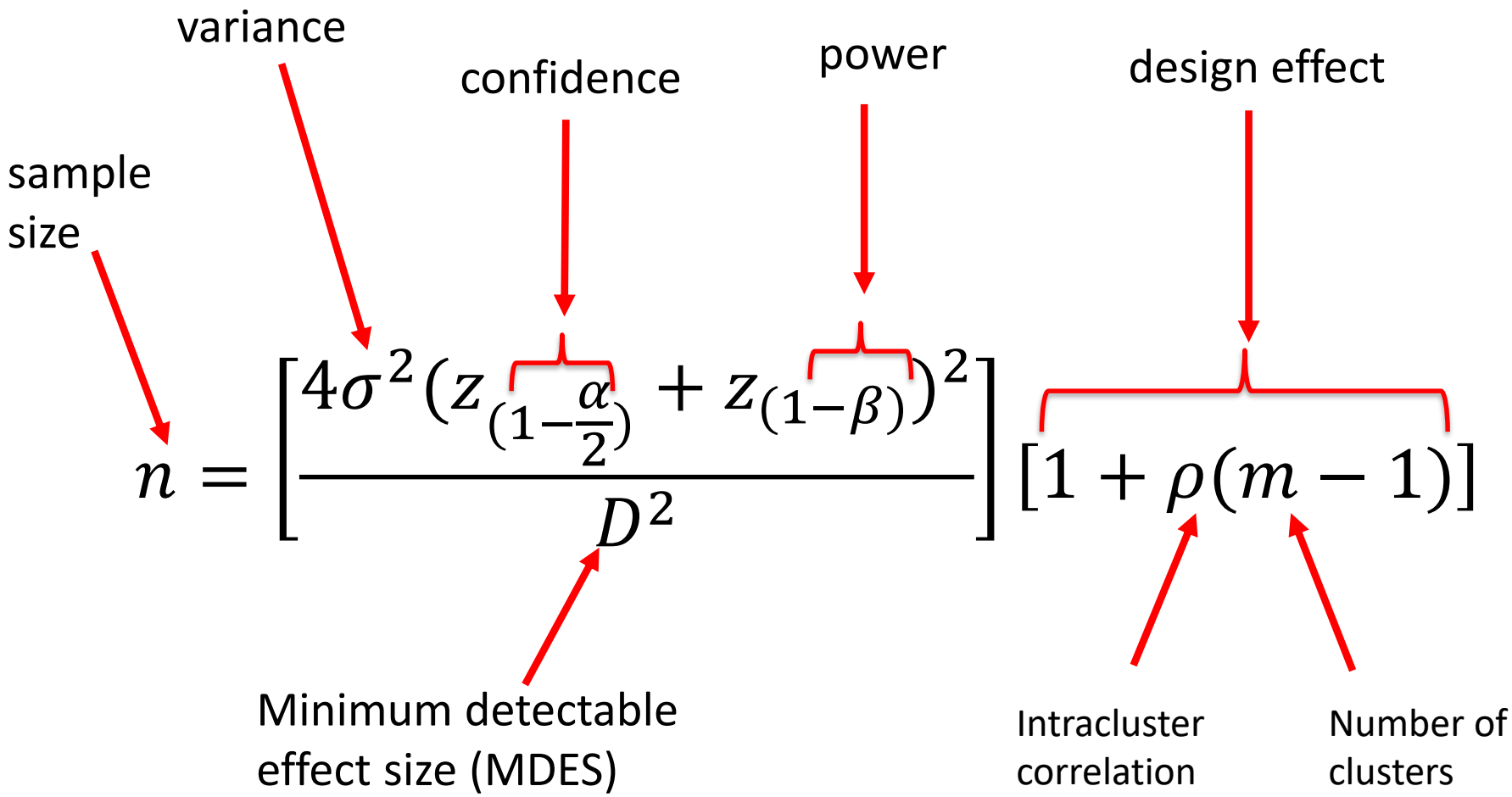
- Learn the key parameters in sample size equation
- Understand how each impacts sample size
- Discuss other factors that (indirectly) affect statistical power

If time allows…

- Sampling in challenging contexts: 2 case studies

a quick caveat …

the key equation



types of power calcs

- Three options
  - Compute sample size given power and effect size
  - Compute power given sample size and effect size
  - Compute effect size given power and sample size

- For IE, typically solve for sample size or MDES
  - Often sample size is pre-determined (IE design / population / budget), so solve for effect size
  - If sample size is flexible, plug in effect size based on context and solve for sample size

confidence and power

$$n = \left[\frac{4\sigma^2\left(z_{(1-\frac{\alpha}{2})} + z_{(1-\beta)}\right)^2}{D^2}\right][1 + \rho(m-1)]$$

confidence and power

- Typically use standard assumptions for these parameters
- type I error: false positive
  - detect an effect when no effect is present
  - α (alpha) = probability of a type I error
    - statistical confidence : (1 − α)
  - standard assumption: 95% confidence
- type II error: false negative
  - fail to detect an effect when an effect is present
  - β (beta) = likelihood of making a type II error
    - power: (1 − β)
  - standard assumption: 80% power

effect size

$$n = \left[\frac{4\sigma^2\left(z_{\left(1-\frac{\alpha}{2}\right)} + z_{(1-\beta)}\right)^2}{D^2}\right][1 + \rho(m-1)]$$

Effect size

- What is minimum detectable effect?

