



# Introduction to IPA Philippines

Nassreena Sampaco-Baddiri  
November 23, 2015  
Executive Education Course, Manila

A man with short dark hair, wearing a patterned grey zip-up shirt, is shown in profile from the chest up. He is holding a piece of food, possibly a piece of bread or a vegetable, in his right hand. The background is a dimly lit shop or market stall. Shelves behind him are filled with various items, including several green glass bottles and blue and white cans. There are also some bags hanging from the shelves. The overall atmosphere is that of a busy, everyday environment.

# OUR VISION

More evidence, less poverty

A photograph of a classroom scene. A teacher, a woman with her hair in a ponytail wearing a blue and white striped shirt, is leaning over a wooden desk. She is interacting with a young student wearing a green shirt. Other students in orange shirts are visible in the background, some sitting at desks. The classroom has posters on the wall and a window with a grid pattern on the left.

# OUR MISSION

**To discover and promote  
effective solutions to global  
poverty problems.**

# THE PROBLEM



**WASTED MONEY, ENDURING POVERTY**

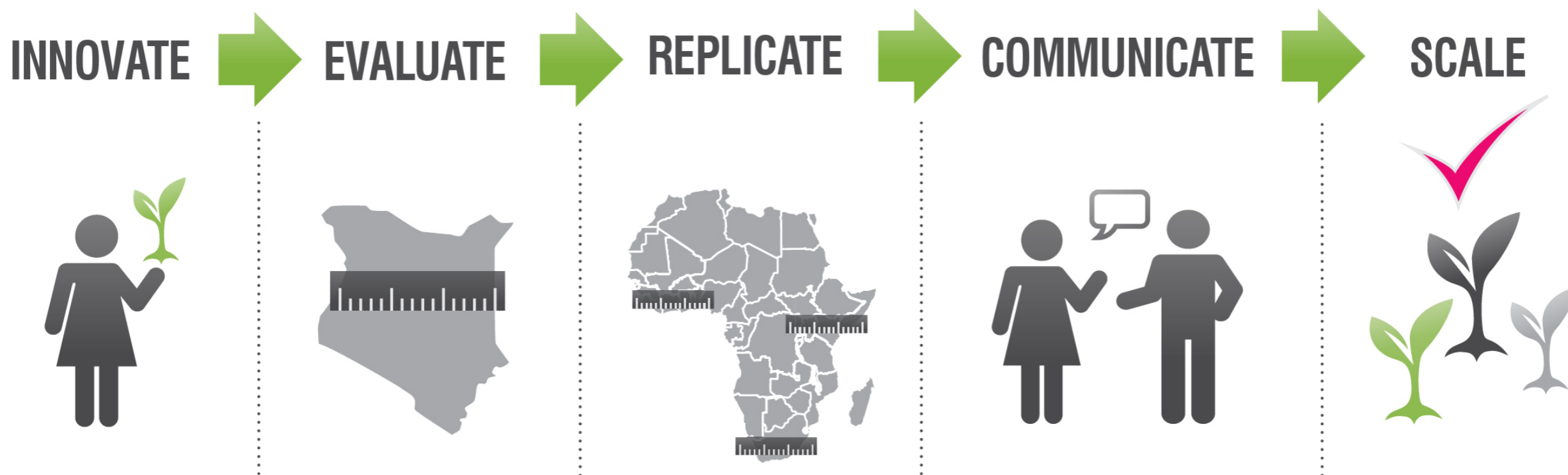
# THE SOLUTION



**MORE EVIDENCE, LESS POVERTY**

# IPA's Approach

*We generate insights on what works and what does not through randomized evaluations, and ensure that those findings will be useful to, and used by practitioners and policy makers.*

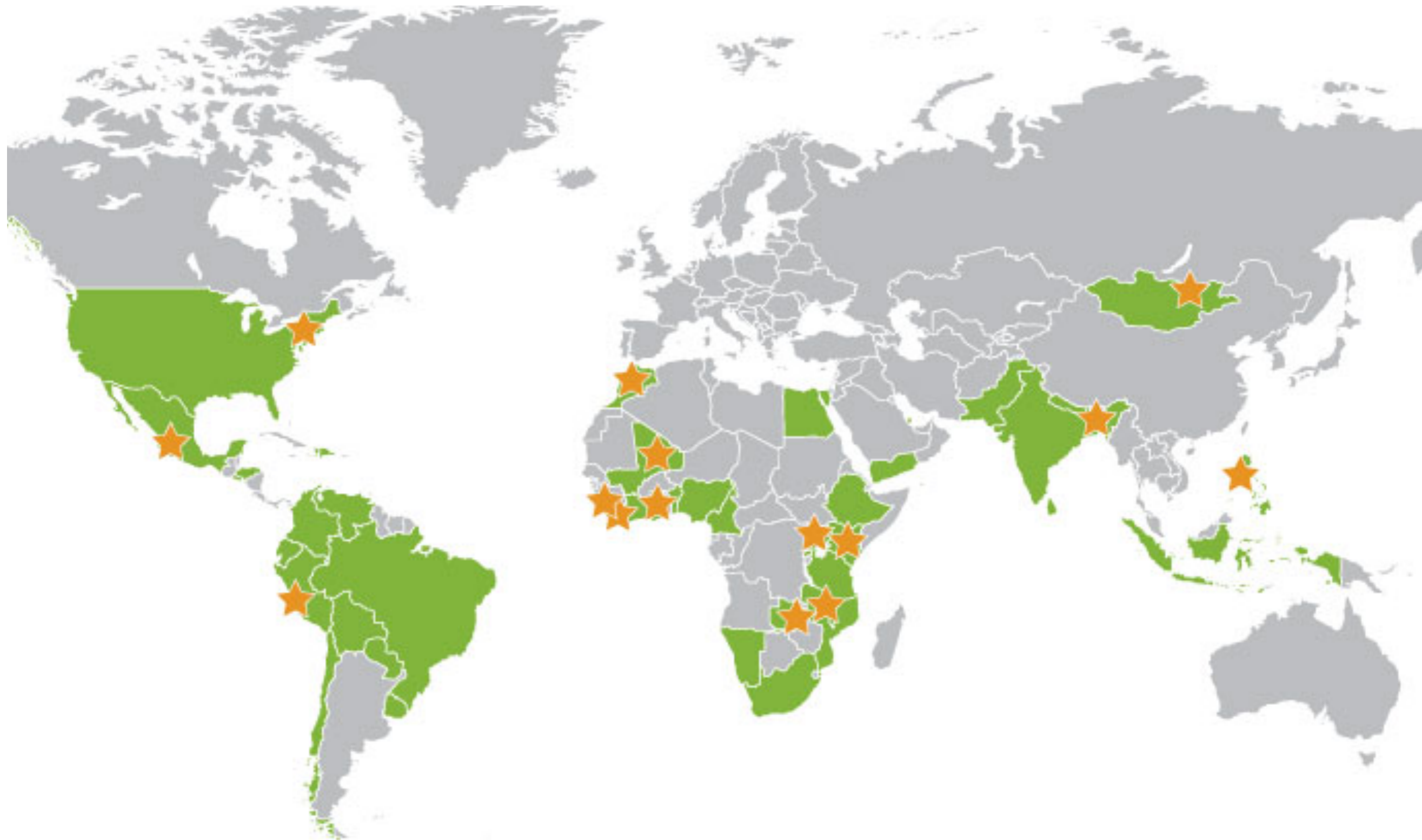


# IPA Stats

- 200 completed projects
- 235 active projects
- 250+ leading academics
- 400+ partner organizations

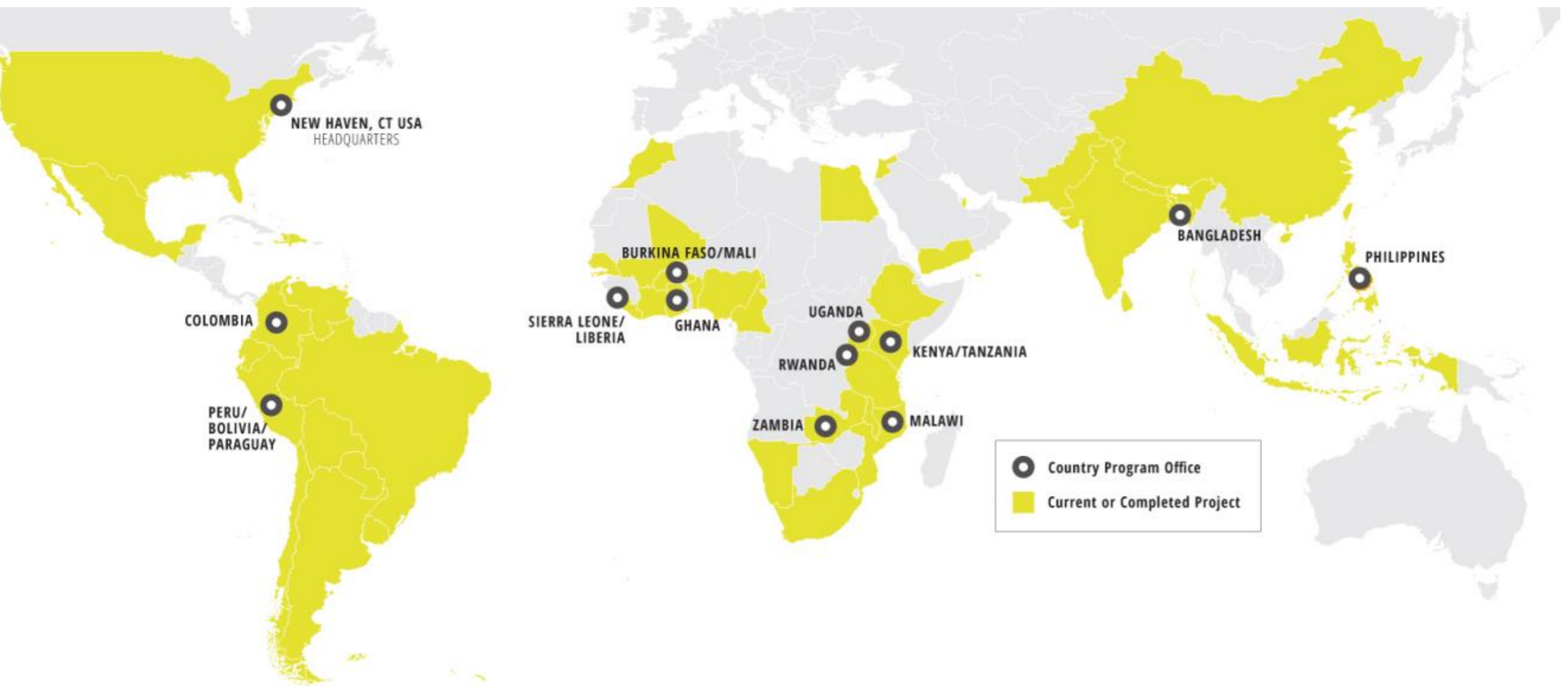


# IPA EVALUATES AND COMMUNICATES WHAT WORKS





# Where We Work



51 countries

12 country offices supporting 17 countries

# WE WORK ACROSS SECTORS



**Financial services**



**Health**



**Agriculture**



**Education**



**Governance & Democracy**



**Small & Medium  
Enterprises**

# IPA Philippines

We have partnered with:

- Department of Labor and Employment
- Department of Agrarian Reform
- Department of Social Welfare and Development
- Development Bank of the Philippines
- Philippine Crop Insurance Corporation
- Bank of the Philippine Islands
- University of the Philippines School of Economics
- Philippine Institute for Development Studies
- Asian Institute of Management
- And many NGOs: (NWTF, ASKI, PALFSI, ICM)

# About IPA and J-PAL

- IPA and J-PAL are complementary organizations that work together towards the common goal of reducing poverty by ensuring that policy is based on scientific evidence
  - **IPA** is an international non-profit research organization that has a strong presence in 14 countries through its country programs.
  - **J-PAL** is a network of 88 affiliated professors working through seven research centers based at leading universities around the world.

# About IPA and J-PAL

- The two organizations work on three core activities, dividing up the work based on their respective strengths and local presence:
  - **Research:** IPA and J-PAL conduct randomized evaluations of social programs in countries where their offices are located
  - **Capacity Building:** J-PAL builds policymakers' capacity to understand and conduct randomized evaluations through its Executive Education course. IPA specializes in training our research staff to implement randomized evaluations.
  - **Policy Outreach:** J-PAL analyzes the results of evaluations to draw out practical implications for social policy and engages policymakers at the global and regional level. Capitalizing on J-PAL's policy analyses, IPA uses its country programs to build relationships with local policymakers and practitioners and to inform their

Salamat po and now, with the  
introductions...

# Executive Education Course: Evaluating Social Programs

Manila, November 2015



ABDUL LATIF JAMEEL  
Poverty Action Lab

J-PAL SOUTHEAST ASIA

# Course Structure

1. Lectures
2. Case Studies
3. Exercises
4. Group Work

	Monday November 23	Tuesday November 24	Wednesday November 25	Thursday November 26	Friday November 27
8:00 – 8:30	Registration				
8:30 – 9:00	Welcoming Remarks 8:30-10:00	Lecture 3: Why Randomize?	Lecture 5: Sampling and Sample Size	Lecture 6: Threats and Analysis	Lecture 8: Cost effectiveness Analysis
9:00 – 10:30	Participation Survey Lecture 1: What is Evaluation				
10:30 – 10:45	Coffee Break	Coffee Break	Coffee Break	Coffee Break	Coffee Break
10:45 – 12:15	Case Study 1: Theory of Change: Reforming School Monitoring  Decision on group projects	Case study 2: Why Randomize: Vocational Training	Group Exercise : Random Sampling & Randomization Mechanics	Case Study 4: Threats and Analysis: Deworming in Kenya	Feedback survey Group work on Presentation: Finalize presentations
12:30 – 1:30	Lunch 12:00 – 1:00	Lunch	Lunch	Lunch	Lunch
1:30 – 3:00	Lecture 2: Outcomes, Indicators, and Measuring Impact	Lecture 4: How to Randomize	Group Exercise C: Sample Size Estimation	Lecture 7: Randomized Evaluation: Start-to- Finish	Group presentations
3:00 – 3:30	Coffee Break	Coffee Break		Coffee Break	Coffee Break
3:30 – 5:00	Group work on Presentation: Theory of Change, research question	Case Study 3: How to Randomize Combatting Corruption  Group work on Presentation: Randomization Design	Coffee Break  Group work on Presentation: Power and sample size	Group work on Presentation (Threats and analysis)	Group presentations Closing remarks  Distribution of certificates
5:00 – 8:00	Cocktail hour				



# Course Cast

## Lecturers

- Nassreena Sampaco-Baddiri (IPA-Philippines)
- Dr. Ryoko Sato (NUS)
- Faith McCollister (IPA-New York)
- Hector Salazar Salame (J-PAL SEA)
- Dr. Aniceto Orbeta (UP)
- Lina Marliani (J-PAL SEA)

## TAs

1. Sharanya Chandran (J-PAL SA)
2. Peter Srouji (IPA-Philippines)
3. Agnese Carrera (IPA-Philippines)
4. Lina Marliani (J-PAL SEA)
5. Neil Mirochnick (IPA-Philippines)

# Course Lecture Overview

1. What is evaluation? (Nassreena Sampaco-Baddiri)
2. Measurement (Dr. Ryoko Sato)
3. Why randomize? (Faith McCollister)
4. How to randomize? (Hector Salazar Salame)
5. Sampling and sample size (Dr. Aniceto Orbeta)
6. Threats and Analysis (Lina Marliani)
7. Cost Effectiveness Analysis (Dr. Aniceto Orbeta)
8. RCT: Start to Finish (Faith McCollister)

# Course Handbook

## Table of Contents

Table of Contents	3
Agenda	4
Case Study 1 Reforming School Monitoring – Measuring Impact of a School Monitoring Program <i>Thinking about measurement and outcomes</i>	6
Case Study 2 Why Vocational Training for Disadvantaged Youth? <i>Assessing Different Impact Evaluation Methods</i>	13
Exercise A Understanding random sampling and the law of large numbers	21
Exercise B How to do Random Assignment using MS Excel	22
Case Study 3 Reducing Inefficiencies in Road Construction <i>How to Randomize?</i>	31
Exercise C How to do Power Calculations in Optimal Design Software	37
Case Study 4 Deworming in Kenya <i>Addressing Threats to Experimental Integrity</i>	54
Appendix A Checklist For Reviewing a Randomized Controlled Trial of a Social Program or Project, To Assess Whether It Produced Valid Evidence	62
Appendix B Evaluation Glossary	69

# Course Clickers

▪

- Everyone gets one
- Collect after last lecture

# Course "Clickers"

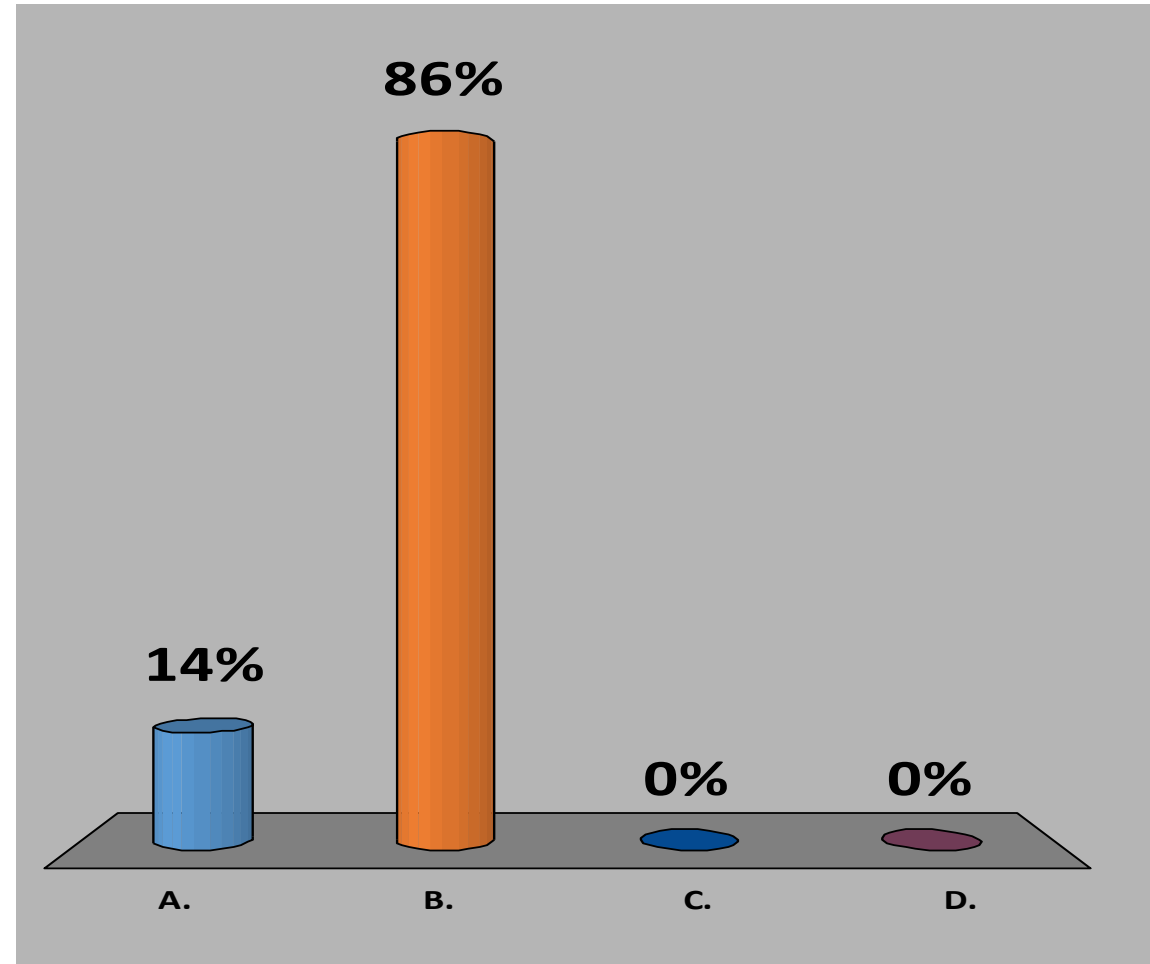
## Have you used this before?

A. Yes

B. No

C. Something similar

D. Something different



# Things to Consider

- Each day will begin promptly at 8:30am
- Registration will occur twice a day
- Please keep cell phones on silent mode
- Laptops needed during group exercises
- Buzzword: “tayo”
- Baseline quiz

# What is Evaluation?

Nassreena Sampaco-Baddiri

*Country Director, IPA Philippines*

Manila, November 2015



ABDUL LATIF JAMEEL  
Poverty Action Lab

J-PAL SOUTHEAST ASIA

# Course Overview

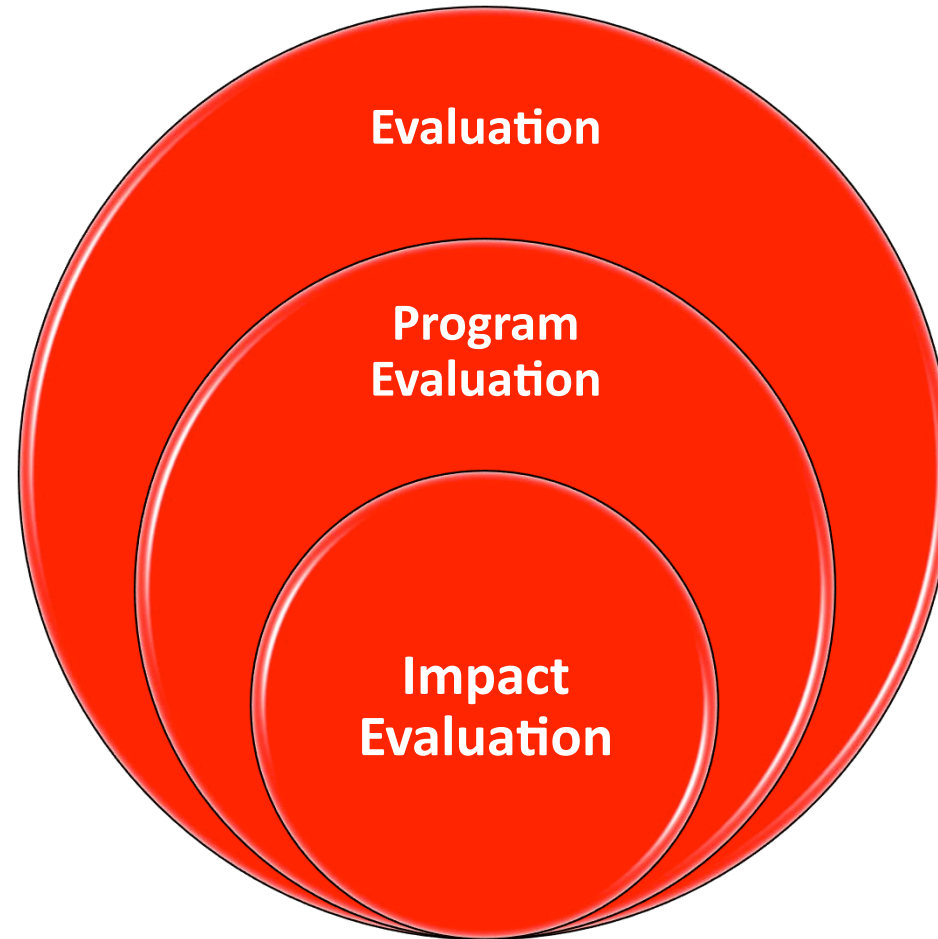
1. What is evaluation?
2. Measuring impacts (outcomes, indicators)
3. Why randomize?
4. How to randomize?
5. Sampling and sample size
6. Threats and Analysis
7. RCT: Start to Finish
8. Cost Effectiveness Analysis and Scaling Up



# Course Overview

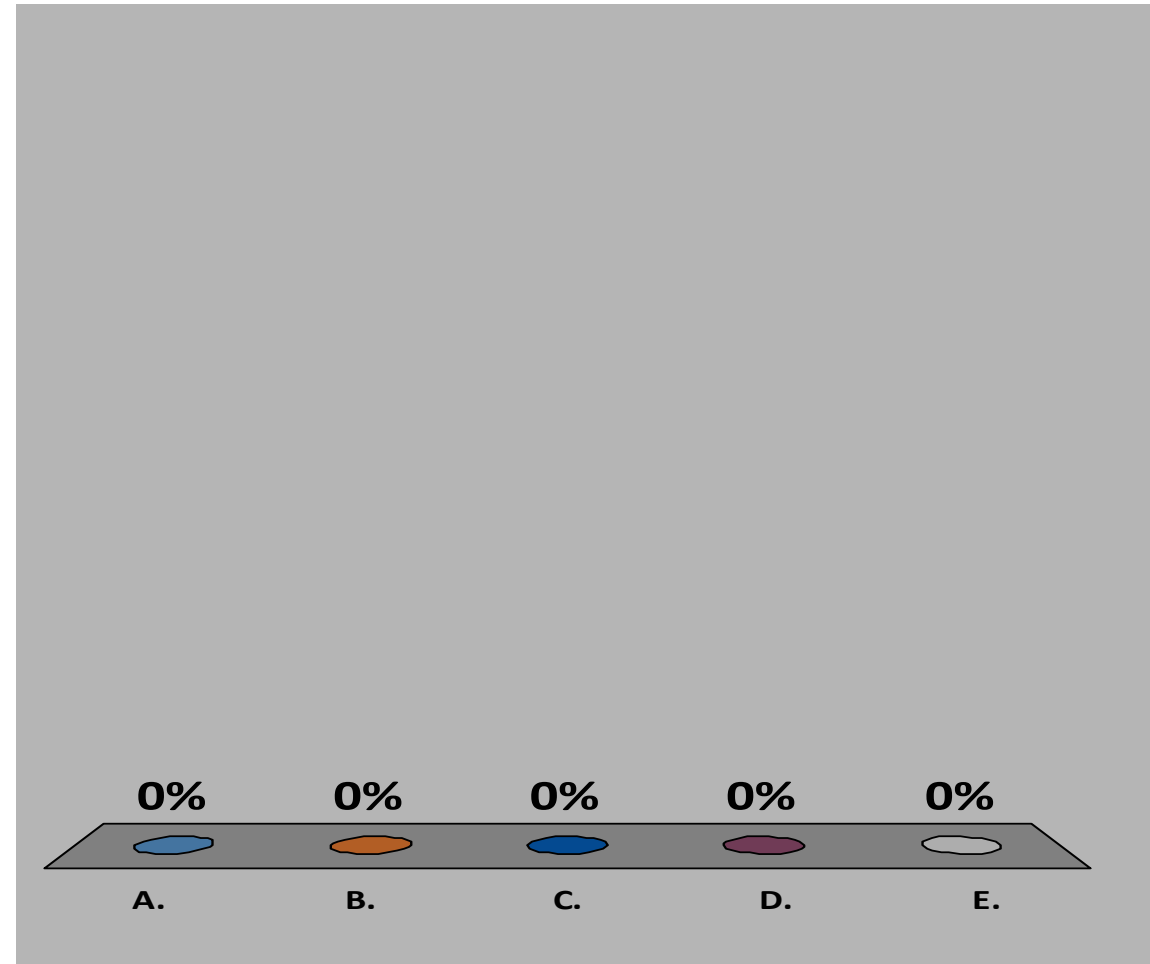
1. **What is evaluation?**
2. Measuring impacts (outcomes, indicators)
3. Why randomize?
4. How to randomize?
5. Sampling and sample size
6. Threats and Analysis
7. RCT: Start to Finish
8. Cost Effectiveness Analysis and Scaling Up

# What is Evaluation?

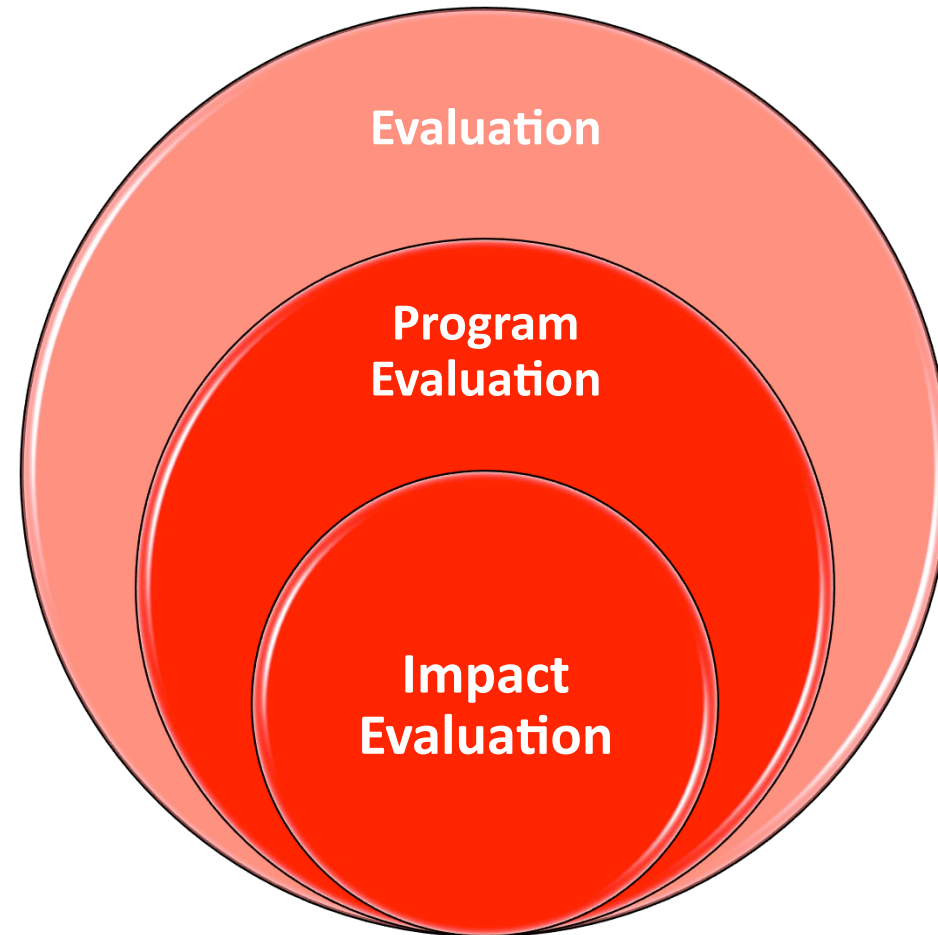


# *What's the difference between: Monitoring and Evaluation?*

- A. Nothing. They are different words to describe the same activity
- B. Monitoring is conducted internally, Evaluation is conducted externally
- C. Monitoring is for management, Evaluation is for accountability
- D. Don't know
- E. Other



# Program Evaluation

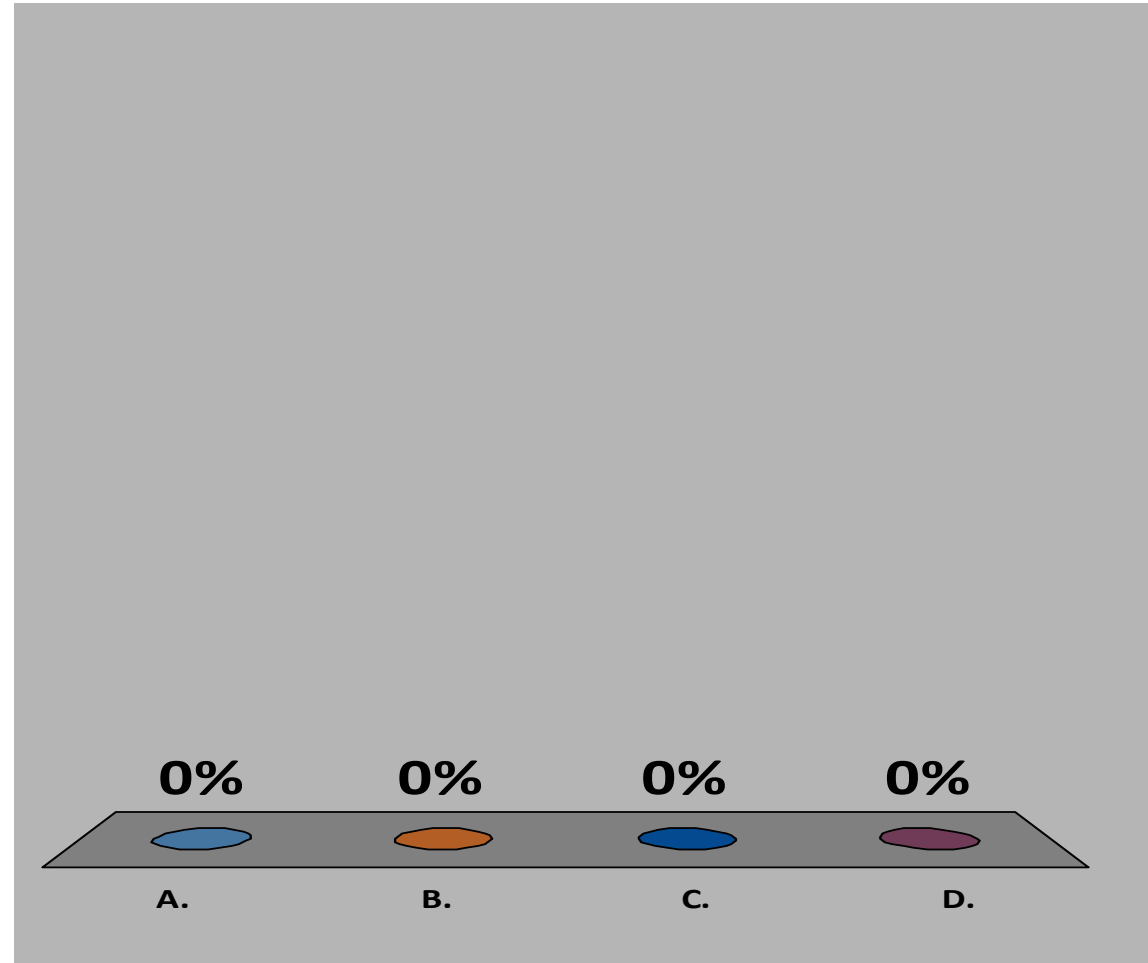


# Components of Program Evaluation

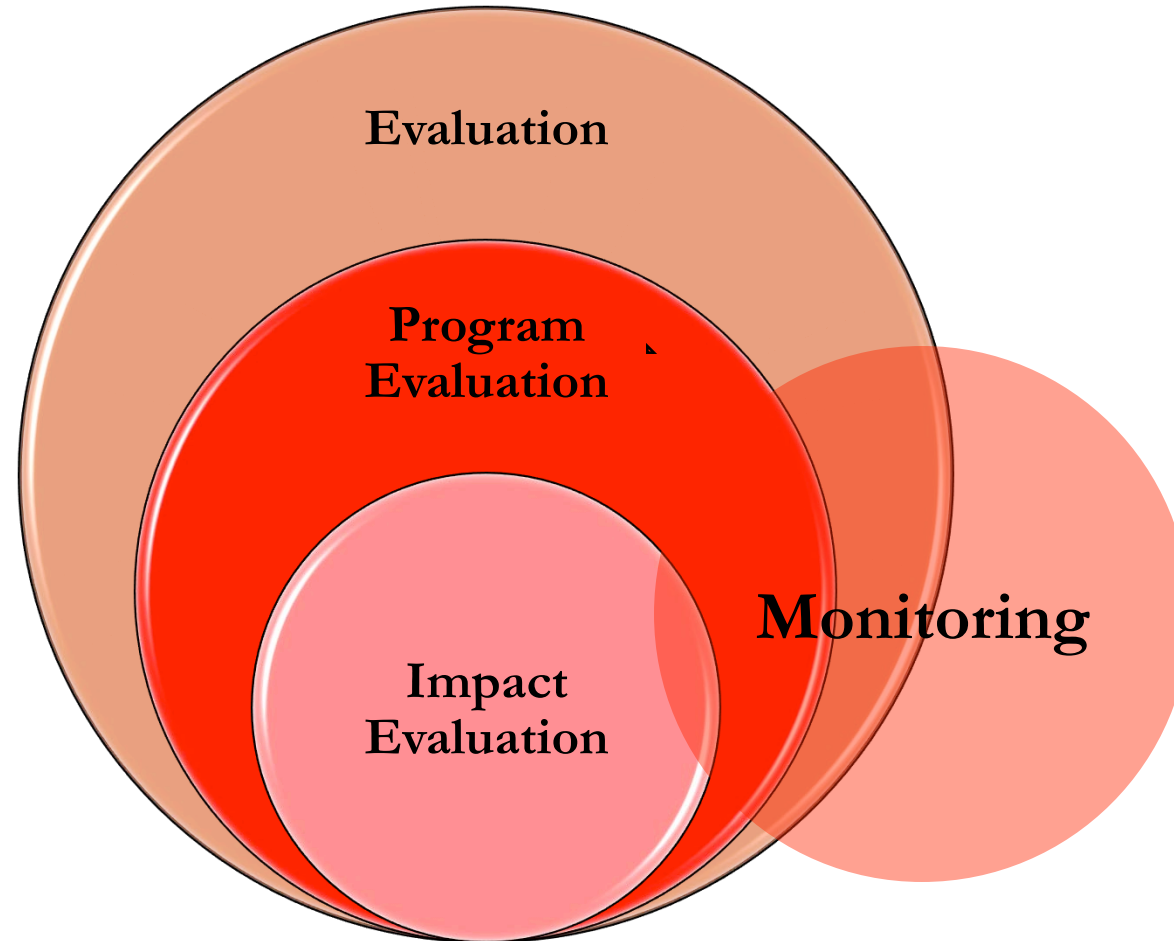
- Needs Assessment
  - What is the problem?
- Program Theory Assessment
  - How, in theory, does the program fix the problem?
- Process Evaluation
  - Does the program work as planned?
- Impact Evaluation
  - Were its goals achieved?  
The magnitude?
- Cost Effectiveness
  - Given magnitude and cost, how does it compare to alternatives?

# Evaluations should usually be conducted:

- A. Externally and independent from the implementers of the program being evaluated
- B. Externally and closely integrated with program implementers
- C. Internally
- D. Don't know



# Monitoring and Evaluation



# What Makes a Good Program Evaluation?

- Ask the right questions
  - For accountability
  - For organizational learning
- Answers “the right” questions in unbiased and definitive way
- Establish a model of how you think the program works, which includes assumptions about why it works the way it does
  - Includes making clear the need being addressed (i.e., “Needs Assessment”) & charting out the “program theory” (e.g., theory of change)

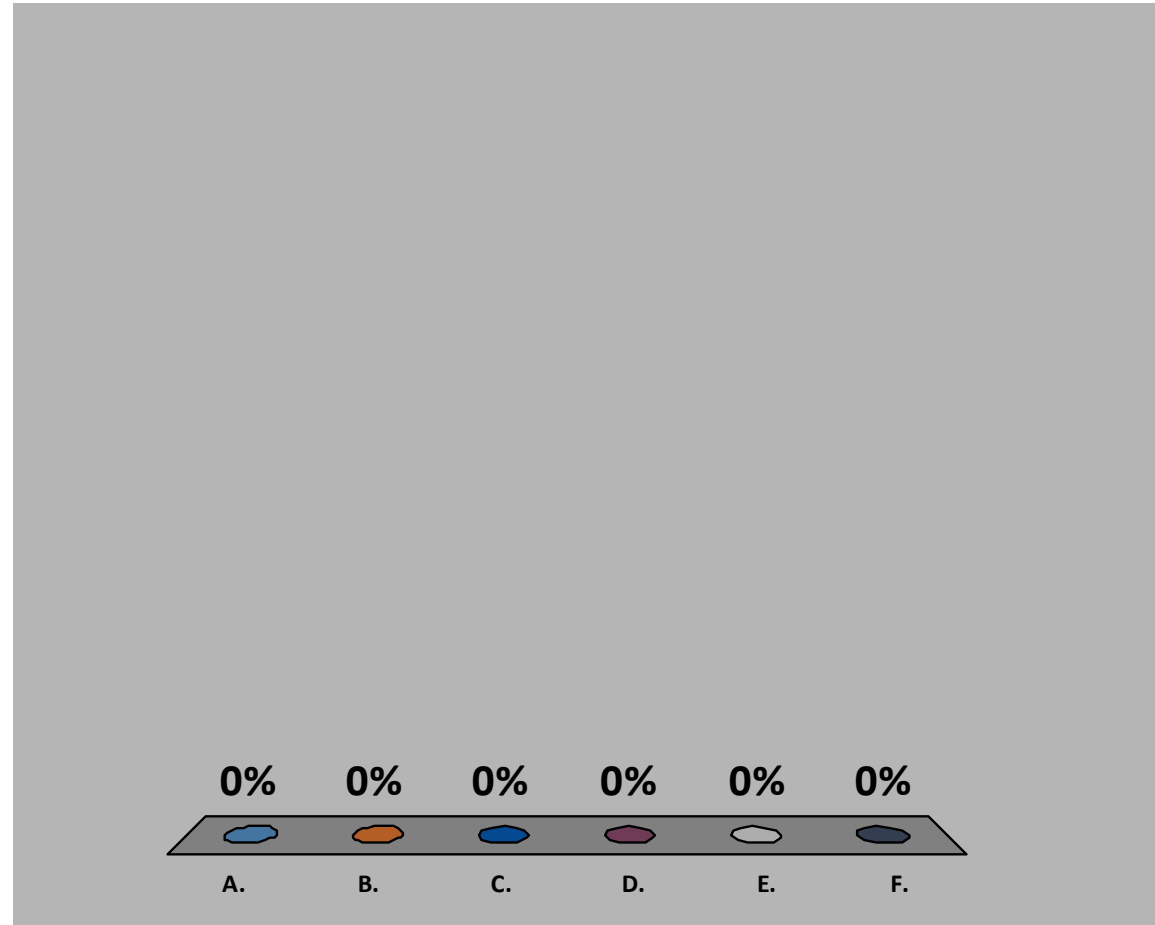


# Who is this evaluation for?

- Academics
- Donors
  - Their Constituents
- Politicians / policymakers
- Technocrats
- Implementers
- Proponents, Skeptics
- Beneficiaries

# Who is your *most important audience* for evaluation?

- A. Agency leadership
- B. Donor Politicians / policymakers
- C. Donor Constituents
- D. Academics
- E. Country Politicians / policymakers
- F. Technocrats
- G. Implementers
- H. Proponents, Skeptics
- I. Beneficiaries



# How can impact evaluations help us?

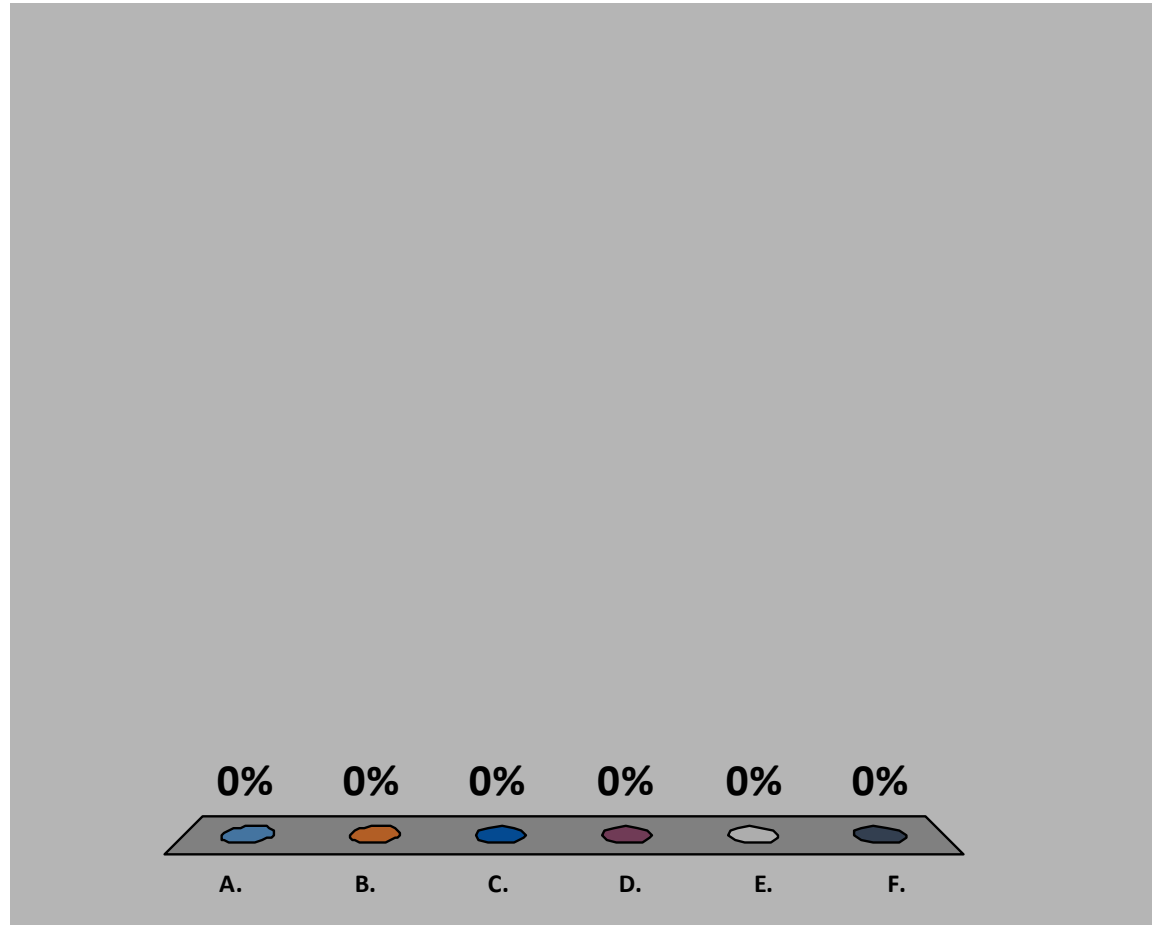
- Surprisingly little hard evidence on what works
- Can do more with given budget with better evidence
- If people knew money was going to programs that worked, could help increase pot for anti-poverty programs
- Instead of asking “do aid/development programs work?” should be asking:
  - Which programs work best, why and when?
  - Which concepts work, why and when?
  - How can we scale up what works?
- Add to our body of evidence
  - part of a well-thought out evaluation (research) strategy

Case Study in Kenya

# Methods to Improve Water Quality

# What do you think is the most cost-effective way to reduce diarrhea?

- A. Develop piped water infrastructure
- B. Improve existing water sources
- C. Increase supply of and demand for chlorine
- D. Education on sanitation and health
- E. Improved cooking stoves for boiling water
- F. Improve sanitation infrastructure



# 1. NEEDS ASSESSMENT

Identifying the problem

# The Need

- Nearly 2 million children die each year from diarrhea
- 20% all child deaths (under 5 years old) are from diarrhea

# The Likely Problem

- 13% of world population do not have access to “good quality water source”
- Lack of access to clean water
  - People’s reported value for clean water translates to willingness to pay nearly \$1 per averted diarrhea episode, \$24 per DALY (Kremer et al 2009)



# The Goal

- MDG: “reduce by half the proportion of people without access to sustainable drinking water”



# The Solution(s)



# Really the Problem?

- *Quantity* of water is a better determinant of health than *quality* of water (Curtis et al, 2000)
- Water quality helps little without hygiene (Esrey, 1996)
  - 42% live without a toilet at home
- Nearly 2.6 billion people lack any improved sanitation facilities ([WHO](#))
- People are more willing to pay for convenient water than clean water
- Chlorine is very cheap,
  - In Zambia, \$0.18 per month for a family of six
  - In Kenya, \$0.30 per month
- Yet less than 10% of households purchase treatment
- 25% of households reported boiling their drinking water the prior day

# Alternative Solution(s)?



# Devising a Solution

- What is the theory behind your solution?
- How does that map to your theory of the problem?

## 2. PROGRAM THEORY ASSESSMENT

Blueprint for Change

# Program Theory Assessment

- Theory of Change
- Logical Framework (LogFrame, LFA)



# Theory of Change

**WATER SOURCE  
MODIFICATION (to prevent  
contamination)**

*Assumption:* water source is preserved well and there must be water supply

**CLEAN WATER PRESERVED AT  
HOME**

*Assumption:* water is kept inside a consistently-clean container.

**FAMILY MEMBERS  
CONSUME CLEAN WATER**

*Assumption:* all people in the village consume water from a clean water source, not from a contaminated one

**HEALTH OUTCOMES INCREASE (i.e.  
decrease in number of diarrhea  
incidence)**

*Assumption:* the new behavior as a result of the intervention is sustained over time

An assumption is an enabling factor, it facilitates or enables the linkages between the various components.

# Program Theory Assessment

- How will the program address the needs put forth in your needs assessment?
  - What are the prerequisites to meet the needs?
  - How and why are those requirements currently lacking or failing?
  - How does the program intend to target or circumvent shortcomings?
  - What services will be offered?