- How should you determine a reasonable MDES?

MDES

- D: the smallest effect size that, if it were any smaller, the intervention would not be worth the effort
  - a.k.a **Minimum Detectable Effect Size** (MDES)
- The smaller the effect you want to be able to detect, the larger the sample you will need
  - larger sample ➔ more precise measuring device
- Very common to solve for MDES when doing IE power calculations

What's a reasonable MDES?

- Depends entirely on project context: how large of an impact is expected?
  - As part of project appraisal, teams create a Results Framework that documents their anticipated impacts for key indicators

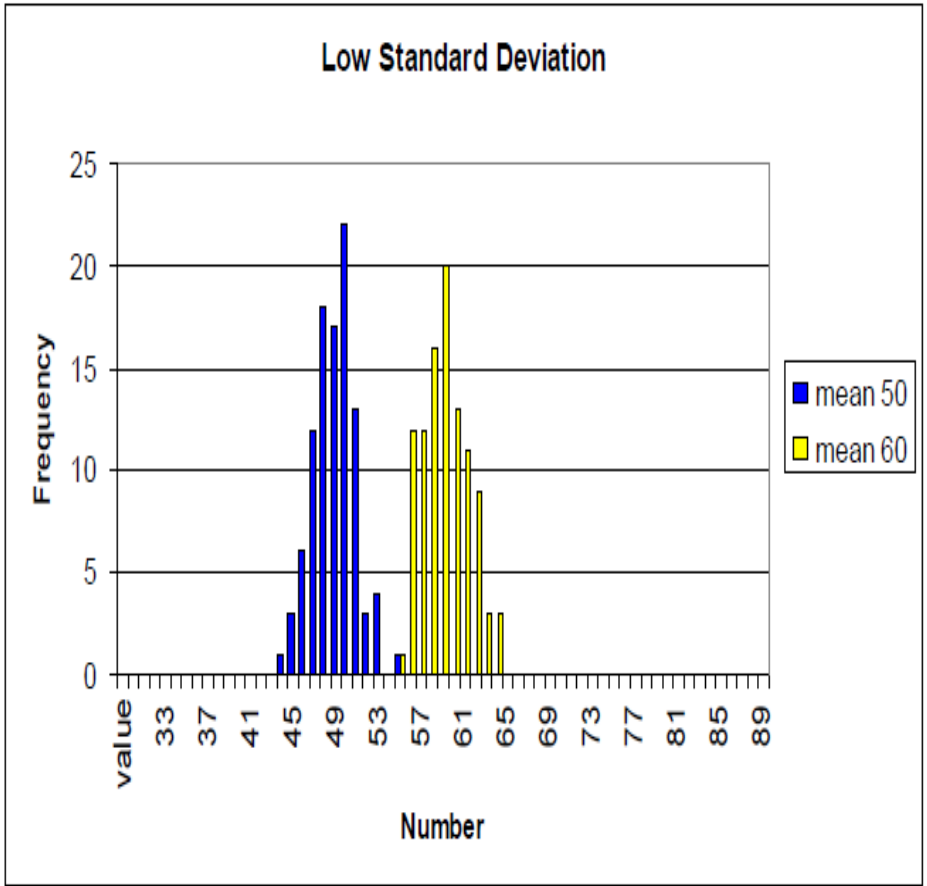
- Example: an irrigation project anticipates a 10% increase in staple crop yield by mid-term and a 20% increase in yield by project closure
  - What if, given initial sample size assumption, MDES is 15%?
  - What if it's 5%?