# Logical Framework

	Objectives Hierarchy	Indicators	Sources of Verification	Assumptions / Threats
<b>Impact (Goal/ Overall objective)</b>	Lower rates of diarrhea	Rates of diarrhea	Household survey	Waterborne disease is primary cause of diarrhea
<b>Outcome (Project Objective)</b>	Households drink cleaner water	( $\Delta$ in) drinking water source; E. coli CFU/ 100ml	Household survey, water quality test at home storage	Shift away from dirty sources. No recontamination
<b>Outputs</b>	Source water is cleaner; Families collect cleaner water	E. coli CFU/ 100ml;	Water quality test at source	continued maintenance, knowledge of maintenance practices
<b>Inputs (Activities)</b>	Source protection is built	Protection is present, functional	Source visits/ surveys	Sufficient materials, funding, manpower

**Needs assessment**



**Impact evaluation**



**Process evaluation**



# 3. PROCESS EVALUATION

Making the program work

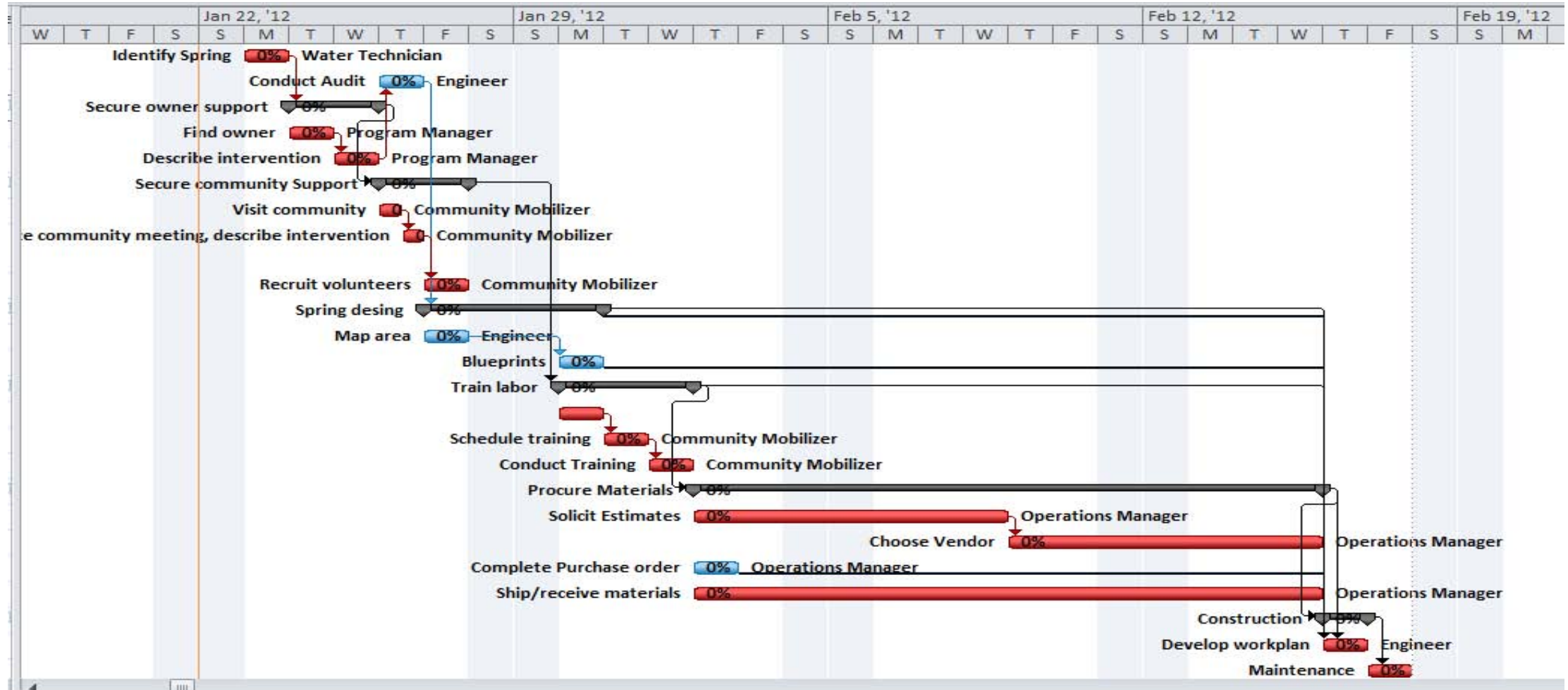
# Process Evaluation

- Supply Side
  - Logistics
  - Management
- Demand Side
  - Assumption of knowledge, preferences
  - Assumptions of response

# Process Evaluation: Logistics

- Construction
  - Construct spring protection
  - Install fencing
  - Install drainage
- Maintenance
  - Patch concrete
  - Clean catchment area
  - Clear drainage ditches

# Process Evaluation: Supply Logistics



# Process Evaluation: Demand-side

- Do households collect water from improved source?
- Does storage become re-contaminated?
- Do people drink from “clean” water?



# Process Evaluation

- Are basic tasks being completed?
- Are the services being delivered?
- Is the intervention reaching the target population?
- Is the intervention being completed well or efficiently and to the beneficiaries' satisfaction?

# Process vs. Impact Evaluation



- Process Evaluation / Monitoring
  - Accountability
  - Did we do what we said we would do?
- Impact Assessment
  - Which program/policies work and which don't?
  - Which was most effective?
  - How did a successful program change behavior?

# Process was okay, so....

- What happened to diarrhea?

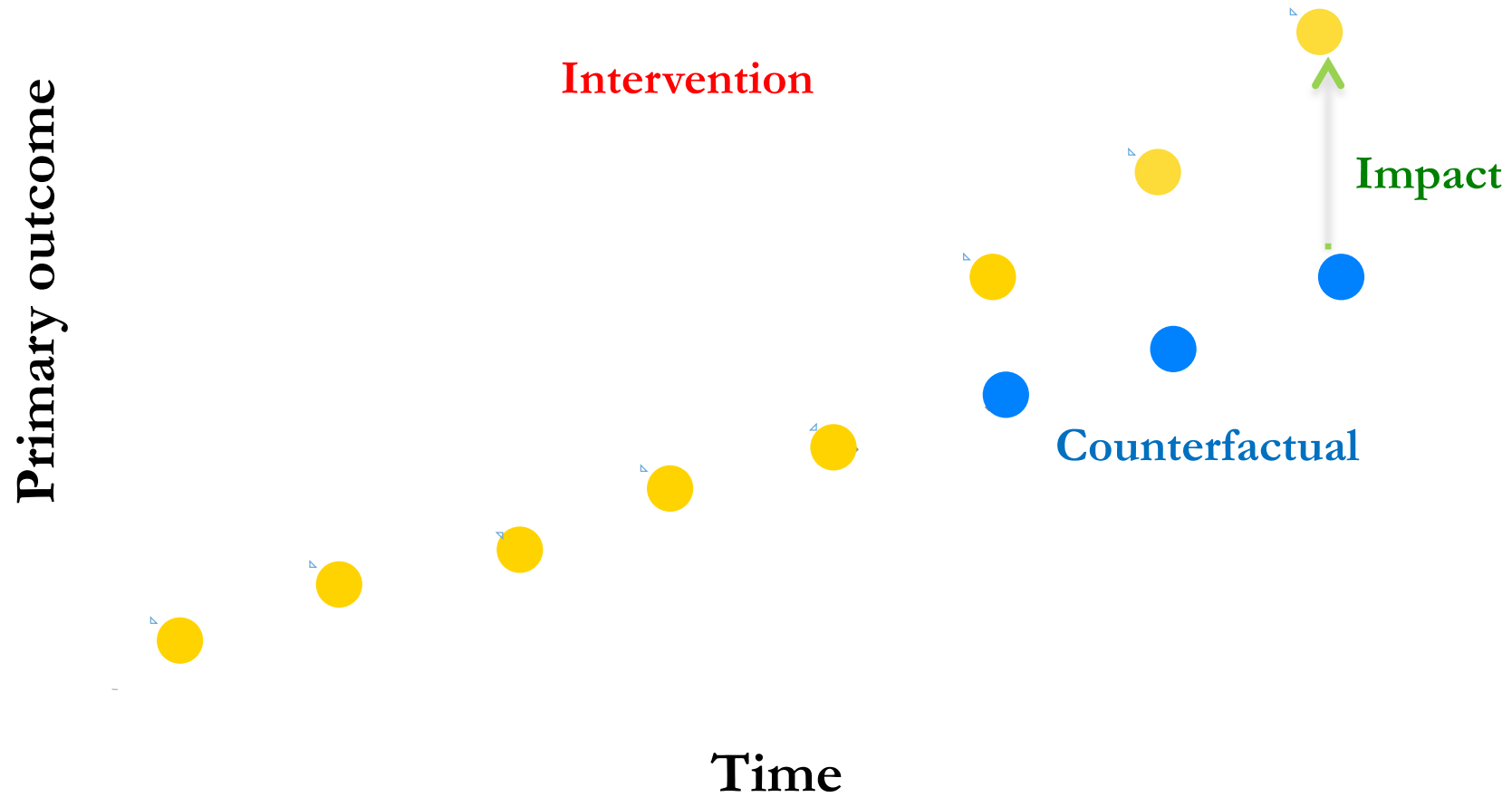
## 4. IMPACT EVALUATION

Measuring how well it worked

# Did we achieve our goals?

- Primary outcome (impact): did spring protection reduce diarrhea?
- Also distributional questions: what was the impact for households with good v. bad sanitation practices?

# What is impact?



# How to measure impact?

- What would have happened in the absence of the program?
- Take the difference between
  - what happened (with the program) ...and
  - what would have happened (without the program)
  - = IMPACT of the program

# Constructing the Counterfactual

- Counterfactual is often constructed by selecting a group not affected by the program
- Randomized:
  - Use random assignment of the program to create a control group which mimics the counterfactual.
- Non-randomized:
  - Argue that a certain excluded group mimics the counterfactual.



# How impact differs from process?

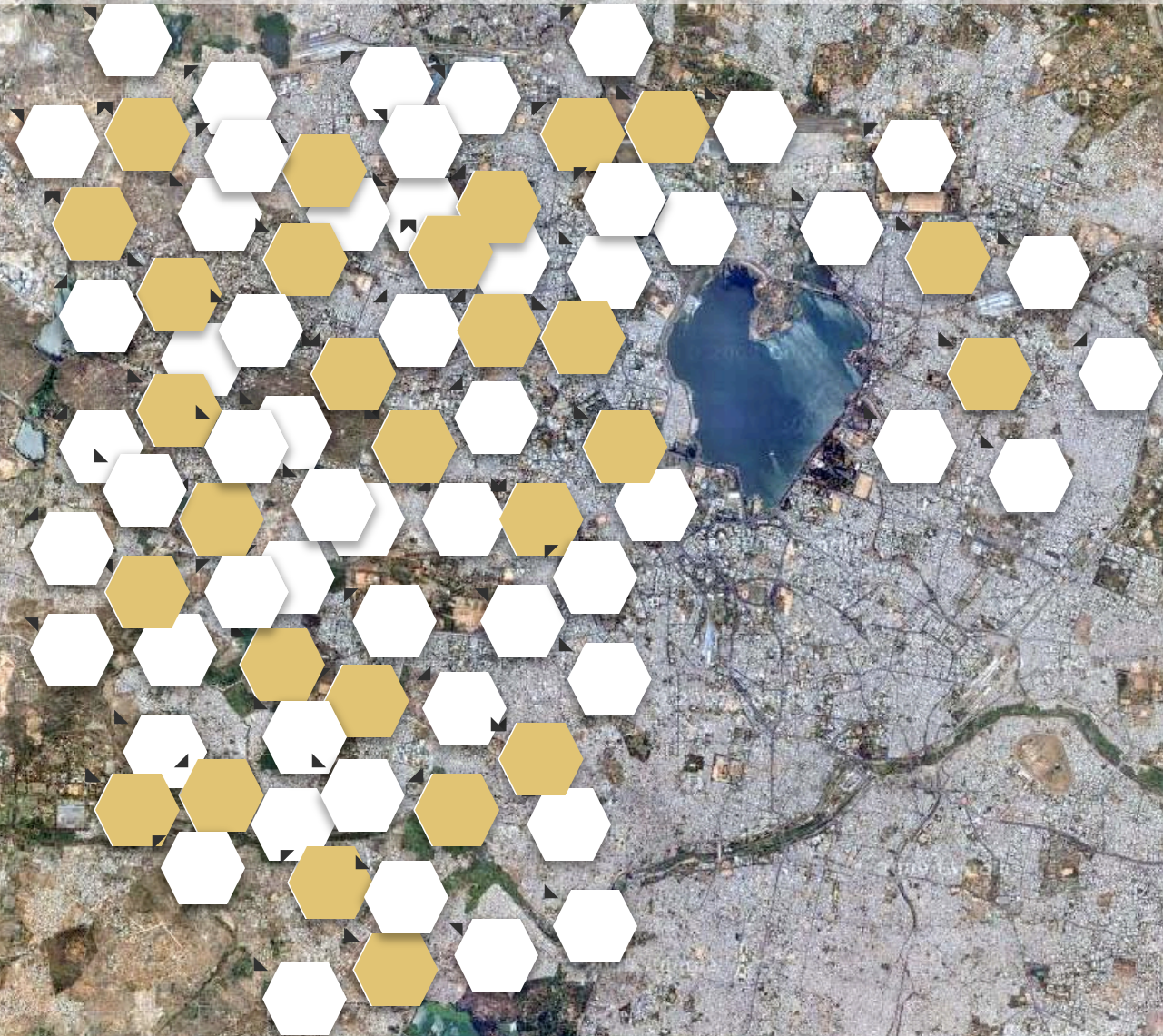
- When we answer a process question, we need to describe what happened.
- When we answer an impact question, we need to compare what happened to what would have happened without the program

# 5. RANDOMIZED EVALUATIONS

The “gold standard” for Impact Evaluation

# Random Sampling and Random Assignment

Randomly  
*sample*  
from area of  
interest

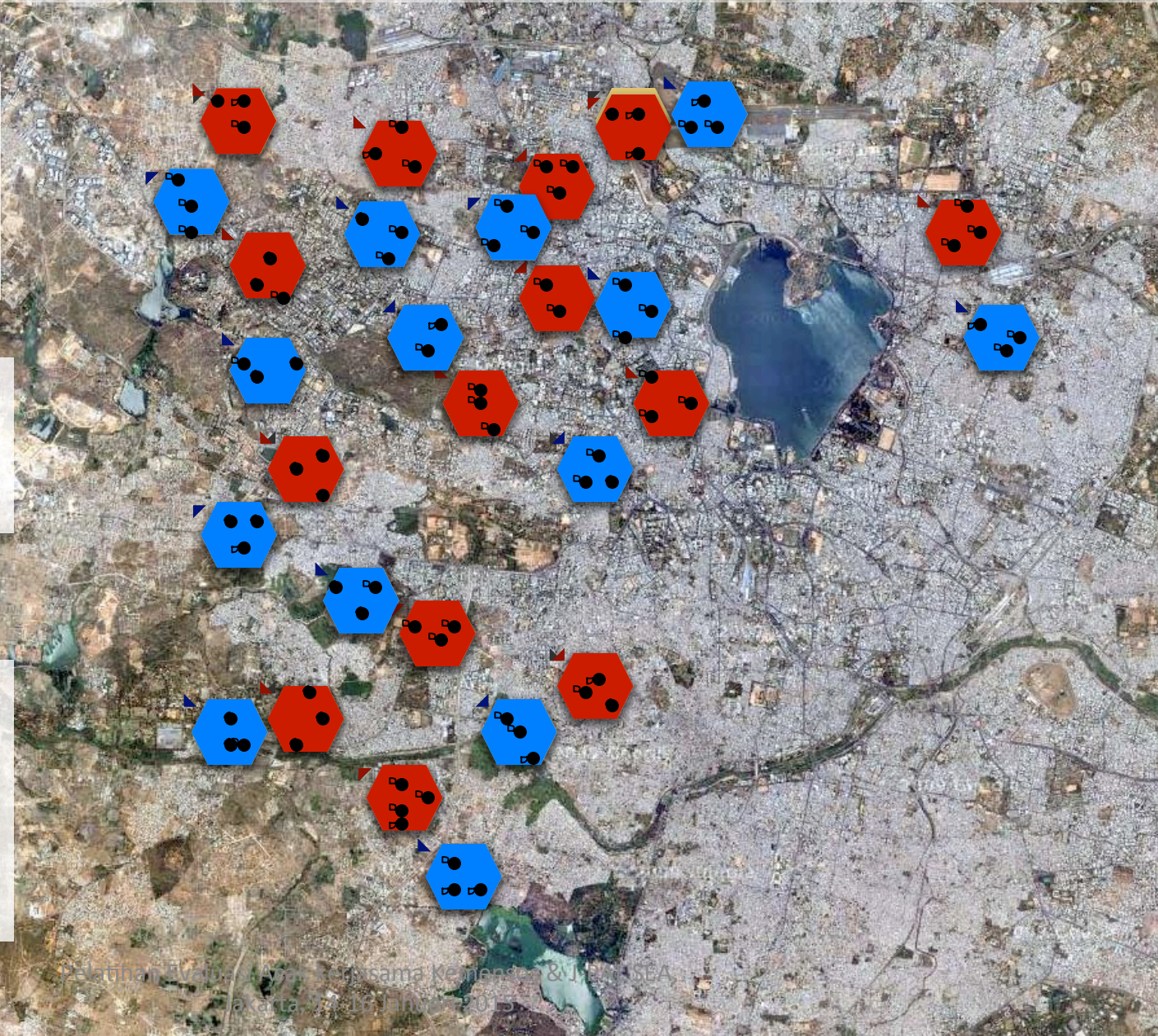


# Random Sampling and Random Assignment

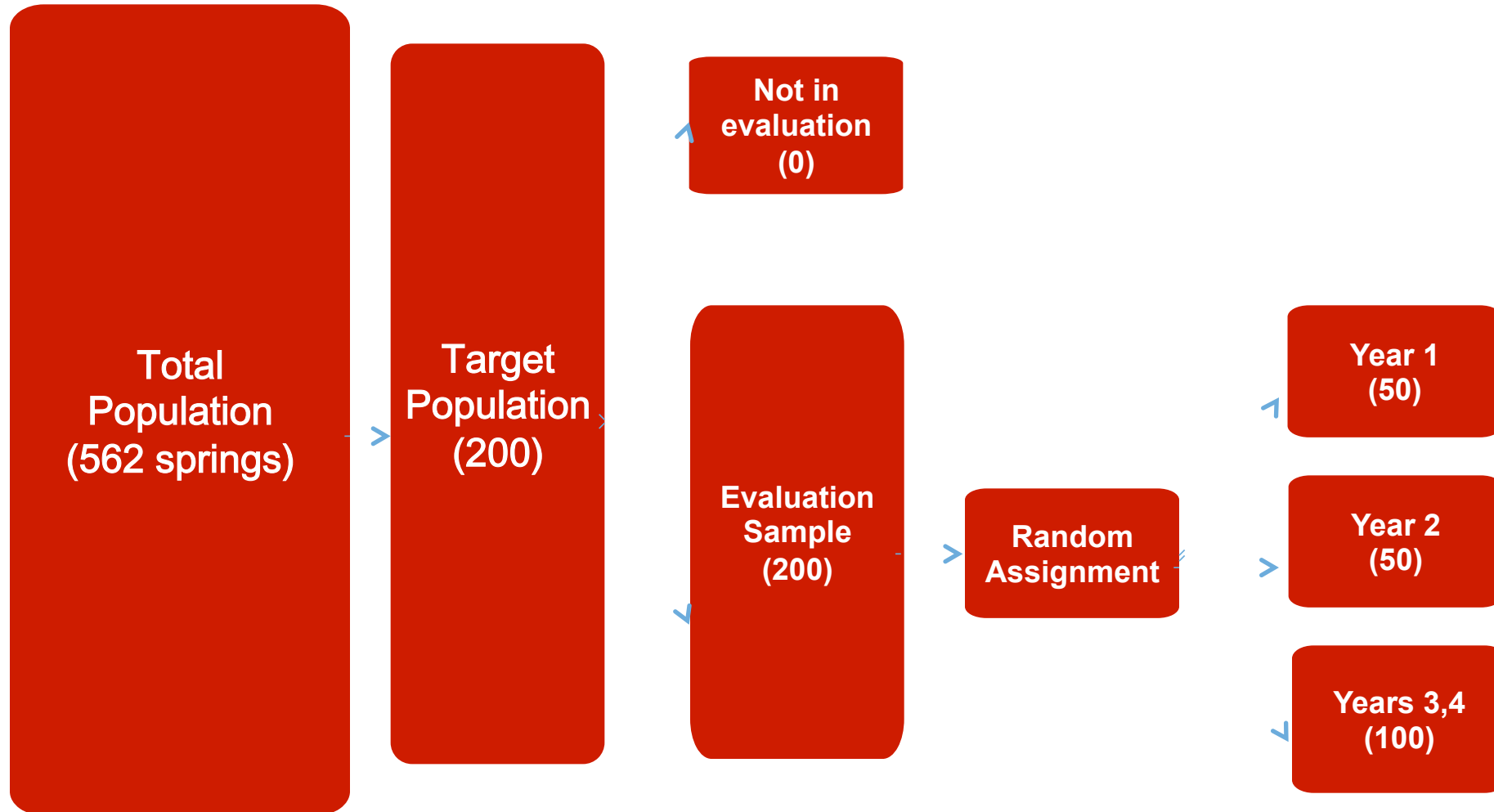
Randomly *sample* from area of interest

Randomly *assign* to **treatment** and **control**

Randomly *sample* from both treatment and control



# Spring Cleaning Sample



# Impact

- 66% reduction in source water e coli concentration
- 24% reduction in household E coli concentration
- 25% reduction in incidence of diarrhea

# Making Policy from Evidence

Intervention	Impact on Diarrhea
Spring protection (Kenya)	25% reduction in diarrhea incidence for ages 0-3

# Making Policy from Evidence

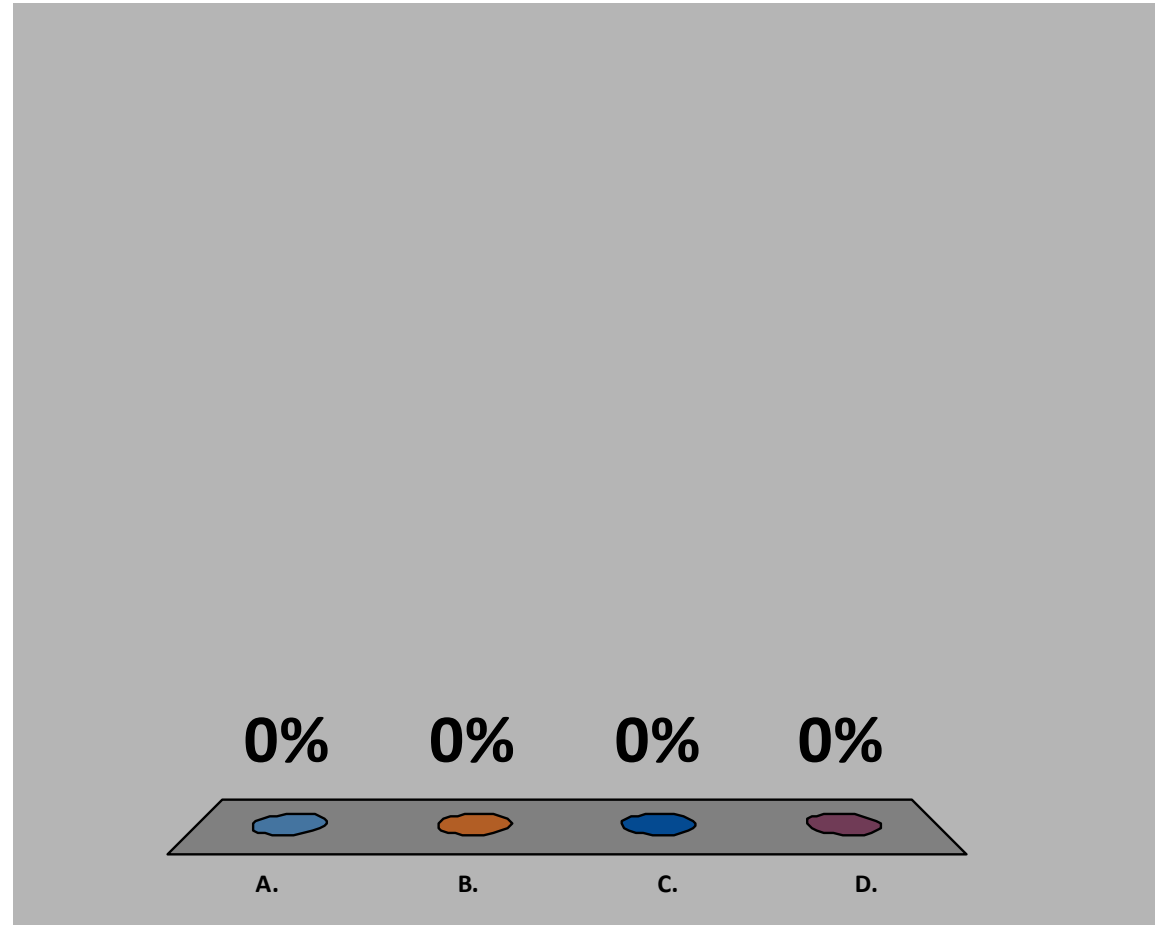
Intervention	Impact on Diarrhea
Spring protection (Kenya)	25% reduction in diarrhea incidence for ages 0-3
Source chlorine dispensers (Kenya)	20-40% reduction in diarrhea
Home chlorine distribution (Kenya)	20-40% reduction in diarrhea
Hand-washing (Pakistan)	53% drop in diarrhea incidence for children under 15 years old
Piped water in (Urban Morocco)	0.27 fewer days of diarrhea per child per week



When to do a randomized evaluation?

# When is a good time to do a randomized evaluation?

- A. After the program has begun and you are not expanding it elsewhere
- B. When a positive impact has been proven using rigorous methodology
- C. When you are rolling out a program with the intension of taking it to scale
- D. When a program is on a very small scale e.g one village with treatment and one without



# When to do a randomized evaluation?

- When there is an important question you want/need to know the answer to
- Timing--not too early and not too late
- Program is representative not gold plated
  - Or tests a basic concept you need tested
- Time, expertise, and money to do it right
- Develop an evaluation plan to prioritize

# When NOT to do an RE

- When the program is premature and still requires considerable “tinkering” to work well
- When the project is on too small a scale to randomize into two “representative groups”
- If a positive impact has been proven using rigorous methodology and resources are sufficient to cover everyone
- After the program has already begun and you are not expanding elsewhere

# In what circumstances can the RE fail?

- No one asks the question that is answered by the study
- RE measures wrong results
- Too many unanswered important questions
- RE generates a biased result

# Program and Evaluation: Where and how to start?

## Intervention

- Start with a problem
- Verify whether the problem really takes place
- Create a theory to explain why such a problem exists
- Design the program
- Determine whether the solution is cost-effective

## Program Evaluation

- Start with a question
- Verify the question hasn't been answered
- State a hypothesis
- Design the evaluation
- Determine whether the value of the answer is worth the cost of the evaluation

# Components of Program Evaluation

- Needs Assessment
  - Program Theory Assessment
  - Process Evaluation
  - Impact Evaluation
  - Cost Effectiveness
- What is the problem?
  - How, in theory, does the program fix the problem?
  - Does the program work as planned?
  - Were its goals achieved?  
The magnitude?
  - Given magnitude and cost, how does it compare to alternatives?

Questions?



Salamat po!

[nbaddiri@poverty-action.org](mailto:nbaddiri@poverty-action.org)

# Measurement

---

Ryoko Sato

National University of Singapore

Manila. November 2015



ABDUL LATIF JAMEEL  
Poverty Action Lab

J-PAL SOUTHEAST ASIA

# Course Overview

---

1. What is evaluation?
2. Measurement
3. Why randomize?
4. How to randomize?
5. Sampling and sample size
6. Threats and Analysis
7. Cost Effectiveness Analysis
8. RCT: Start to Finish

# Course Overview

---

1. What is evaluation?
- 2. Measurement**
3. Why randomize?
4. How to randomize?
5. Sampling and sample size
6. Threats and Analysis
7. Cost Effectiveness Analysis
8. RCT: Start to Finish

# Lecture Overview

---

- What to Measure
  - Identifying mechanisms of WHY it works or not
    - Case study review (Theory of Change)
- How to measure it (well)
  - Validity, Reliability
  - How to measure the immeasurable
  - Sources of data
  - Data collection

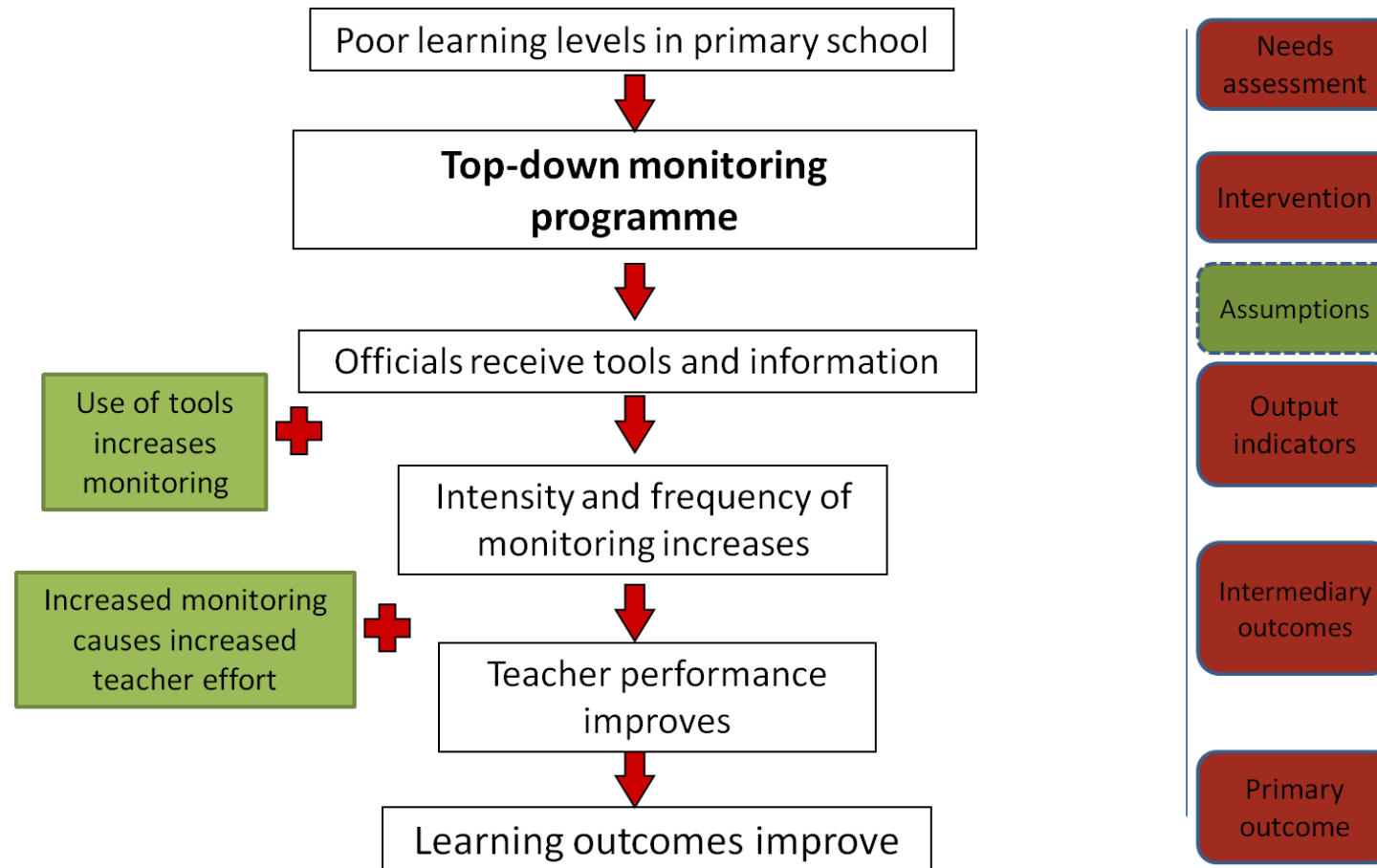
# Case Study

---

- Reforming School Monitoring in Madagascar

# Theory of Change

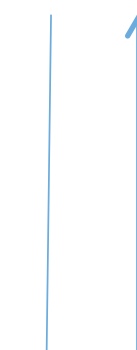
---



# Logical Framework

	Objectives Hierarchy	Indicators	Sources of Verification	Assumptions / Threats
<b>Impact (Goal/ Overall objective)</b>	Improving learning level in primary schools	Test scores	Report cards	Scores are appropriately measured and weighted.
<b>Outcome (Project Objective)</b>	Teacher performance improves	Attendance, lesson plan, frequency and quality of evaluation.	Administrative data: Education statistics	Schools might not be fully operative; poor management.
<b>Outputs</b>	Intensity and frequency of monitoring increases	Number of visits to school, allocation of time & budget	Administrative records	Schools are visited
<b>Inputs (Activities)</b>	Officials receive tools & information	Self-reported receipt and usage rates	Teacher log books	Tools were received

Needs assessment



Impact evaluation



Process evaluation





# Data used

---

Sources of Measurement	Indicators
Teacher log books	<ul style="list-style-type: none"><li>• Daily lesson plan checklist</li></ul>
School administrative data	<ul style="list-style-type: none"><li>• Student attendance rate</li><li>• Student repetition rate</li><li>• Student dropout rate</li></ul>
Enumerator unannounced visit	<ul style="list-style-type: none"><li>• Teacher absenteeism rate</li><li>• Teacher performing all tasks</li><li>• Directors performing all tasks</li></ul>
Report cards	<ul style="list-style-type: none"><li>• Student test scores</li></ul>
Country educational statistics	<ul style="list-style-type: none"><li>• Number of primary schools</li></ul>

# Tasks considered by Malagasy Educators to be Essential for Teachers and School Directors

---

Teachers	School Directors
Takes daily roll call	Signs off daily roll call
Prepares daily lesson plan	Follows up with teachers on lesson planning
Prepared bimonthly lesson plans	
Has tested students during the past two months	Review student test results
Discusses student learning issues with school director	Keeps a register of enrollments
	Informs subdistrict or district administrator of teacher absences

# Results

---

	Non-Treatment School	Treatment school	Difference
<b>Impact on service provider behaviors</b>			
Percentage of teachers performing all tasks	42	63	21**
Percentage of schools with teachers performing all tasks	24	43	19**
<b>Some impacts on Students' Schooling and Learning Outcomes:</b>			
Repetition rate	22.6	17.5	-5.1**
Dropout rate	6.1	5.5	-0.6
<b>Some Test Scores:</b>			
French	29.9	30.5	0.6
Malagasy	49.8	52.1	2.3

# Identifying why an intervention worked (or not)

---

- “Intervention does not improve students’ learning”
- Why?
  - Intervention successfully improved teachers’ teaching behavior
    - Maybe low quality of teacher?
  - Teachers’ behavior → Students’ learning?
- Ruling out mechanisms

# Identifying why an intervention worked (or not)

---

- “Conditional cash transfer program (CCT) to increase antenatal care visits did not reduce infant mortality rate”
- Why?
  - CCT did not increase the antenatal care visit?
  - Antenatal care is not good enough
    - Why?
      - Supply side problem?
- Mechanisms

# Identifying why an intervention worked (or not)

---

- PROGRESA
  - government social assistance program in Mexico 1997-
  - Conditional Cash Transfer Program
    - regular school attendance
    - health clinic visits
    - nutritional support
  - Decreased poverty
    - Why?
      - Mechanisms?

# Identifying why an intervention worked (or not)

- Why don't women receive a vaccine?
- Hypothesis: Because of high psychic costs of vaccination (fear of needles, fear of side effects)
- Intervention
  - T1: CCT if clinic visit
  - T2: CCT if clinic visit + vaccination
- Output measure: clinic attendance
  - T1: Costs of clinic attendance = transportation costs & opportunity costs
  - T2: Costs of clinic attendance = transportation costs & opportunity costs & **psychic costs of vaccine**

# Identifying why an intervention worked (or not)

---

- Results
  - No difference in clinic attendance between T1 and T2
- Why?
  - No psychic costs
  - Women did not understand the condition under which they can receive money?
    - SHOULD have asked to rule out this possibility...
- Minimize “Should have asked” measurement...



# Identifying mechanisms

---

- Output/outcome measurement
- Subjective measure (self-report)
  - Did you do....
- Objective measure
  - Actual visit to health clinic

# Measurement problem:

---

- Conditional Cash Transfer Program
  - Cash if antenatal care visit
  - Cash if receive a vaccine
- Outcome measure
  - Subjective(self-report)
    - Did you go to antenatal care?
      - Incentive to lie in the treatment group
  - Objective
    - Actual visit (record at clinic)

# Measurement problem:

---

- Difficult to measure in objective way
  - Diarrhea
  - Handwashing
  - Perception
- Self report: Incentive to tell lies?
  - Cash incentives to comply
  - Can we check with objective measure
    - Handwashing
      - Censor
    - Perception
      - Heart rate

# How to Measure it (Well)

---

- The basics

# Census Example

# The Basics

---

- Data that should be easy?
  - E.g. Age, # of rooms in house, # in hh
- What is the survey question identifying?
  - E.g. Are hh members people who are related to the household head? People who eat in the household? People who sleep in the household?
- Pre-test questions in local languages

# When the obvious is not so obvious...

---

- Let's think about the people who eat from the same pot in the household where you usually stay. There are how many adults, adolescents, and children? Adults are age 18 and older, adolescents are ages 13 to 17, and children are ages 12 and younger.
  - So in total how many people are there in the household where you usually stay? DON'T ADD TOTAL FOR RESPONDENT.

# Validity, Reliability

---

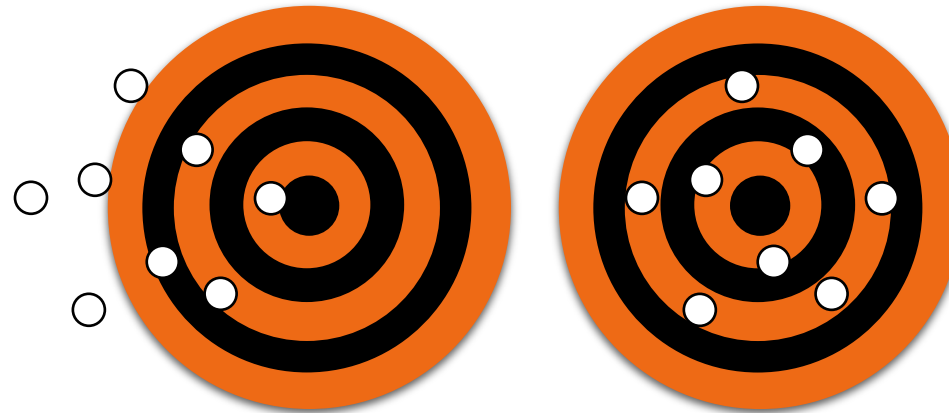
- How to measure it (well)



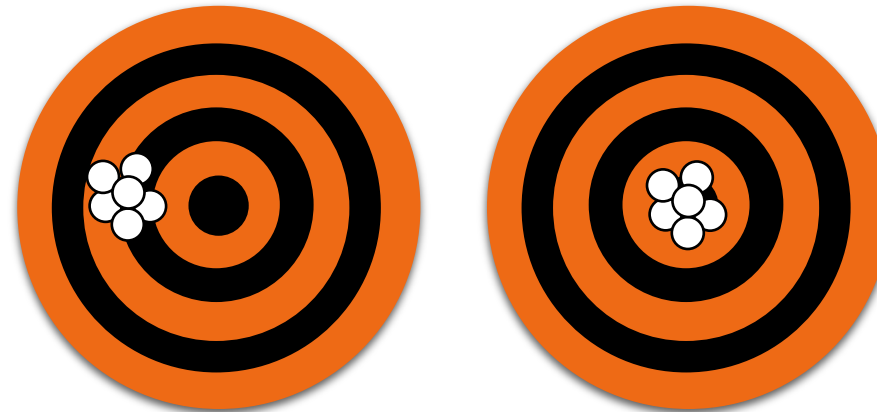
# The main challenge in measurement

---

- Accuracy
  - Right question?



- Precision
  - Right answer?



# The main challenge in measurement

---

- Validity



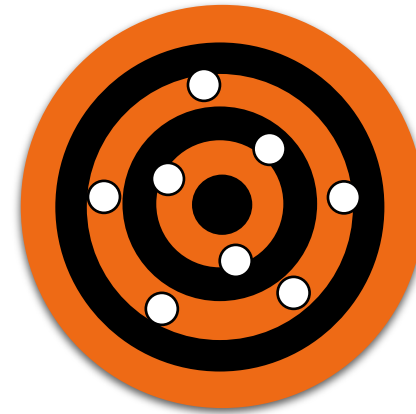
- Reliability



# The biggest challenge in measurement

---

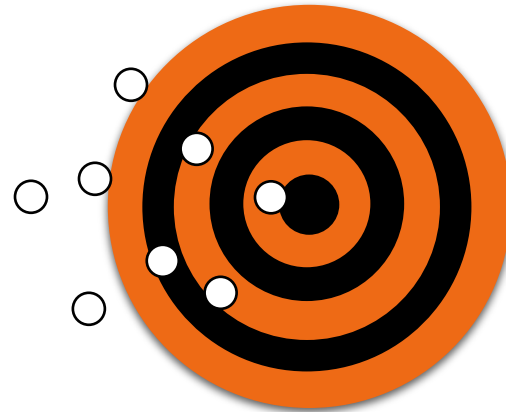
- Random error (noise)



# Systematic Error

---

- Bias+noise



- Bias



# *Is this a random or a systematic error?*

One surveyor did not really follow one of the instructions for a survey question:

Surveyor – “You feel a little unsafe in this environment, don’t you?”

Respondent – “Hmmm.. Sometimes so.”

Surveyor – “So the answer is yes?”

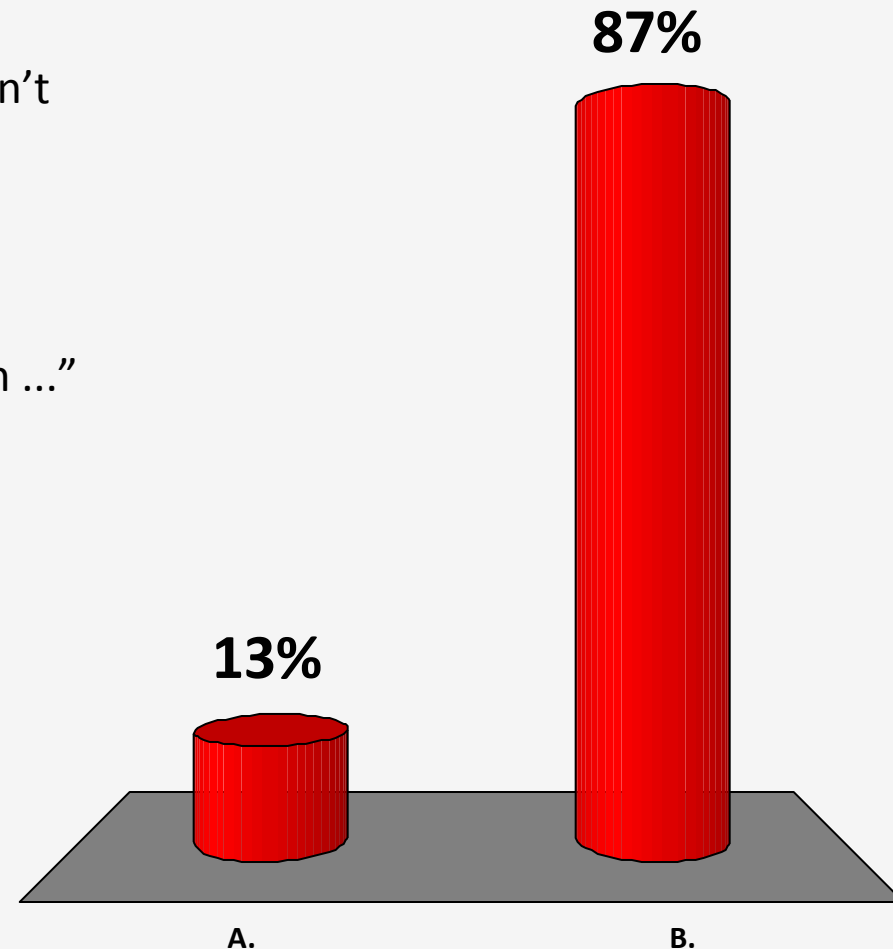
Respondent – “Yeah sure.”

Surveyor – “Alright, thanks for the response. Next question ...”

***Answer code 1 = yes***

A. Random Error

B. Systematic Error



# Validity

---

- In theory:
  - How well does the indicator map to the outcome?
    - e.g. intelligence → IQ tests
    - fear → heart rate
- In practice:
  - Are your survey questions unbiased (in a systematic way)?
  - Potential biases:
    - Social desirability bias
    - Demand bias (response bias)
    - Framing effect
    - Recall bias
    - Anchoring bias

# Reliability

---

- In theory:
  - The measure is consistent, precise, but answer not necessarily reliable
- In practice:
  - Length, fatigue
  - “How much did you spend on sweets last week?” (as a measure of annual consumption of sweets)
  - Ambiguous wording (definitions, relationships, recall period)
  - Answer choice (open/closed)

# General Noise

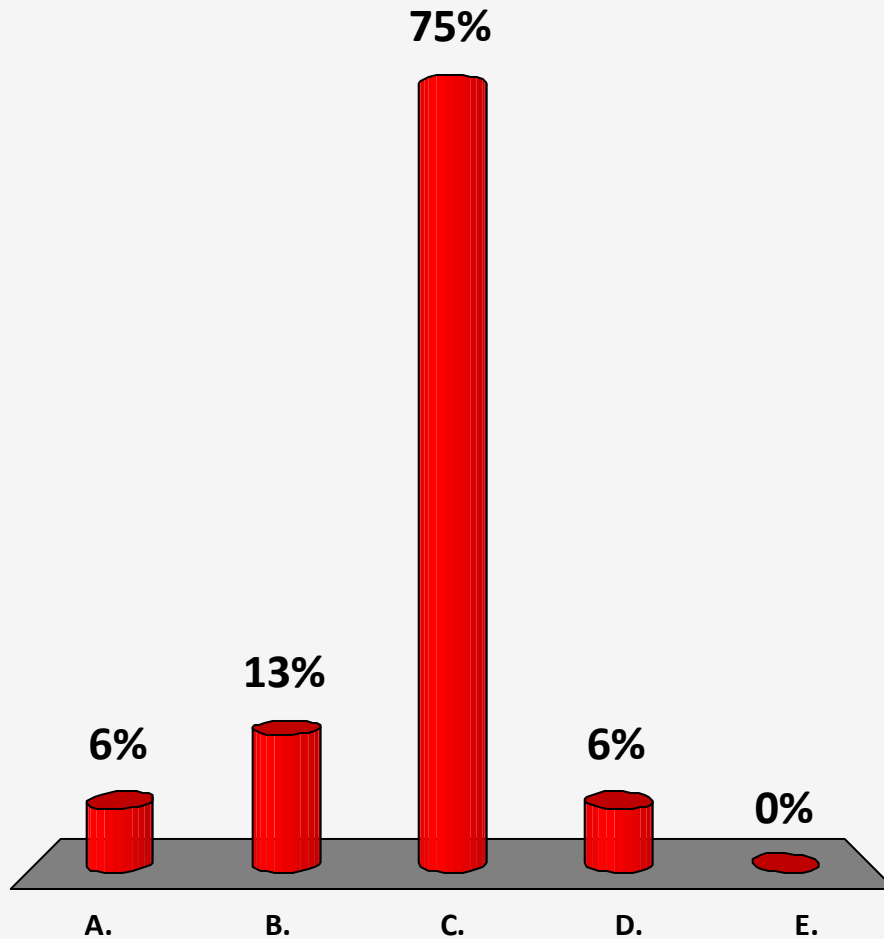
---

- Surveyor training/quality
- Data entry
- Poor translation
- How do you generalize from certain questions?



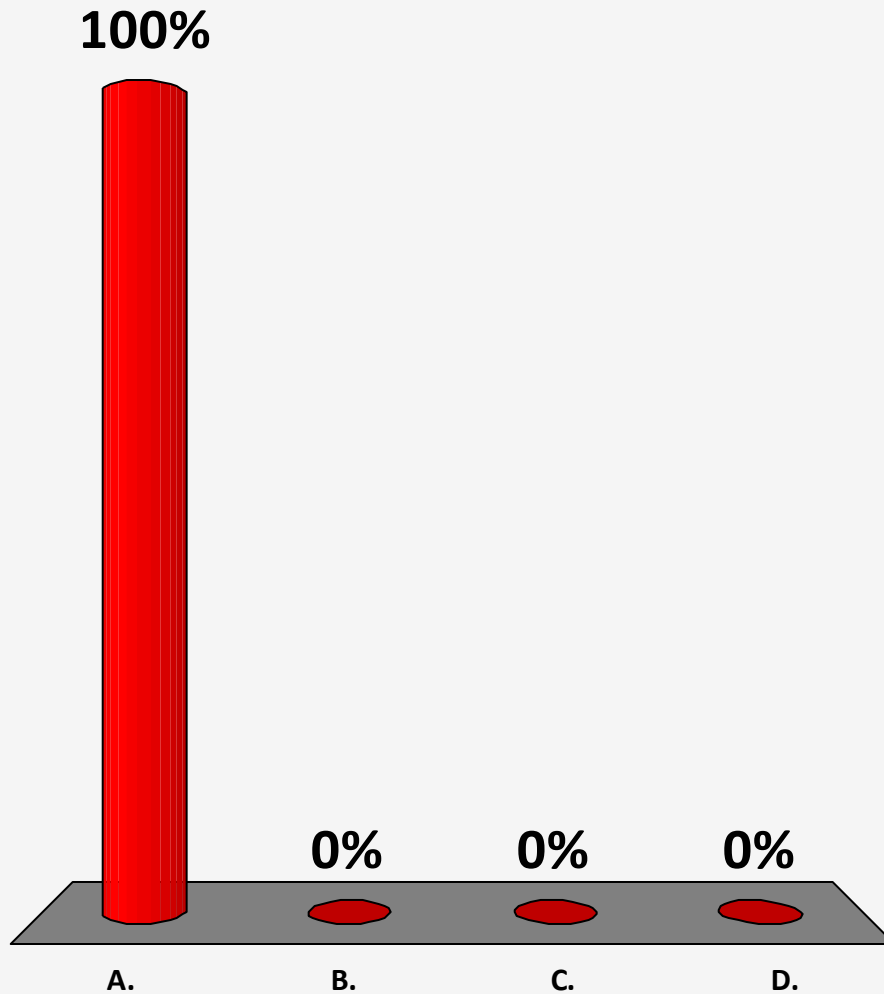
# Which is worse?