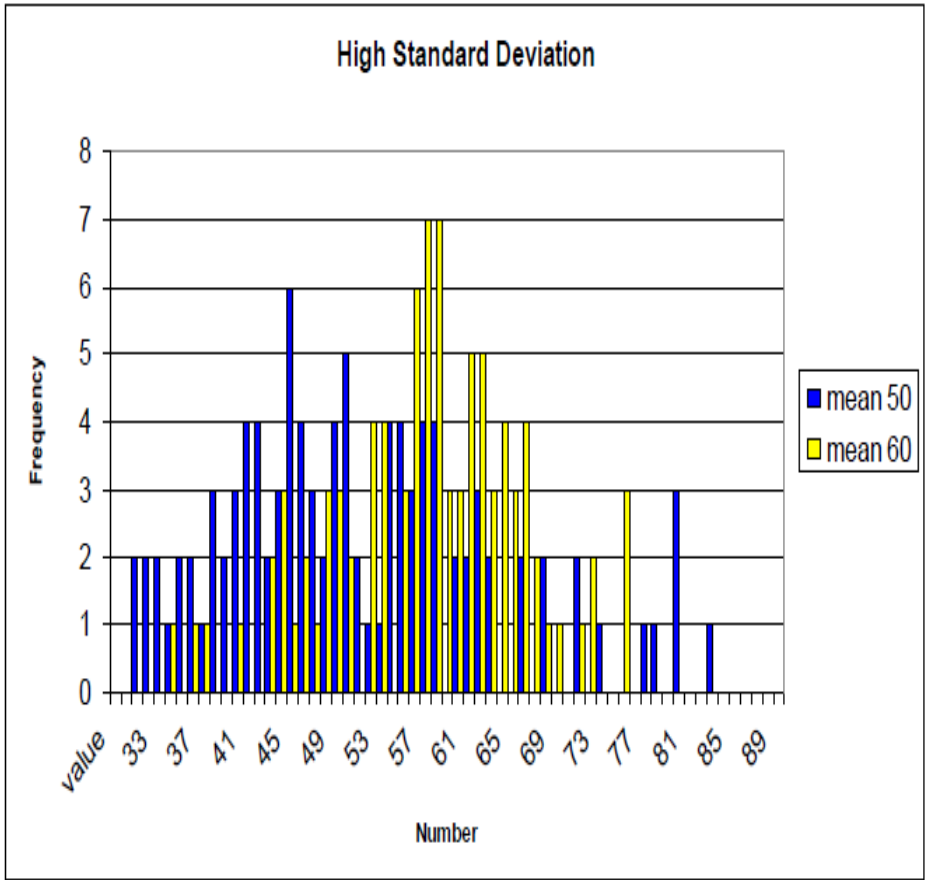
variance of outcomes

$$n = \left[\frac{4\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{D^2}\right][1 + \rho(m-1)]$$

QUIZ

An intervention increases employment by 10% for treatment group on average in two different populations. Would you expect a difference in sample size needed to detect the effect in the two populations?





variance of outcomes

- σ = variance of outcome of interest for study population

- More underlying variance (heterogeneity)
  - → more difficult to detect difference
  - → need larger sample size

- **Tricky:** How do we know about heterogeneity *before* we decide our sample size and collect our data?
  - Ideal: pre-existing data … but often non-existent
  - Can use pre-existing data from a *similar* population
    - Example: LSMS, data routinely collected by govt, satellite imagery
  - Common sense

clustering (aka “design effect”)

$$n = \left[\frac{4\sigma^2\left(z_{\left(1-\frac{\alpha}{2}\right)} + z_{(1-\beta)}\right)^2}{D^2}\right][1 + \rho(m-1)]$$

QUIZ

Which sampling strategy is likely to give you more statistical power?

A. 400 classrooms, 5 students per classroom = 2,000 students
B. 50 classrooms, 40 students per classroom = 2,000 students
C. Both should give you similar statistical power
D. Don’t know

clustering

- Unit for sample size calculation depends on :
  - Level of randomization (intervention) **AND**
  - Level of measured impacts
- In a simple randomized control trial (RCT), unit of randomization and expected impact is the same.
  - Example: subsidies awarded by individual-level lottery
- In a cluster randomized control trial, units are different
  - Example: infrastructure investment at village level, interested in impacts on HHs
    - Randomly assign villages to treatment / control
    - Sample household within villages

clustering

- Level of intervention (“cluster”) most important for sample size calculation

- If few clusters, precision will be limited, regardless of number of HHs sampled

QUIZ

Which sampling strategy is likely to give you more statistical power?

A. 100 villages, 5 HHs per village = 2,000 HHs
B. 100 villages, 50 HHs per village = 2,000 HHs
C. Both should give you similar statistical power
D. Don’t know