- A. Poor Validity
- B. Poor reliability
- C. Equally bad
- D. Depends
- E. Don't know/can't say



# Biased measurements will bias impact evaluations

- A. Yes
- B. No
- C. Depends
- D. Don't know



# Measuring the immeasurable

---

- How to measure it (well)

# What is hard to measure?

---

- (1) Things people do not know very well
- (2) Things people do not want to talk about
- (3) Abstract concepts
- (4) Things that are not (always) directly observable
- (5) Things that are best directly observed

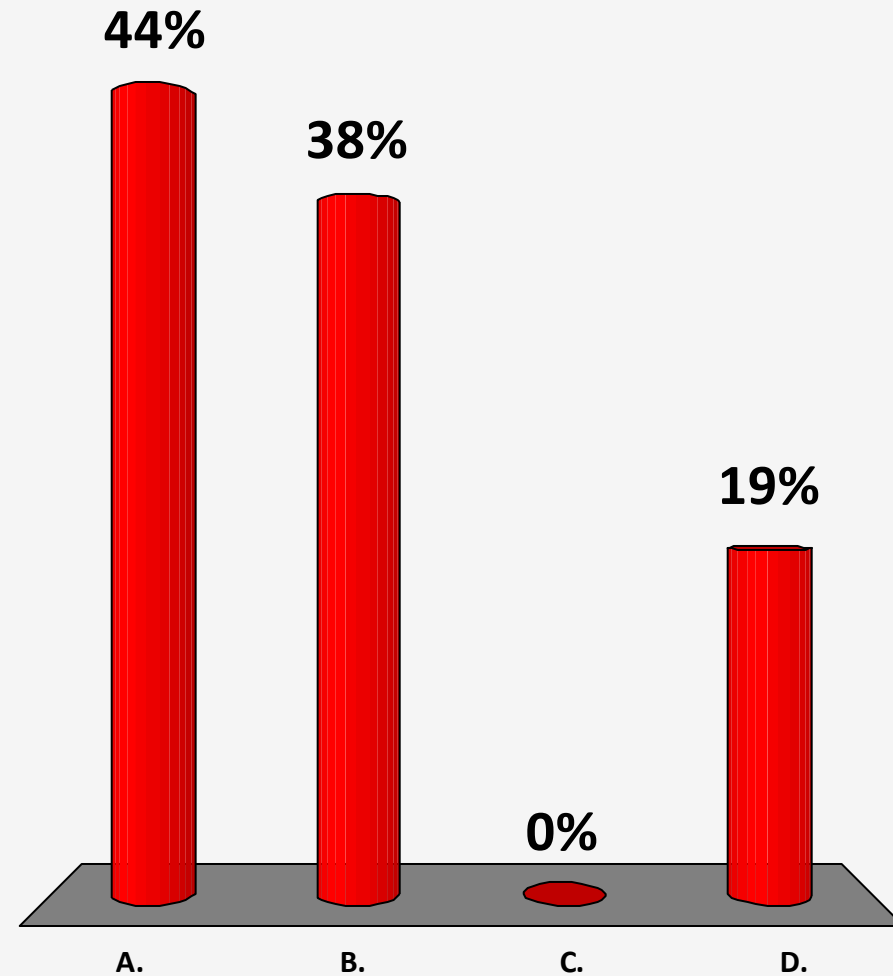
# What can we do with measurement and without?

---

- Missing key characteristics that:
  - interact with policies to change their impact
  - help us tailor policies and programs to better reach stated objectives
- If we can't measure it, we can't evaluate its importance

# How much juice did you drink last month?

- A. <2 liters
- B. 2-5 liters
- C. 6-10 liters
- D. >11 liters



# 1. Things people do not know very well

---

**What: Anything to estimate**, particularly across time. Prone to recall error and poor estimation

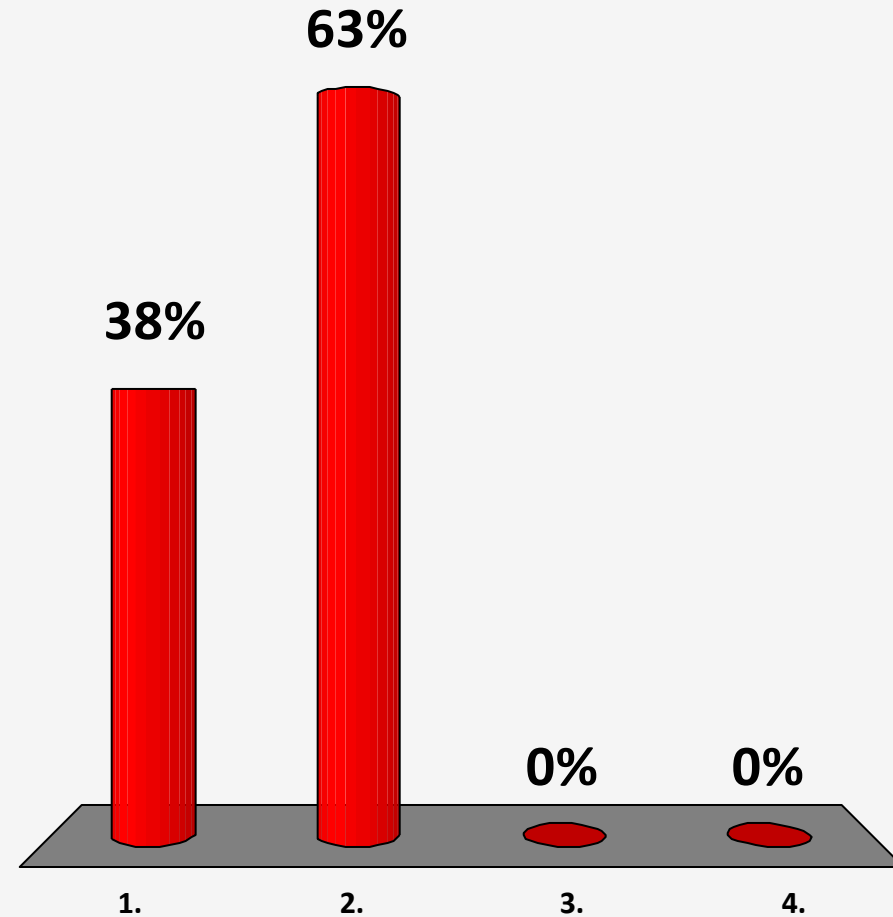
- **Examples:** distance to health center, profit, consumption, income, plot size

## **Strategies:**

- Consistency checks – How much did you spend in the last week on x? How much did you spend in the last 4 weeks on x?
- Multiple measurements of same indicator – How many minutes does it take to walk to the health center? How many kilometers away is the health center?

# How many glasses of juice did you drink yesterday?

- 0
- 1-3
- 4-6
- >6





# What is hard to measure?

---

(1) Things people do not know very well

(2) Things people do not want to talk about

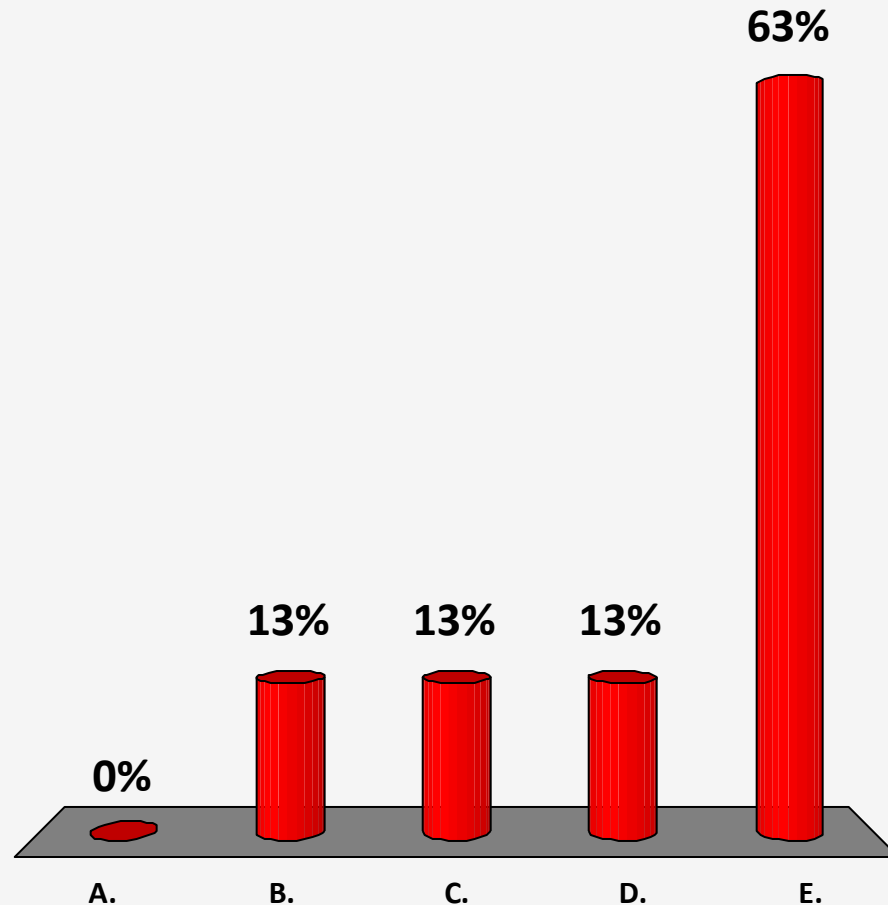
(3) Abstract concepts

(4) Things that are not (always) directly observable

(5) Things that are best directly observed

# How often do you swear at your spouse/partner?

- A. Every day
- B. Sometimes in a week
- C. Once a week
- D. Once a month
- E. Never



## 2. Things people don't want to talk about

---

**What:** Anything socially “risky” or something painful

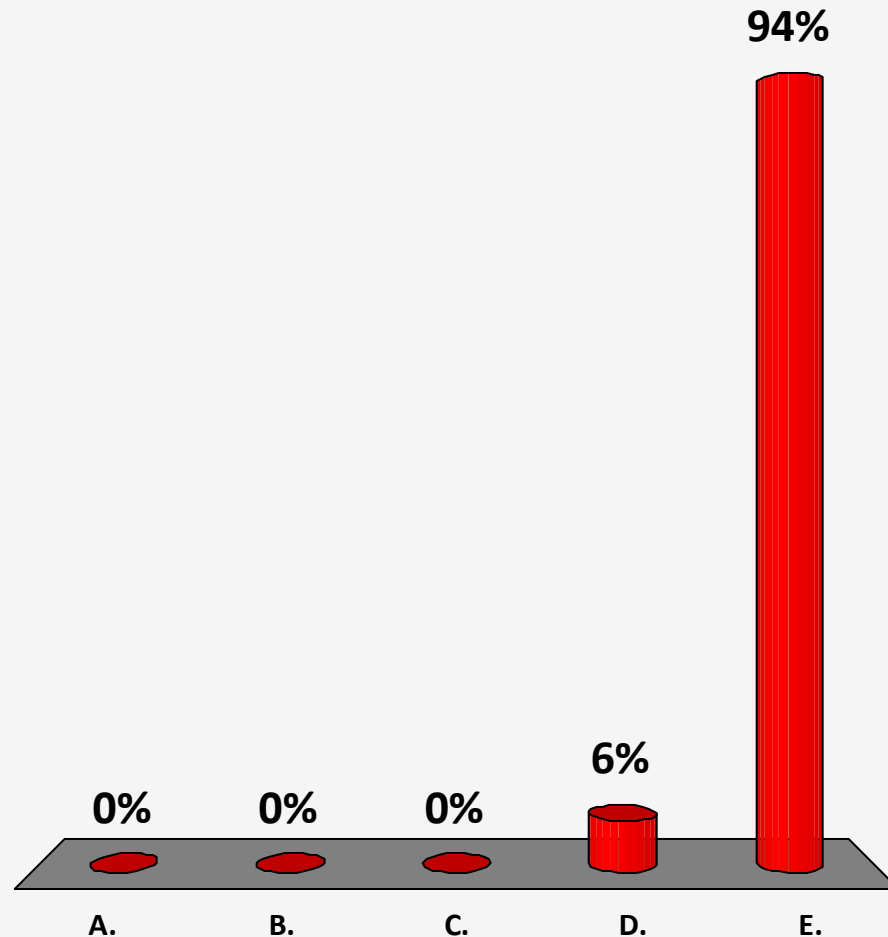
**Examples:** sexual activity, alcohol and drug use, domestic violence, mental health

**Strategies:**

- Don't start with the hard stuff!
- Consider asking question in third person
- Always ensure comfort and privacy of respondent

# How often does your spouse/partner swear at you?

- A. Every day
- B. Sometimes in a week
- C. Once a week
- D. Once a month
- E. Never



# What is hard to measure?

---

(1) Things people do not know very well

(2) Things people do not want to talk about

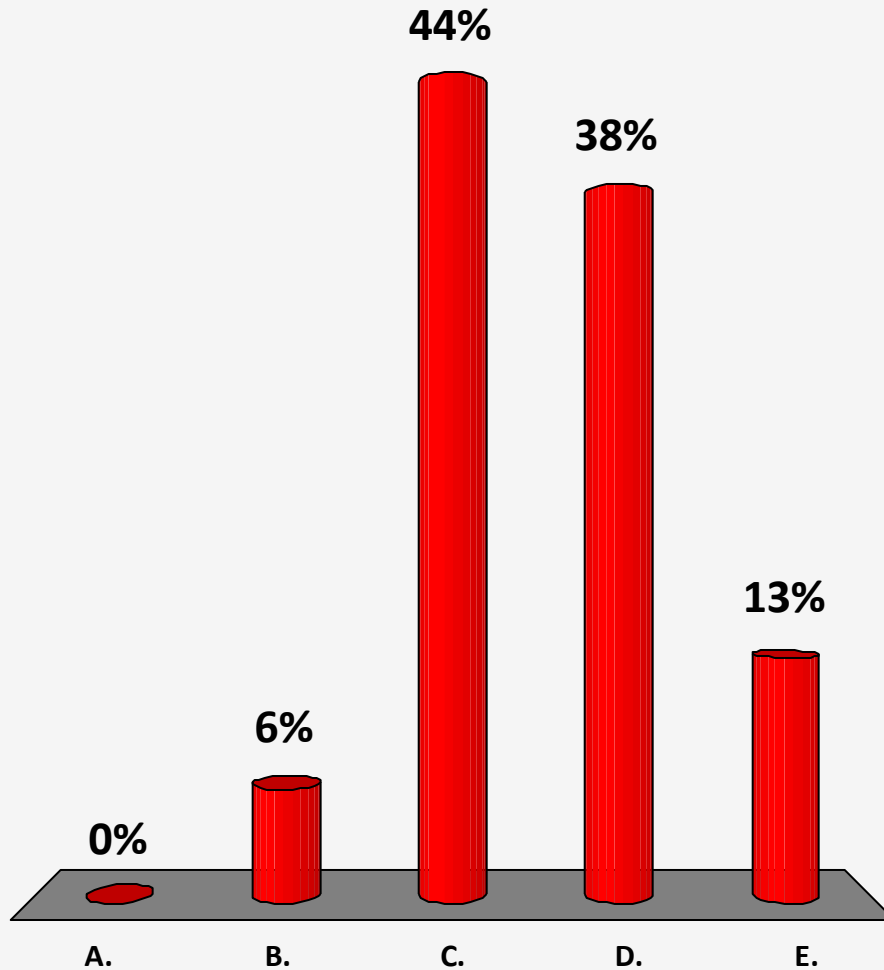
(3) Abstract concepts

(4) Things that are not (always) directly observable

(5) Things that are best directly observed

# “I am a better person compared to last year”

- A. Strongly disagree
- B. Disagree
- C. Neutral
- D. Agree
- E. Strongly agree



# 3. Abstract concepts

---

**What:** Potential indicator type which is challenging but interesting to measure

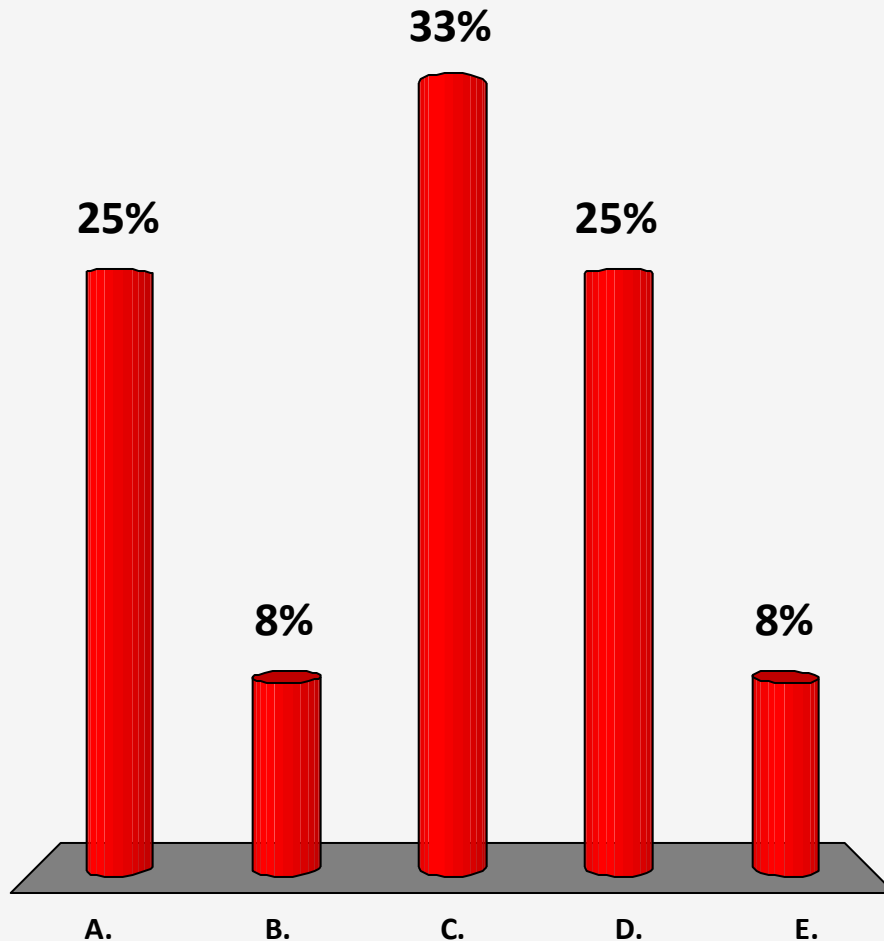
- **Example:** empowerment, negotiating skills, risk preferences

- **Strategies:**

- Three important steps to successfully measure abstract concepts:
  - Understand what you mean by abstract concepts.
  - Choose the results you want to use to measure your concepts.
  - Design good questions to address these measures.
- Often choice between choosing a self-reported measure and a behavioral measure – both can add value!

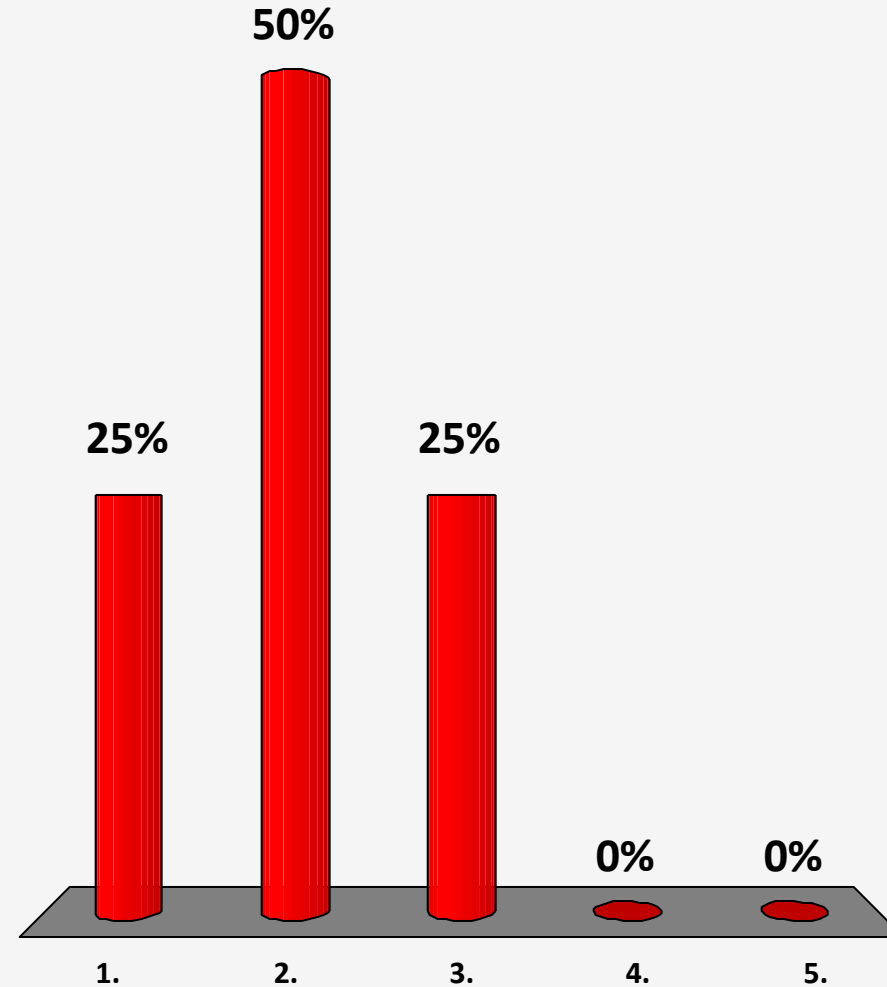
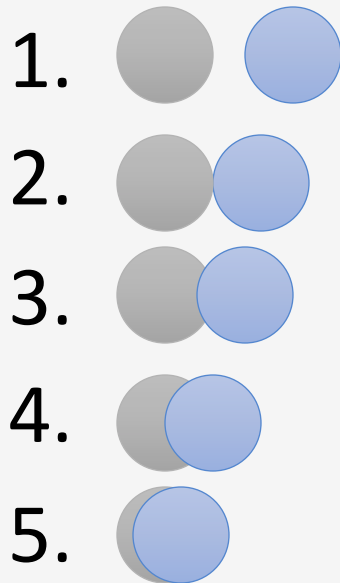
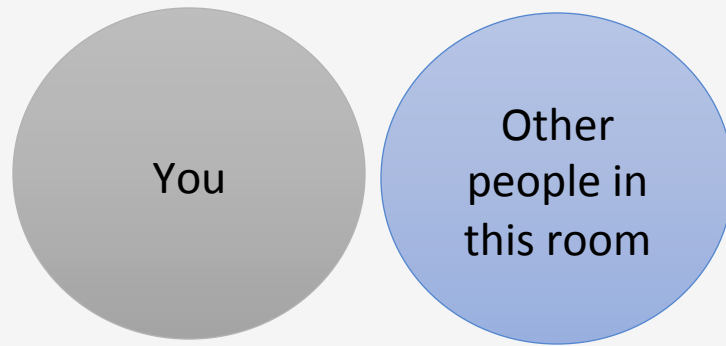
“I was involved in the decision of which schools I sent my kids to”

- A. Strongly disagree
- B. Disagree
- C. Neutral
- D. Agree
- E. Strongly agree



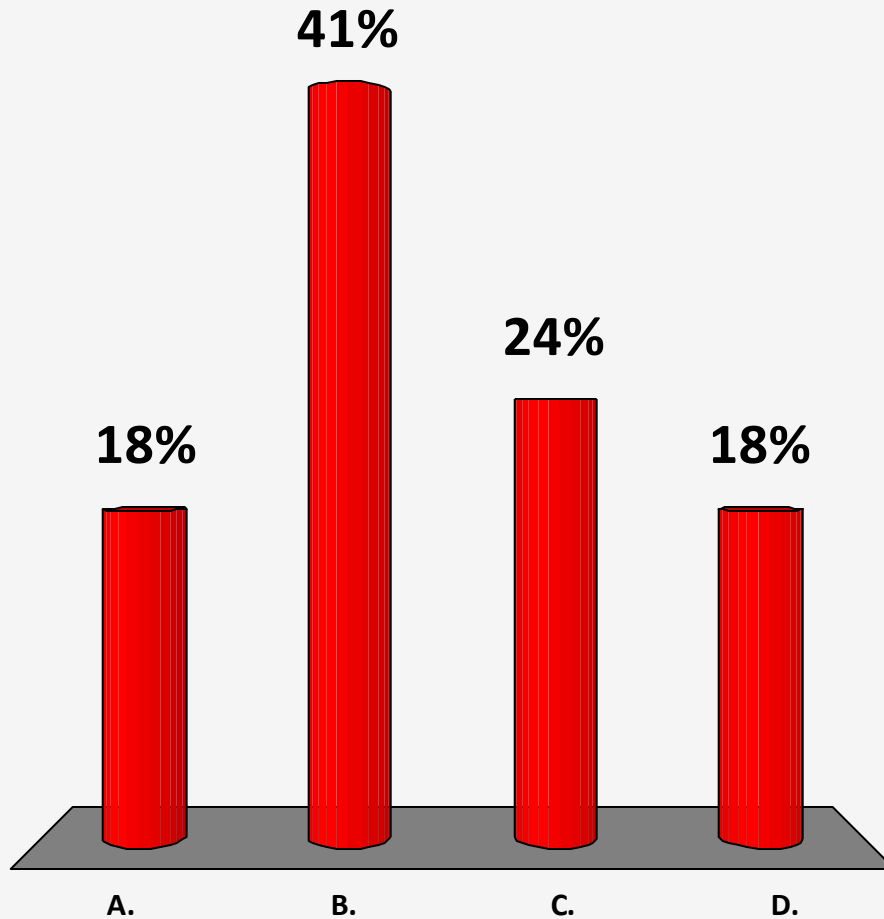


# How “connected” are you with other people in this room?



How likely are you to take a taxi with a driver you don't know after dark?

- A. Very unlikely
- B. Unlikely
- C. Likely
- D. Very likely



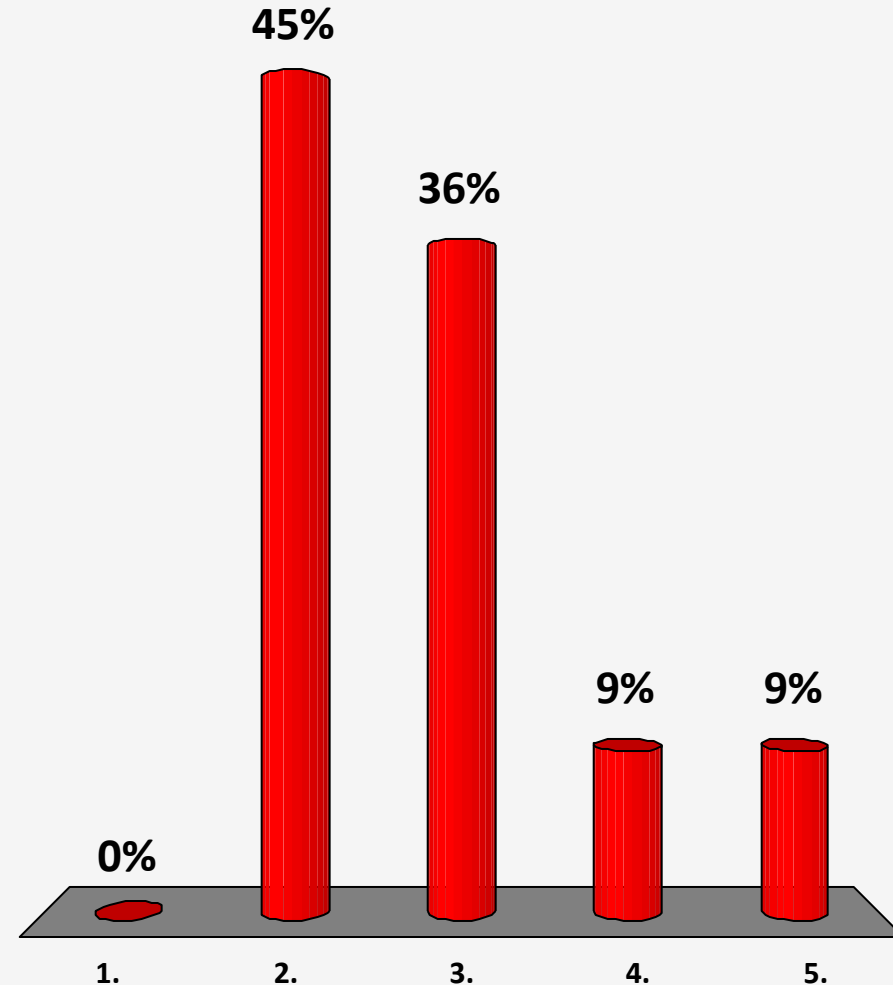
# What is hard to measure?

---

- (1) Things that other people might not well-understand
- (2) Things that are infrequently discussed
- (3) Abstract concepts
- (4) Things that are not (always) observed
- (5) Things which are better directly observed

# What's the proportion of African American in the U.S. that is not granted work due to discrimination

1. 0%
2. 1-20%
3. 21-40%
4. 41-60%
5. >60%



# 4. Things that cannot be directly observed

---

**What:** You want to perhaps measure things that cannot be directly observed.

- **Examples:** corruption, misconduct, discrimination

## **Strategies:**

- You have to be intellectually creative sometimes...
- Don't worry – There are more creative people than you out there – do literature review!

# 5. Things that are best directly observed

---

**What:** Behavioral preferences, anything that is more believable when done than said.

**Strategies:**

- Develop a more detailed protocol
- Ensure behavioral data collection is done in unvarying conditions across individuals

# 5. Things that are best directly observed

---

- Willingness to pay
  - How much would you like to pay for CFL bulb?
  - BDM method
    - How much would you like to pay for CFL bulb?
    - If the amount you stated is higher than the random number  $X$ , then you can actually pay the amount and get the CFL bulb

# Sources of Data

---



# Where can we get data from?

---

- Administrative data
  - Government census
  - Anonymous voting data
  - Cellphone use
- Other secondary data
  - World Bank/UN/IFPRI
- Primary data
  - Your survey results

# Primary Data Collection

---

- Surveys which are reported on your own
- Exams, tests, etc.
- Games
- Direct observations
- Daily notes

# Why collect your own data?

---

-Standards of an RCT usually comprise:

- Baseline
- Throughout intervention
- Endline
- Scale-up, intervention

-Pros vs. cons of own data collection

- Scale, cost
- Focus of study
  - I can get the data I want

# Considerations during Data Collection

---

- Quality monitoring
- Training for surveyors
- Surveyor gender composition
- Human subjects
- Data security
- Electronic vs. paper
- Cost

# Don't forget...

---

- Etiquette

# Why Randomize?

---

Manila, November 2015



ABDUL LATIF JAMEEL  
Poverty Action Lab

J-PAL SOUTHEAST ASIA

# Course Overview

---

1. What is evaluation?
2. Measuring impacts (outcomes, indicators)
3. Why randomize?
4. How to randomize?
5. Threats and Analysis
6. Sampling and sample size
7. RCT: Start to Finish
8. Cost Effectiveness Analysis and Scaling Up

# Course Overview

---

1. What is evaluation?
2. Measuring impacts (outcomes, indicators)
- 3. Why randomize?**
4. How to randomize?
5. Threats and Analysis
6. Sampling and sample size
7. RCT: Start to Finish
8. Cost Effectiveness Analysis and Scaling Up



# Presentation Overview

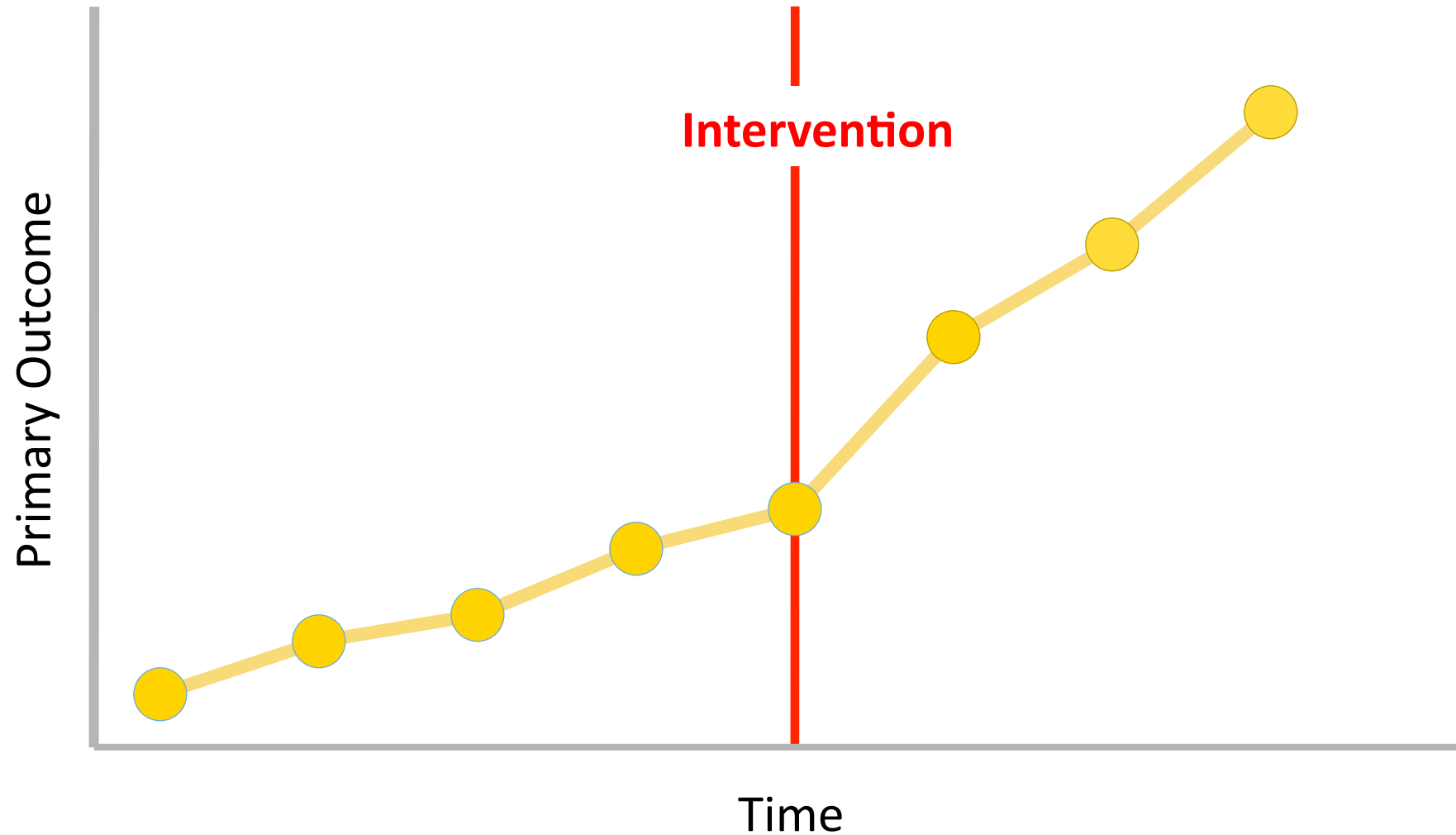
---

1. Background
2. What is a randomized experiment?
3. Why randomize?

# I – Background

# What is the impact of this program?

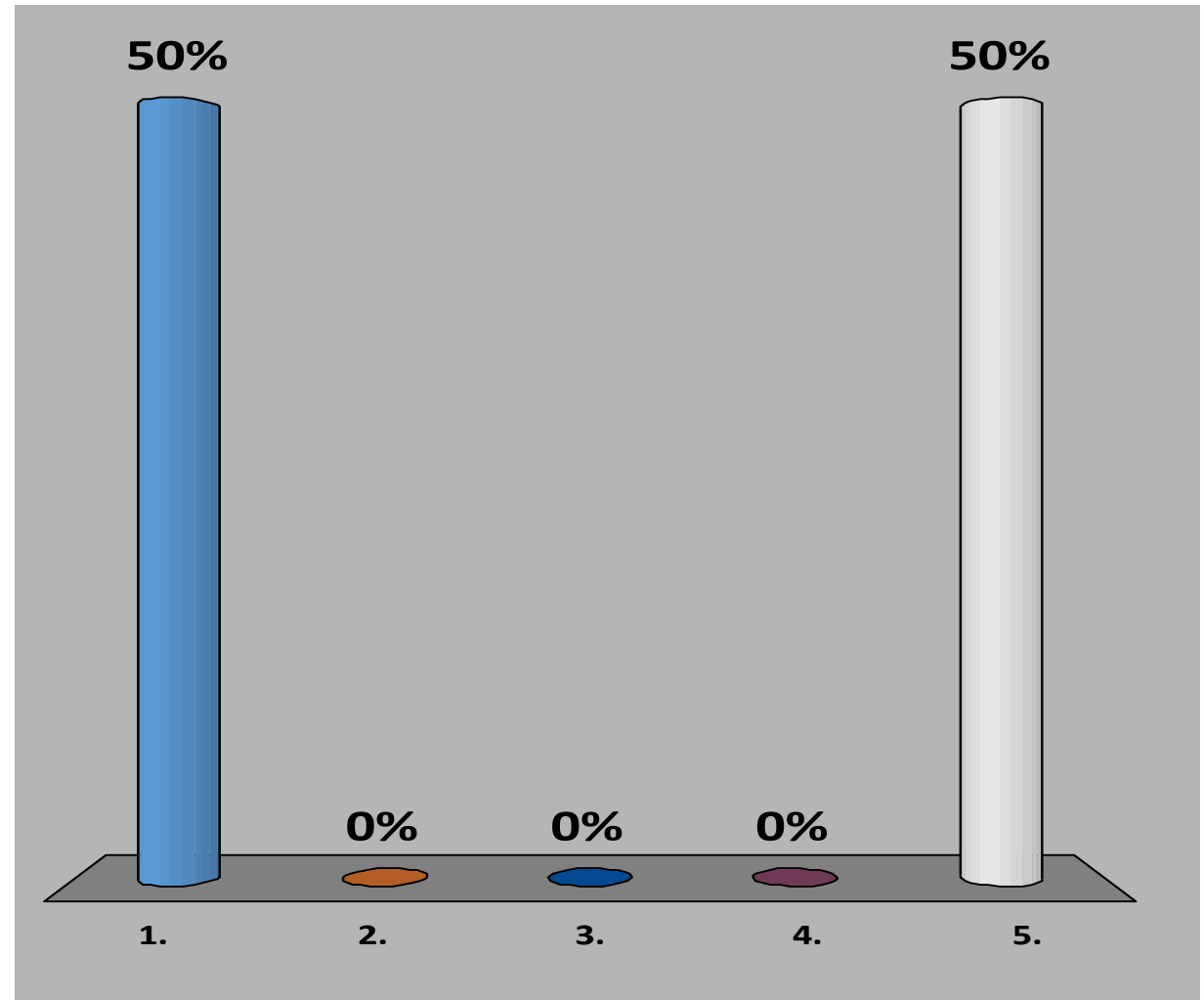
---



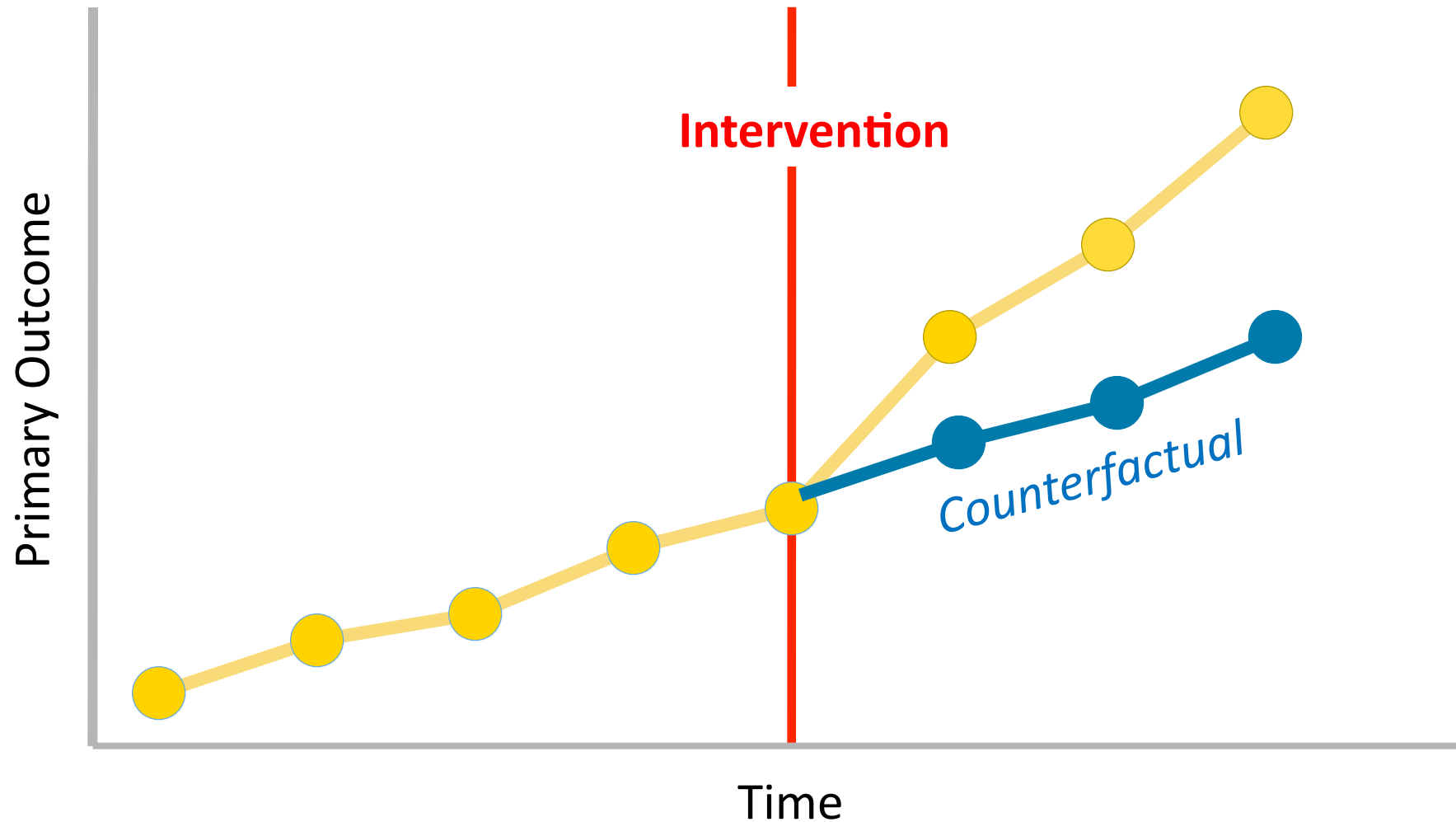
# What is the impact of this program?

---

1. Positive
2. Negative
3. Zero
4. I don't know
5. Not enough info



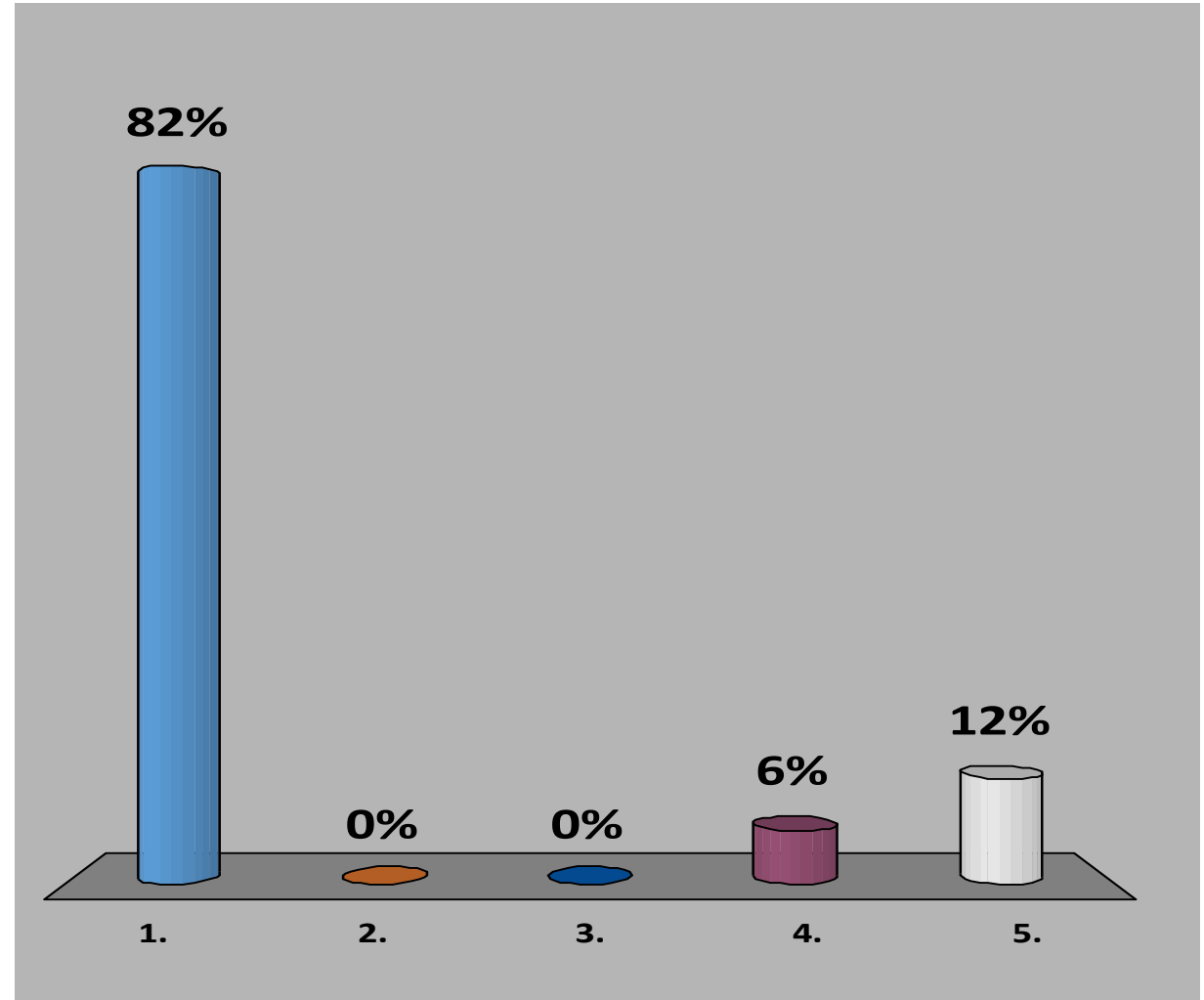
# What is the impact of this program?