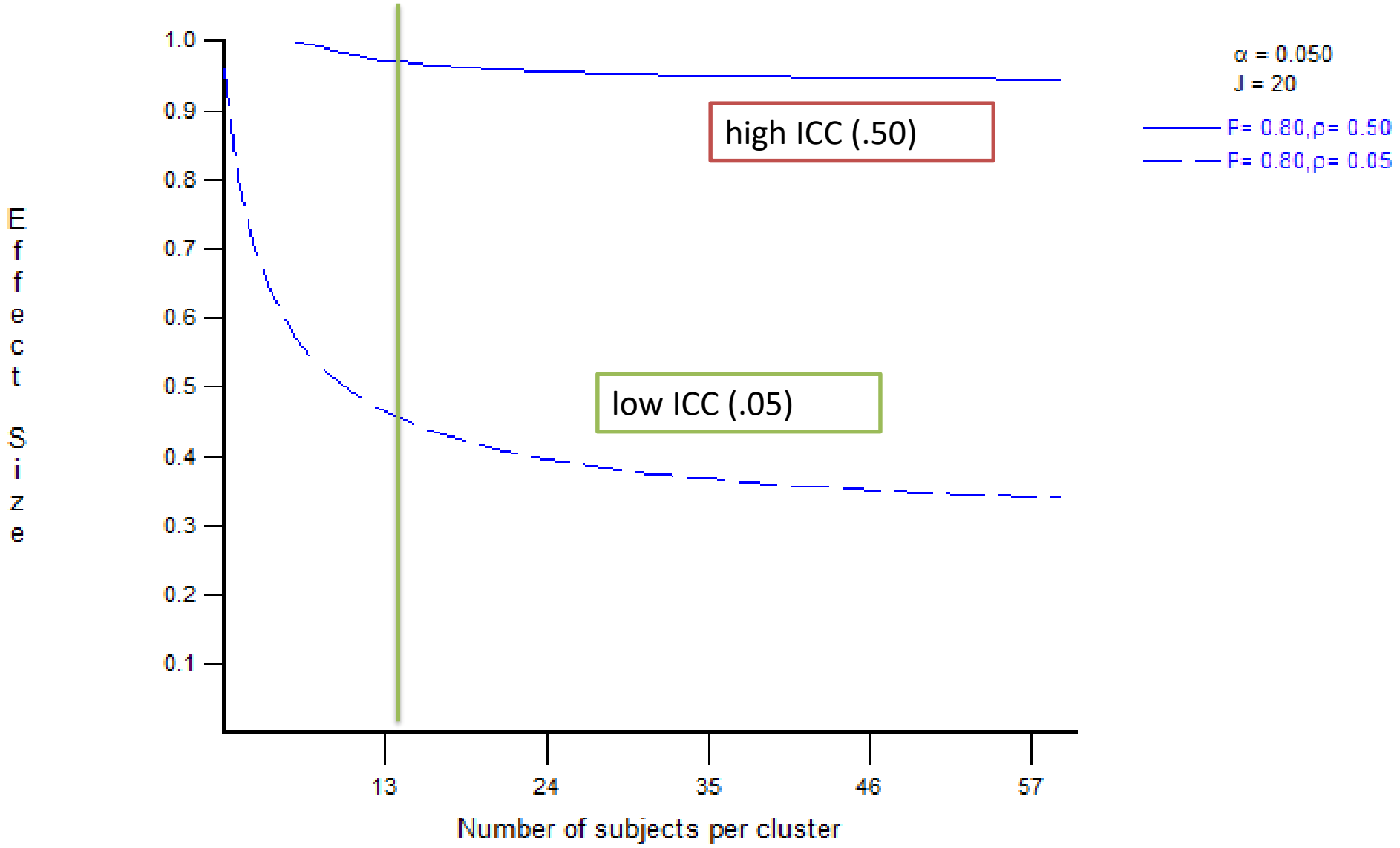
clustering

- **Intracluster correlation (ICC):** similarity of units within clusters

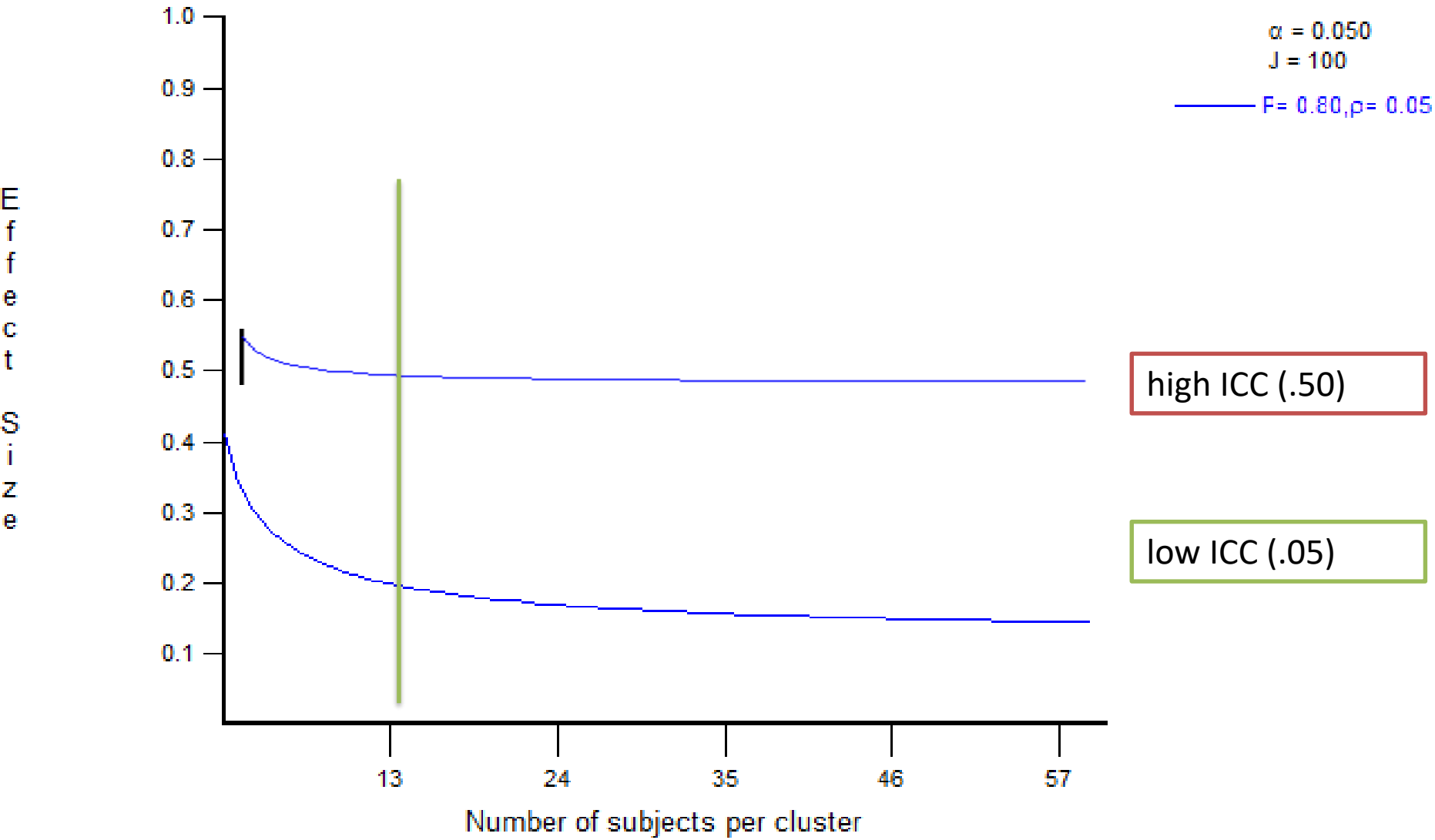
- Is the variation in outcome of interest coming mostly from differences *within* villages (low ICC), or *between* villages (high ICC)?
  - If HHs in village A are similar to each other, but different from HHs in village B, high ICC
  - If HHs in village A are similar to HHs in village B, low ICC

- If ICC = 0, no design effect (same as individual-level randomization)

20 clusters


α = 0.050
J = 20
P= 0.80,ρ= 0.50
P= 0.80,ρ= 0.05
high ICC (.50)
low ICC (.05)
Effect Size
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
13
24
35
46
57
Number of subjects per cluster

100 clusters



clustering

Takeaway

High *intra-cluster correlation* (HHs in same cluster similar)

lower marginal value per extra sampled unit in the cluster

More clusters needed

Other factors that affect power

- Take-up
- Attrition
- Compliance
- Data Quality

QUIZ

- You do power calculations and decide you need a sample of 1,000 HHs for an impact evaluation. The project starts. 6 months in, a monitoring survey shows that take-up of the intervention is 50%.

- What effect will this have on statistical power, given the sample size of 1,000 HHs?

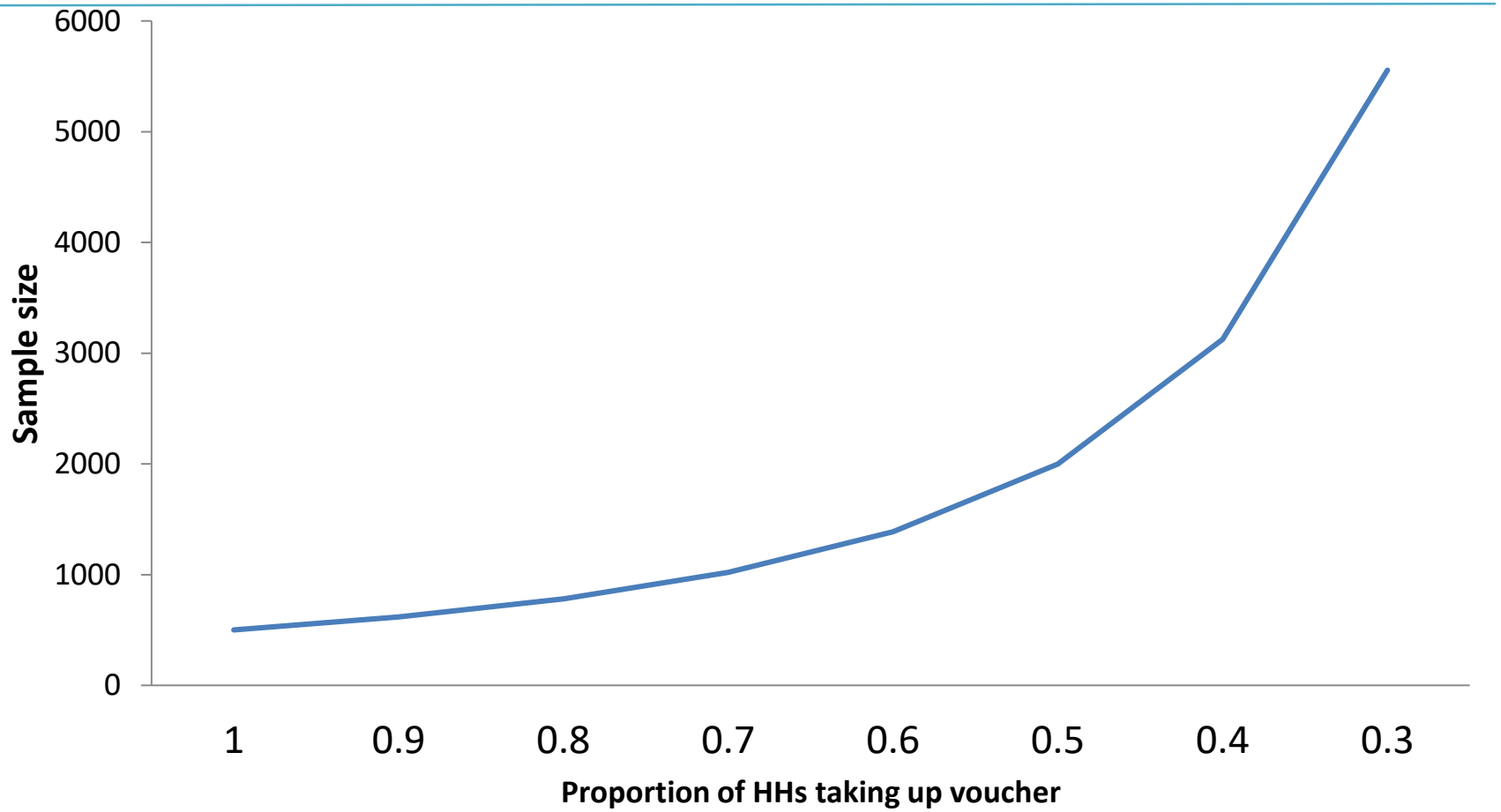
- What could you do to improve power, if increasing the number of HHs is not feasible?

take-up

- Low take-up (rate) for intervention lowers precision
  - Effectively decreases sample size / increases minimum detectable effect
  - Can only detect an effect if it is really large