# What is the impact of this program?

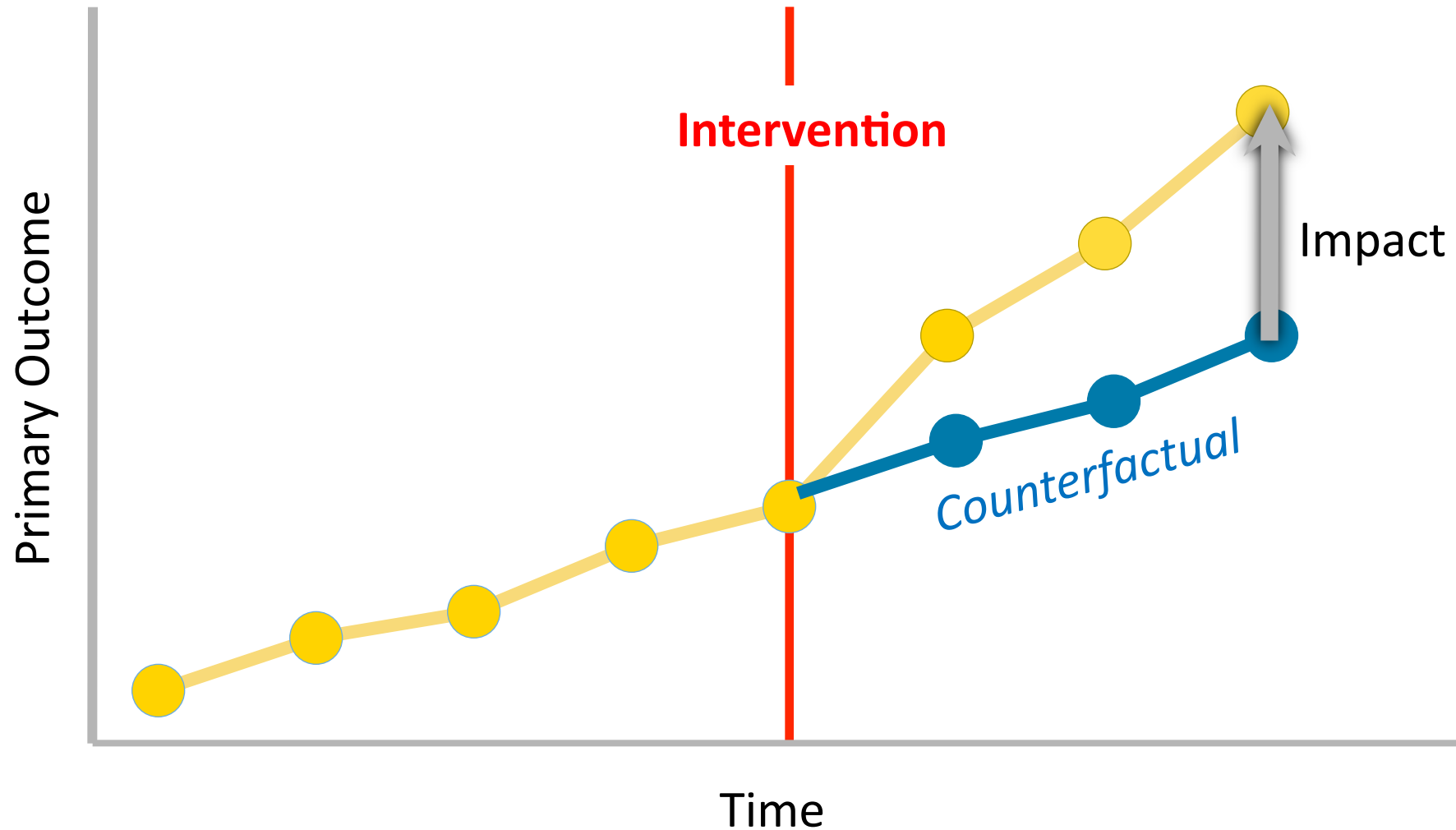
---

1. Positive
2. Negative
3. Zero
4. I don't know
5. Not enough info

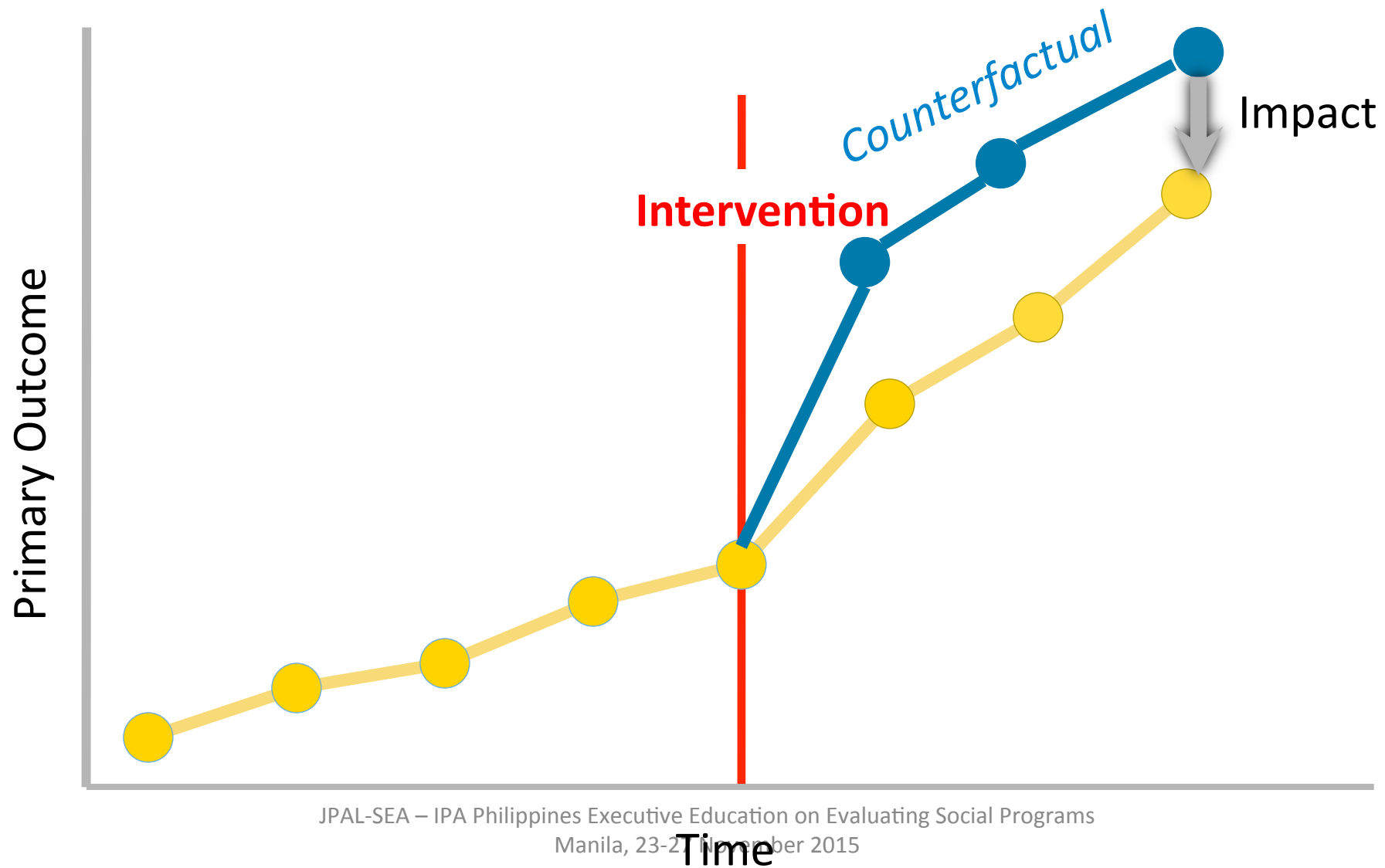


# Impact: What is it?

---



# Impact: What is it?





# How to measure impact?

---

***Impact*** is defined as a comparison between:

1. The outcome some time after the program has been introduced
2. The outcome at that same point in time had the program not been introduced (***the “counterfactual”***)

# Counterfactual

---

The **counterfactual** represents the state of the world that program participants would have experienced in the absence of the program (i.e. had they not participated in the program)

**Problem:** Counterfactual cannot be observed

**Solution:** We need to “mimic” or construct the counterfactual

# Impact evaluation methods

---

## 1. Randomized Experiments

- Also known as:
  - Random Assignment Studies
  - Randomized Field Trials
  - Social Experiments
  - Randomized Controlled Trials (RCTs)
  - Randomized Controlled Experiments

# Impact Evaluation Methods

---

## 2. Non- or Quasi-Experimental Methods

- a. Pre-Post
- b. Simple Difference
- c. Differences-in-Differences
- d. Multivariate Regression
- e. Statistical Matching
- f. Interrupted Time Series
- g. Instrumental Variables
- h. Regression Discontinuity

# II – What is a randomized experiment?

# The Basics

---

Start with simple case:

Take a sample of program applicants

***Randomly*** assign them to either:

**Treatment Group** – is offered treatment

**Control Group** - not allowed to receive treatment (during the evaluation period)

# Key advantage of experiments

---

Because members of the groups (treatment and control) **do not differ systematically** at the outset of the experiment,

any difference that subsequently arises between them can be **attributed** to the program rather than to other factors.

# Evaluation of “Cash Transfers, Behavioral Changes, and Cognitive Development in Early Childhood”: Treatment vs. Control villages at *baseline*

Variable	Control group	Treatment group	Difference
Vocabulary introduction test	5.37	6.23	0,860 (0,290)
Weight-for-Age z-score	-0. 88	-1.06	-0.18 (0.094)
Height-for-Age z-score	-1,08	-1.27	-0.19 (0.109)
Birth weight	6.79	6.75	0.04 (0.947)

*P-value* is inside bracket.

\*/\*\*/\*\*\*: Statistically significant at 10% / 5% / 1% level

Sumber: Macours, Norbert Schady and Renos Vakis (2012)



# Some variations of the basics

---

- Assigning to multiple treatment groups
- Assigning to units other than individuals or households
  - Health Centers
  - Schools
  - Local Governments
  - Villages

# III – Why Randomize?

# Why randomize? – Conceptual Argument

---

If properly designed and conducted, randomized experiments provide **the most credible** method to estimate the impact of a program

# Why “most credible”?

---

Because members of the groups (treatment and control) **do not differ systematically** at the outset of the experiment,

any difference that subsequently arises between them can be **attributed** to the program rather than to other factors.

# Example #1: Balsakhi Program

---



# Balsakhi Program: Brief Background

---

- Implemented by **Pratham**, an NGO from India
- This program **provides additional teachers** (tutor/Balsakhi) to help at-risk children acquire the basic skills they need to participate fully in the classroom
- In Vadodara, Balsakhi is implemented in public schools from **2002 to 2003**
- **Teachers will decide** which children will receive Balsakhi

# Balsakhi: Outcomes

---

- Children were tested at the beginning of the school year (Pretest) and at the end of the year (Post-test)
- **QUESTION:** How can we estimate the impact of the balsakhi program on test scores?

# Methods to estimate impacts

---

- Let's look at different ways of estimating the impacts using the data from the schools that got a Balsakhi
  1. Pre – Post (Before vs. After)
  2. Simple difference
  3. Difference-in-difference
  4. Other non-experimental methods
  5. Randomized Experiment



# 1 - Pre-post (Before vs. After)

---

- Look at average change in test scores over the school year for the balsakhi children



# 1 - Pre-post (Before vs. After)

---

Average <u>post-test</u> score for children with a balsakhi	51.22
Average <u>pretest</u> score for children with a balsakhi	24.80
<b>Difference</b>	<b>26.42</b>

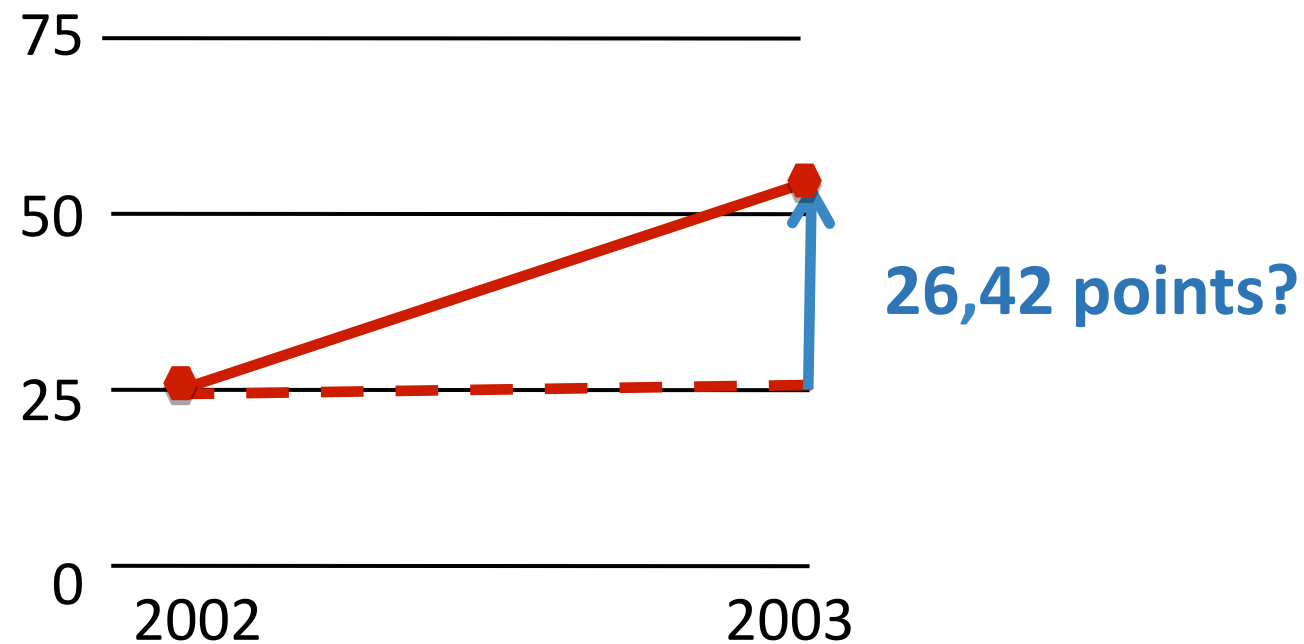
**QUESTION:** Under what conditions can this difference (26.42) be interpreted as the impact of the balsakhi program?

# What would have happened without balsakhi?

---

Method 1: Before vs. After

Impact = 26.42 points?



# 2 - Simple difference

---

Compare test scores of...



Children who **got**  
balsakhi

With test scores of...



Children who **did not get**  
balsakhi

## 2 - Simple difference

---

Average score for children with a balsakhi	51.22
Average score for children without a balsakhi	56.27
Difference	-5.05

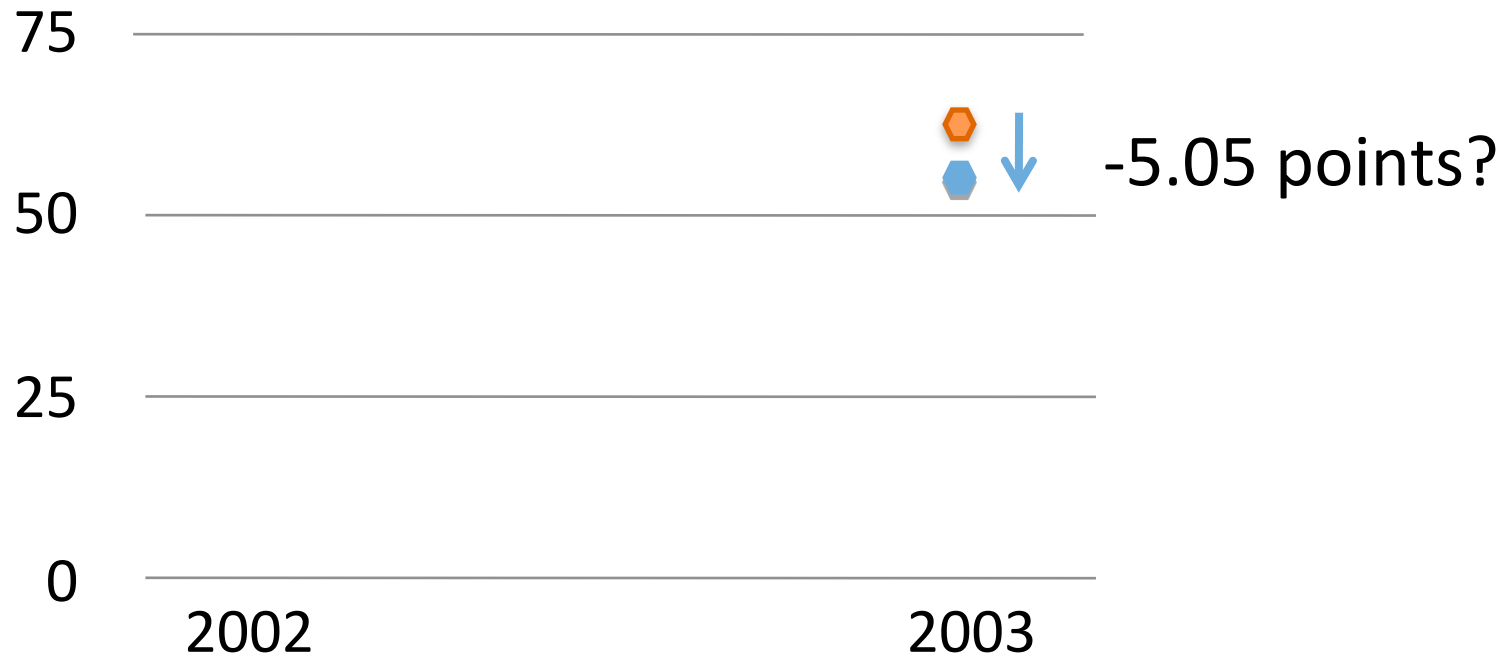
**QUESTION:** Under what conditions can this difference (-5.05) be interpreted as the impact of the balsakhi program?

# What would have happened without balsakhi?

---

Method 2: Simple Comparison

Impact = -5.05 points?



# Selection

---

- Any characteristic that is related to how each group/sample is treated and independently affects the treatment outcome could result in a **bias**
  - Education background (treatment) and income (outcomes)
  - Water source improvement (treatment) and diarrhea incidence (outcomes)

# 3 – *Difference-in-Differences*

---

Compare **gains** in test scores of...



Children who **got**  
balsakhi

With gains in test  
scores of...



Children who **did not** get  
balsakhi



# 3 – *Difference-in-Differences*

---

	Pretest	Post-test	Difference
Average score for children with a balsakhi	24.80	51.22	26.42

# 3 – *Difference-in-Differences*

---

	Pretest	Post-test	Difference
Average score for children <b>with</b> a balsakhi	24.80	51.22	26.42
Average score for children <b>without</b> a balsakhi	36.67	56.27	19.60

# 3 – *Difference-in-Differences*

---

	Pretest	Post-test	Difference
Average score for children <b>with</b> a balsakhi	24.80	51.22	26.42
Average score for children <b>without</b> a balsakhi	36.67	56.27	19.60
<b>Difference</b>			<b>6.82</b>

- **QUESTION:** In what condition the 6.82 can be interpreted as the impact from Balsakhi program?

# 4 – Other Methods

---

- There are more sophisticated non-experimental methods to estimate program impacts:
  - Regression
  - Matching
  - Instrumental Variables
  - Regression Discontinuity
- These methods rely on being able to “mimic” the counterfactual **under certain assumptions**
- **Problem:** Assumptions are not testable

# 5 – Randomized Experiment

---

- Suppose we evaluated the Balsakhi program using a randomized experiment
- **QUESTION #1:** What would this entail? How would we do it?
- **QUESTION #2:** What would be the advantage of using this method to evaluate the impact of the balsakhi program?

# Impact of Balsakhi - Summary

---

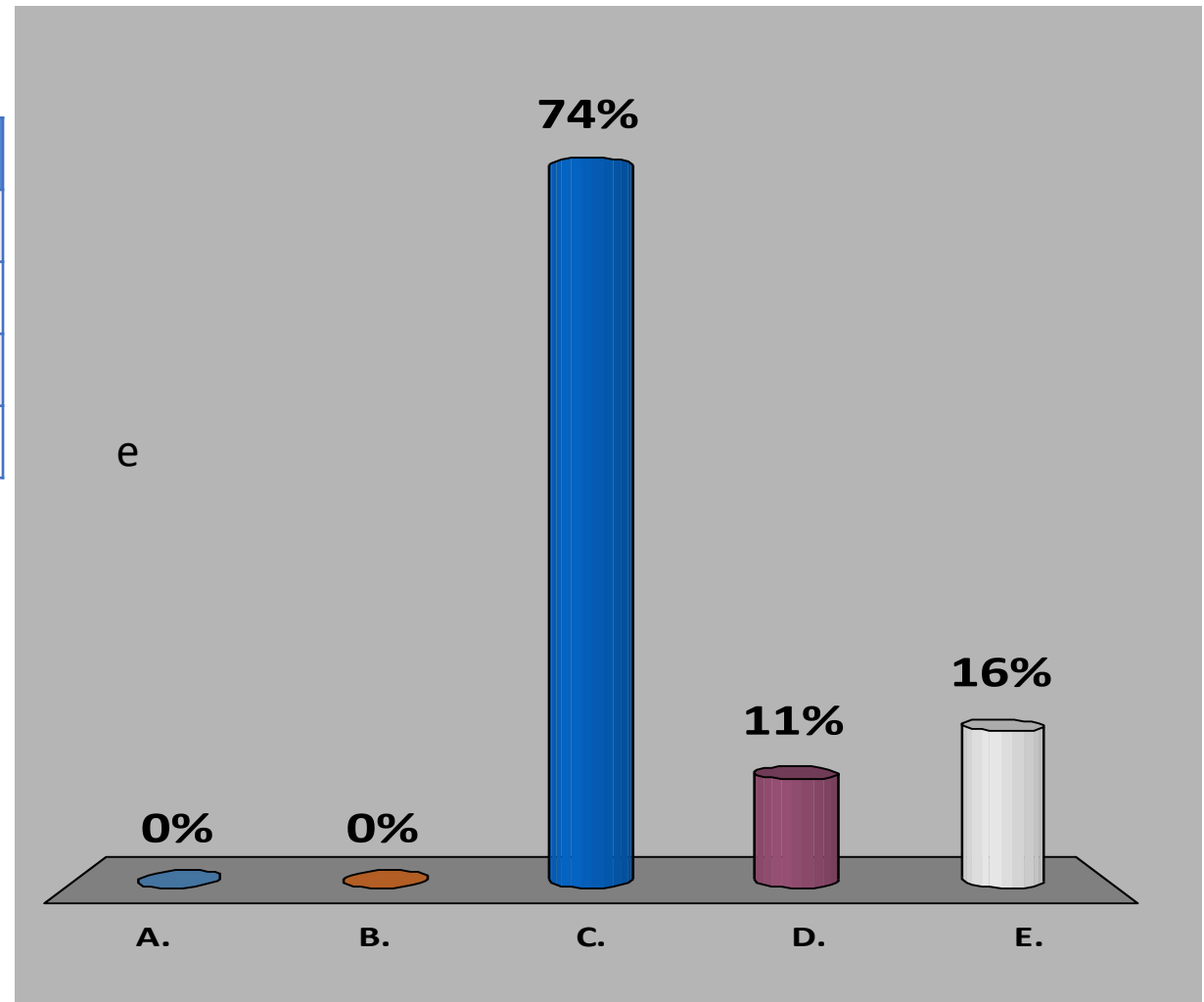
Method	Impact Estimate
(1) Pre-post	26.42*
(2) Simple Difference	-5.05*
(3) Difference-in-Difference	6.82*
(4) Regression	1.92
<b>(5) Randomized Experiment</b>	<b>5.87*</b>

\*: Statistically significant at the 5% level

# Which of these methods do you think is closest to the truth?

Method	Impact Estimate
(1) Pre-post	26.42*
(2) Simple Difference	-5.05*
(3) Difference-in-Difference	6.82*
(4) Regression	1.92

- A. Pre-Post
- B. Simple Difference
- C. Difference-in-Differences
- D. Regression
- E. Don't know



# Impact of Balsakhi - Summary

---

Method	Impact Estimate
(1) Pre-post	26.42*
(2) Simple Difference	-5.05*
(3) Difference-in-Difference	6.82*
(4) Regression	1.92
<b>(5) Randomized Experiment</b>	<b>5.87*</b>

\*: Statistically significant at 5% level



# Impact of Balsakhi - Summary

---

Method	Impact Estimate
(1) Pre-post	26.42*
(2) Simple Difference	-5.05*
(3) Difference-in-Difference	6.82*
(4) Regression	1.92
(5) Randomized Experiment	5.87*

\*: Statistically significant at 5% level

**Bottom Line: Which method we use matters!**

# Example #2 – *Microfinance* in South Africa

---

Methods	Impact Estimation
(1) Pre-post	2384*
(2) Simple Difference	1838*
(3) Difference-in-Difference	1068*
(4) Regression	1412
<b>(5) Randomized Experiment</b>	

\*: Statistically significant at 5% level

# Example #2 – *Microfinance* in South Africa

---

Methods	Impact Estimation
(1) Pre-post	2384*
(2) Simple Difference	1838*
(3) Difference-in-Difference	1068*
(4) Regression	1412
<b>(5) Randomized Experiment</b>	<b>292*</b>

\*: Statistically significant at 5% level

# Example #3 –Read India program in Pratham

---



# Example #3 –Read India program in Pratham

---

Methods	Impact Estimation
(1) Pre-post	0,60*
(2) Simple Difference	-0,90*
(3) Difference-in-Difference	0,31*
(4) Regression	0,06
<b>(5) Randomized Experiment</b>	

\*: Statistically significant at 5% level

# Example #3 –Read India program in Pratham

---

Methods	Impact Estimation
(1) Pre-post	0,60*
(2) Simple Difference	-0,90*
(3) Difference-in-Difference	0,31*
(4) Regression	0,06
<b>(5) Randomized Experiment</b>	<b>0.88*</b>

\*: Statistically significant at 5% level

WHEN TO DO A RANDOMIZED EVALUATION?

# When to do a randomized evaluation?

---

- When there is an important question you want/need to know the answer to
- Timing--not too early and not too late
- Program is representative not gold plated
  - Or tests an basic concept you need tested
- Time, expertise, and money to do it right
- Develop an evaluation plan to prioritize



# When NOT to do an RE?

---

- When the program is premature and still requires considerable “tinkering” to work well
- When the project is on too small a scale to randomize into two “representative groups”
- If a positive impact has been proven using rigorous methodology and resources are sufficient to cover everyone
- After the program has already begun and you are not expanding elsewhere

# How can an RE not succeed?

---

- No one asks questions answered by the study
- The evaluation measures the wrong result
- It leaves too many unanswered important questions
- It produce a biased result

# IV – Conclusion

# Conclusion – Why Randomize?

---

- There are **many methods** to estimate the impact of certain program
- This training discusses one of the methods: **Randomized evaluation**
  - **Conceptual argument:** If designed and implemented well, an RE can provide the most credible method to evaluate the impact of certain program
  - **Empirical argument:** Different methods can produce different impact estimation

# How to Randomize?

---

Manila, November 2015



ABDUL LATIF JAMEEL  
Poverty Action Lab

J-PAL SOUTHEAST ASIA

# Course Overview

---

1. What is evaluation?
2. Measuring impacts (outcomes, indicators)
3. Why randomize?
4. How to randomize?
5. Threats and Analysis
6. Sampling and sample size
7. RCT: Start to Finish
8. Cost Effectiveness Analysis and Scaling Up

# Course Overview

---

1. What is evaluation?
2. Measuring impacts (outcomes, indicators)
3. Why randomize?
- 4. How to randomize?**
5. Threats and Analysis
6. Sampling and sample size
7. RCT: Start to Finish
8. Cost Effectiveness Analysis and Scaling Up

# Lecture Overview

---

- ***Unit and method of randomization***
- Real-world constraints
- Revisiting unit and method
- Variations on simple treatment-control



# Unit of Randomization: Options

---

1. Randomizing at the individual level
  2. Randomizing at the group level  
“Cluster Randomized Trial”
- Which level to randomize?

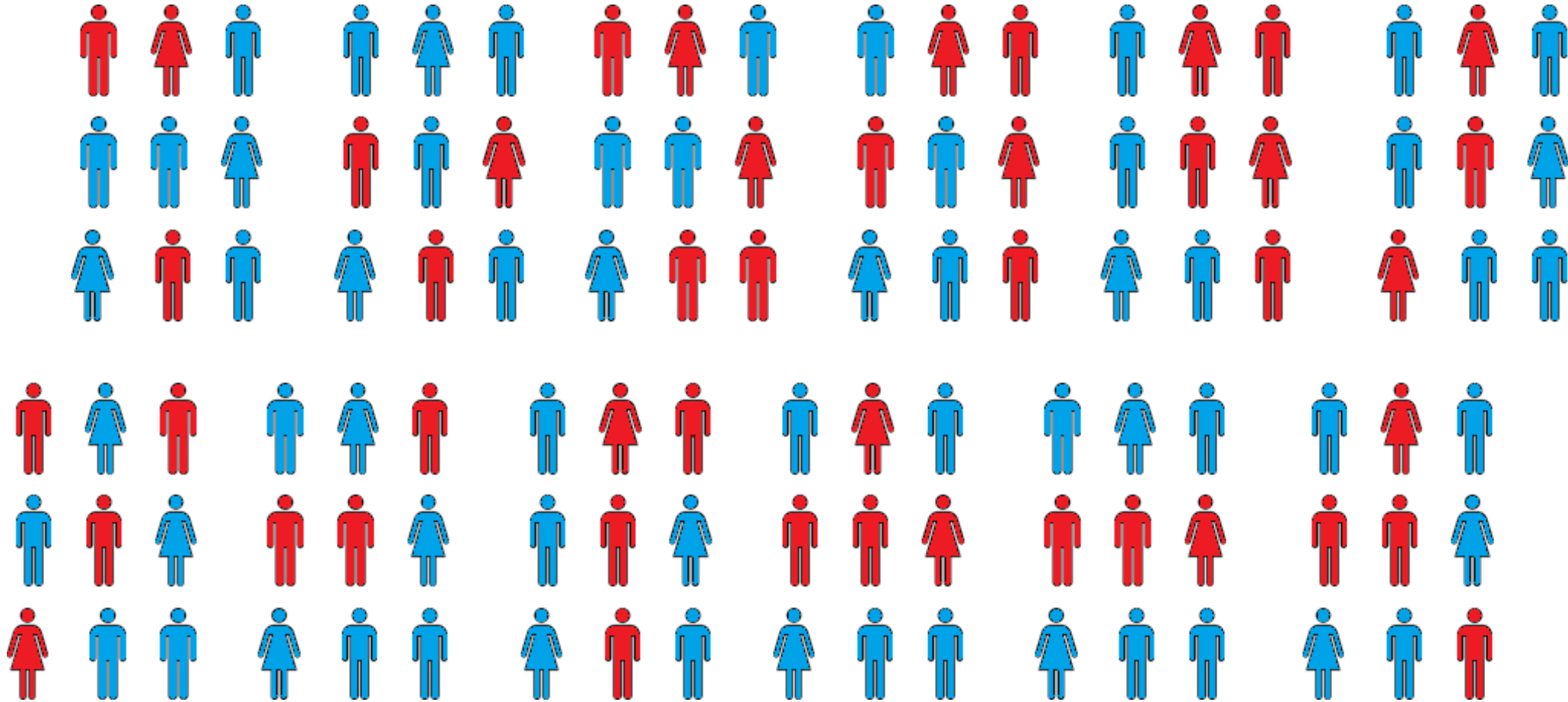
# Unit of Randomization: Individual?

---



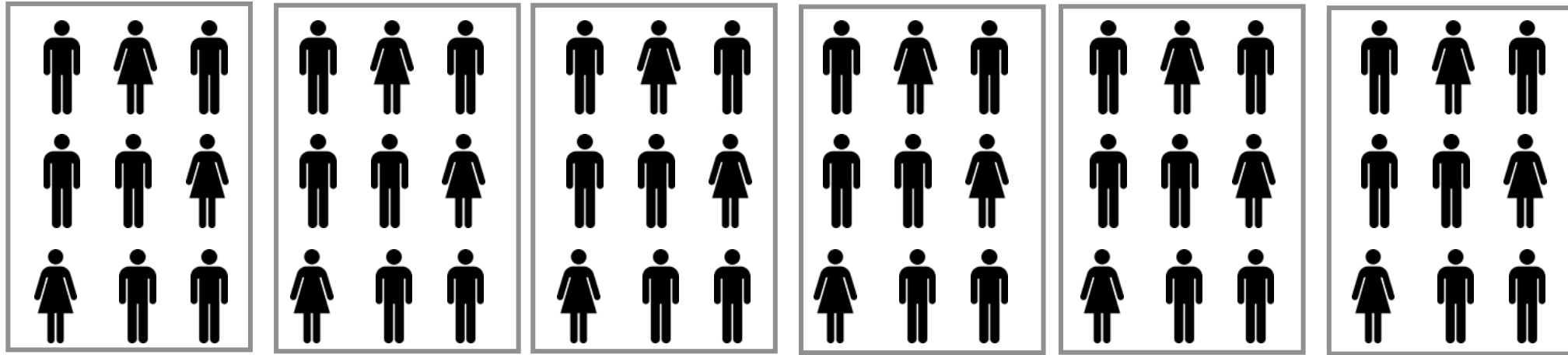
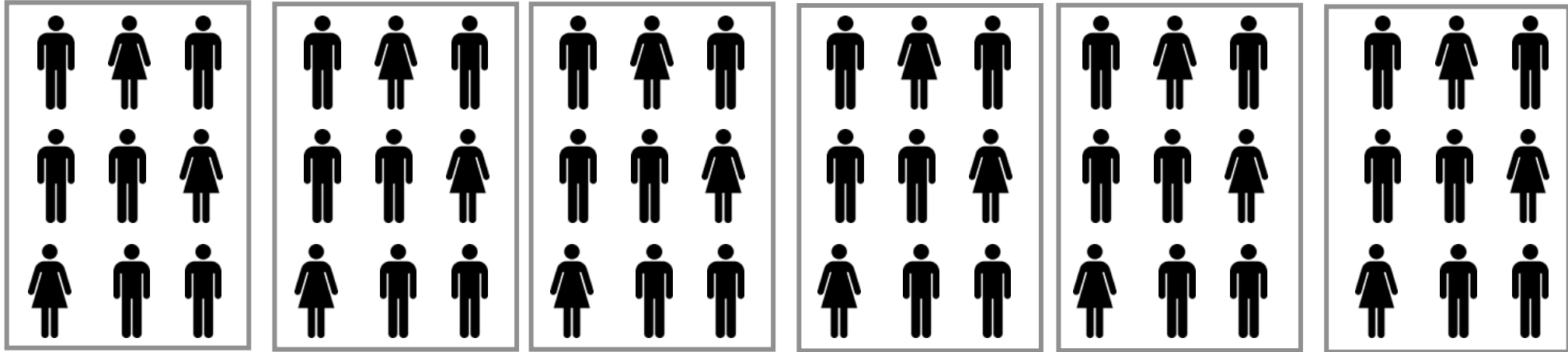
# Unit of Randomization: Individual?

---

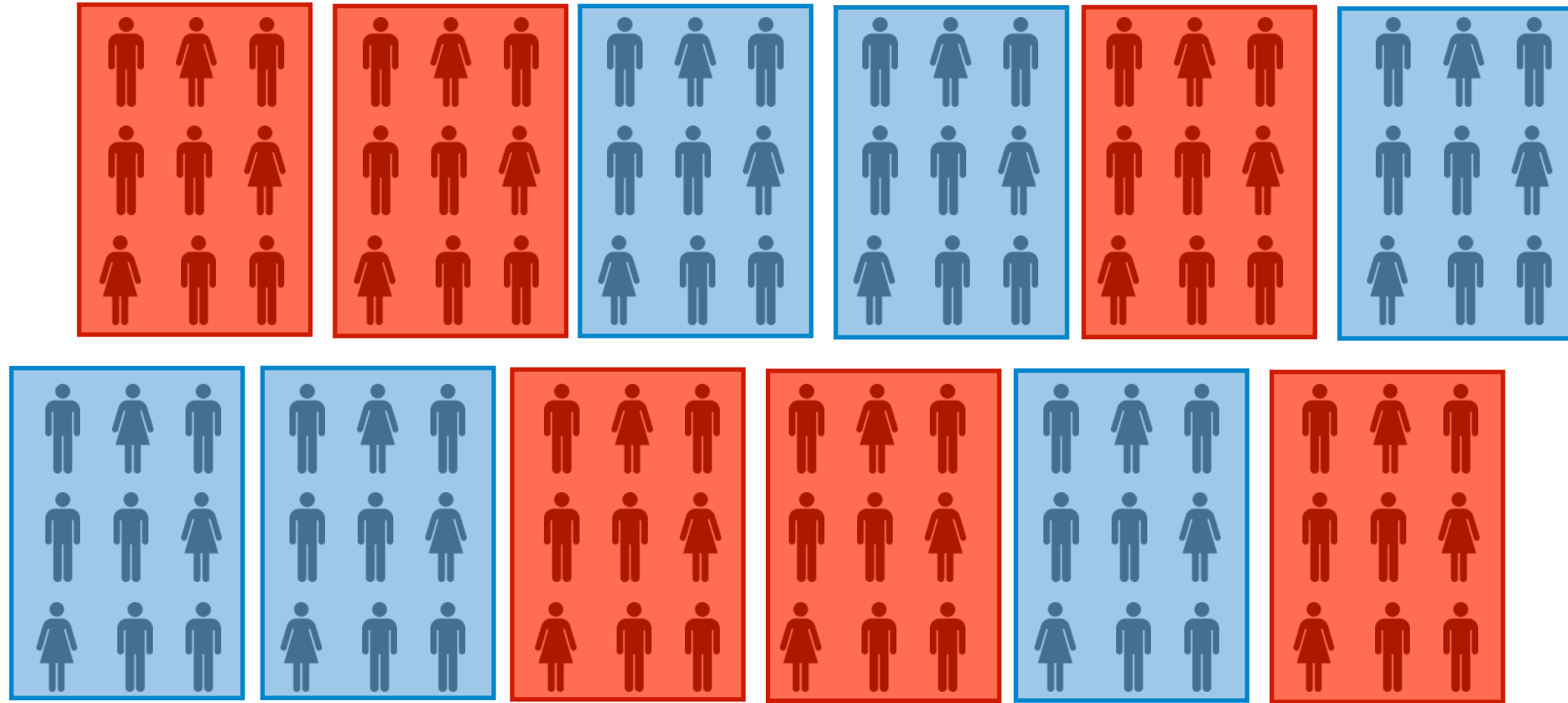


# Unit of Randomization: Class?

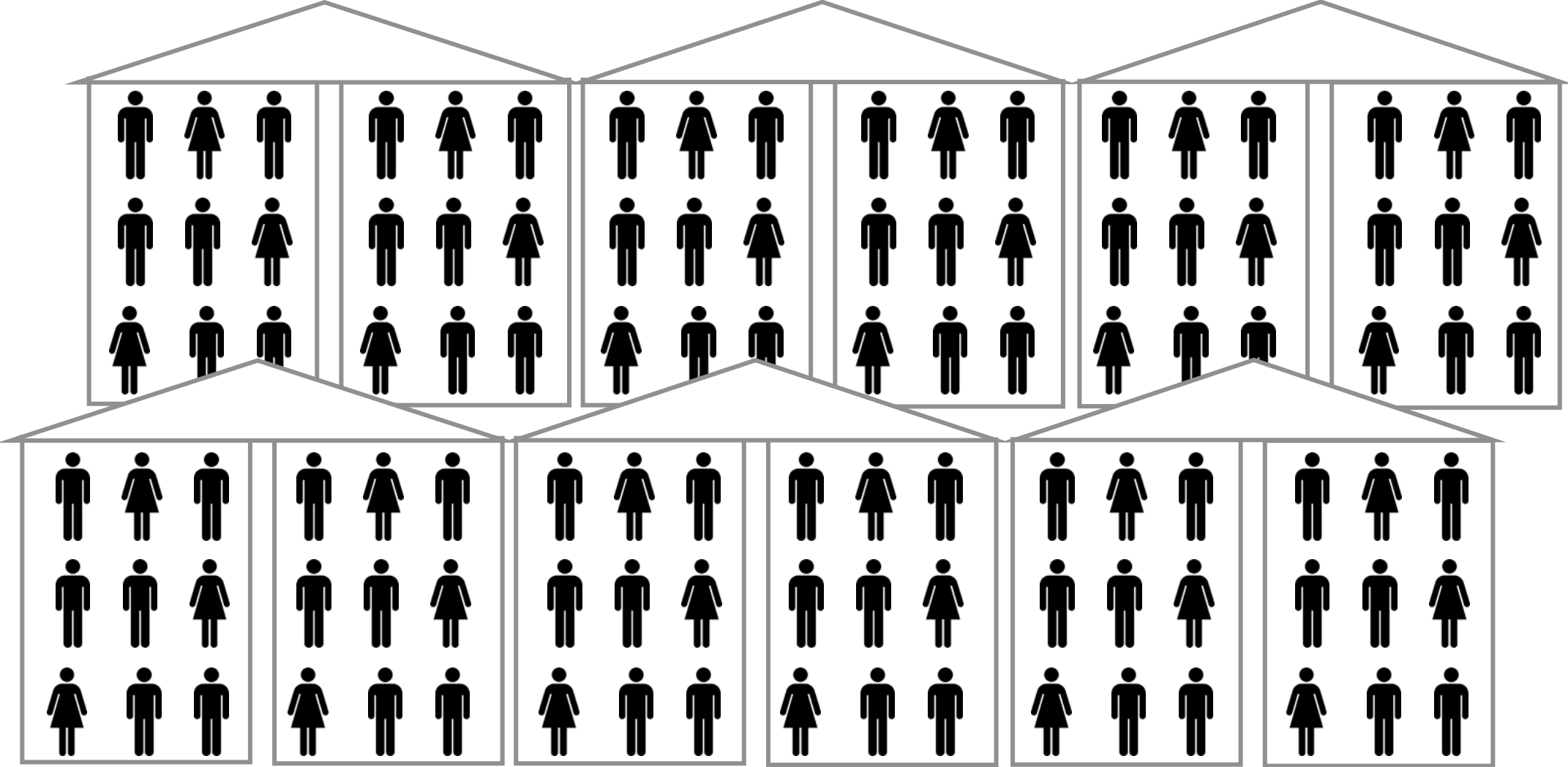
---



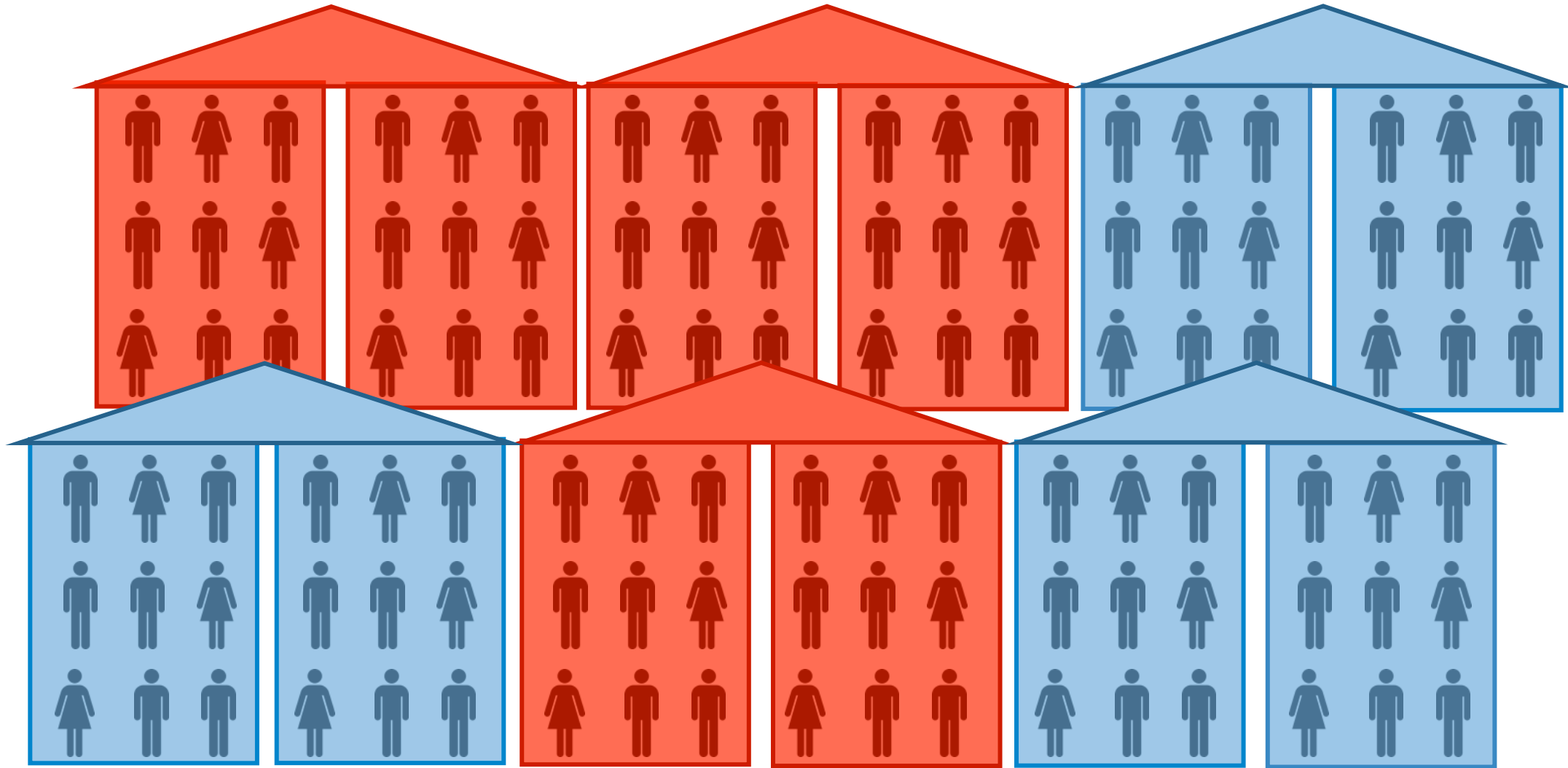
# Unit of Randomization: Class?



# Unit of Randomization: School?



# Unit of Randomization: School?



# How to Choose the Level?

---

- Nature of the Treatment
  - How is the intervention administered?
  - What is the catchment area of each “unit of intervention”
  - How wide is the potential impact?
- Aggregation level of available data
- Power requirements
- Generally, best to randomize at the level at which the treatment is administered.



# Lecture Overview

---

- Unit and method of randomization
- ***Real-world constraints***
- Revisiting unit and method
- Variations on simple treatment-control

# Constraints: Resources

---

- Most programs have limited resources
  - Vouchers, Farmer Training Programs
- Results in more eligible recipients than resources will allow services for
- Limited resources can be an evaluation opportunity

# Constraints: Political Advantages

---

- Not as severe as often claimed
- Lotteries are simple, common and transparent
- Randomly chosen from applicant pool
- Participants know the “winners” and “losers”
- Simple lottery is useful when there is no a priori reason to discriminate
- Perceived as fair
- Transparent

# Constraints: contamination Spillovers/Crossovers

---

- Remember the counterfactual!
- If control group is different from the counterfactual, our results can be biased
- Can occur due to
  - Spillovers
  - Crossovers

# Constraints: logistics

---

- Need to recognize logistical constraints in research designs.
- E.g. individual de-worming treatment by health workers
  - Many responsibilities. Not just de-worming.
  - Serve members from both T/C groups
  - Different procedures for different groups?

# Constraints: fairness, politics

---

- Randomizing at the child-level within classes
- Randomizing at the class-level within schools
- Randomizing at the community-level

# Constraints: sample size

---

- The program is only large enough to serve a handful of communities
- Primarily an issue of statistical power
- Will be addressed tomorrow

# Lecture Overview

---

- Unit and method of randomization
- Real-world constraints
- ***Revisiting unit and method***
- Variations on simple treatment-control



# Possible randomization methods

---

- I. Randomization on who will get the program
  1. Basic Lottery
  2. Randomization Variation: “the bubble”
- II. Randomization on when will the beneficiary get the program
  3. Phase-in
  4. Rotation
- III. Randomization on which beneficiary will get the encouragement to participate in the program

All methods mentioned above can be combined.