- Low take-up massively increases sample size requirement (to stay at given MDES and power)

take up vs. sample size



take-up

- To account for take-up rate of 50%, have to increase sample size by factor of 4

- To account for take-up of 10%, increase sample size by factor of 100

attrition

- Effectively the same problem as take-up
- Especially serious if cluster-level
- This is why careful tracking of respondents during follow-up surveys is extremely important

compliance

- Compliance: treatment group receives treatment, control group does not
- Non-compliance: some of the control group receives treatment, *or* some of the treatment group does not
- Power problem is similar to take-up or attrition: if 100% of treated group take-up but 25% of control group take-up, equivalent to take-up of 75%
- Important to avoid this through careful field monitoring of study implementation in real time (if you find out during the follow-up survey, it's too late)

data quality

- Missing data (e.g. skipped questions) → effectively equivalent to attrition

- Measurement or input errors → effectively equivalent to increased variance

Take-aways

- All the following reduce power :
  - Incomplete take-up
  - Sample attrition*
  - Non-compliance with study design*
  - Poor data quality*

** Likely to also introduce bias. Avoid these by careful monitoring, do not modify sample size calculation to account for them.*

Now you know how many people to sample ….

How do you identify them?

Sampling in practice

- Best case scenario: complete sampling frame already exists
- Most often not the case as impact evaluations typically focus on specific subpopulation
- Typically first conduct a listing, then sample
- However, that may not be entirely straightforward, as the two case studies show

Case Study 1 -

Market Listing & Trader Survey

Context

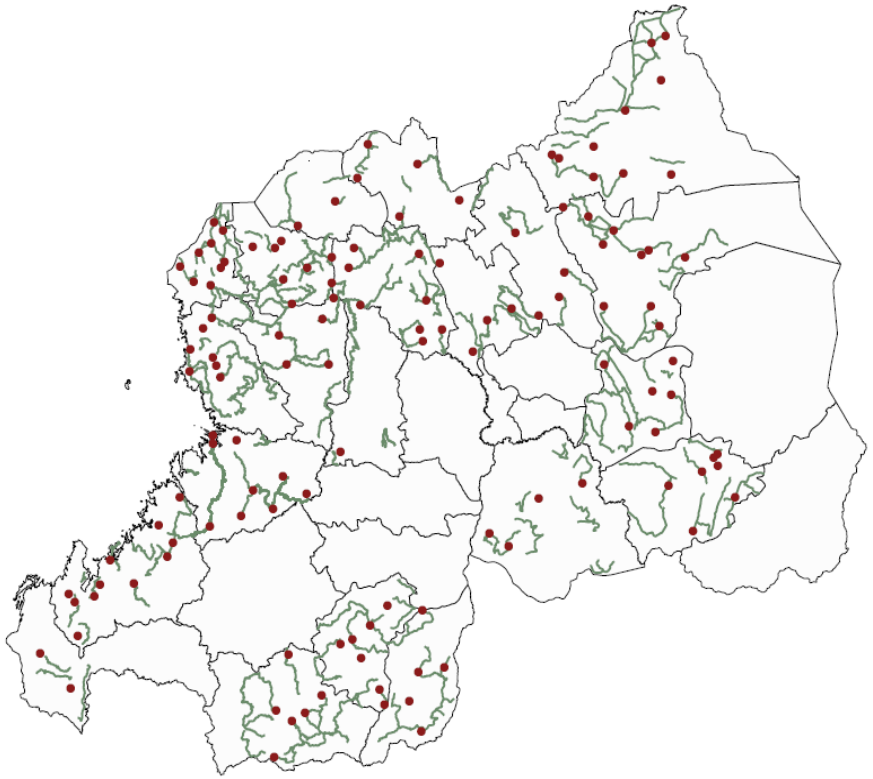
Rural Feeder Roads:

Does improved connectivity change lives?

Market survey - setup

- Understand how market structure and composition has changed over time through
  - a visual listing of all traders present in the market
  - conducting a short trader survey for a sub-sample of traders listed in each market

Market sample



Sampling for trader survey

- Based on power calculations and field practicalities, research team designed a sampling strategy in which the number of traders to survey per market depends on total market size

Market size	Sample size
<=30	100%
31-100	50%
101-400	33.3%
401-500	25%
501-600	20%
601-700	16.67%
701-800	14.28%
801-900	12.5%
901-1000	11.11%
...so on	

Sampling for trader survey

- The trader survey has to be conducted on the same day as the listing
  - for quicker completion as markets do not meet everyday
  - to avoid attrition / confusion as traders are identified by location, clothing, and type of goods, which will change between market days

How can the sample be dynamically selected as soon as listing is complete?

Selecting the sample – method 1

- Allow the enumerators to select the traders to interview

- Pros:
  - Easiest for the enumerator
- Cons:
  - Enumerators are likely to chose the traders they can find easily
  - No guarantee of representation of all types of traders

Selecting the sample – method 2

- Provide a walking skip pattern based on market size for enumerators to follow

Market size	Sample size	Skip pattern to follow
<=30	100%	every **trader** will be interviewed
31-100	50%	every **2nd trader** will be interviewed
101-400	33.3%	every **3rd trader** will be interviewed
401-500	25%	every **4th trader** will be interviewed
501-600	20%	every **5h trader** will be interviewed
601-700	16.67%	every **6th trader** will be interviewed
701-800	14.28%	every **7th trader** will be interviewed
801-900	12.5%	every **8th trader** will be interviewed
901-1000	11.11%	every **9th trader** will be interviewed
...so on		

Selecting the sample – method 2

- Pros
  - There is some form of randomness
  - All trader types are likely to be represented
- Cons:
  - Enumerators have to do a lot of mental math!
  - Very hard to verify whether sampling pattern was followed

Selecting the sample – method 3

- Rely on technology - Program the survey form to dynamically pick the traders to survey

- Pros
  - Enumerators just have to locate the stall listed on the tablet screen
  - All trader types are likely to be represented
- Cons:
  - Programming of the randomization might take time
  - Randomization is not replicable if done on SurveyCTO

What would you do?

What was actually done?

- Weighing the pros and cons of each available method, we left the selection to technology (method 3)
- Overestimated the required sample in each market to ensure required number of trader surveys were reached

Case Study 2 -

Irrigation Scheme Farmer Survey

Irrigation impact evaluation

sampling

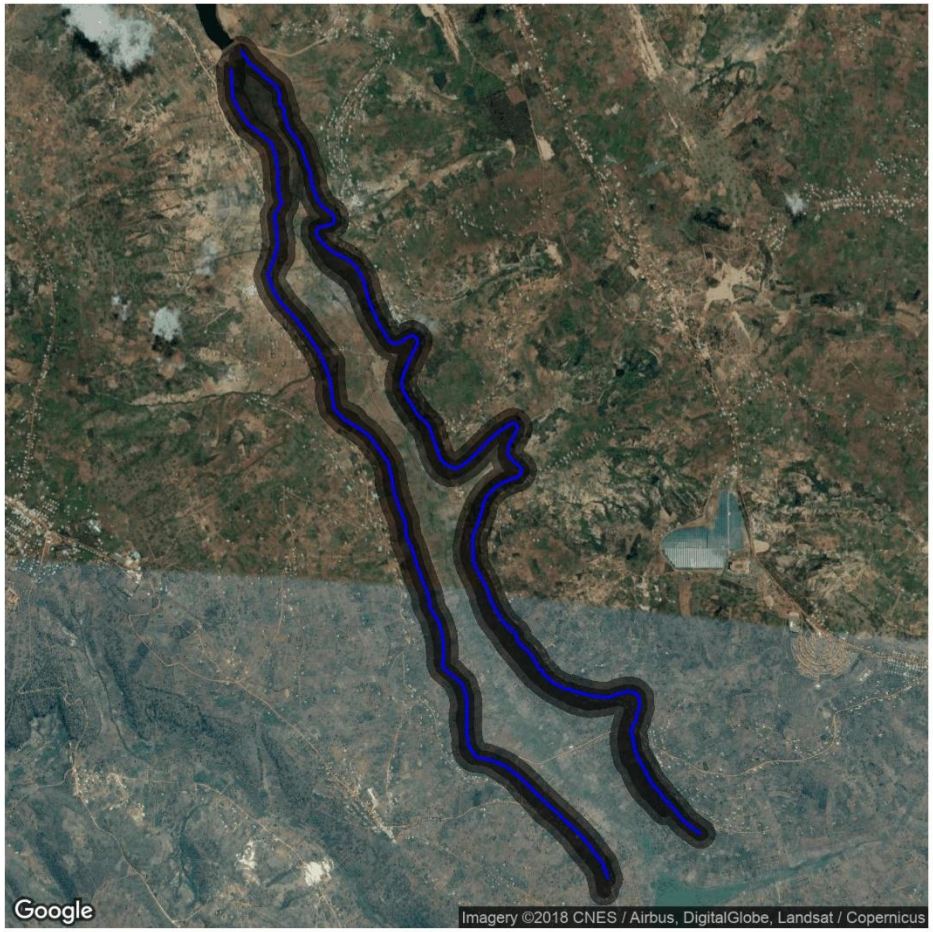
- Spatial regression discontinuity design
  - Compare plots just below irrigation canal to those just above
    - Therefore need to sample plots close to irrigation canal

- How to do that?
  - Listing HHs in the neighboring villages?
    - Plots aren't necessarily close to villages
    - People won't accurately be able to say whether the plot is within 50m of the canal