---

# I. Randomization on who will get the program

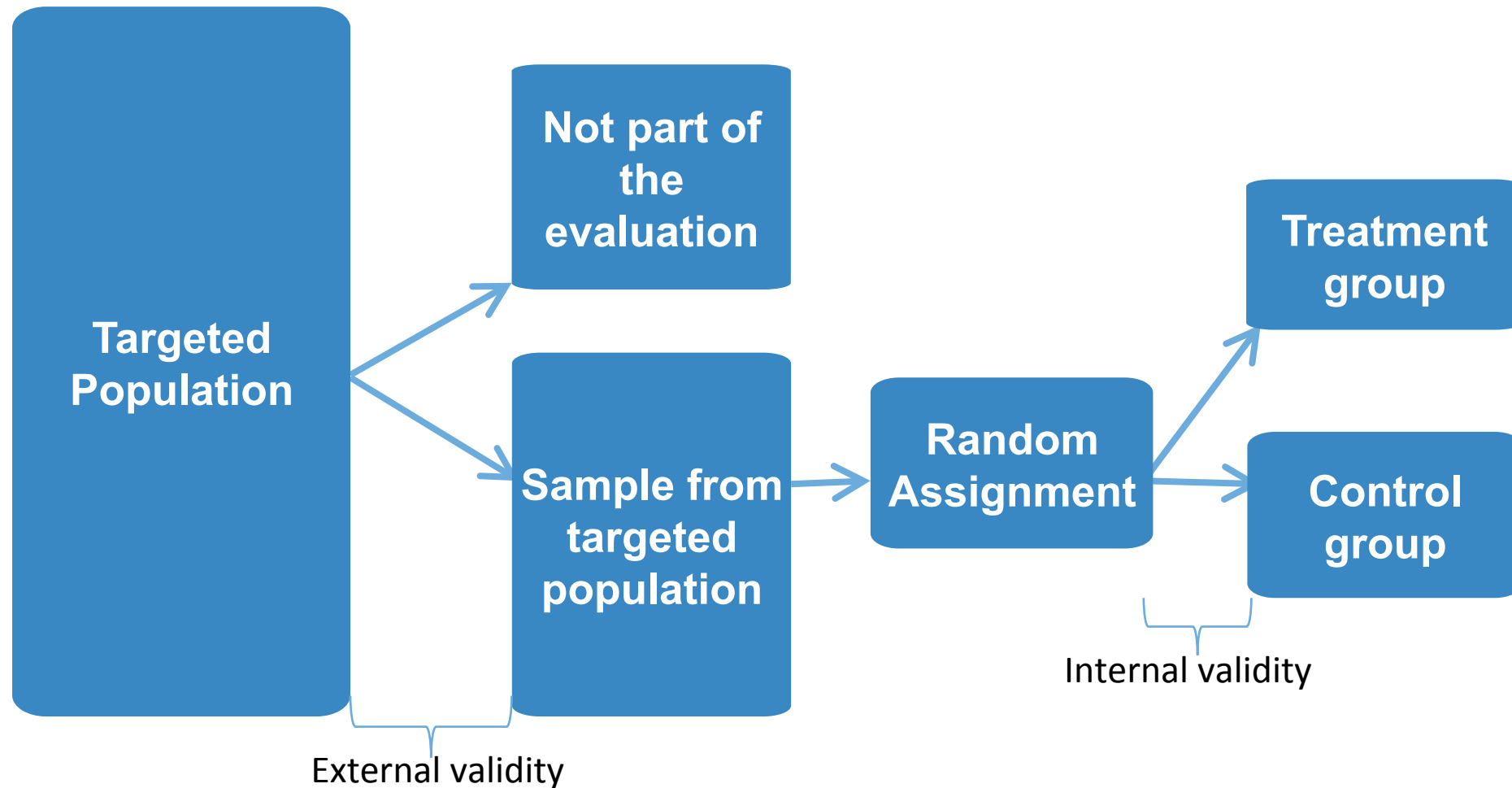
# I. Randomization on who will get the program

---

- Randomly select locations/people to receive or not receive treatment
- Randomization “in the bubble” narrows the sample to a particular subset of subjects

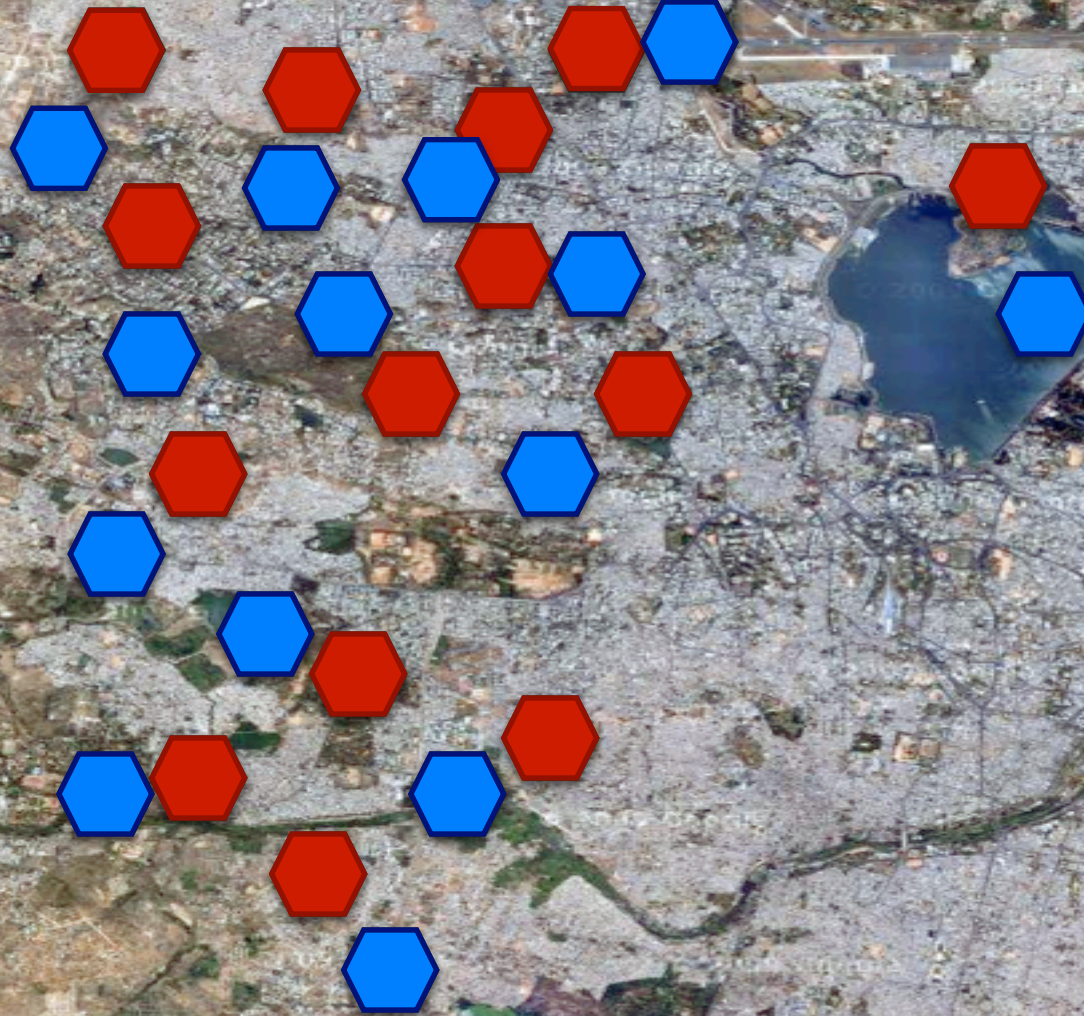
# 1. Basic Lottery RCT structure

---



# Basic Lottery RCT Design

Random  
assignment for  
**treatment** and  
**control group**



# Basic Lottery RCT Impact Measure

---

Take the average outcome measure difference between:

what happened with the program (Treatment Group) ...and  
counterfactual (Control Group)  
**= IMPACT of the Program**

## Data required:

Baseline (depending) and outcome data for control and experimental groups

## 2. Randomization Variation: “the bubble”

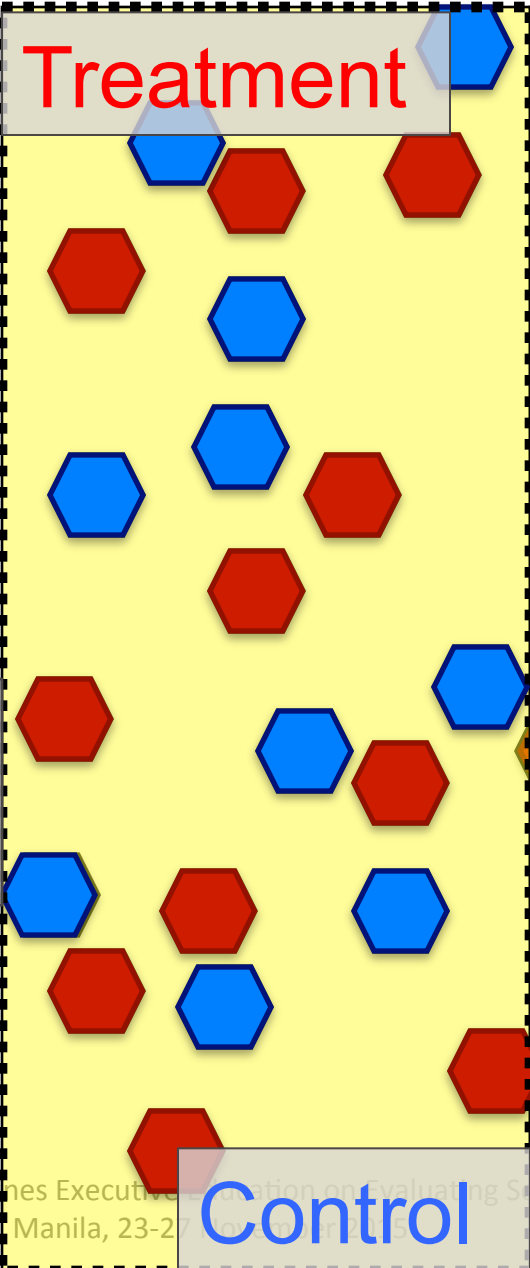
---

- In particular circumstances, evaluation partners may not agree with the randomization of the entire sample which is entitled to the program
- But they may agree to randomization for particular groups (“the bubble”).
- Respondents at the margin/“bubble” are those who may not get treatment or may not be declined treatment.
  - Right above the margin → entitled to treatment, but almost declined
  - Right below the margin → not entitled to treatment, but almost granted
- What treatment effects are we measuring? What are its implications to external validity?

# Randomization Variation: “the bubble”

Within the bubble, compare **treatment** to **control**

Non-participants (scores < 500)



Participants (scores > 700)



---

## II. Randomization on when beneficiaries get the program

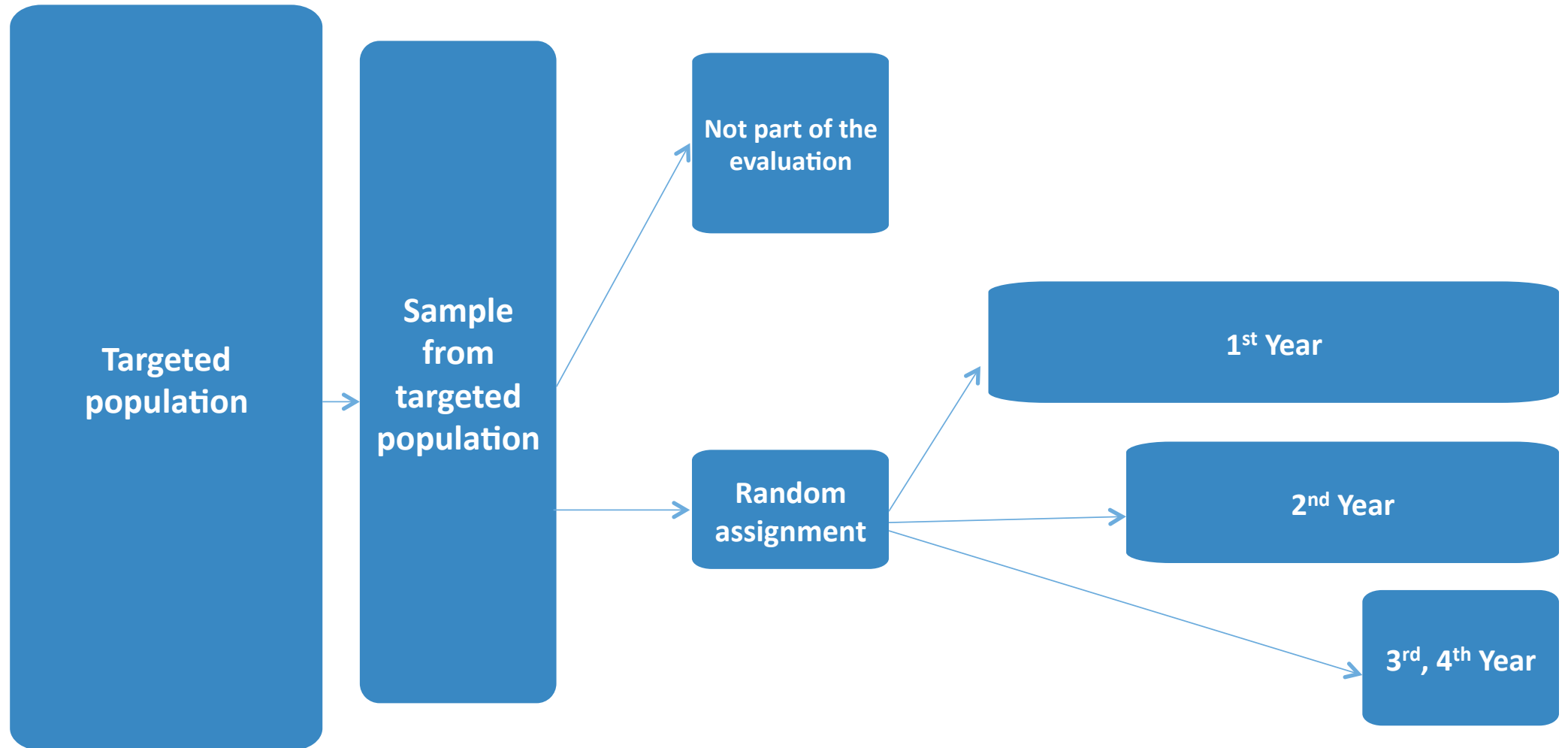
# 3. Phased-In design

---

- Everyone will get the treatment eventually
- This is common method for expanding program with limited resources
- What will decide whether school, branch, etc will receive treatment in certain period?

# Phase-in RCT Structure

---



# Phase-in RCT design

## Round 1

Treatment: 1/3

Control: 2/3

## Round 2

Treatment: 2/3

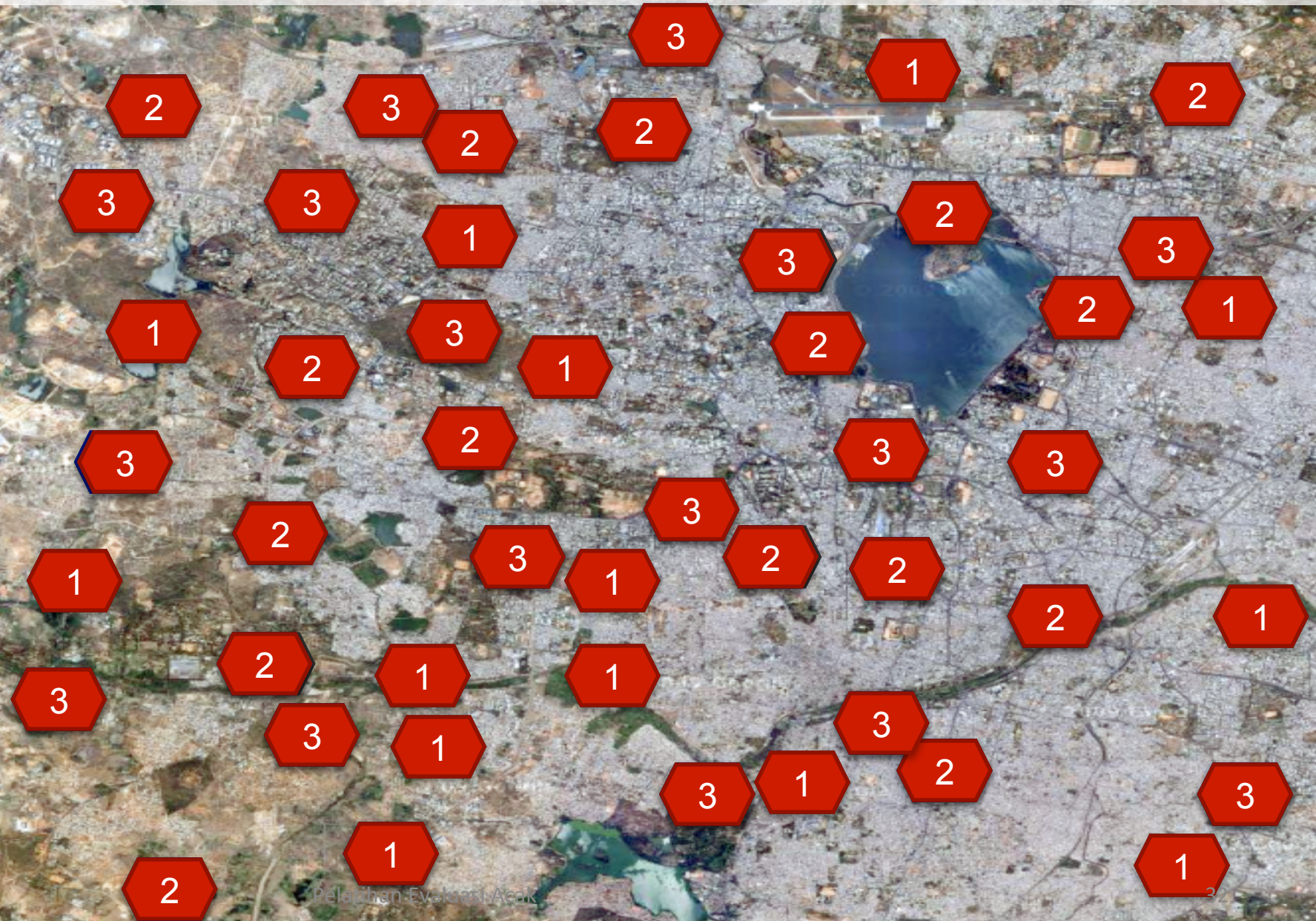
Control: 1/3

Randomized evaluation  
ends

## Round 3

**Treatment:** 3/3

**Control:** 0



# RCT design for phase-in

---

- *Counterfactual:*
  - After 1<sup>st</sup> year, samples which start receiving the intervention in 2<sup>nd</sup> year, 3<sup>rd</sup> year, and so on, will act as “control groups”.
  - After 2<sup>nd</sup> year, samples which start receiving the intervention in 3<sup>rd</sup> year, 4<sup>th</sup> year, and so on, will act as “control groups”.
- Required data
  - Baseline (if necessary) and outcome data
- Considerations:
  - Gradually, control groups will disappear.
  - Control groups will eventually receive the treatment; need further consideration on the impact if control groups already anticipate/expect this.

# RCT design for phase-in

---

After Year 1:

Take the average outcome measure difference between:

what happened with the program (Yr. 1 Treatment Group)

...and counterfactual (groups receiving treatment in Yr. 2,3...)

**= IMPACT of the Program after 1 year of implementation**

After Year 2:

Take the average outcome measure difference between:

Yr. 1 and Yr.2 Treatment Group (pooled)...and

counterfactual (groups receiving treatment in Yr. 3,4...)

**= IMPACT of the Program after 2 years of implementation**

# Phase-in designs

---

## Advantages

- Everyone gets something eventually
- Provides incentives to maintain contact

## Concerns

- Can complicate estimating long-run effects
- Care required with phase-in windows
- Do expectations change actions today?

# 4. Rotation design

---

- Groups get treatment in turns
- Advantages?
- Concerns?



# Rotation design

Round 1

Treatment: 1/2

Control: 1/2

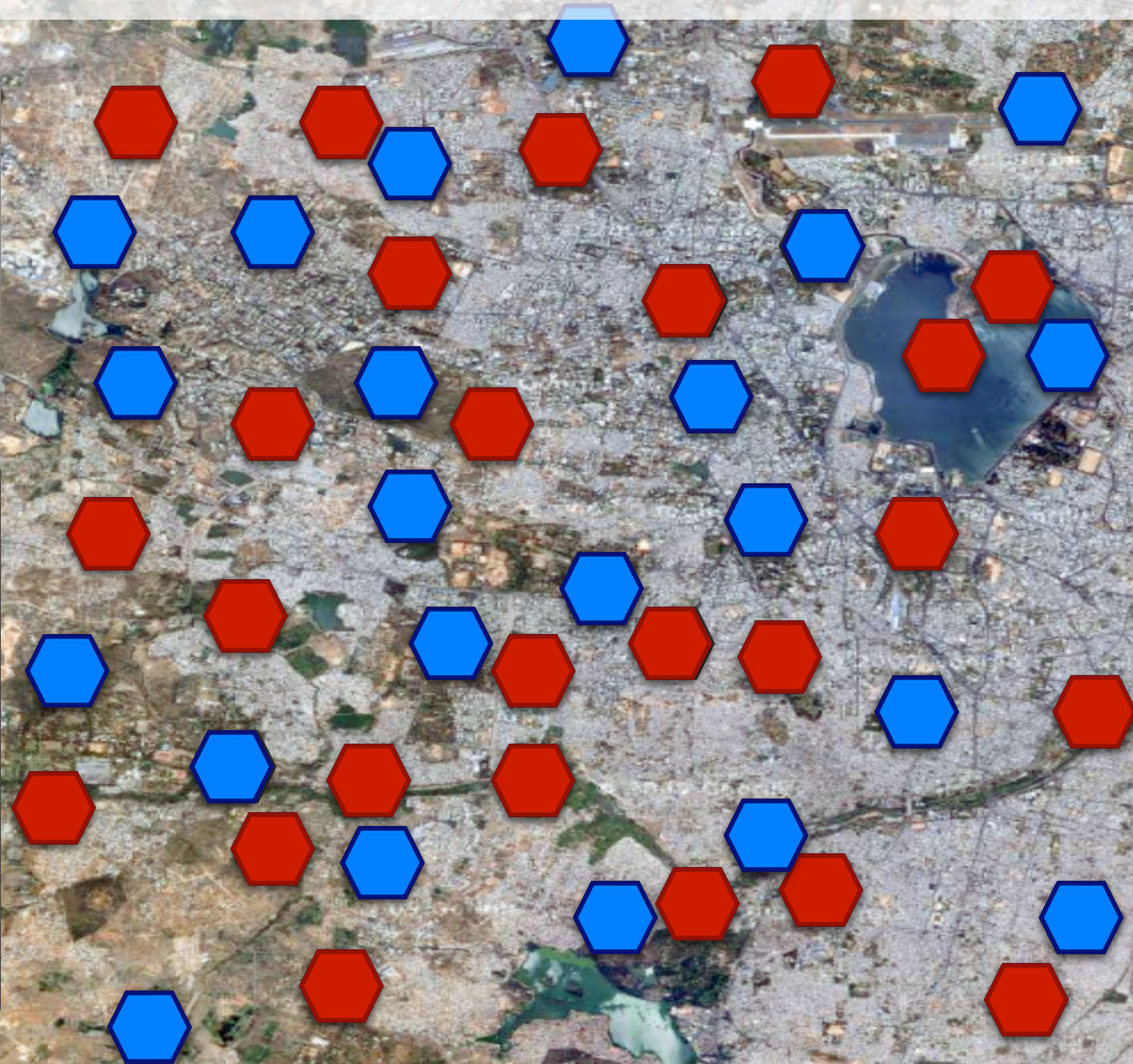
Round 2

Treatment from  
Round 1 →

Control

Control from  
Round 1 →

Treatment



---

# III. Encouragement Design

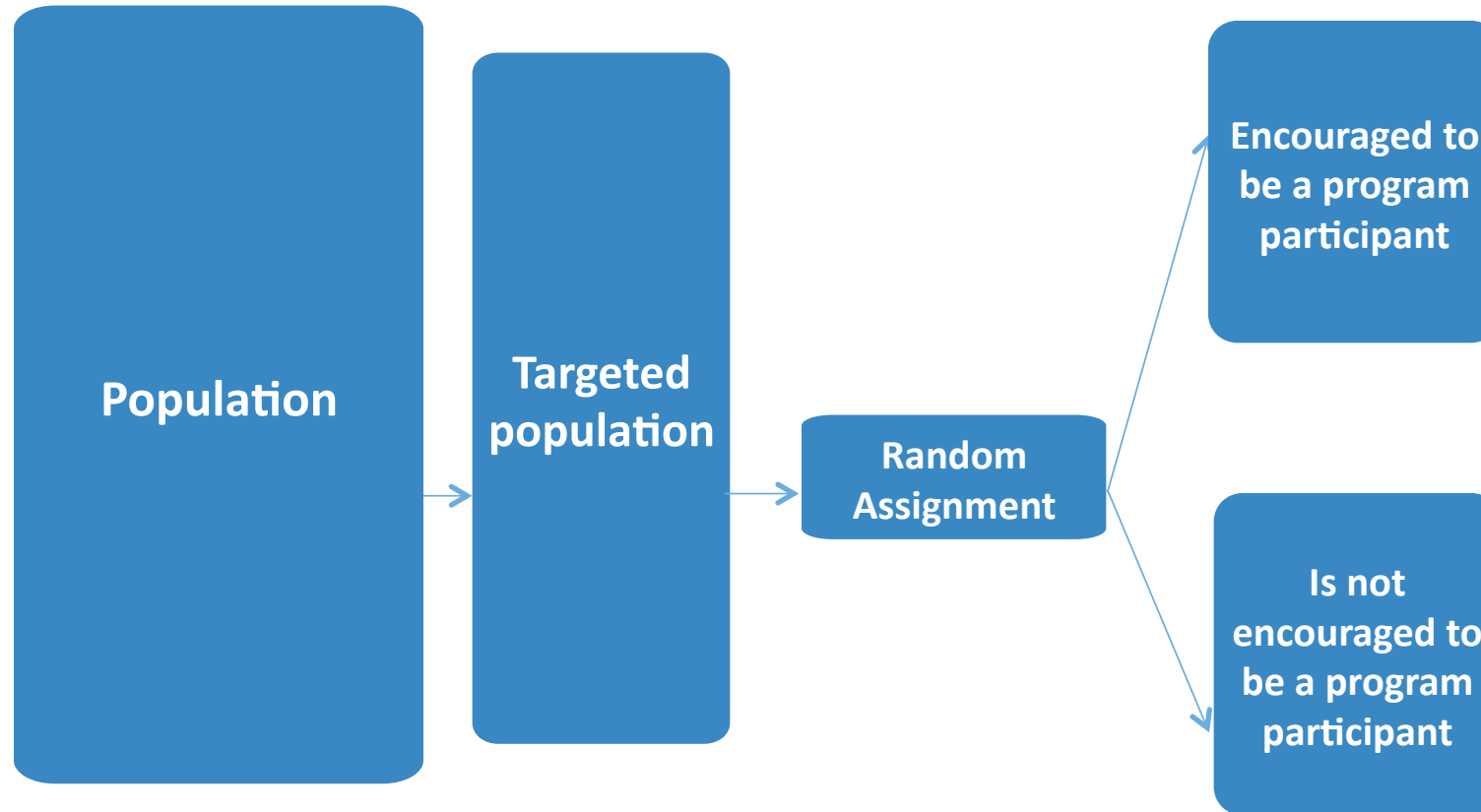
# 5. Encouragement design: What to do when you can't randomize access

---

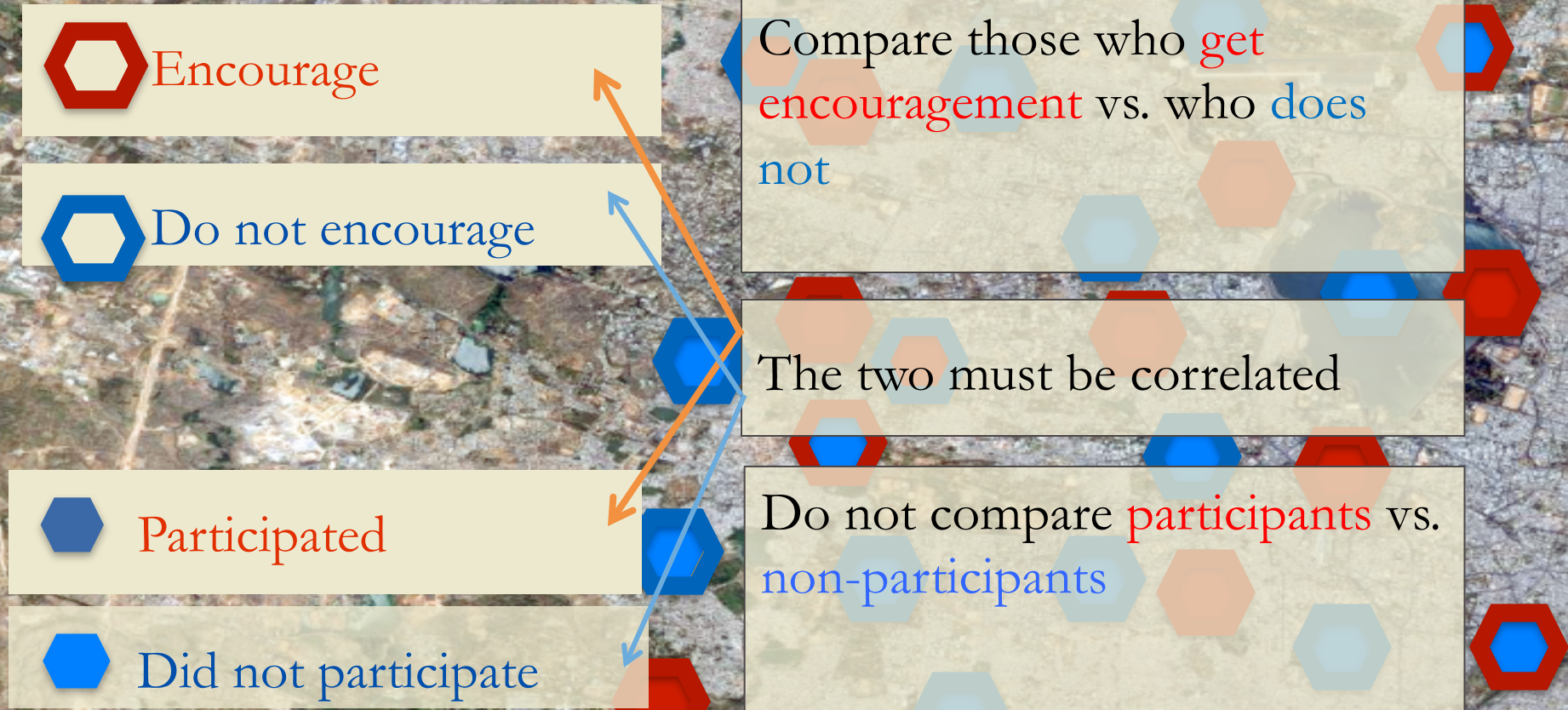
- Sometimes it's practically or ethically impossible to randomize program access
- But most programs have less than 100% take-up
- Randomize encouragement to receive treatment

# Encouragement Design

---



# Encouragement design



# What is “encouragement”?

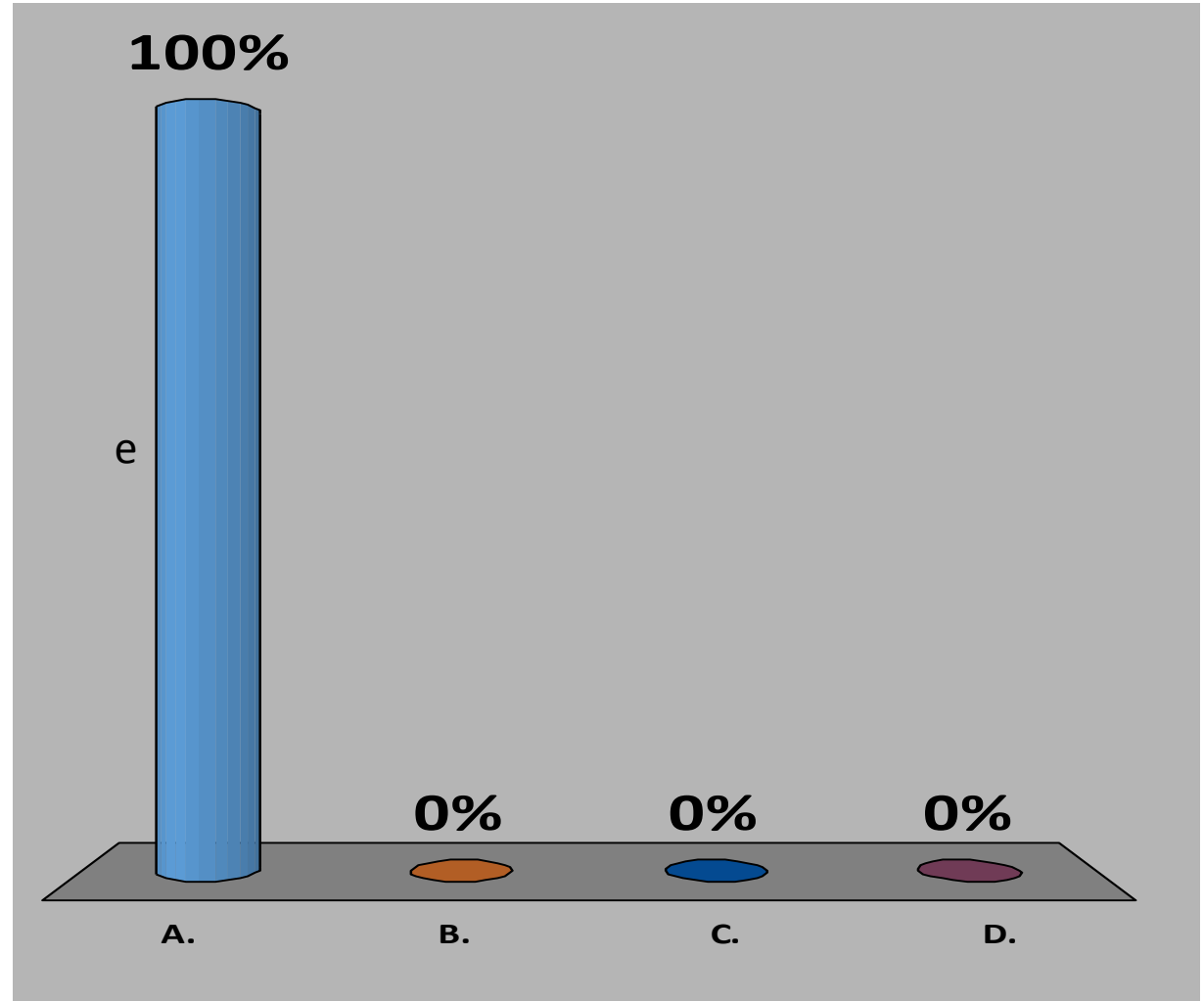
---

- Something that makes some folks more likely to use program than others
- Not itself a “treatment”
- For whom are we estimating the treatment effect?
- Think about who responds to encouragement

# What will you compare for encouragement design?

---

- A. Those who are encouraged vs. who are not
- B. Participant vs. Non-participant
- C. Compliers vs. Non-compliers
- D. Do not know



# Encouragement Design Impact Measure

---

1). Take the difference between:

the average treatment effect w/ encouragement (treatment) ...and  
the average treatment effect w/o encouragement (counterfactual)

2). Divide the difference by the percentage difference of enrollments  
in the encouragement and non-encouragement areas

**= IMPACT of the Program**



# Encouragement Design Impact Measure

---

You launch a universally available job training program and randomly assign certain areas in which individuals receive encouragement to enroll.

You find that the overall percentage of the population that enrolls is 25% higher in encouragement areas. The average income in encouragement areas after one year is \$100; it is \$80 in non-encouragement areas.

The impact of the program is therefore:  $(\$100 - \$80) / .25 = \$25$

---

# Summary of Randomization Methods

# Methods of randomization - recap

Design	Most useful when...	Advantages	Disadvantages
Basic Lottery	<ul style="list-style-type: none"> <li>•Program oversubscribed</li> </ul>	<ul style="list-style-type: none"> <li>•Familiar</li> <li>•Easy to understand</li> <li>•Easy to implement</li> <li>•Can be implemented in public</li> </ul>	<ul style="list-style-type: none"> <li>•Control group may not cooperate</li> <li>•Differential attrition</li> </ul>
Phase-In	<ul style="list-style-type: none"> <li>•Expanding over time</li> <li>•Everyone must receive treatment eventually</li> </ul>	<ul style="list-style-type: none"> <li>•Easy to understand</li> <li>•Constraint is easy to explain</li> <li>•Control group complies because they expect to benefit later</li> </ul>	<ul style="list-style-type: none"> <li>•Anticipation of treatment may impact short-run behavior</li> <li>•Difficult to measure long-term impact</li> </ul>
Rotation	<ul style="list-style-type: none"> <li>•Everyone must receive something at some point</li> <li>•Not enough resources per given time period for all</li> </ul>	<ul style="list-style-type: none"> <li>•More data points than phase-in</li> </ul>	<ul style="list-style-type: none"> <li>•Difficult to measure long-term impact</li> </ul>

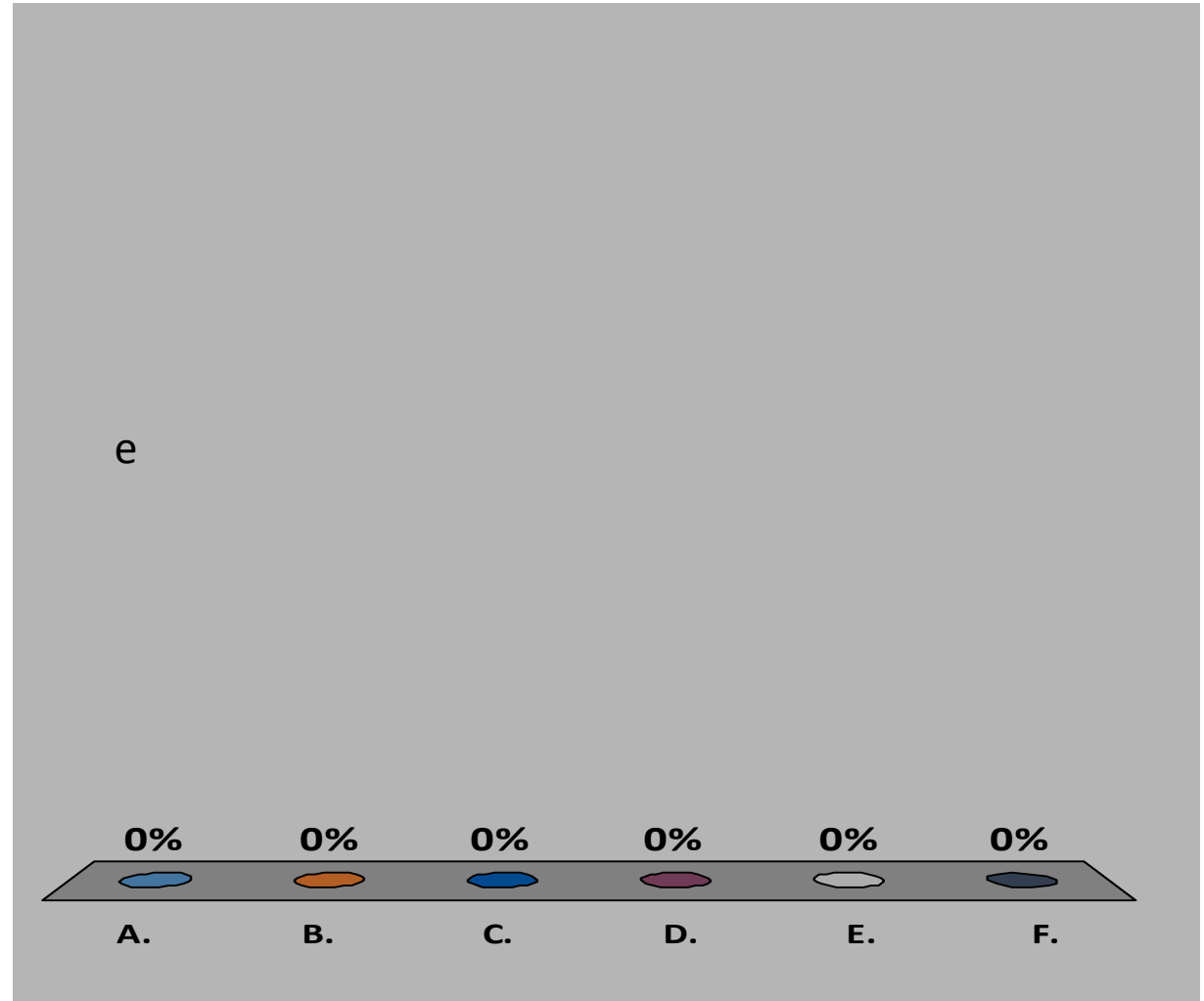
# Methods of randomization - recap

Design	Most useful when...	Advantages	Disadvantages
Bubble	<ul style="list-style-type: none"><li>• Score will determine access to program</li></ul>	<ul style="list-style-type: none"><li>• Flexible assignment for program participants</li></ul>	<ul style="list-style-type: none"><li>• Only measures impact for group within bubble</li></ul>
Encouragement	<ul style="list-style-type: none"><li>• Program has to be open to all comers</li><li>• When take-up is low, but can be easily improved with an incentive</li></ul>	<ul style="list-style-type: none"><li>• Can randomize at individual level even when the program is not administered at that level</li></ul>	<ul style="list-style-type: none"><li>• Measures impact of those who respond to the incentive</li><li>• Need large enough inducement to improve take-up</li><li>• Encouragement itself may have direct effect</li></ul>

What randomization method would you choose if your partner required that everyone receive treatment at some point in time? (Up to 2 responses allowed)

---

- A. Phase-in design
- B. Rotation design
- C. Basic lottery
- D. Randomization in the bubble
- E. Encouragement
- F. Don't know



# Lecture Overview

---

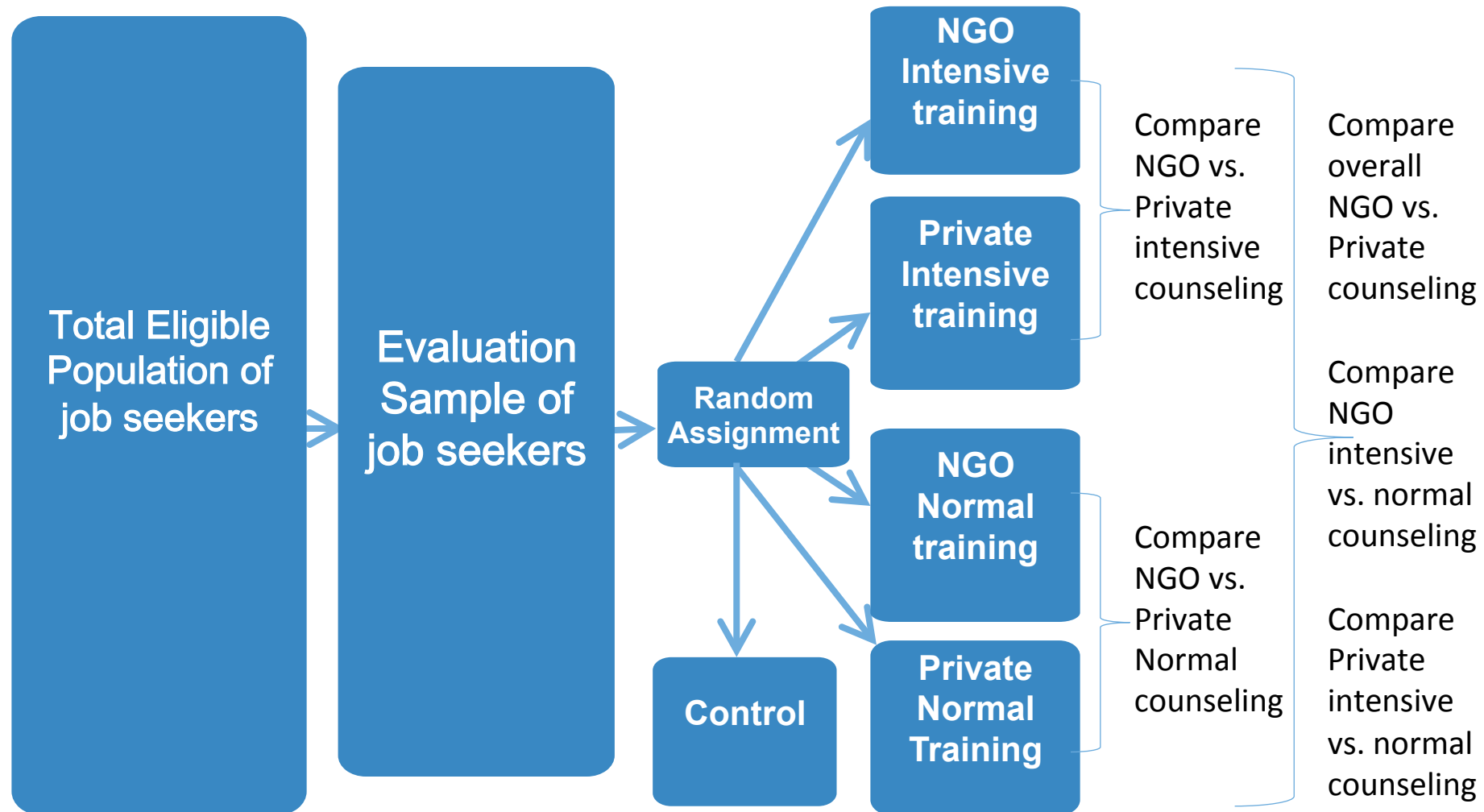
- Unit and method of randomization
- Real-world constraints
- Revisiting unit and method
- ***Variations on simple treatment-control***

# Multiple treatments

---

- Sometimes core question decides among different possible interventions
- You can randomize these programs
- Does this teach us about the benefit of any one intervention?
- Do you have a control group?

# Multi-Arm RCT Structure



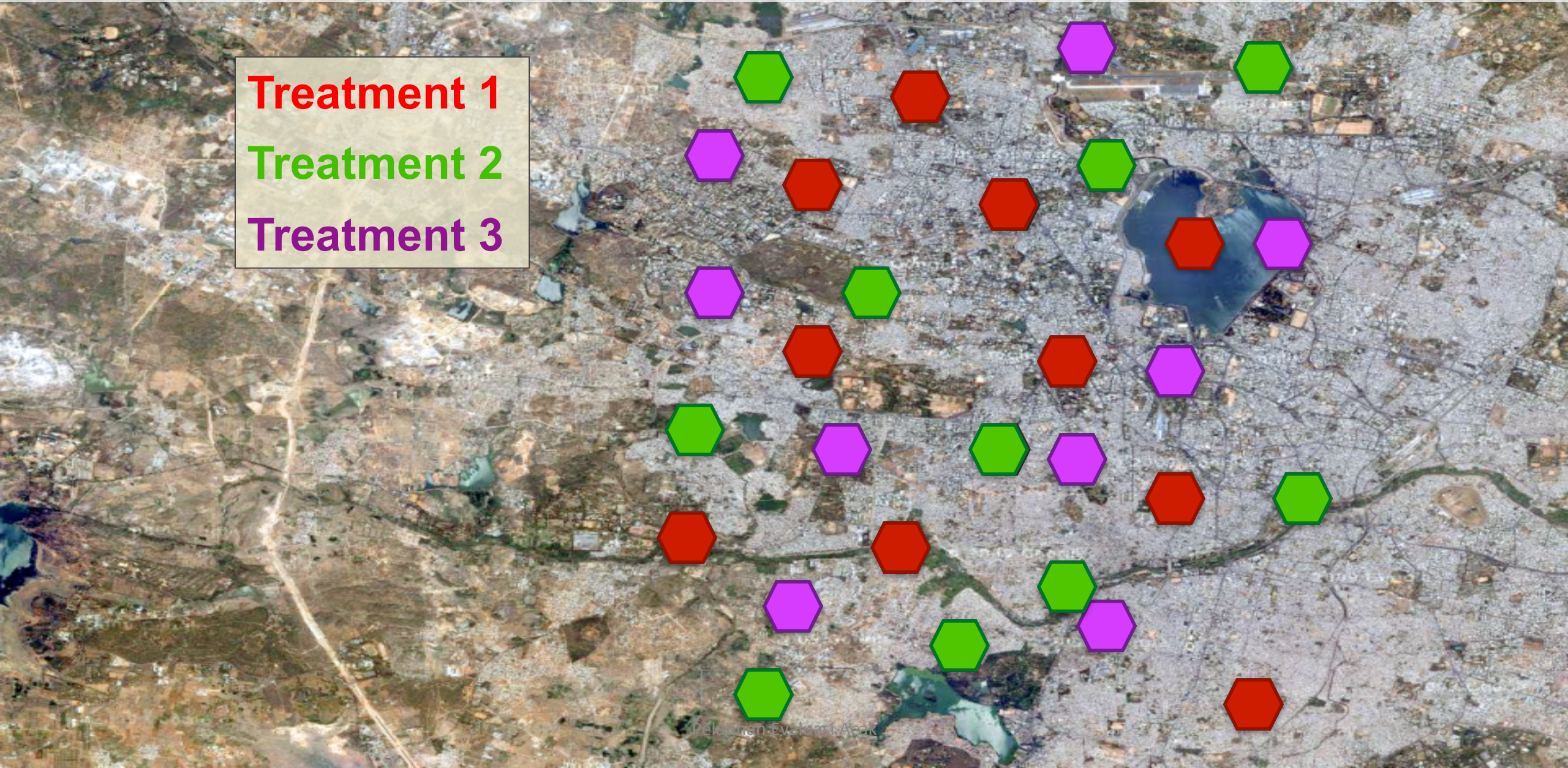


# Multi-Arm RCT Structure

**Treatment 1**

**Treatment 2**

**Treatment 3**



# Multi-Arm Lottery RCT Structure

---

Actor	Intensive Training	Basic Training
NGO	Group 1	Group 2
Private Sector	Group 3	Group 4
	Control (No training)	

# Multi-Arm Lottery RCT Impact Measure

---

Measuring Training vs. No Training

Actor	Intensive Training	Basic Training
NGO	Group 1	Group 2
Private Sector	Group 3	Group 4
	...with THIS	

Compare THIS

# Multi-Arm Lottery RCT Impact Measure

Measuring NGO vs. Private Sector Training

Actor	Intensive Training	Basic Training
NGO	Group 1	Group 2
Private Sector	Group 3	Group 4
	Control (No training)	

**Compare THIS**

**...with THIS**

# Multi-Arm Lottery RCT Impact Measure

---

Measuring Intensive vs. Basic Training

Actor	Intensive Training	Basic Training
NGO	Group 1	Group 2
Private Sector	Group 3	Group 4
	Control (No training)	

**Compare THIS** (Intensive Training groups) **...with THIS** (Basic Training groups)

# Multi-Arm Lottery RCT Impact Measure

Measuring Basic Training NGO vs. Intensive Training Private

Actor	Intensive Training	Basic Training
NGO	Group 1	...with THIS Group 2
Private Sector	Compare THIS Group 3	Group 4
	Control (No training)	

# Cross-cutting treatments

---

- To test:
  - Different components of treatment in different combinations
  - whether components serve as substitutes or compliments
- What is the most cost-effective combination?
- Advantage: win-win for operations, can help answer questions for them, beyond simple “impact”!

# Varying levels of treatment

---

- Some schools are assigned full treatment
  - All kids get pills
- Some schools are assigned partial treatment
  - 50% are designated to get pills
- Testing subsidies and prices



# Stratification

---

- Objective: balancing your sample when you have a small sample
- What is it:
  - dividing the sample into different subgroups
  - selecting treatment and control from each subgroup
- What happens if you don't stratify?

# When to stratify?

---

- Stratify on variables that could have important impact on outcome variable
- Stratify on subgroups that you are particularly interested in (where you may think impact of program may be different)
- Stratification more important with small sample frame
- You can also stratify on index variables you create
- Can stratify closely on one continuous variable or coarsely on multiple
  - Baseline value of Primary Outcome Variable
- Can get complex to stratify on too many variables
- Makes the draw less transparent the more you stratify
- Degrees of freedom

# Mechanics of randomization

---

- Need sample frame
- Pull out of a hat/bucket
- Use random number generator in spreadsheet program to order observations randomly
- Stata program code
- What if no existing list?