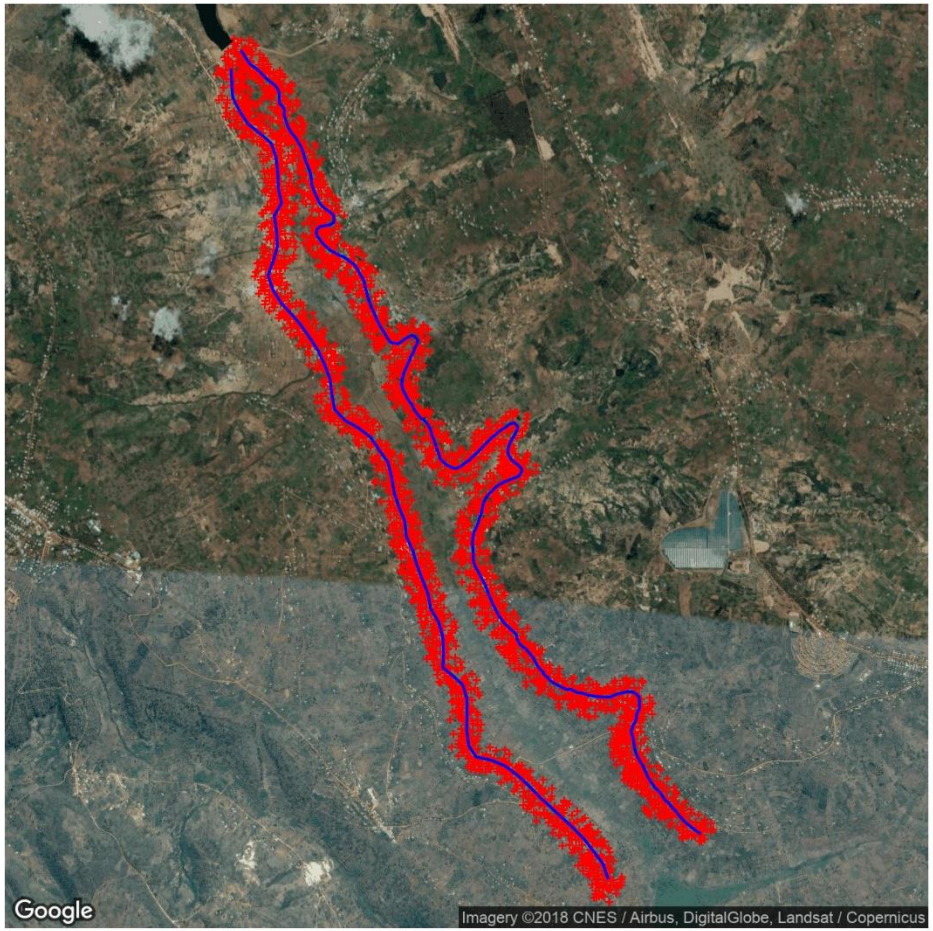
What did we do?

- Dropped uniform grid of points across full site at 2m resolution
- Randomly sampled points, excluding any point within 10m of a point selected
- Enumerators equipped with GPS units visited each sampled point to identify whether the point is agricultural land, and if so find out who cultivates it

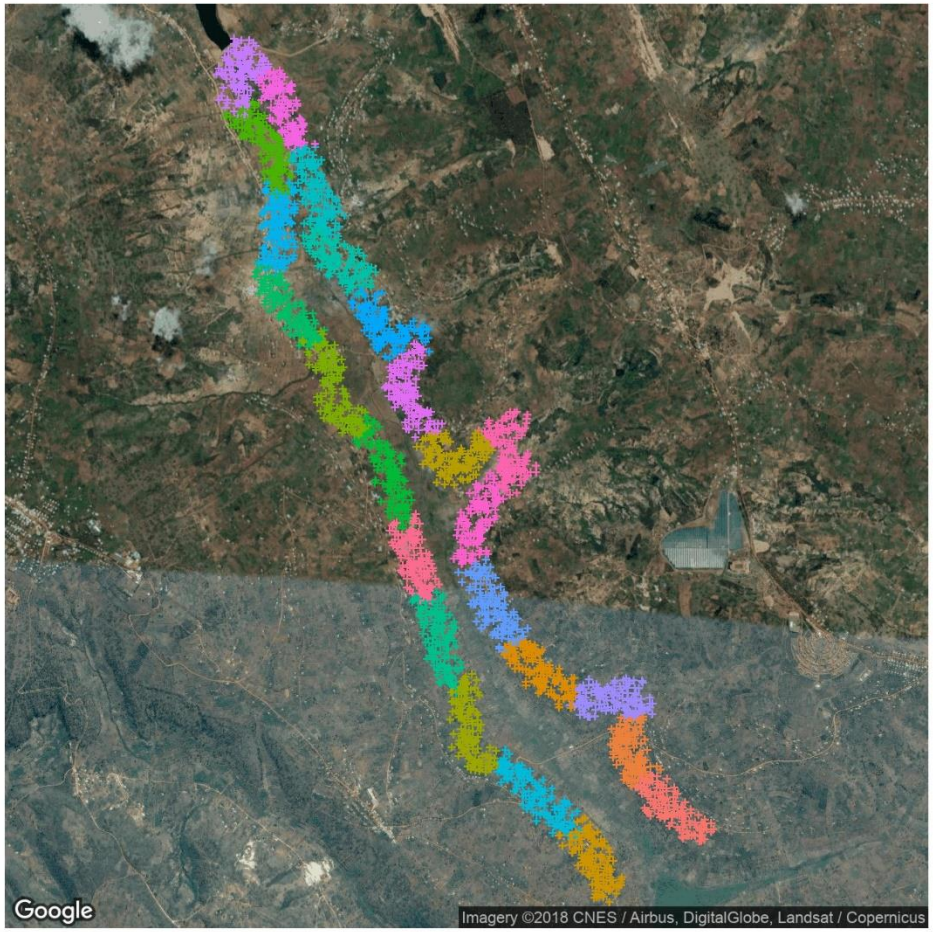
Draw 50m buffer



Drop 3000 points



Assign 24 enumerators to points



outcome

- From 3000 points, 2932 successfully visited and description recorded
- 1,058 distinct village name + cultivator combinations
- Contact village leaders to verify names and remove duplicate households (e.g. husband and wife) → 810 households
- 670 households successfully interviewed (more duplicates discovered during interviews, some names not recognized)
- Once plots were mapped, 76% have sample point within boundaries