Source: Chris Blattman

# Sampling and Sample Size

---

Manila, November 2015



ABDUL LATIF JAMEEL  
Poverty Action Lab

J-PAL SOUTHEAST ASIA

# Course Overview

---

1. What is evaluation?
2. Measuring impacts (outcomes, indicators)
3. Why randomize?
4. How to randomize
5. Threats and Analysis
6. Sampling and sample size
7. RCT: Start to Finish
8. Cost Effectiveness Analysis and Scaling Up

# Course Overview

---

1. What is evaluation?
2. Measuring impacts (outcomes, indicators)
3. Why randomize?
4. How to randomize
5. Threats and Analysis
- 6. *Sampling and sample size***
7. RCT: Start to Finish
8. Cost Effectiveness Analysis and Scaling Up

# Learning Objectives

---

At the end of the presentation, you will:

1. Know the **Central Limit Theorem** and the **Law of Large Numbers**, and why they matter
2. Know the difference between **Type I** and **Type II** error
3. Know what the “**power**” of a study is and why you should care
4. Be ready to tackle the power exercise in the next session

# What's the average result?

---

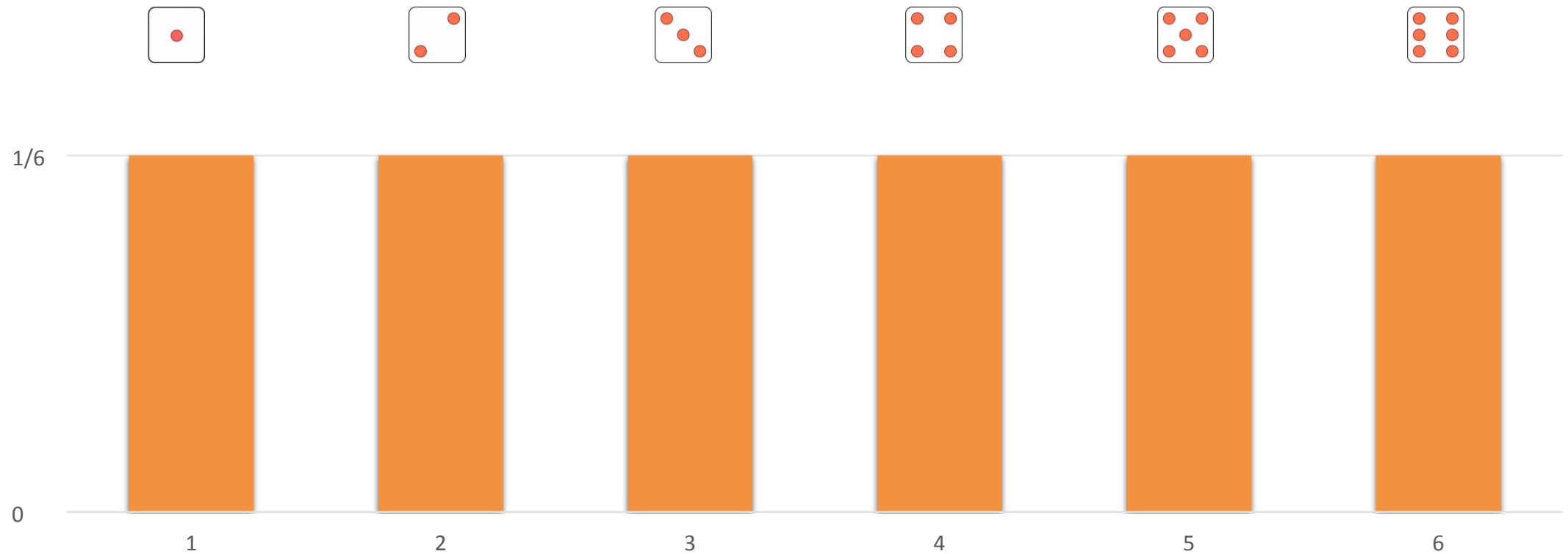
- If you were to roll a die once, what would be the “expected result”? (i.e. the average)





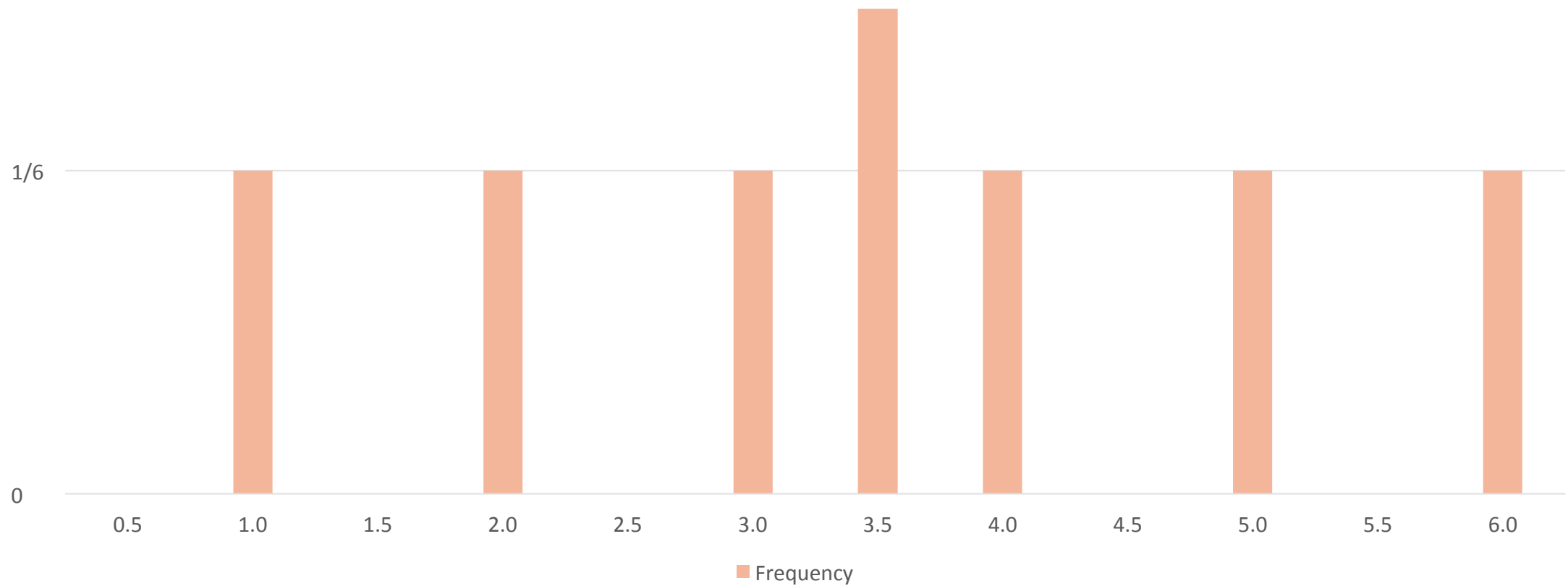
# Possible results & probability: 1 die

---



# Rolling 1 die: possible results & average

---



# What's the average result?

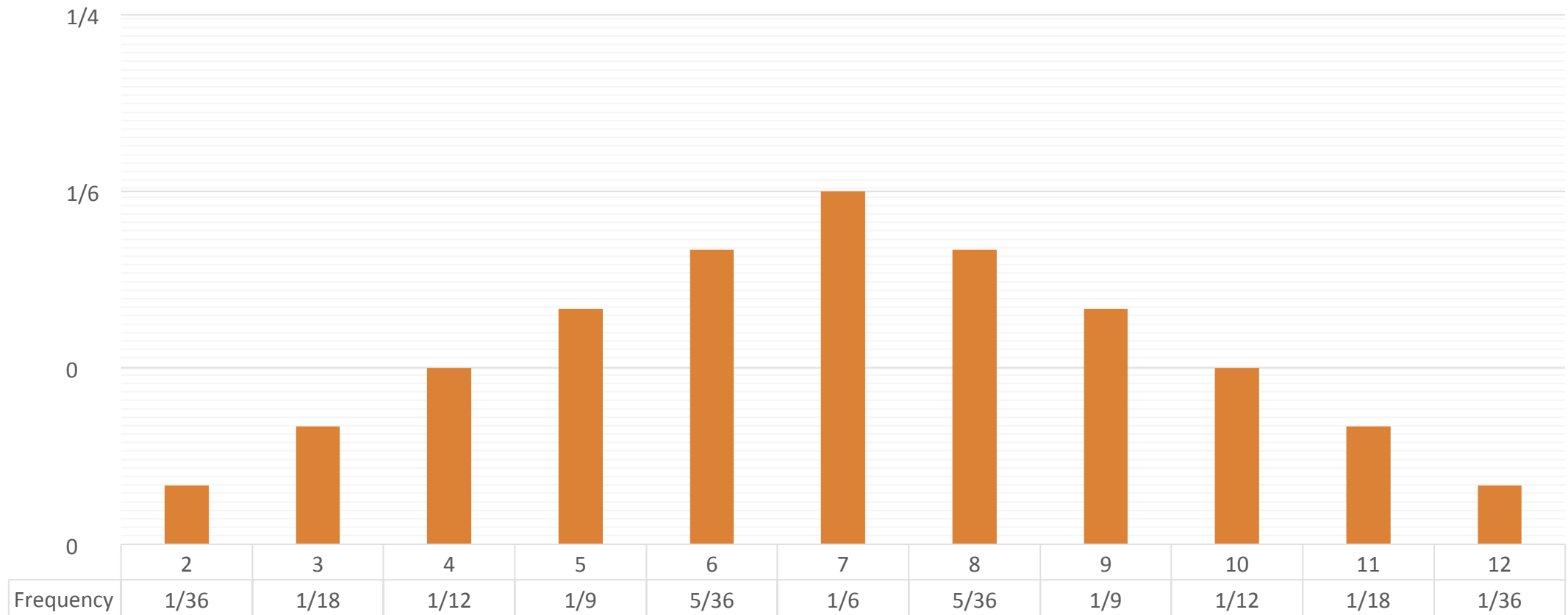
---

- If you were to roll two dice once, what would be the expected average of the two dice?




# Rolling 2 dice: Possible **totals** & likelihood




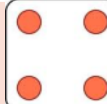

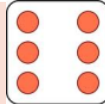






---



# Rolling 2 dice: possible totals

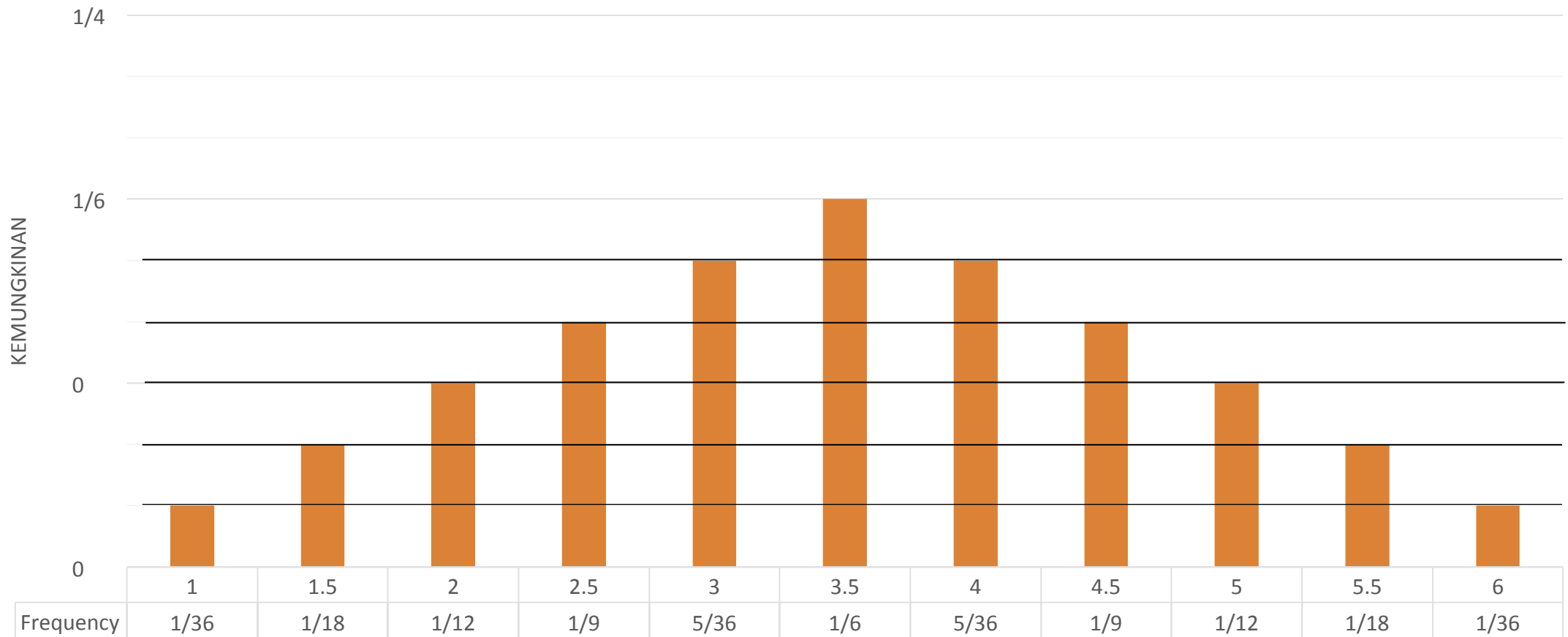
12 possible **totals**, 36 permutations



		Die 1					
							
Die 2		2	3	4	5	6	7
		3	4	5	6	7	8
		4	5	6	7	8	9
		5	6	7	8	9	10
		6	7	8	9	10	11
		7	8	9	10	11	12

# Rolling 2 dice:

## Average score of dice & likelihood



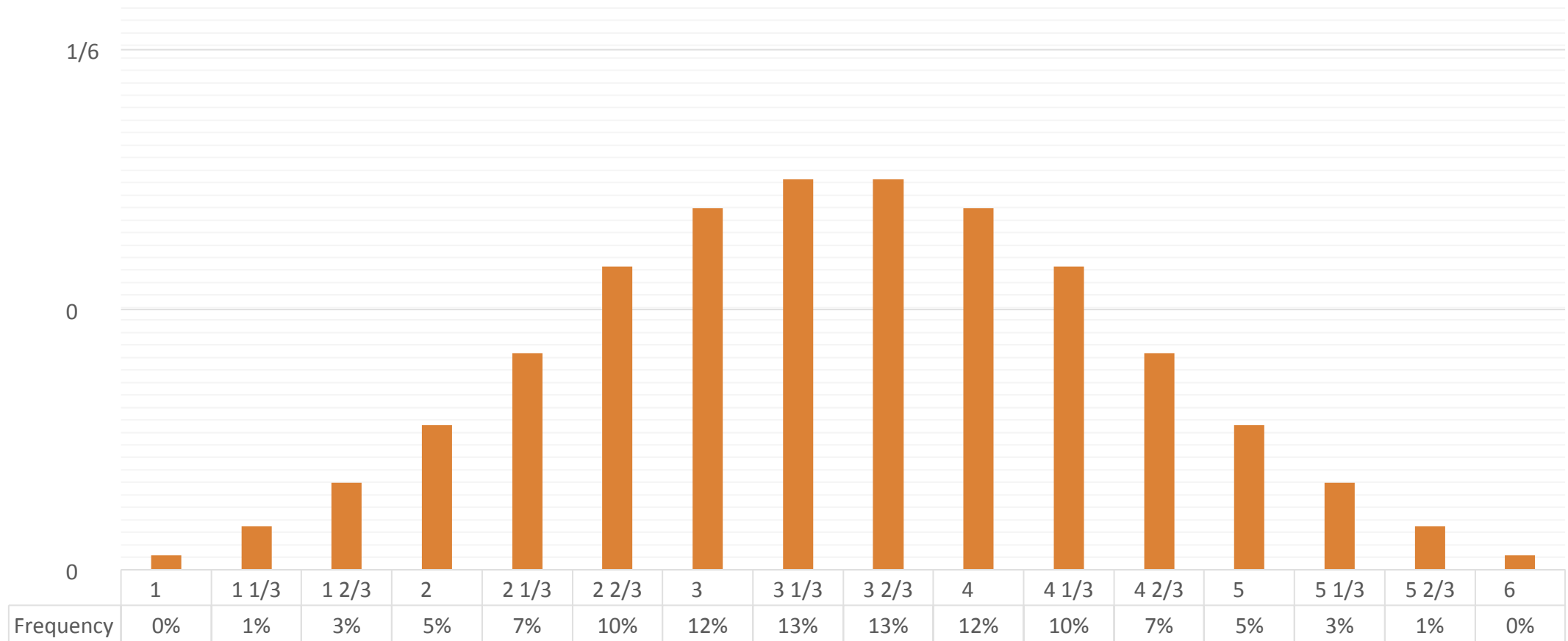
# Outcomes and Permutations

---

- Putting together permutations, you get:
  1. All possible outcomes
  2. The likelihood of each of those outcomes
- Each column represents one possible outcome (average result)
- Each block within a column represents one possible permutation (to obtain that average)

# Rolling 3 dice: 16 results 3→18, 216 permutations

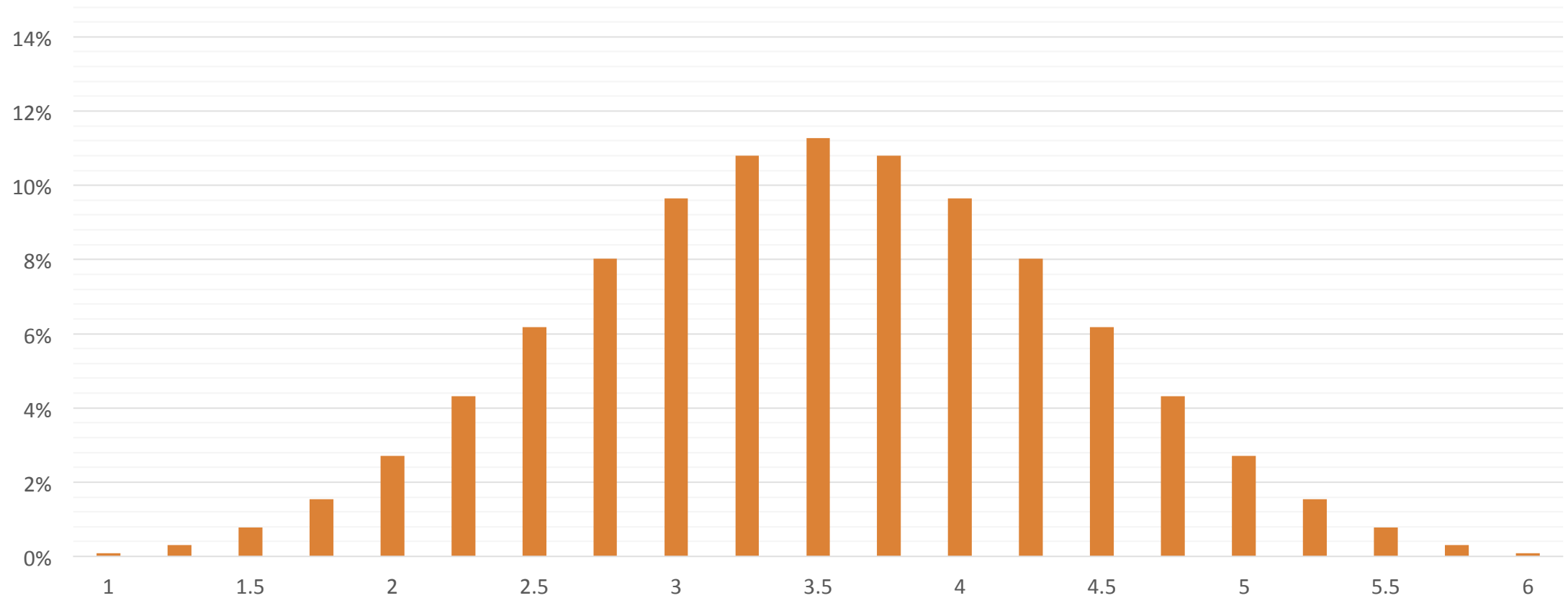
---





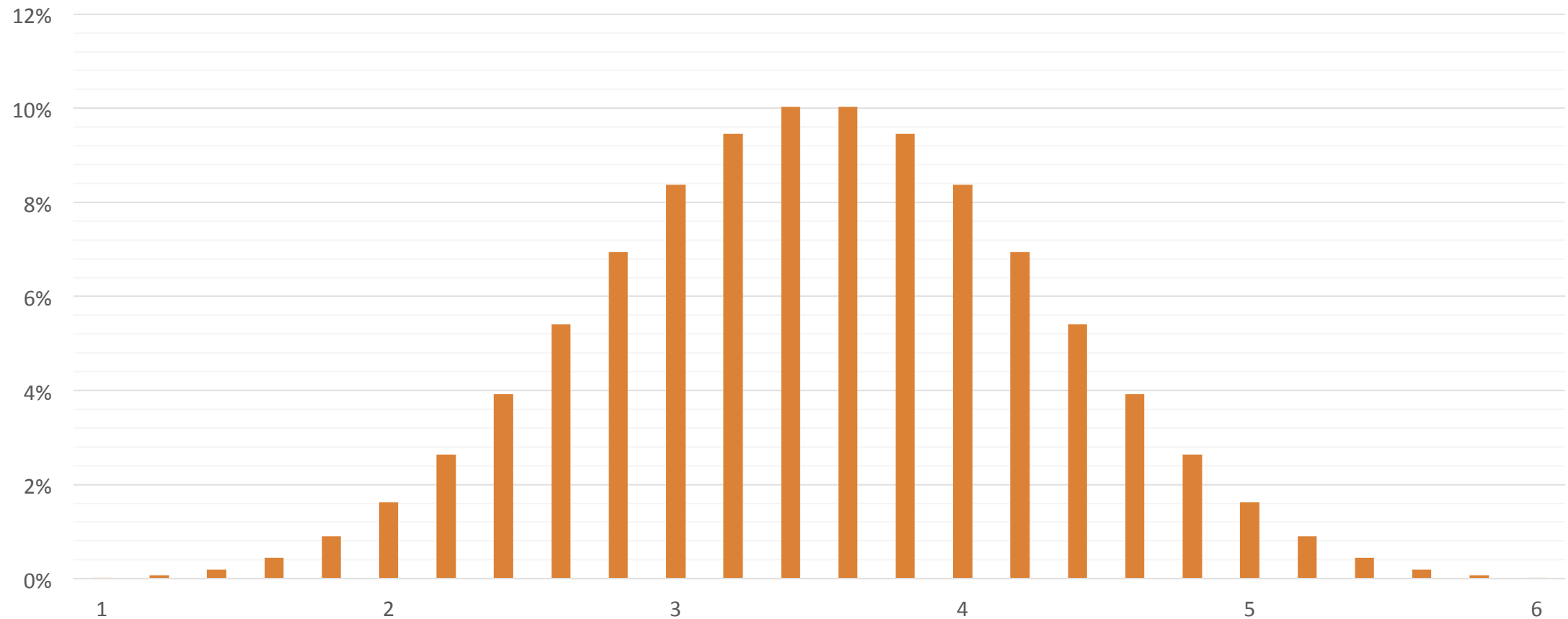
# Rolling 4 dice: 21 results, 1296 permutations

---



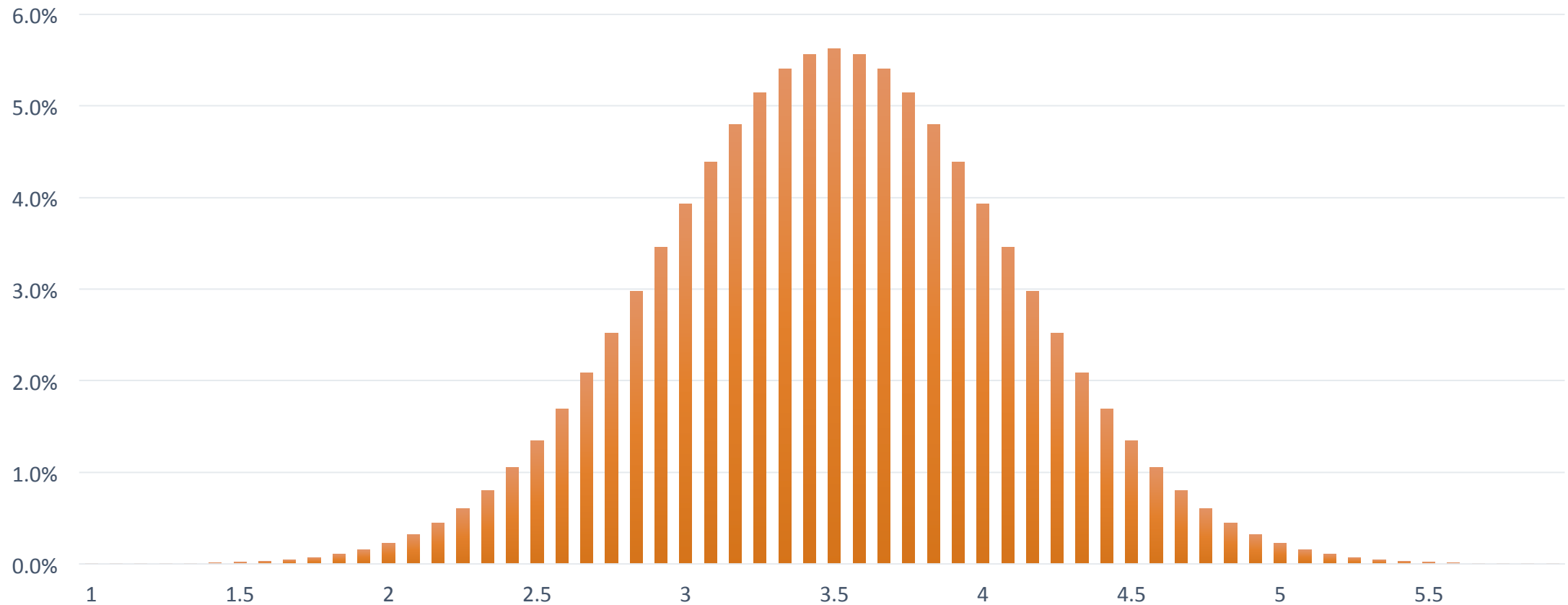
# Rolling 5 dice: 26 results, 7776 permutations

---



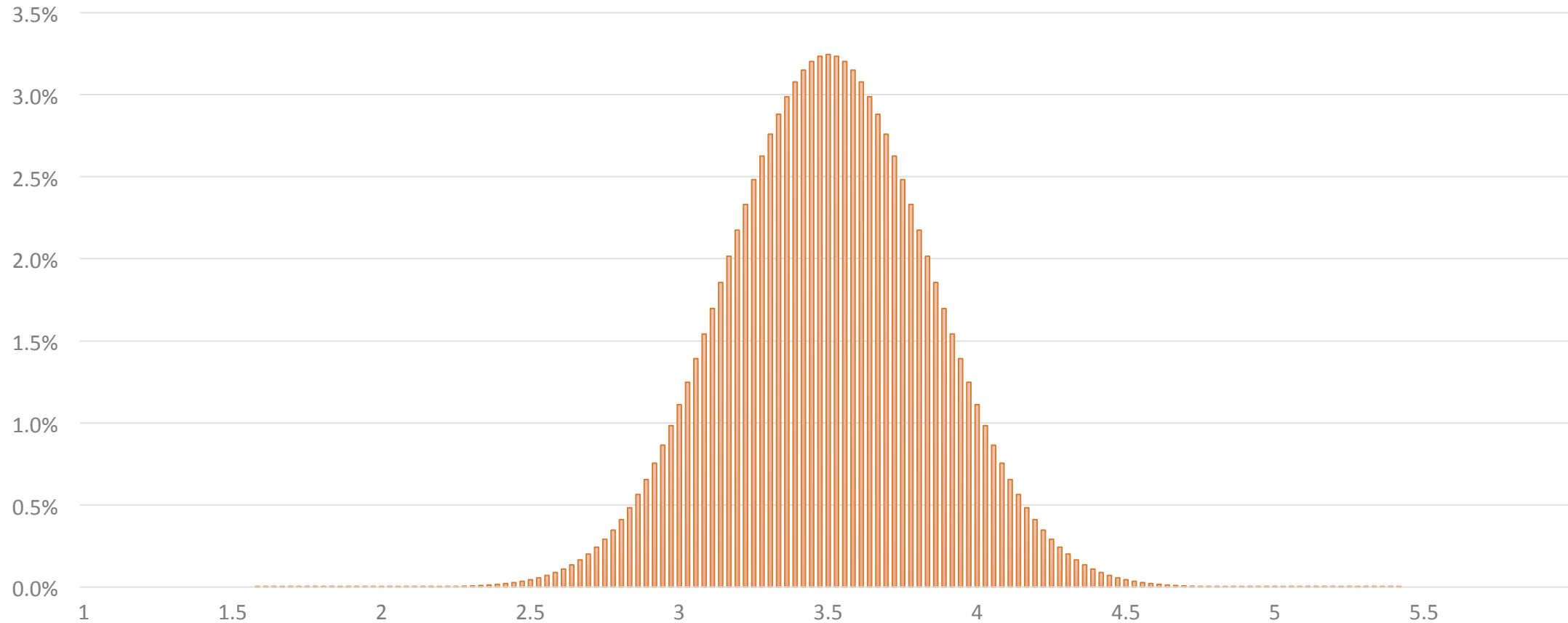
# Rolling 10 dice: 50 results, >60 million permutations

---



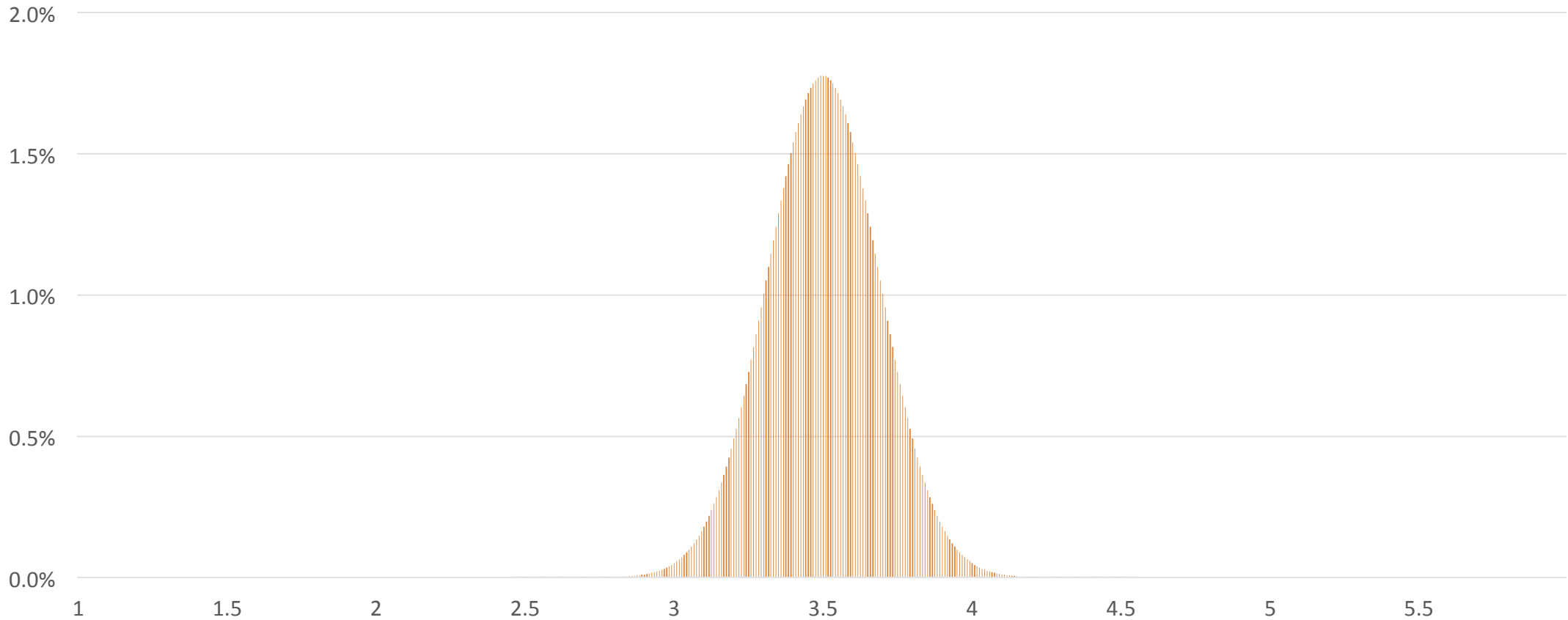
# Rolling 30 dice: 150 results, $2 \times 10^{23}$ permutations

---



# Rolling 100 dice: 500 results, $6 \times 10^{77}$ permutations

---



# Rolling dice: 2 lessons

---

1. The more dice you roll, the closer most averages are to the true average (the distribution gets “tighter”)  
**-THE LAW OF LARGE NUMBERS-**
2. The more dice you roll, the more the distribution of possible averages (the *sampling distribution*) looks like a bell curve (a *normal* distribution)  
**-THE CENTRAL LIMIT THEOREM-**

# So Why Do We Care?

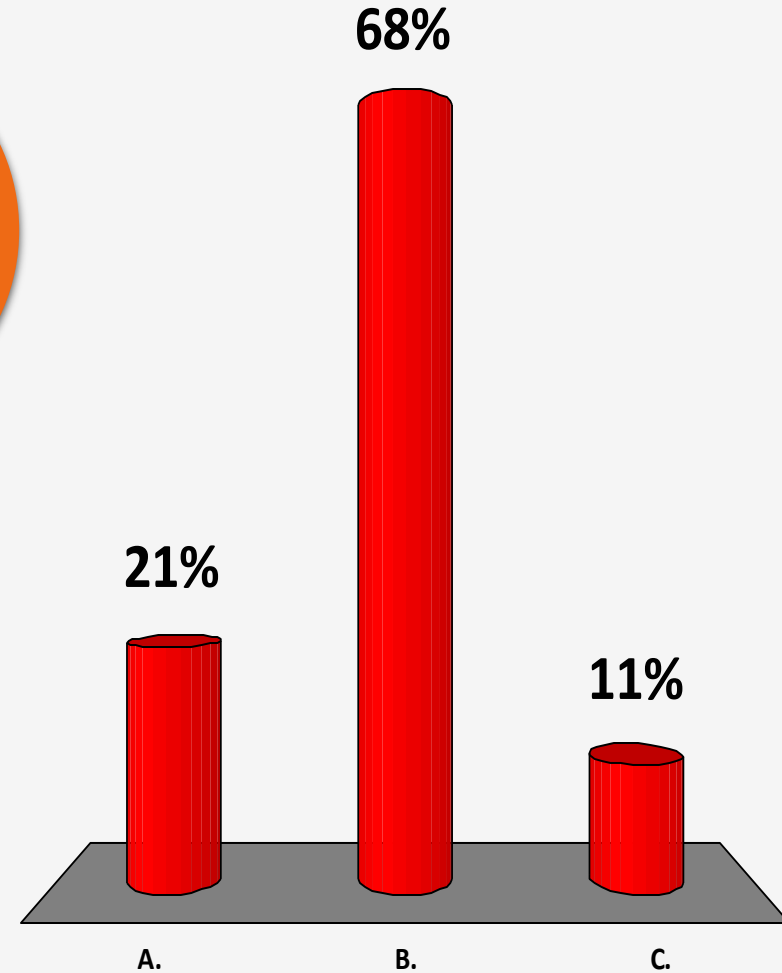
---

- Sampling distribution is probability distributions
- Sampling distribution is a bell curve (irrespective of what underlying distribution is)
- Why does it matter?
- Why do we care if the probability distribution looks like a bell curve?
- **Because we know how to calculate the area underneath!**

# Which of these is more accurate?

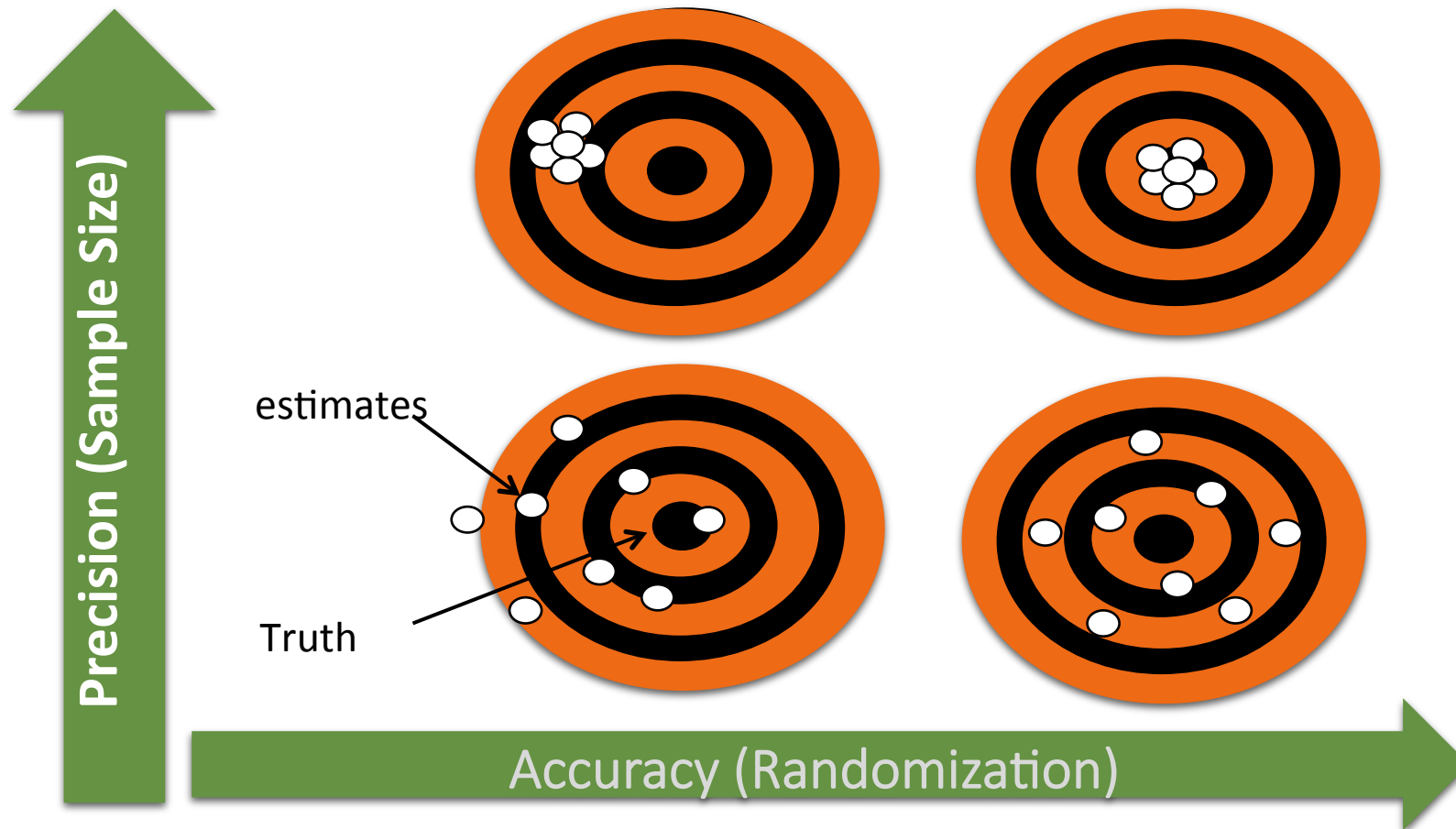


- A. I.
- B. II.
- C. Don't know

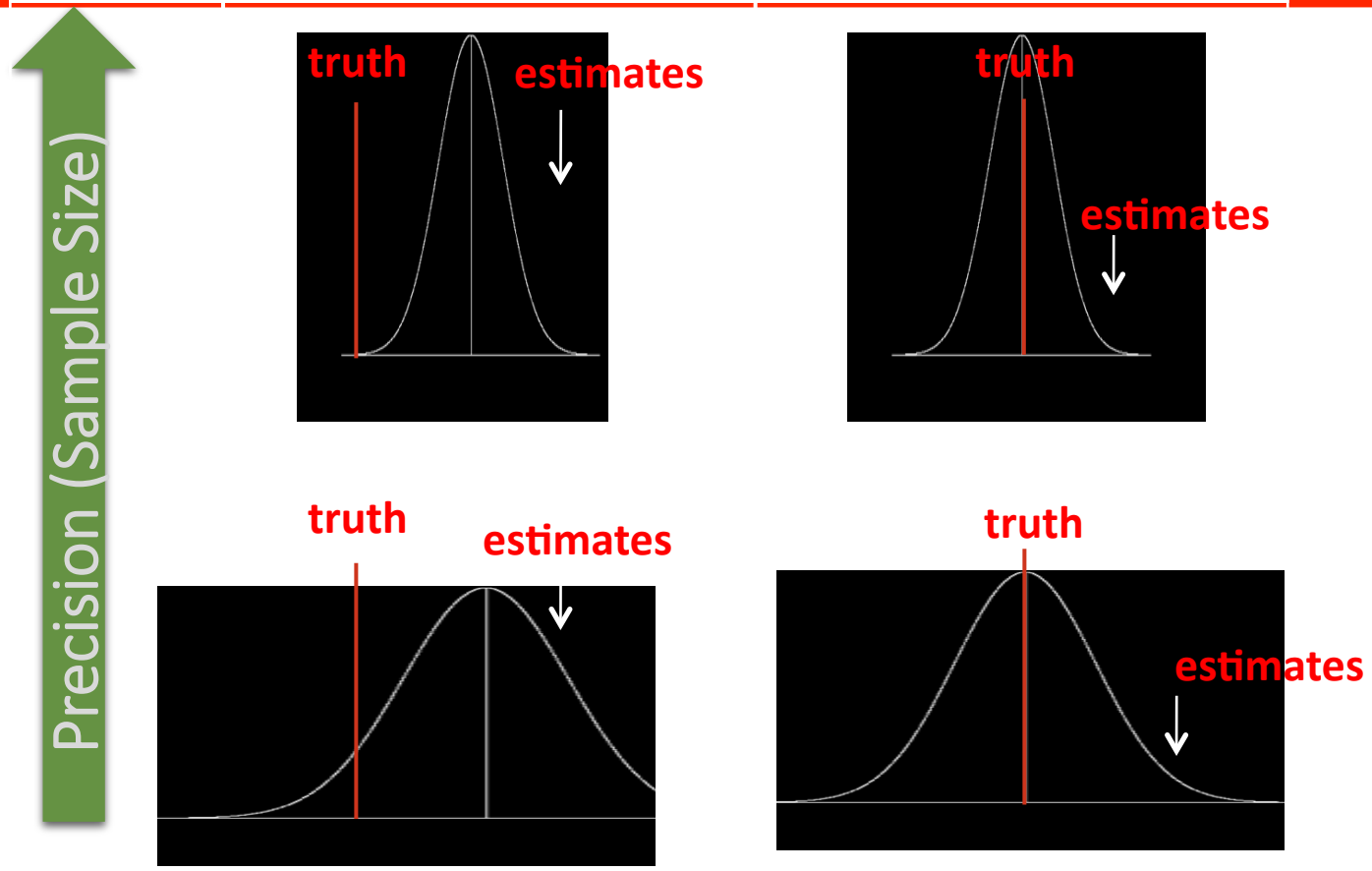




# Accuracy versus Precision



# Accuracy versus Precision



# THE basic questions in statistics

---

- How confident can you be in your results?
  
- → How big does your sample need to be?

# Outline

---

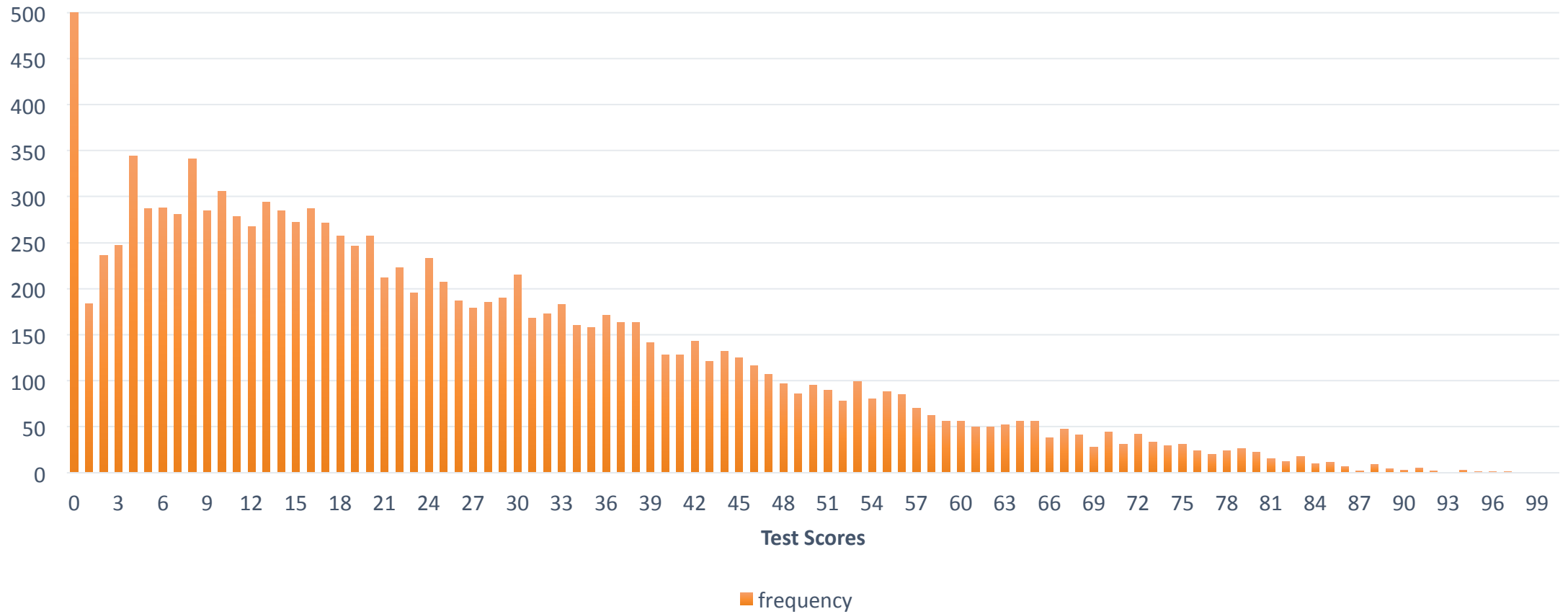
- Sampling distributions
  - Population distribution
  - Sampling distribution
  - Law of large numbers/central limit theorem
  - Standard deviation and standard error
- Detecting impact

# Outline

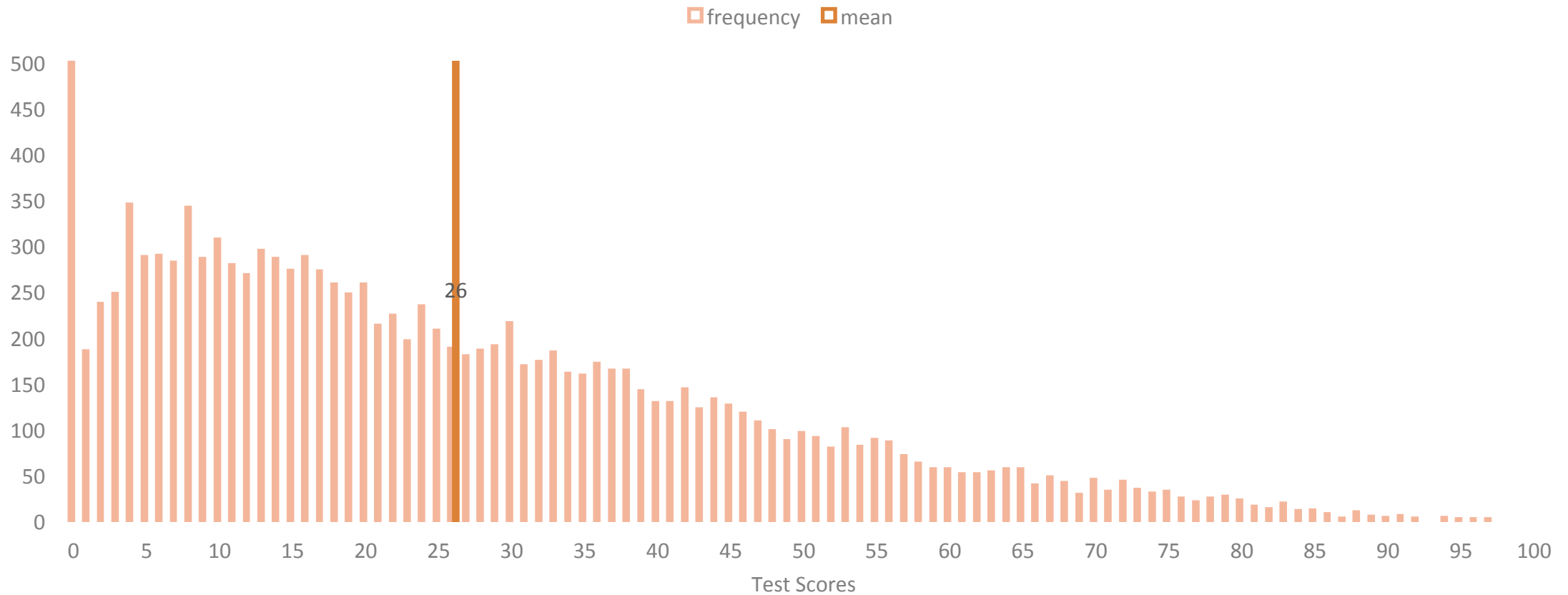
---

- ***Sampling distributions***
  - ***Population distribution***
  - Sampling distribution
  - Law of large numbers/central limit theorem
  - Standard deviation and standard error
- Detecting impact

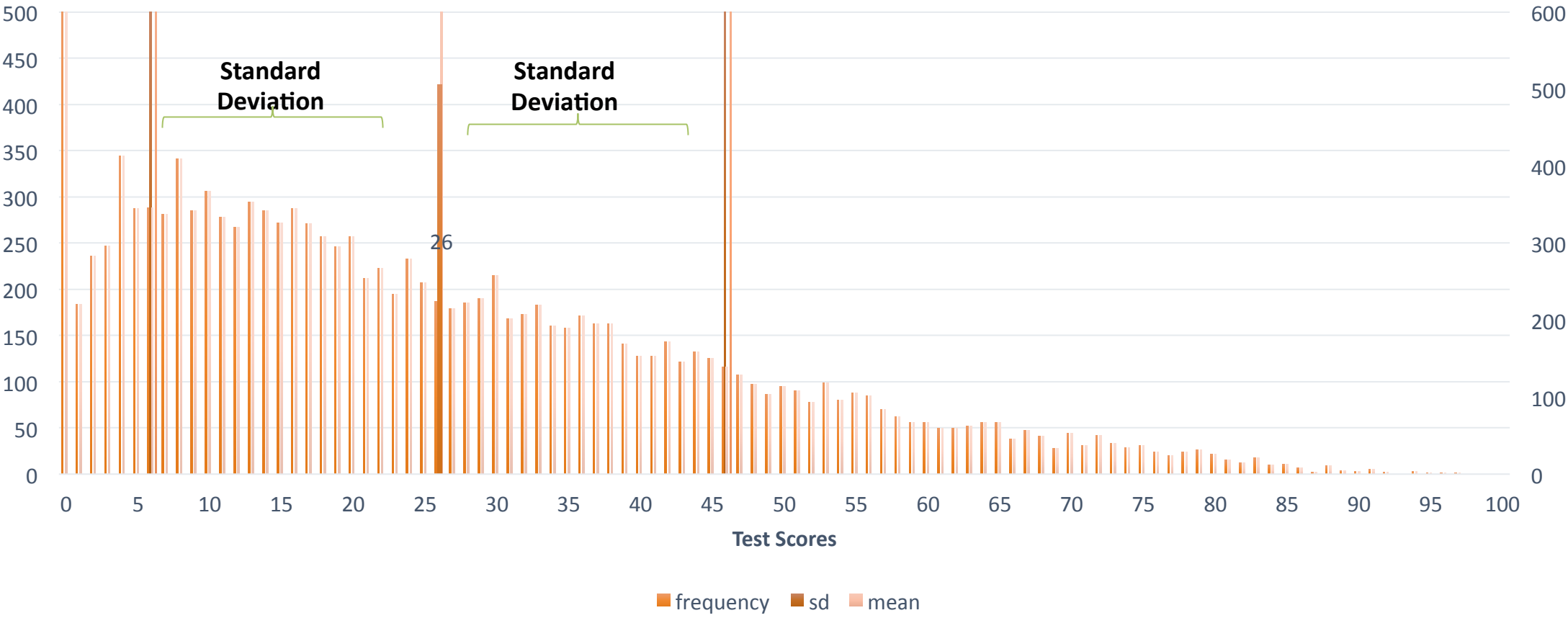
# Baseline Test Scores



# Mean = 26



# Standard Deviation = 20



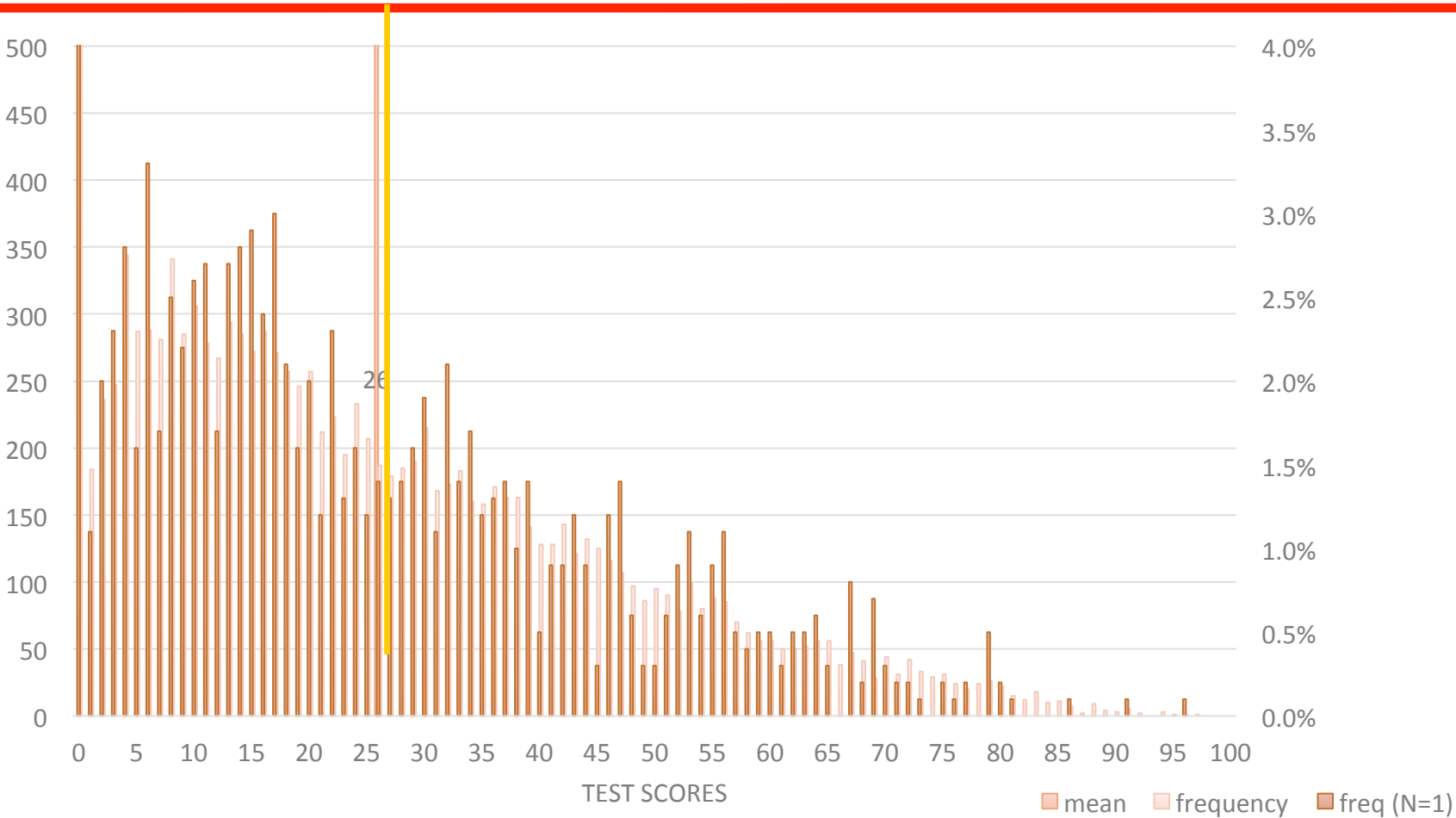


# Let's do an experiment

---

1. Take 1 Random test score from the pile of 16,000 tests
2. Write down the value
3. Put the test back
4. Do these three steps again
5. And again
6. 8,000 times
7. This is like a random sample of 8,000 (*with replacement*)

# What can we say about this sample?



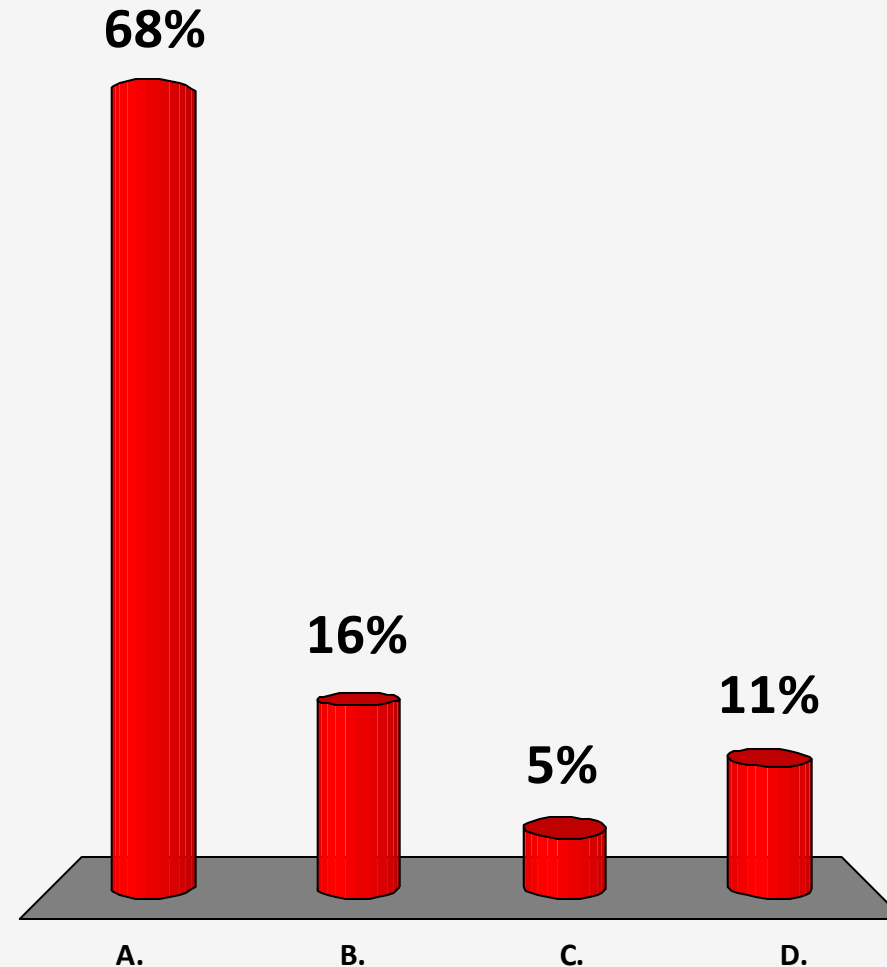
# But...

---

- ... I remember that as my sample goes, up, isn't the sampling distribution supposed to turn into a bell curve?
- (Central Limit Theorem)
- Is it that my sample isn't large enough?

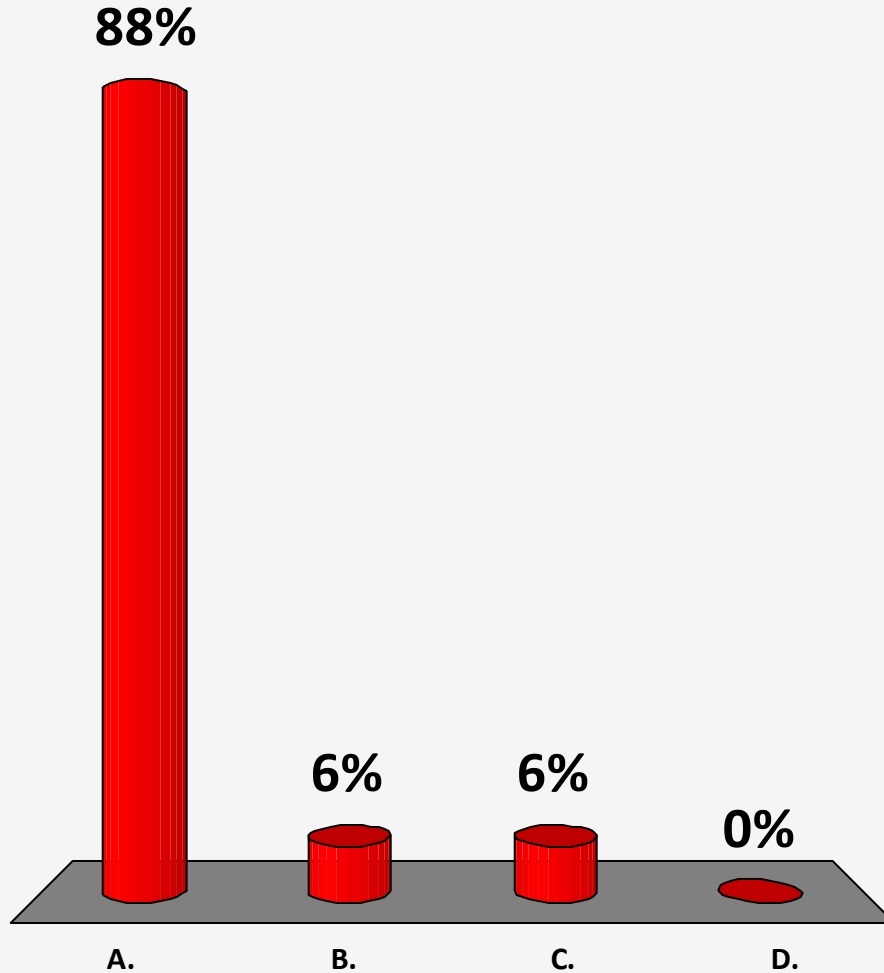
One limitation of statistical theory is that it assumes the population distribution is *normally distributed*

- A. True
- B. False
- C. Depends
- D. Don't know



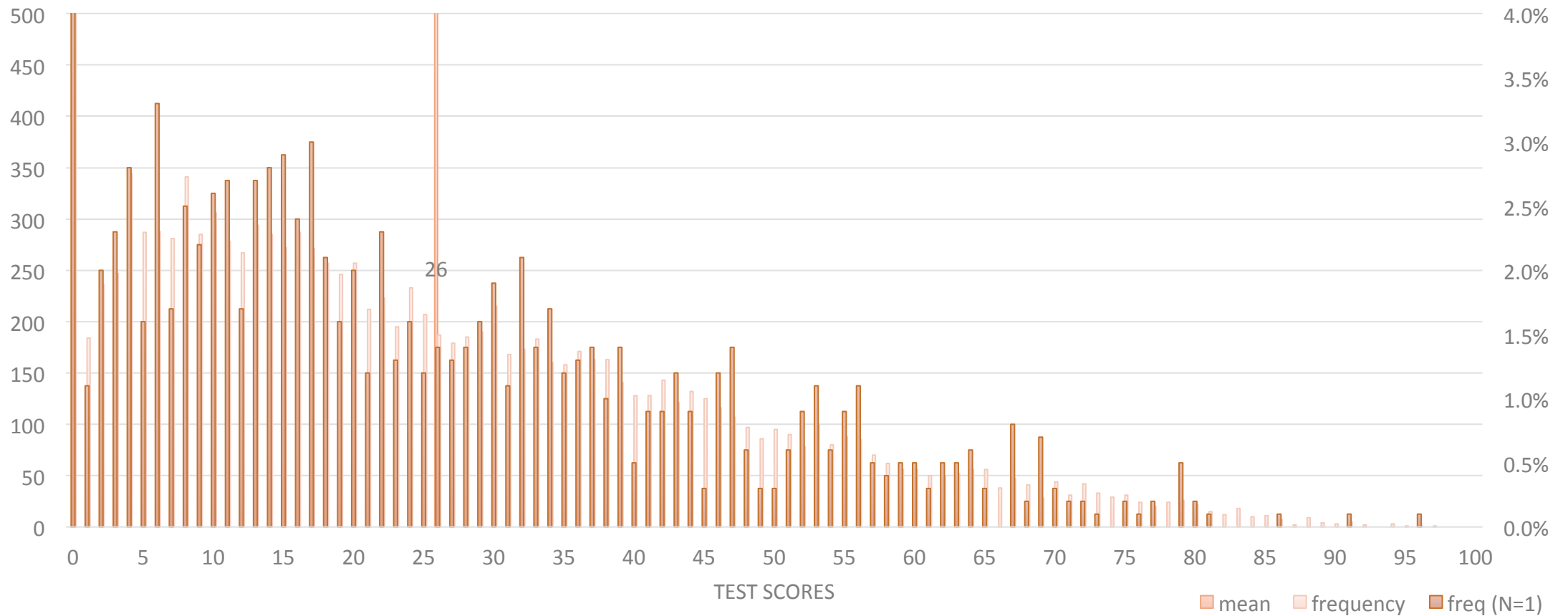
The sampling distribution may not be normal if the population distribution *is skewed*

- A. True
- B. False
- C. Depends
- D. Don't know



# Population vs. sampling distribution

(This is the distribution of my sample of 8,000 students!)

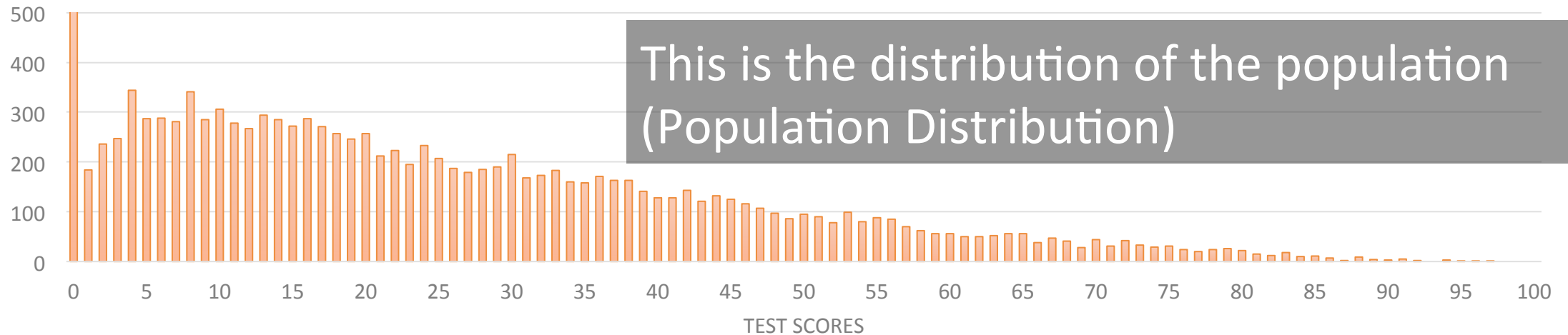


# Outline

---

- ***Sampling distributions***
  - Population distribution
  - ***Sampling distribution***
  - Law of large numbers/central limit theorem
  - Standard deviation and standard error
- Detecting impact

# How do we get from here...



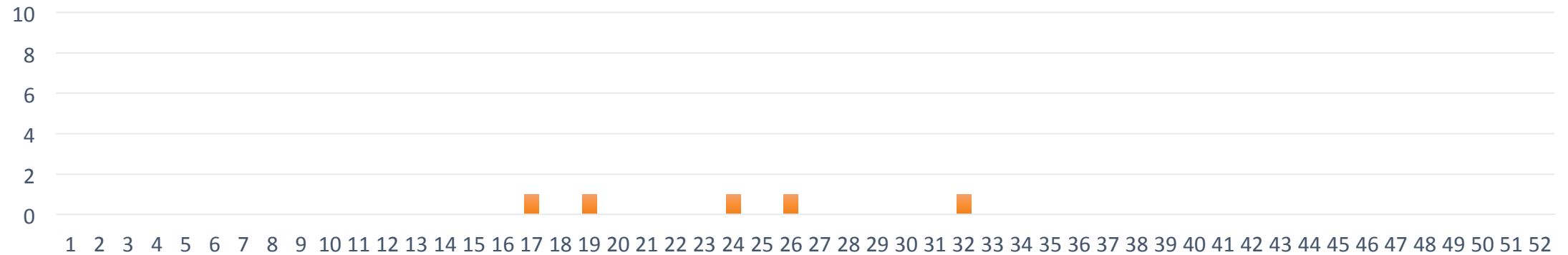
This is the distribution of Means from all Random Samples (Sampling distribution)

The figure shows a smooth, white, bell-shaped curve representing a normal distribution. The curve is centered at approximately 50 on the x-axis. A vertical white line marks the center of the distribution. The background is black, and the text is white.

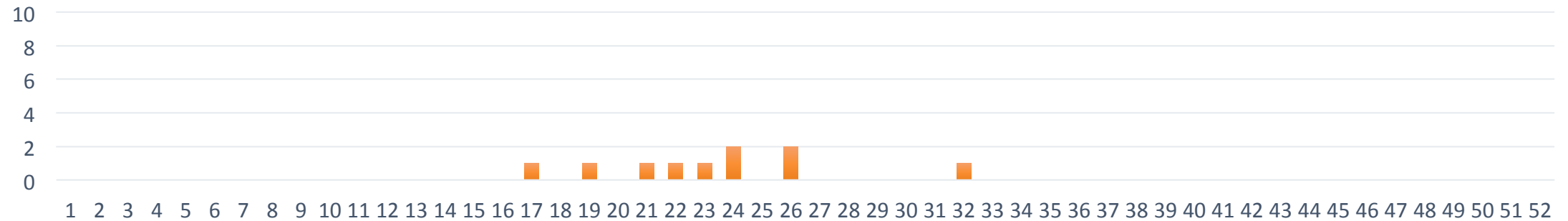


Draw 10 random students, take the average, plot it: Do this 5 & 10 times.

Frequency of Means With 5 Samples

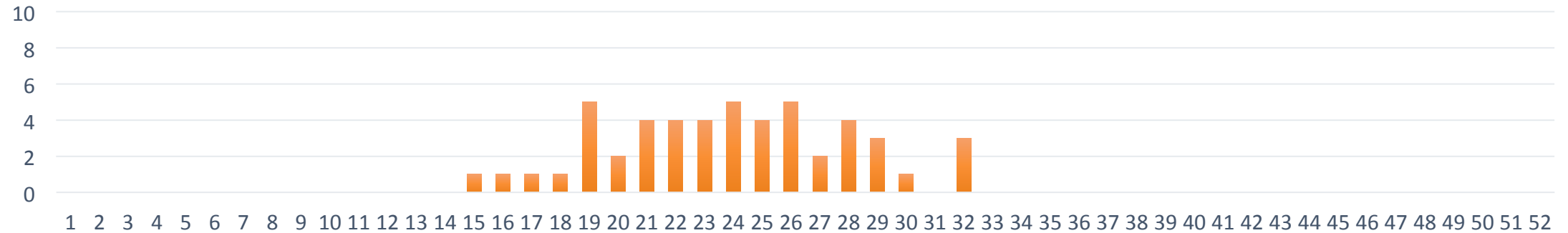


Frequency of Means With 10 Samples

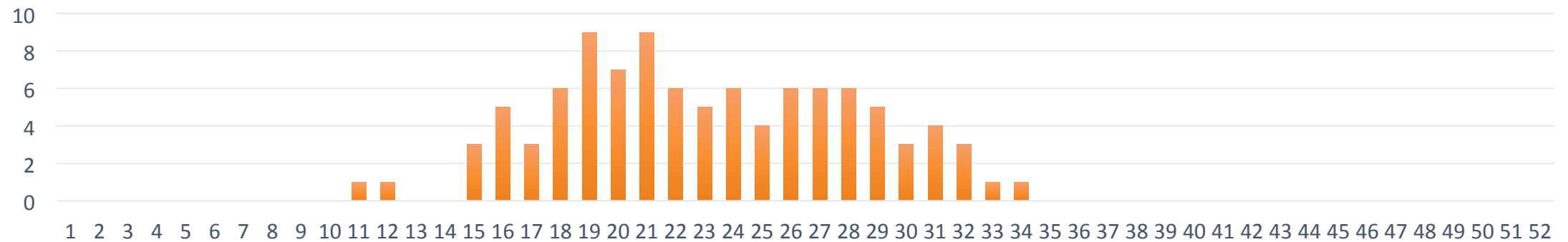


# Draw 10 random students: 50 and 100 times

Frequency of Means With 50 Samples

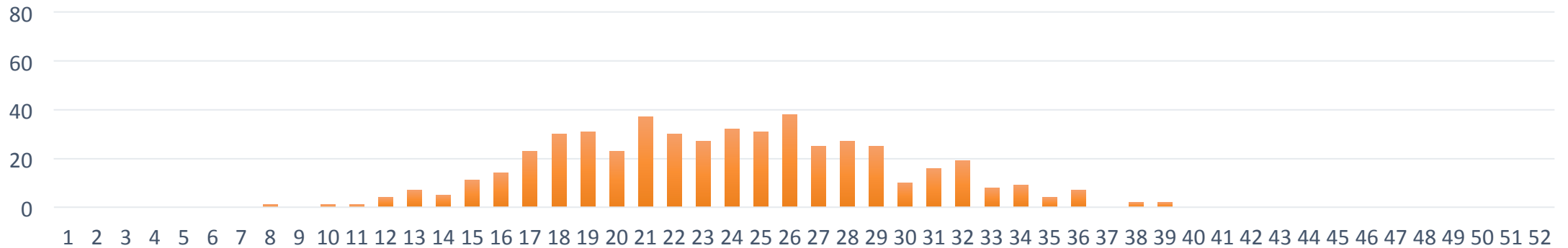


Frequency of Means with 100 Samples

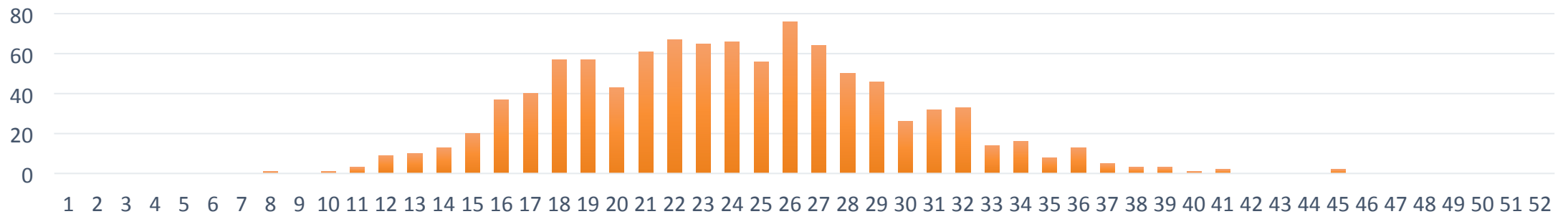


# Draws 10 random students: 500 and 1000 times

Frequency of Means With 500 Samples



Frequency of Means With 1000 Samples



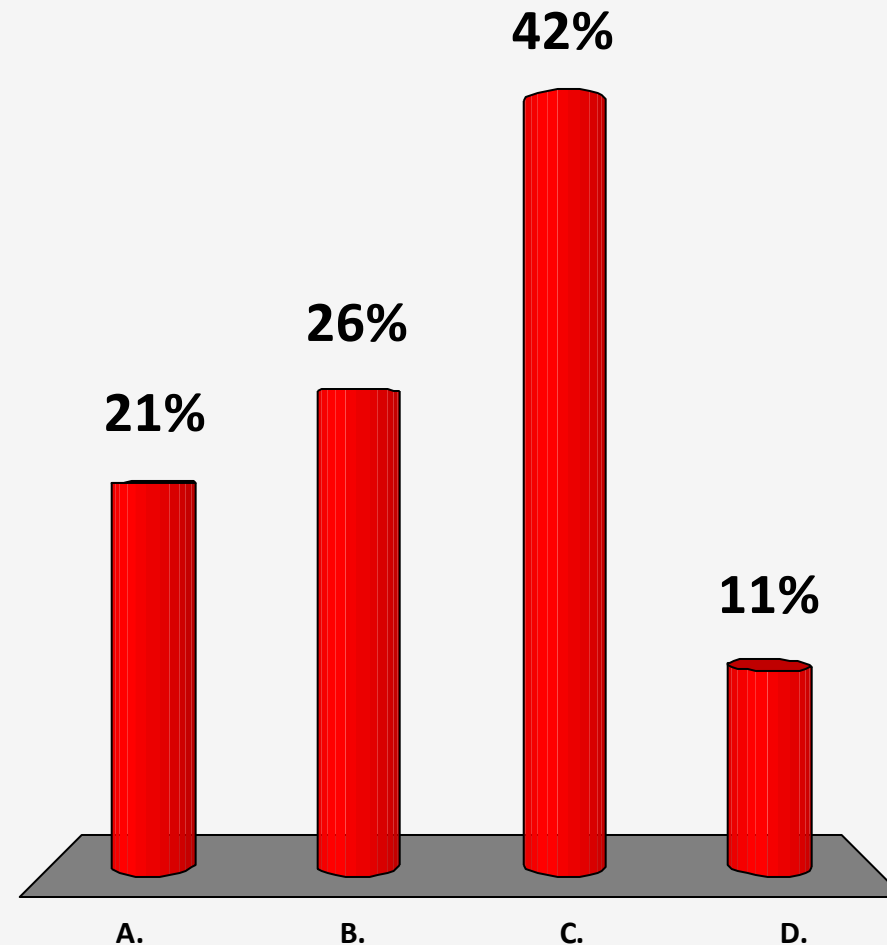
# Draw 10 random students

---

- This is like a sample size of 10
- What happens if we take a sample size of 50?

What happens to the sampling distribution if we draw a sample size of 50 instead of 10, and take the mean (thousands of times)?

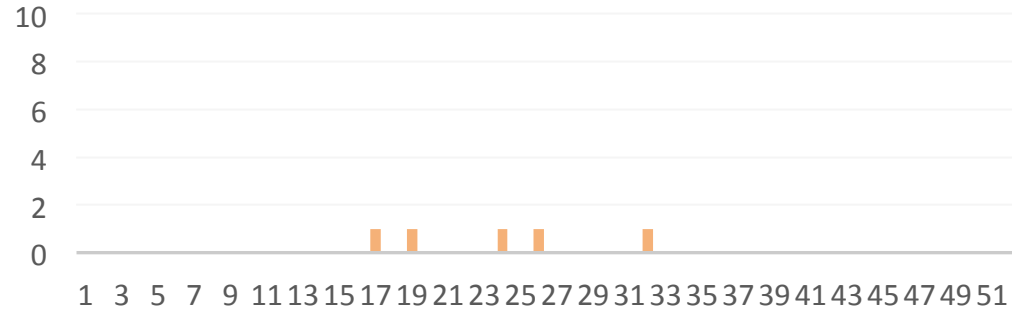
- A. We will approach a bell curve faster (than with a sample size of 10)
- B. The bell curve will be narrower
- C. Both A & B
- D. Neither. The underlying sampling distribution does not change.



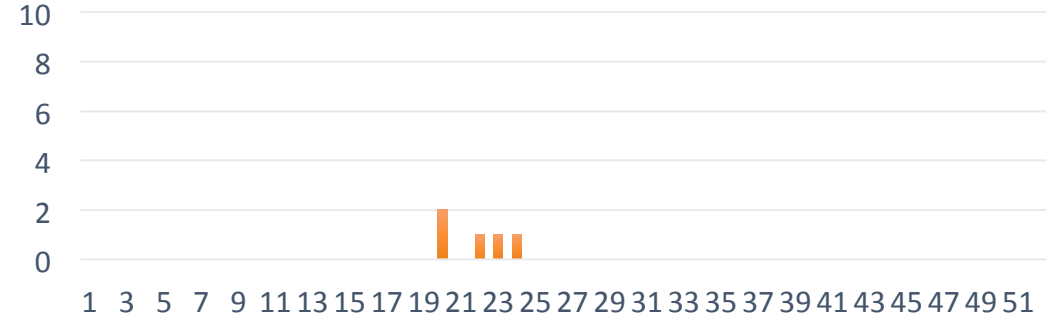
# N = 10

# N = 50

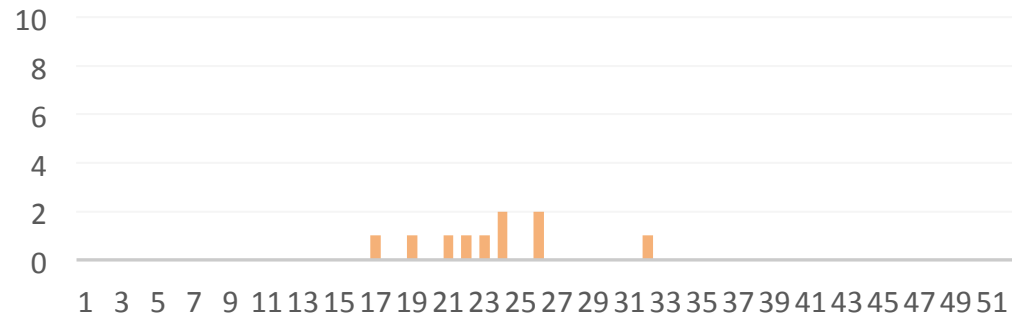
### Frequency of Means With 5 Samples



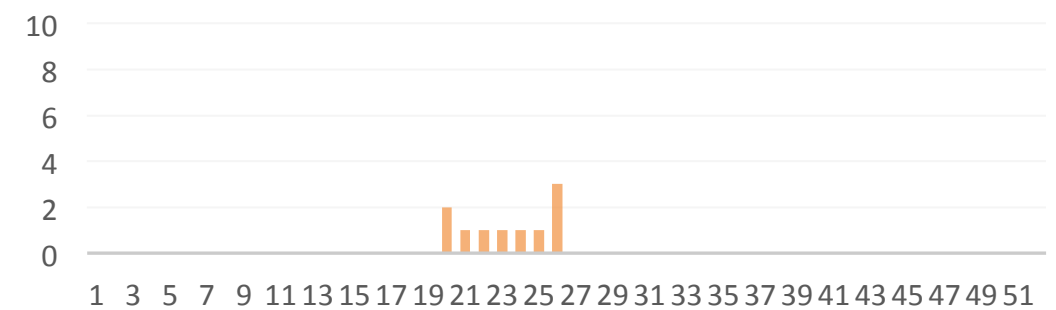
### Frequency of Means With 5 Samples



### Frequency of Means With 10 Samples



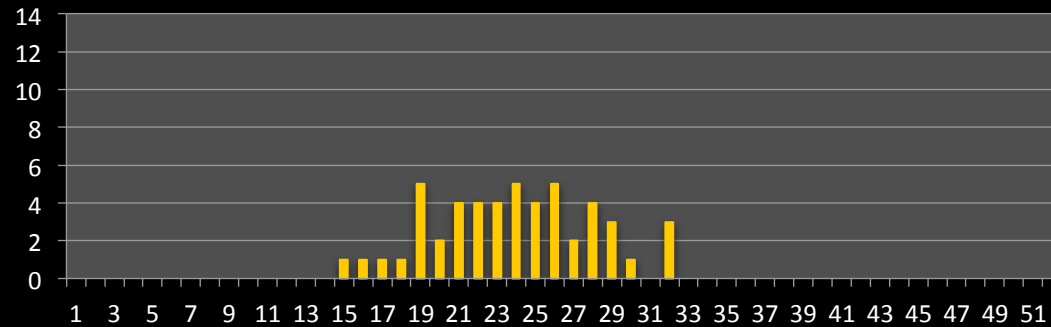
### Frequency of Means With 10 Samples



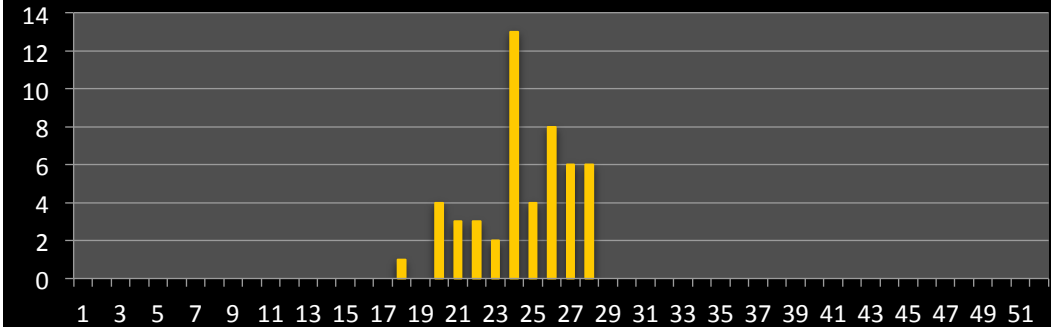
Draws of 10

Draws of 50

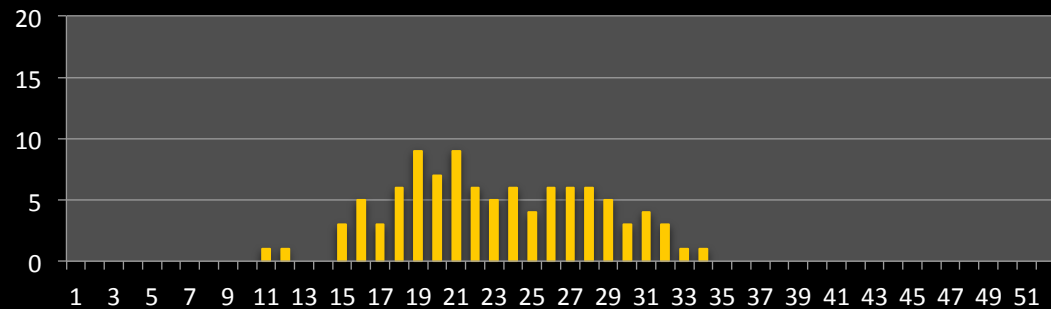
Frequency of Means With 50 Samples



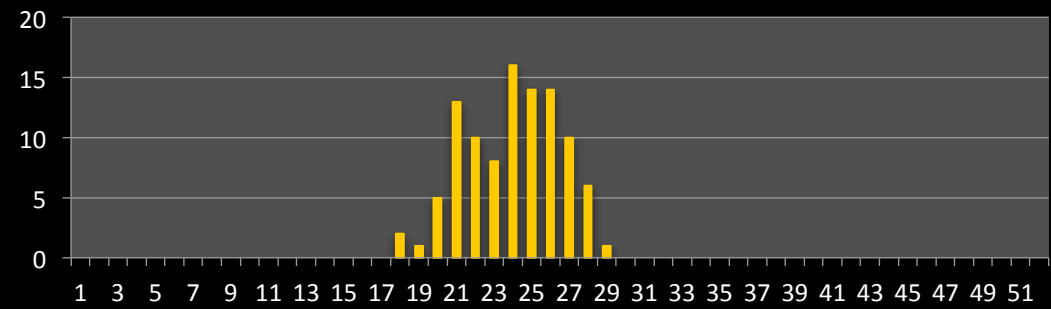
Frequency of Means With 50 Samples



Frequency of Means with 100 Samples



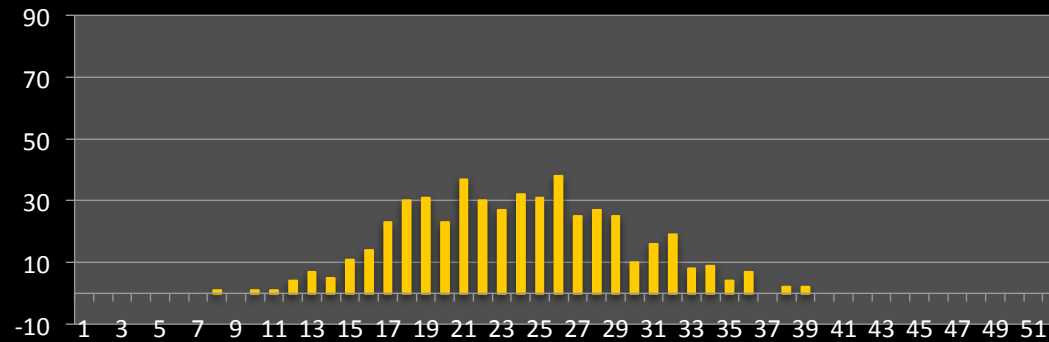
Frequency of Means With 100 Samples



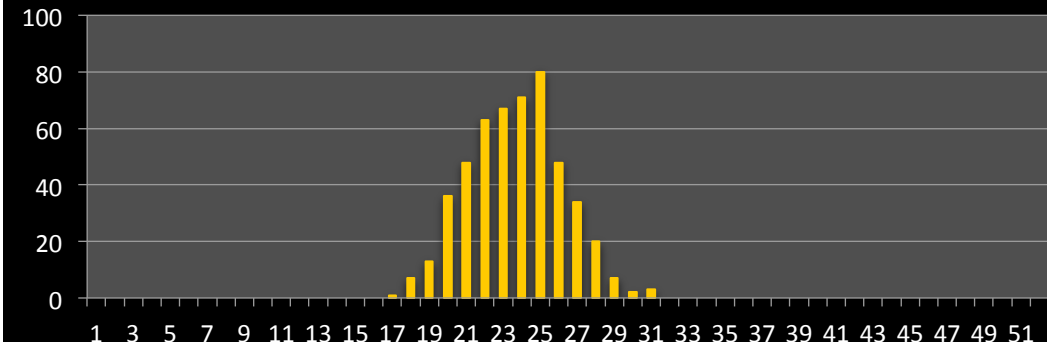
## Draws of 10

## Draws of 50

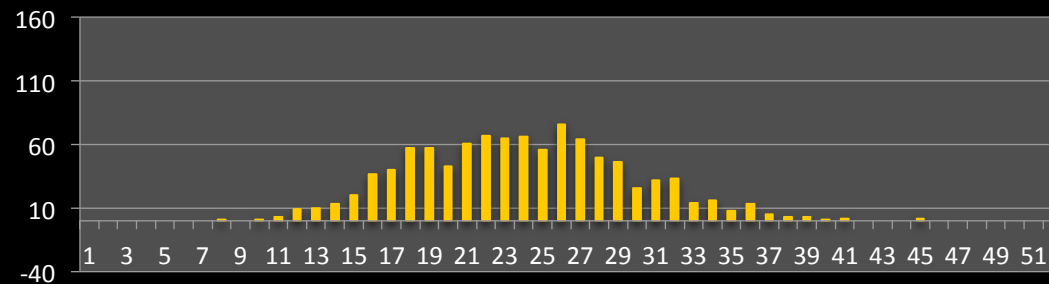
### Frequency of Means With 500 Samples



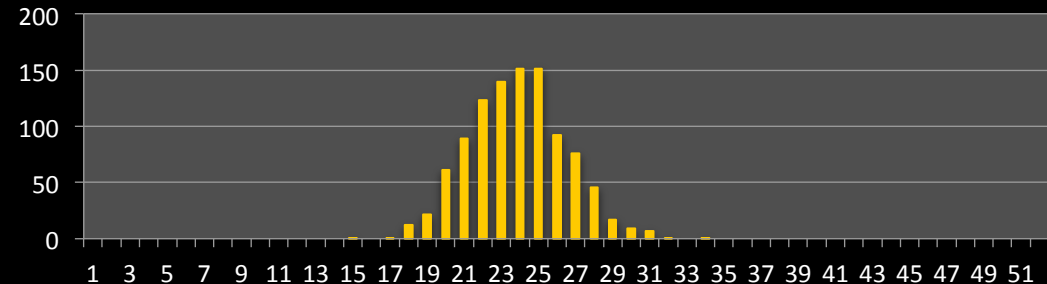
### Frequency of Means With 500 Samples



### Frequency of Means With 500 Samples



### Frequency of Means With 500 Samples



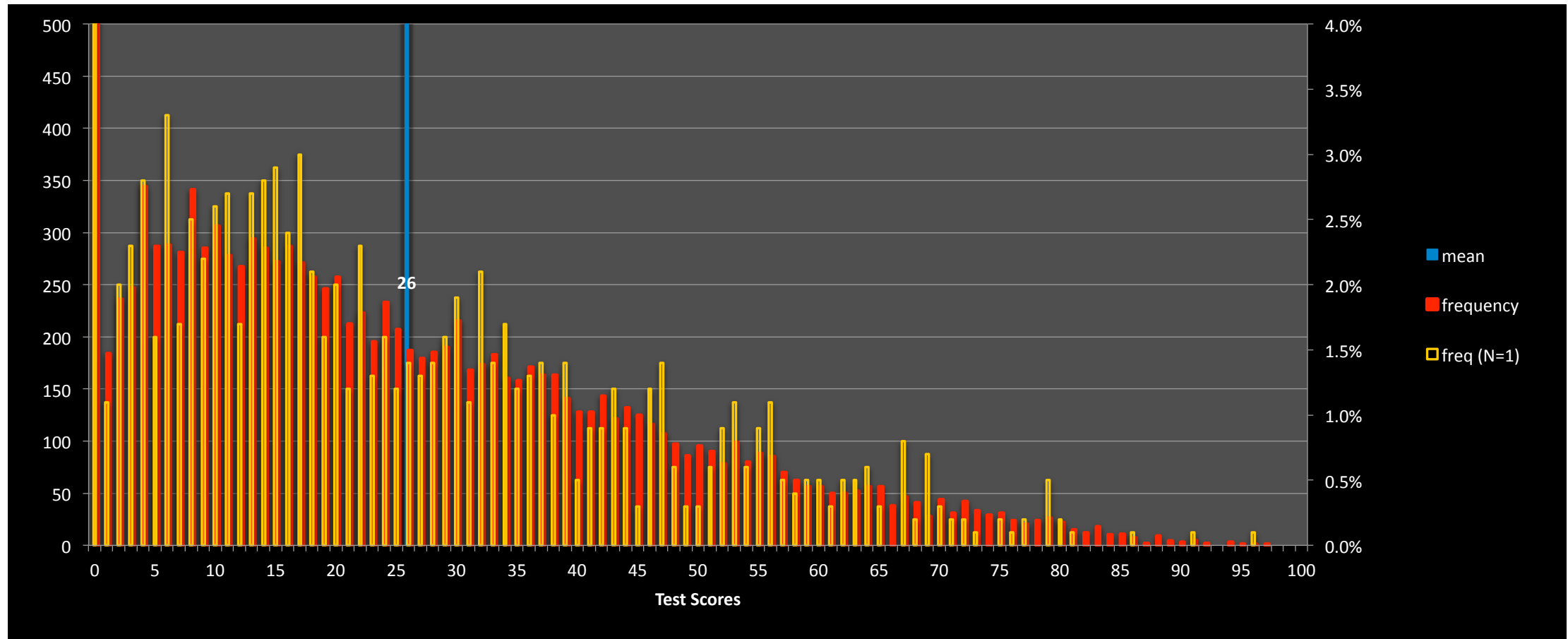


# Outline

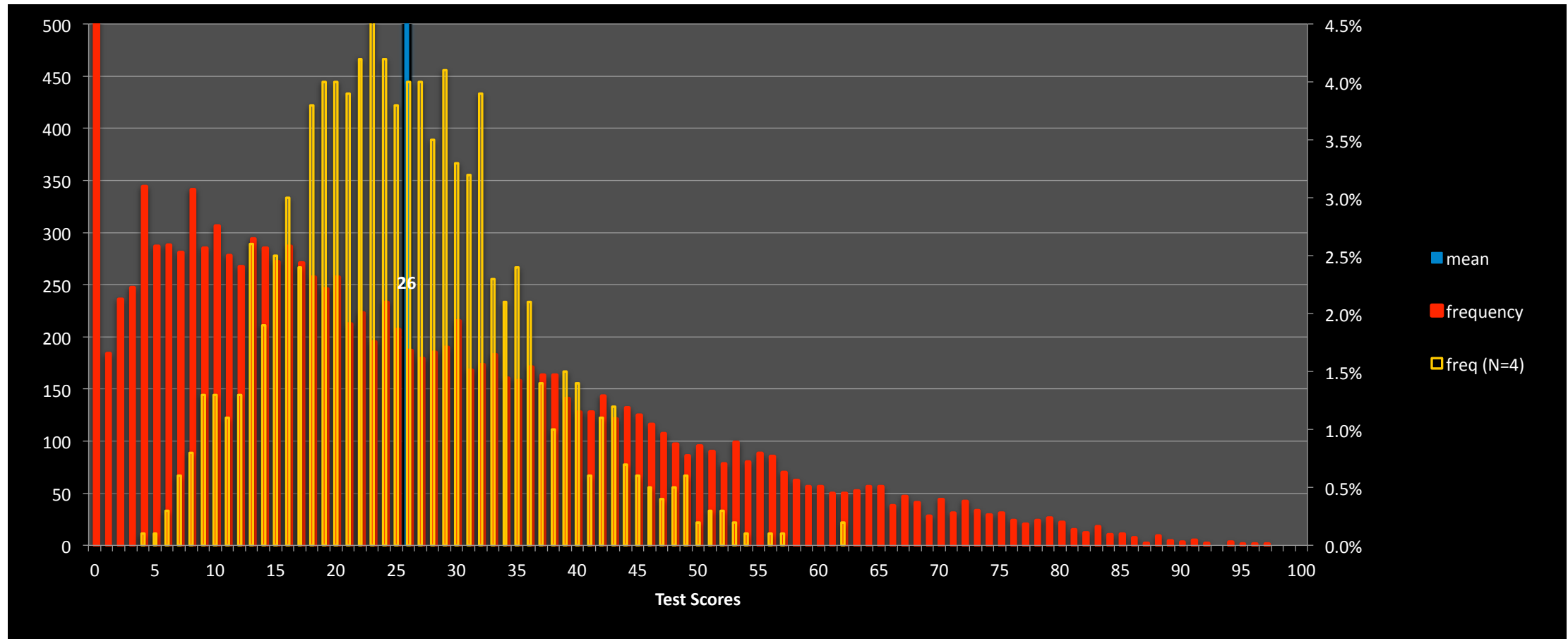
---

- ***Sampling distributions***
  - Population distribution
  - Sampling distribution
  - ***Law of large numbers/central limit theorem***
  - Standard deviation and standard error
- Detecting impact

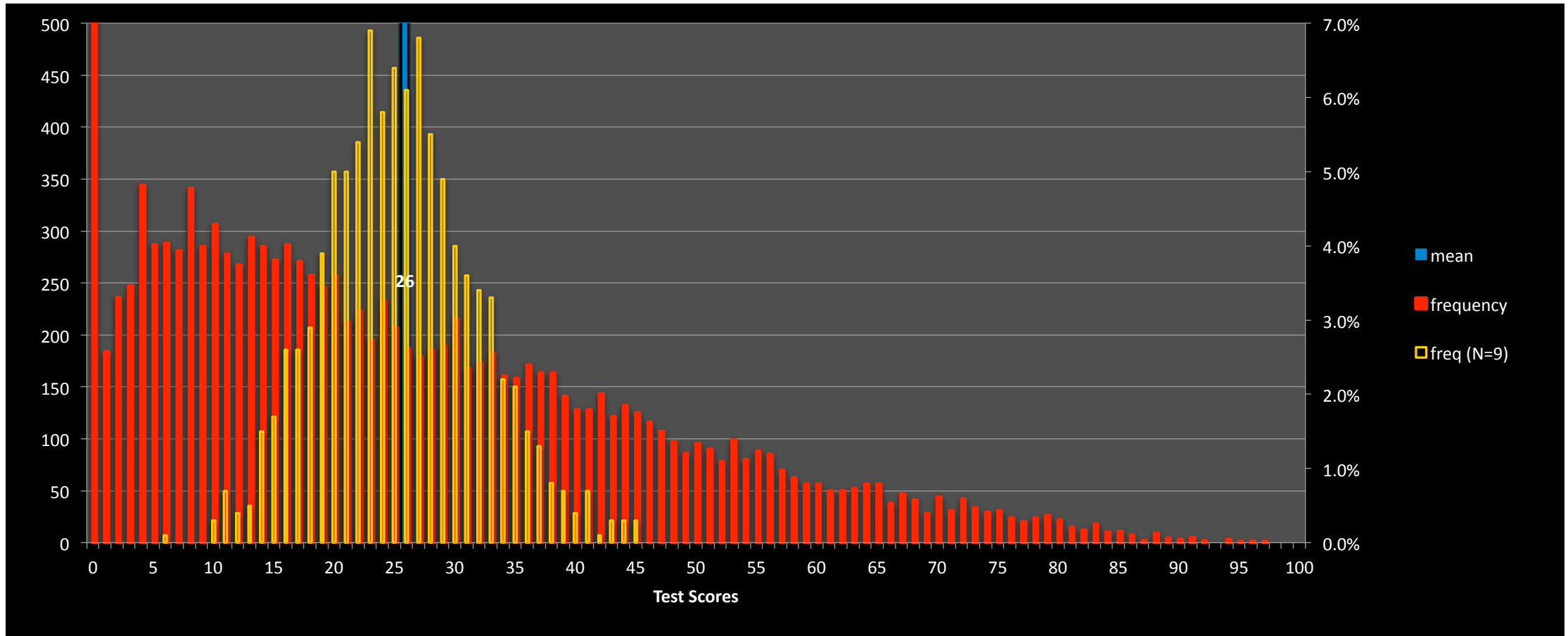
# Population & sampling distribution: Draw 1 random student (from 8,000)



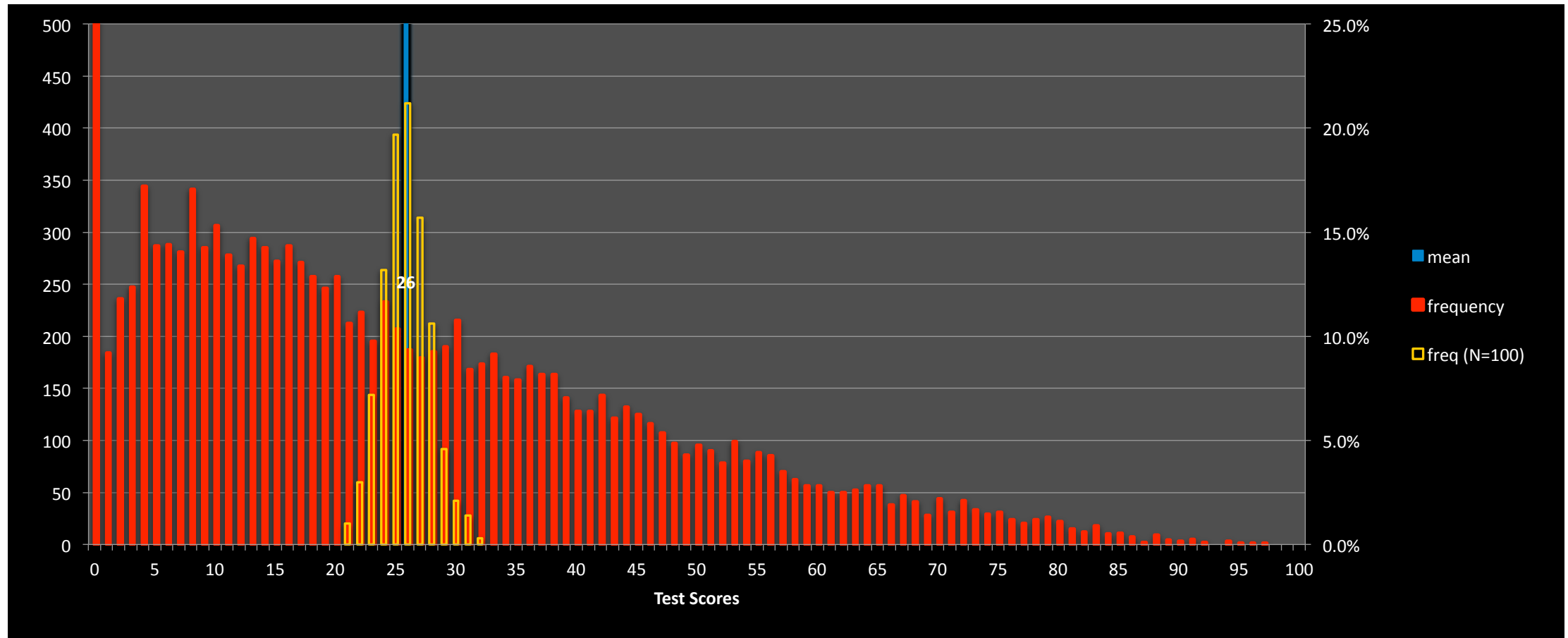
# Sampling Distribution: Draw 4 random students (N=4)



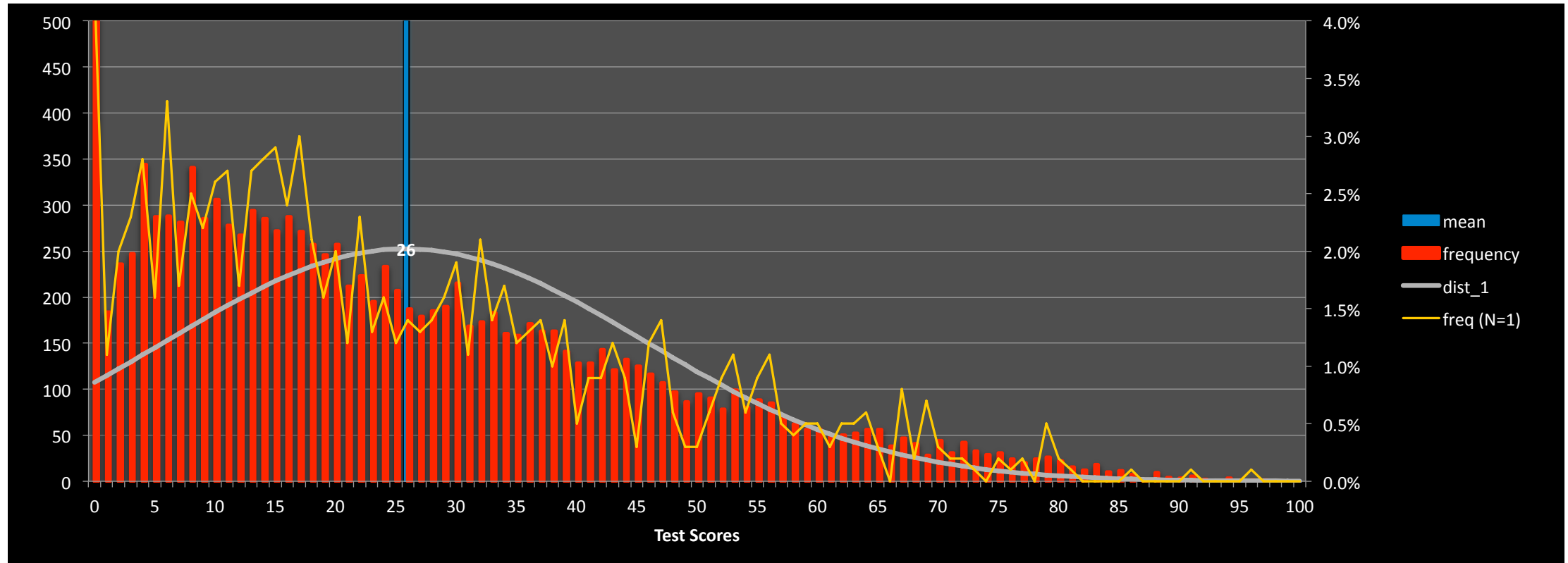
# Law of Large Numbers: N=9



# *Law of Large Numbers : N = 100*

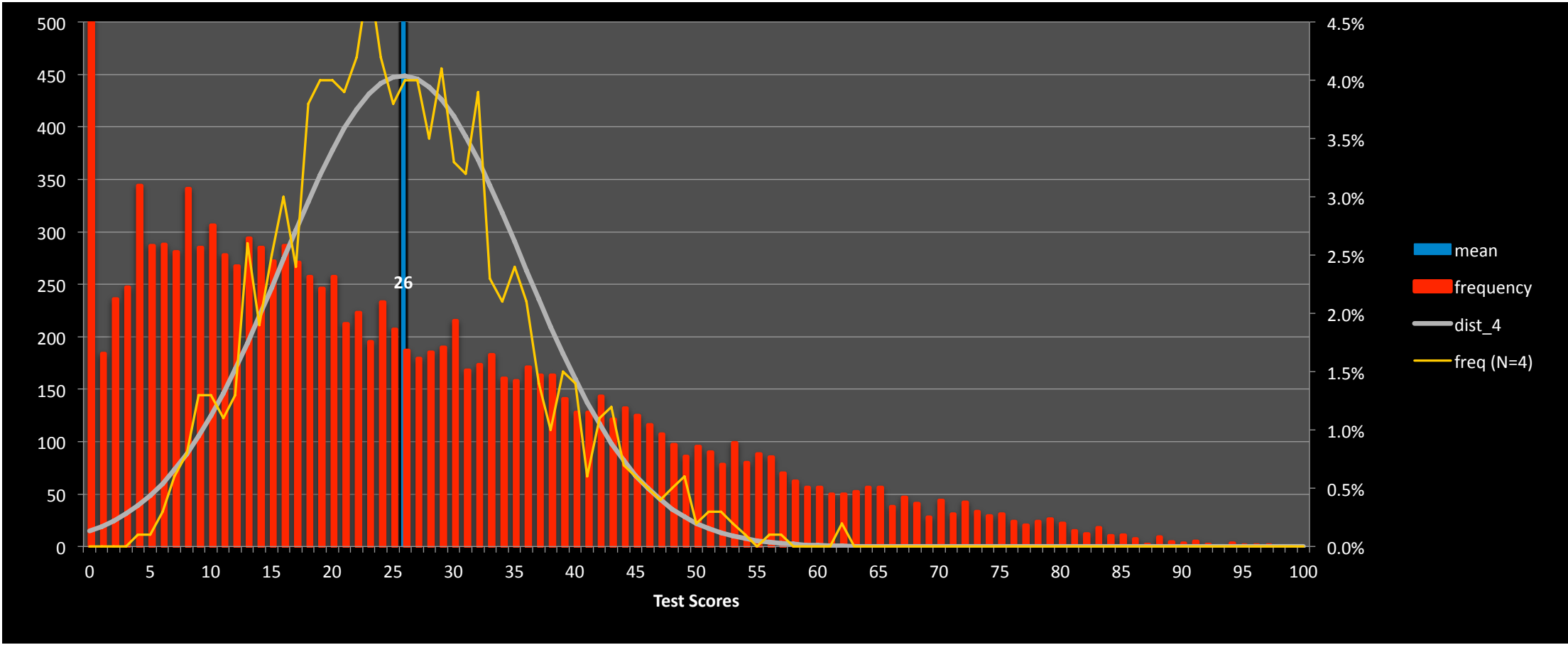


# Central Limit Theorem: N=1

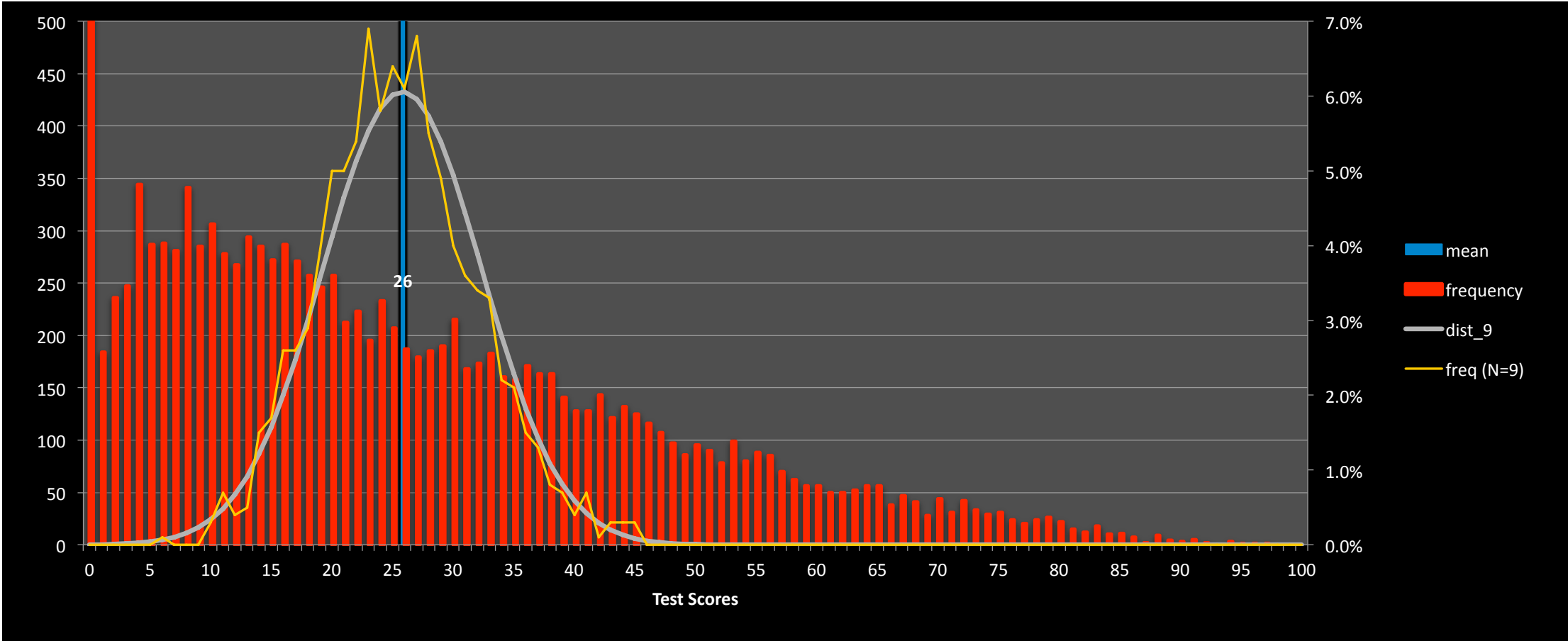


- The white line is a theoretical distribution

# Central Limit Theorem: N=4

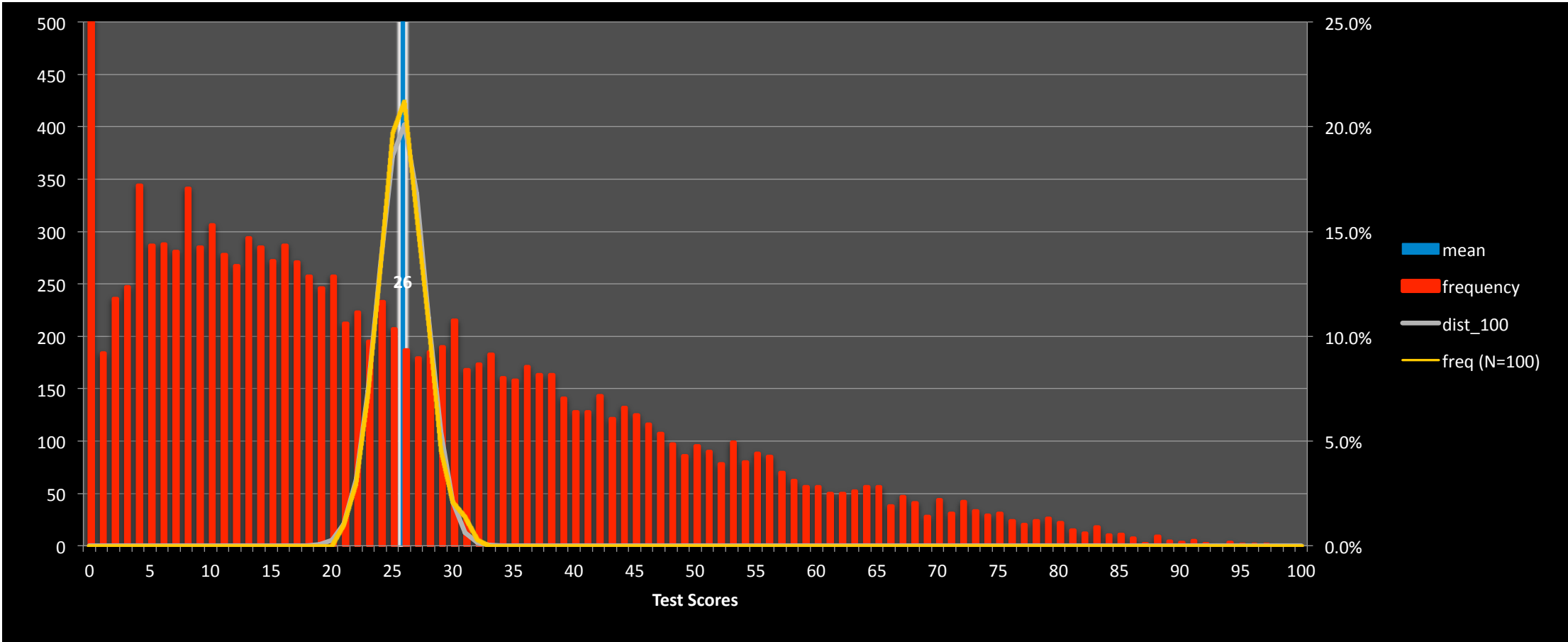


# Central Limit Theorem: N=9





# Central Limit Theorem: N = 100



# Outline

---

- ***Sampling distributions***
  - Population distribution
  - Sampling distribution
  - Law of large numbers/central limit theorem
  - ***Standard deviation and standard error***
- Detecting impact

# Standard deviation/error

---

- What's the difference between the standard deviation and the standard error?
- The standard error = the standard deviation of the sampling distributions

# Variance and Standard Deviation

---

- Variance = 400

$$\sigma^2 = \frac{\sum (Observation\ Value - Average)^2}{N}$$

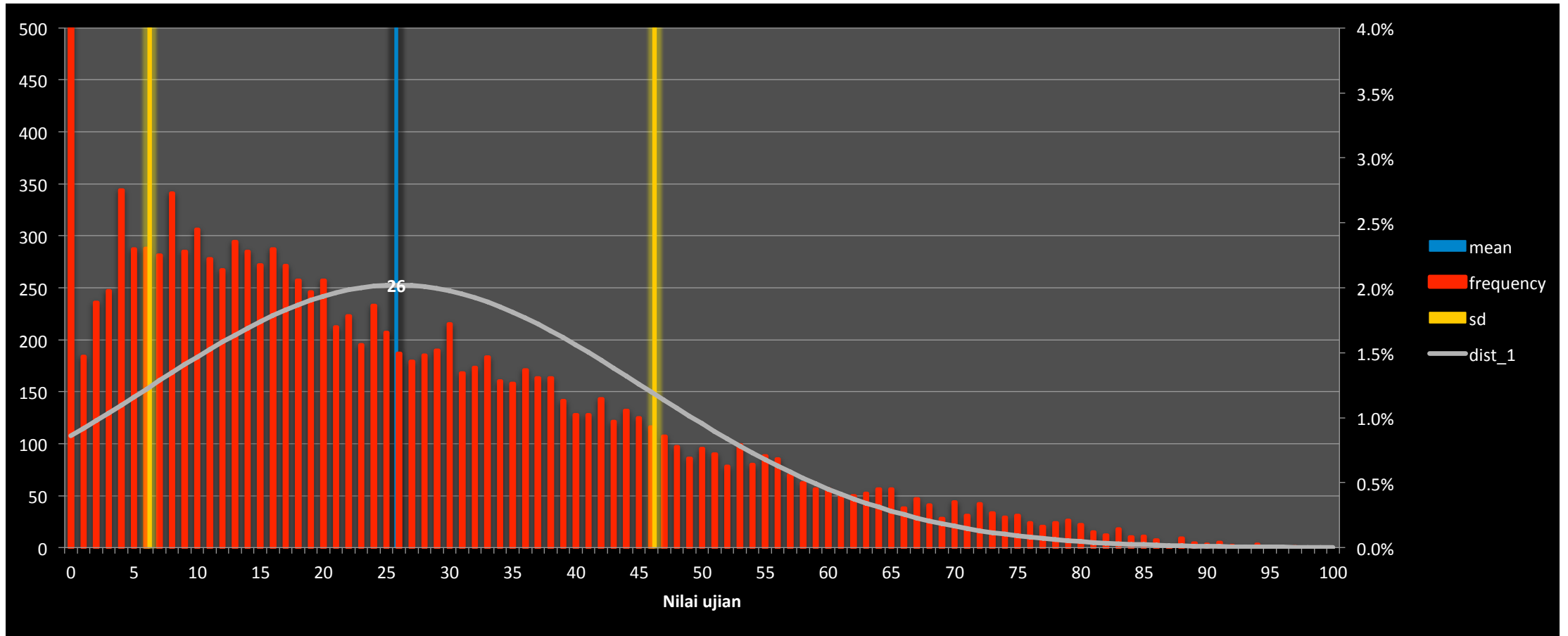
- Standard Deviation = 20

$$\sigma = \sqrt{Variance}$$

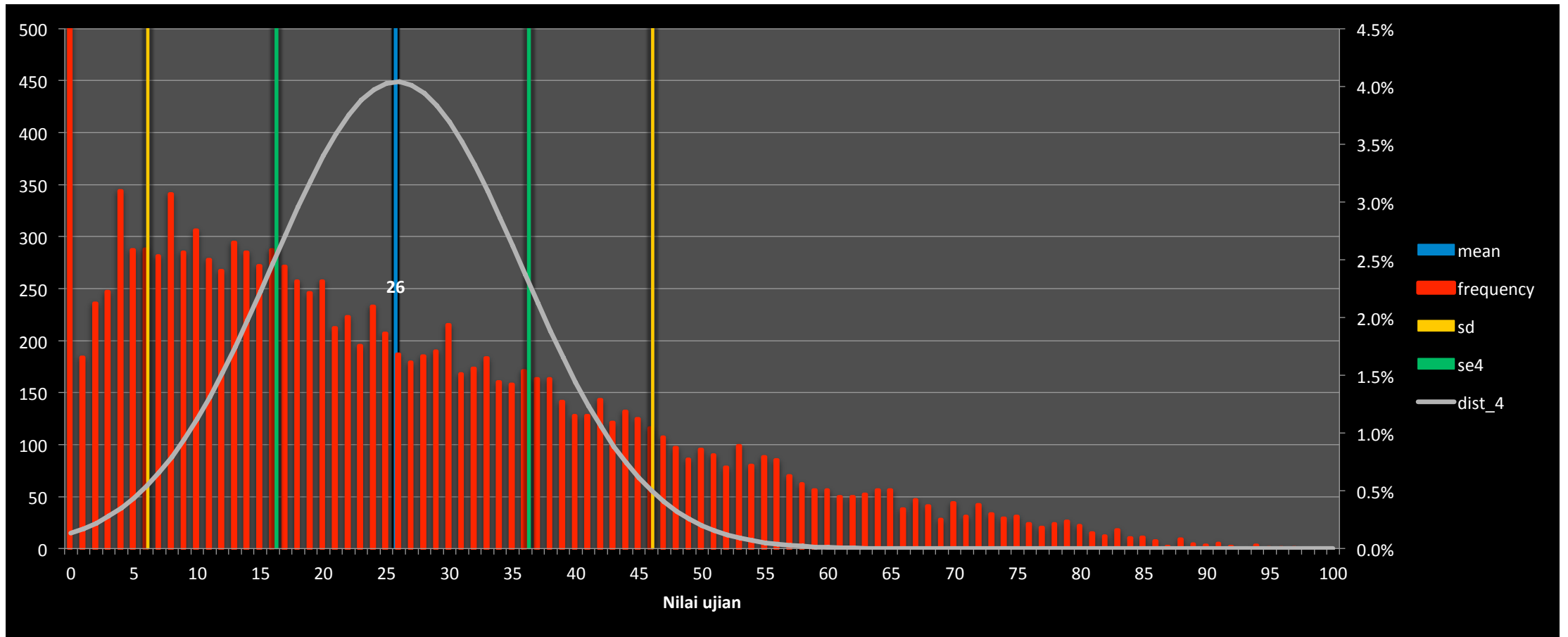
- Standard Error =  $20 / \sqrt{N}$

$$SE = \sigma / \sqrt{N}$$

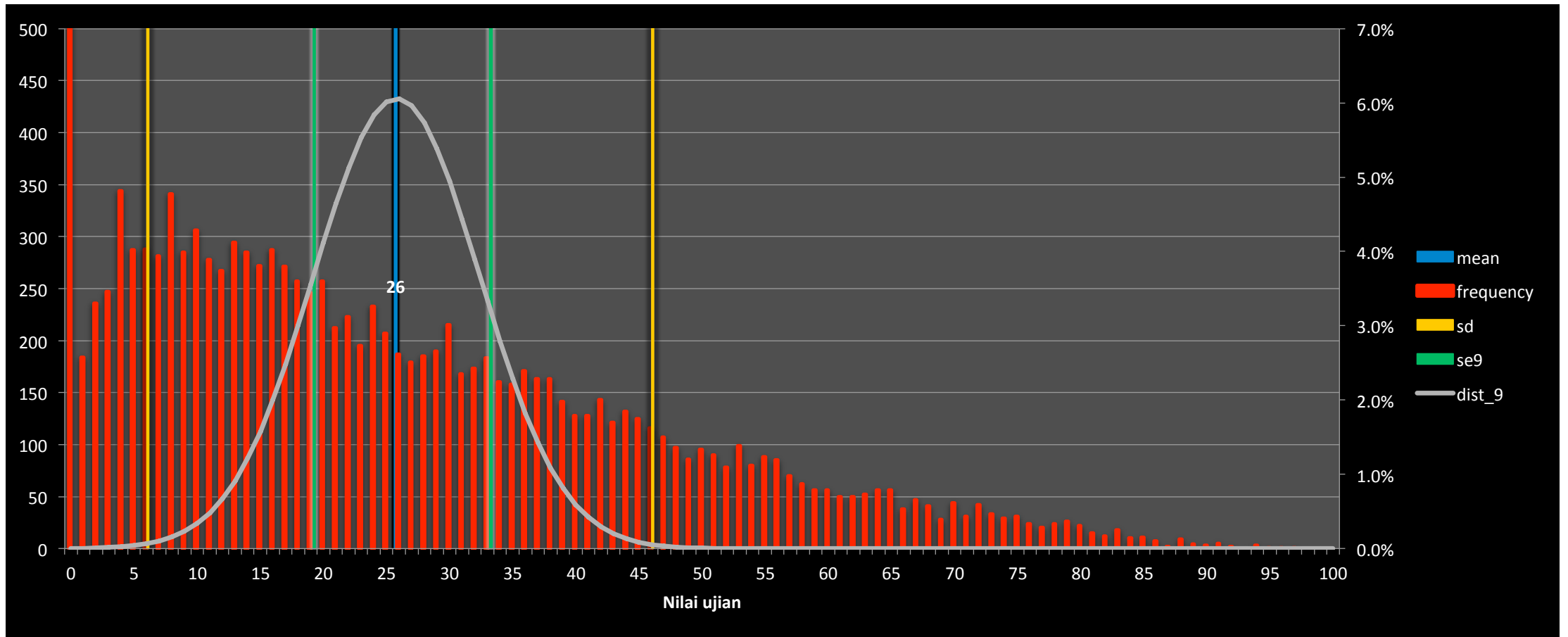
# Standard Deviation/ Standard Error



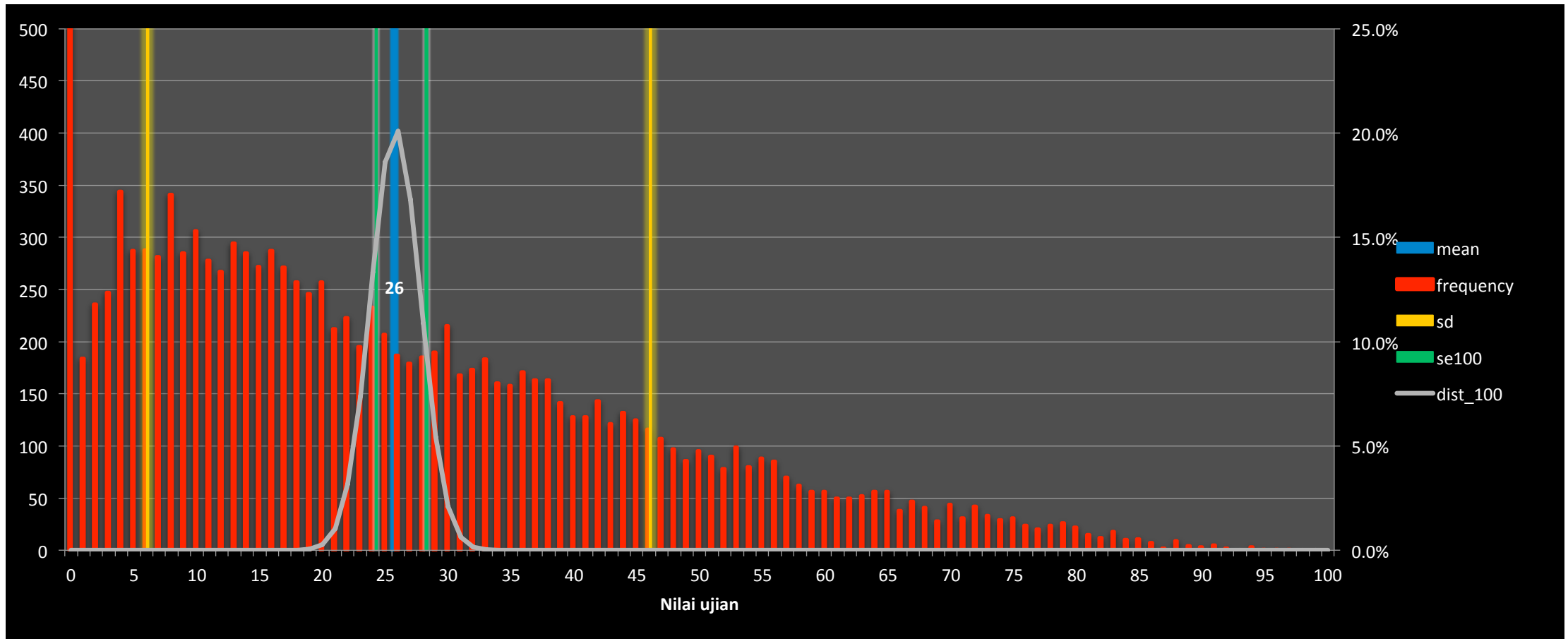
Sample size  $\uparrow$  x4, SE  $\downarrow$   $\frac{1}{2}$



# Sample size $\uparrow$ x9, SE $\downarrow$ ?



# Sample size $\uparrow$ x100, SE $\downarrow$ ?





# Outline

---

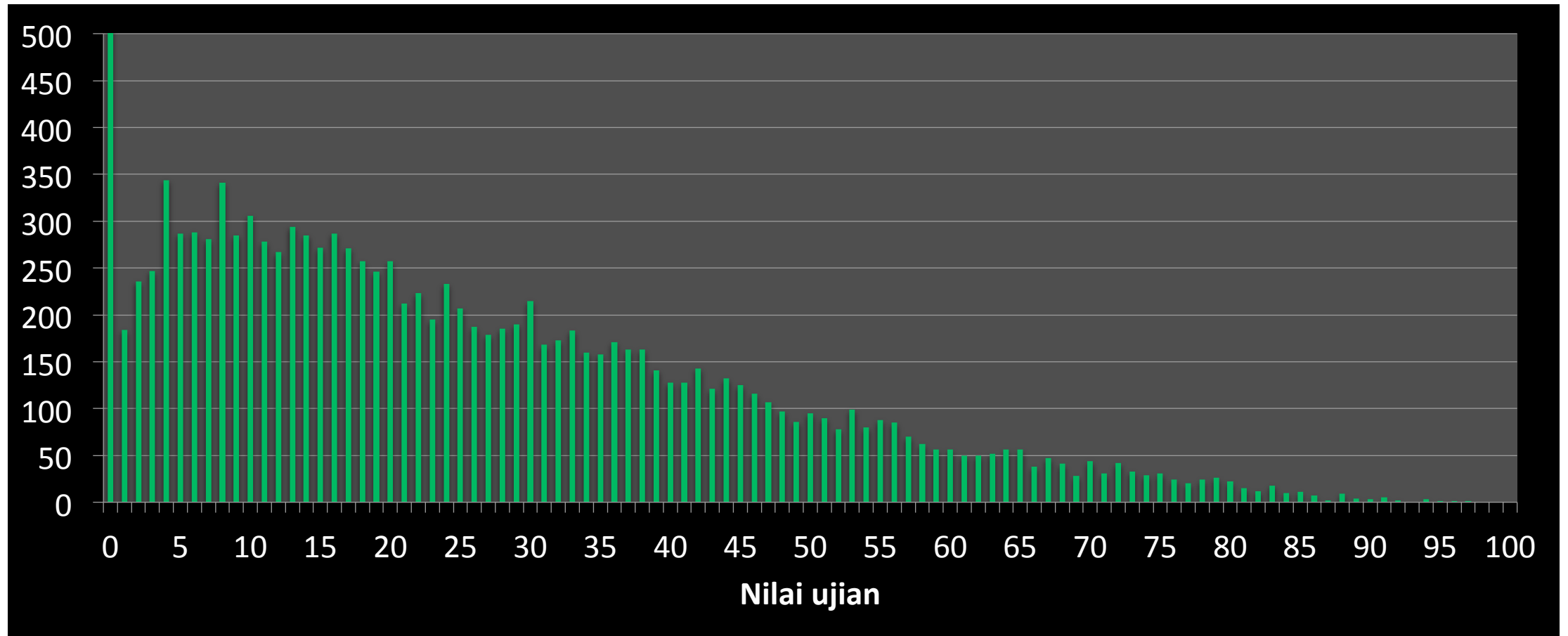
- Sampling distributions
- **Detecting impact**
  - **significance**
  - **effect size**
  - **power**
  - **baseline and covariates**
  - **clustering**
  - **stratification**

# We implement the Balsakhi Program

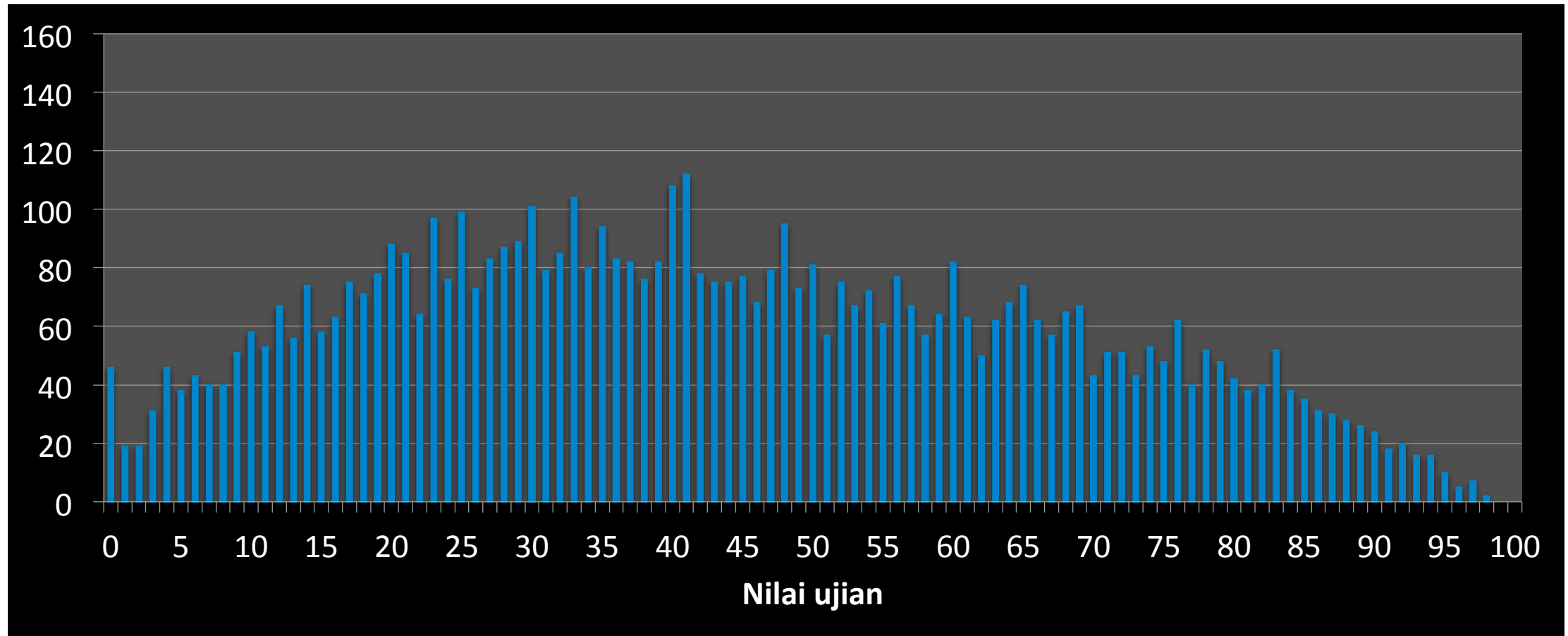
---



# Baseline test scores

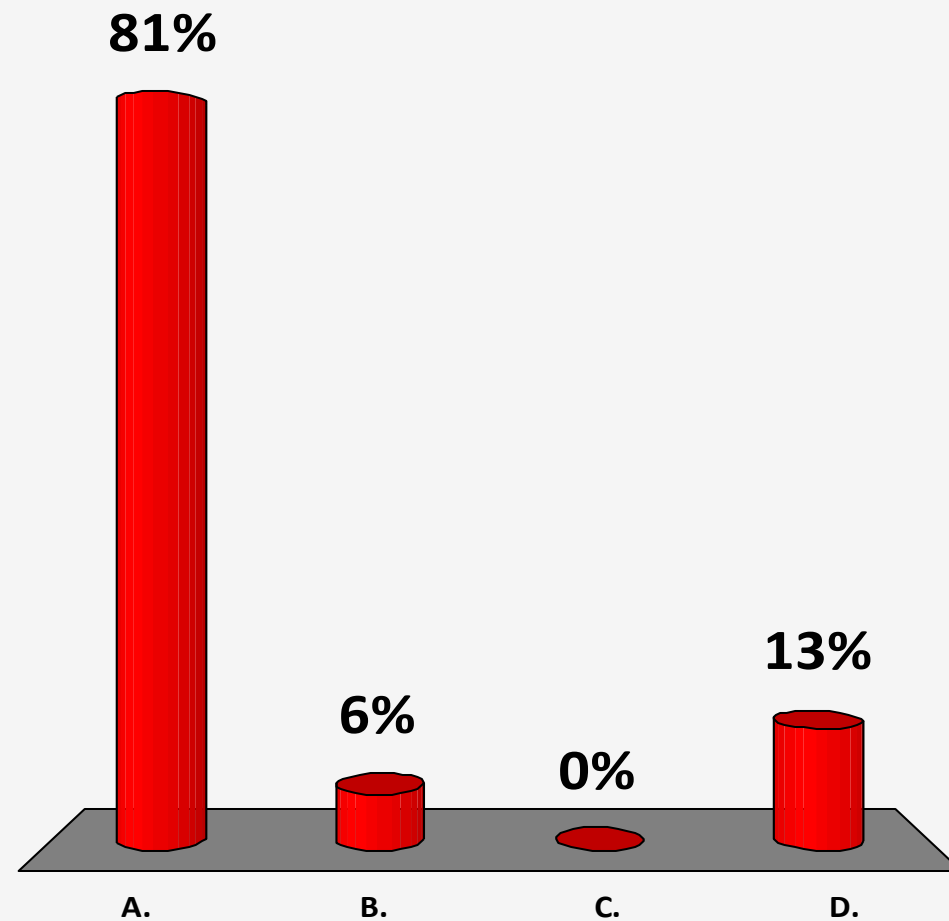


# Endline test scores after Balsakhi program

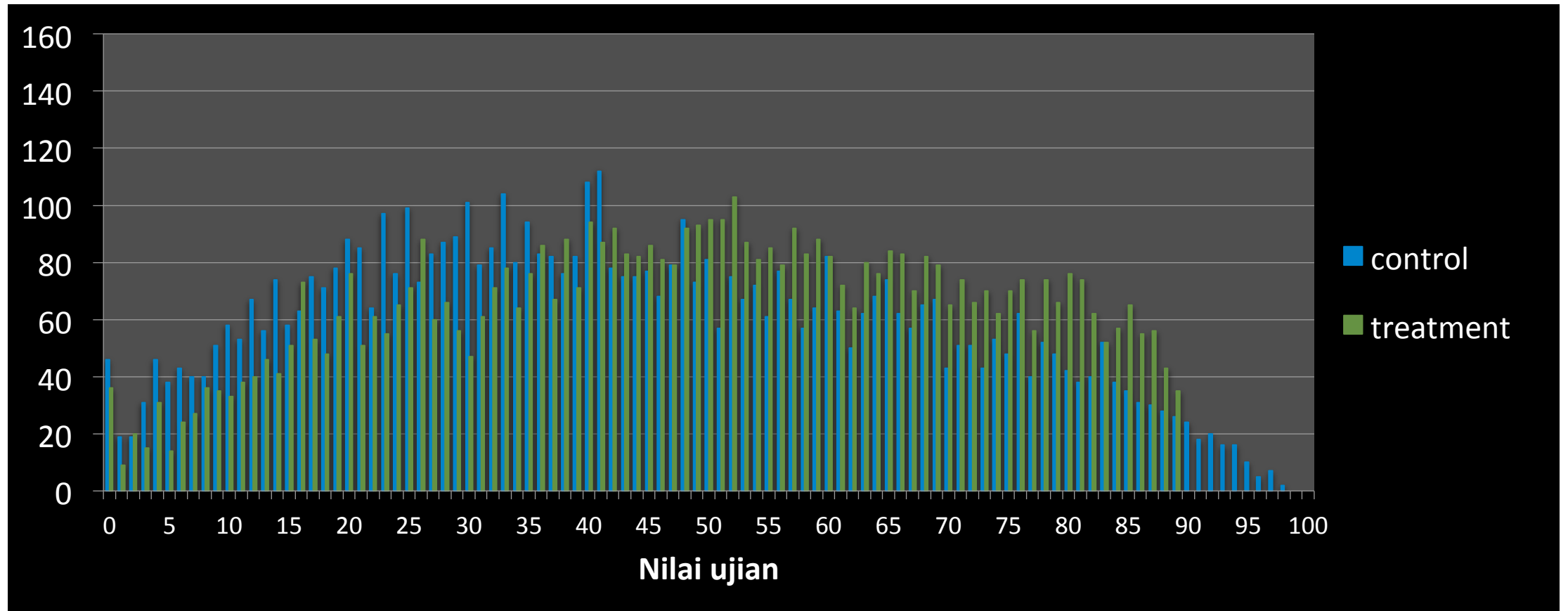


# The impact appears to be?

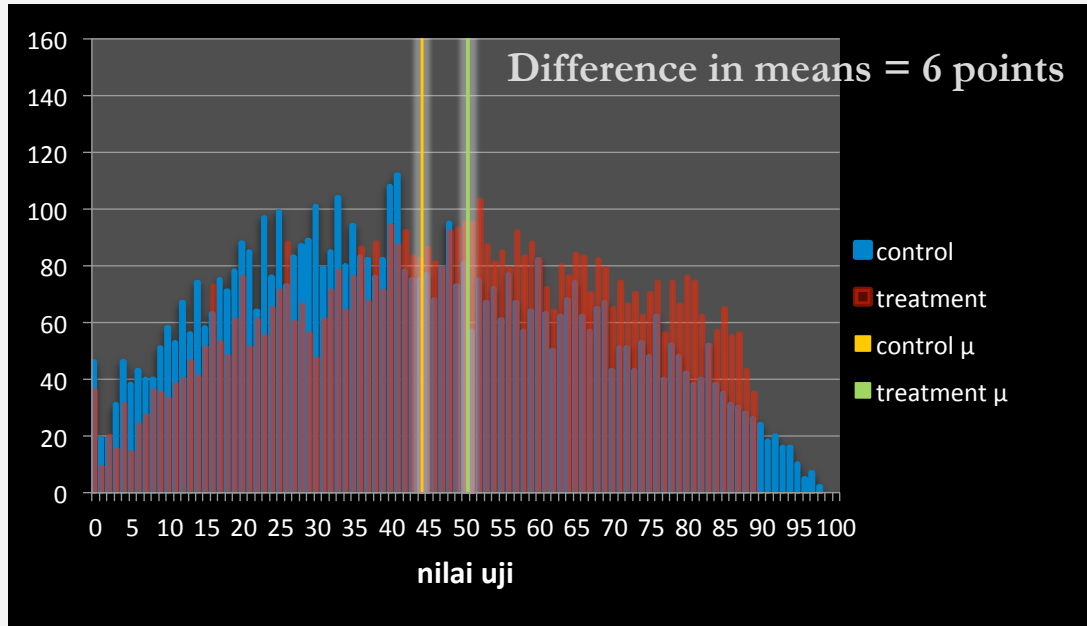
- A. Positive
- B. Negative
- C. No impact
- D. Don't know



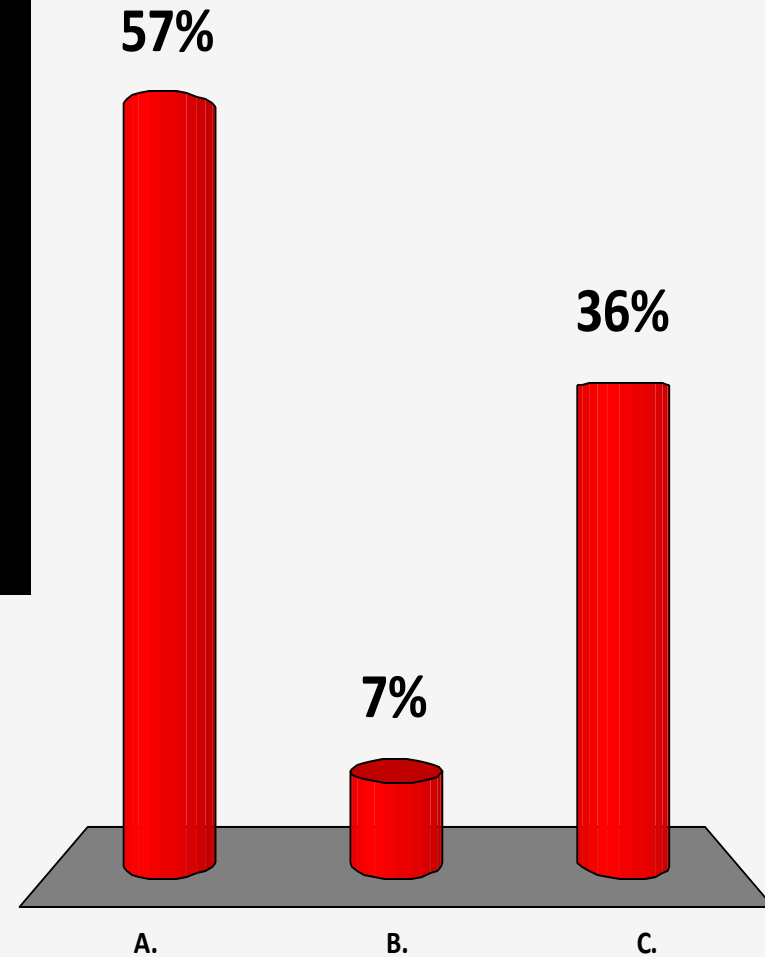
# Post-test: control & treatment



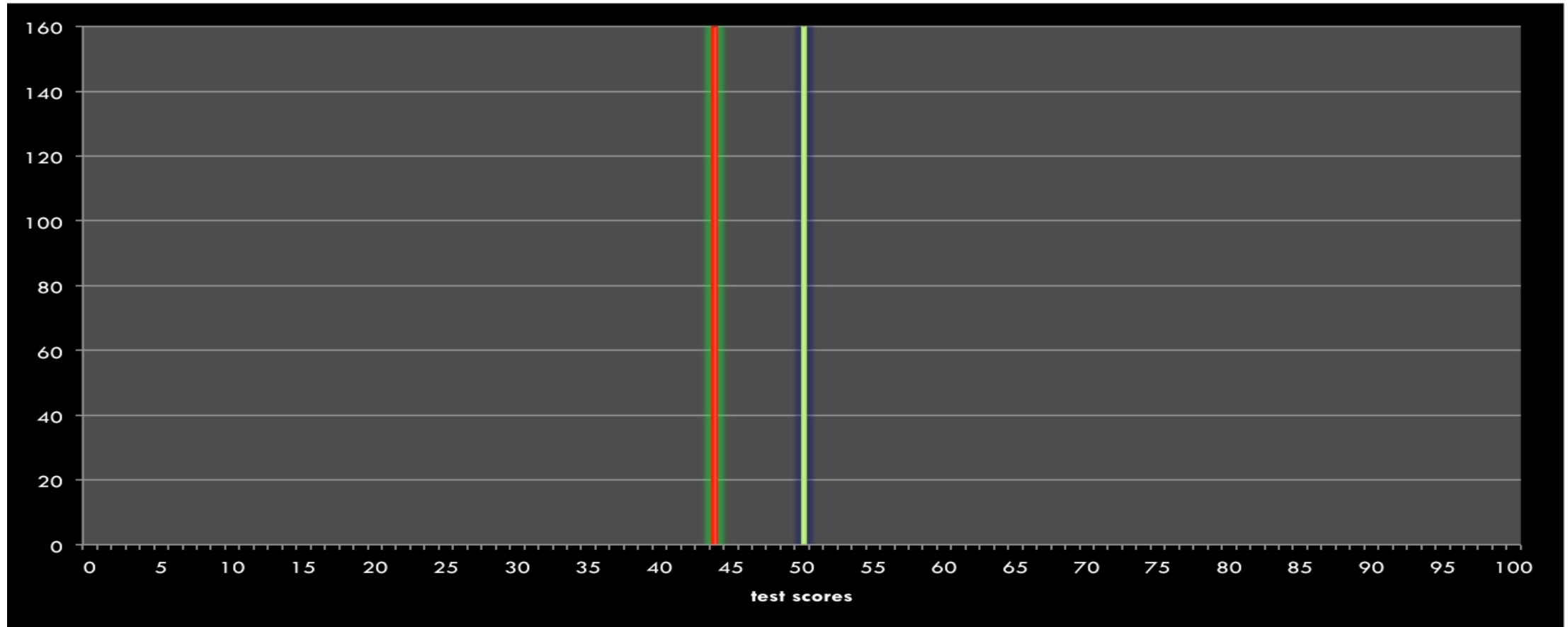
# Is this impact statistically significant?



- A. Yes
- B. No
- C. Don't know

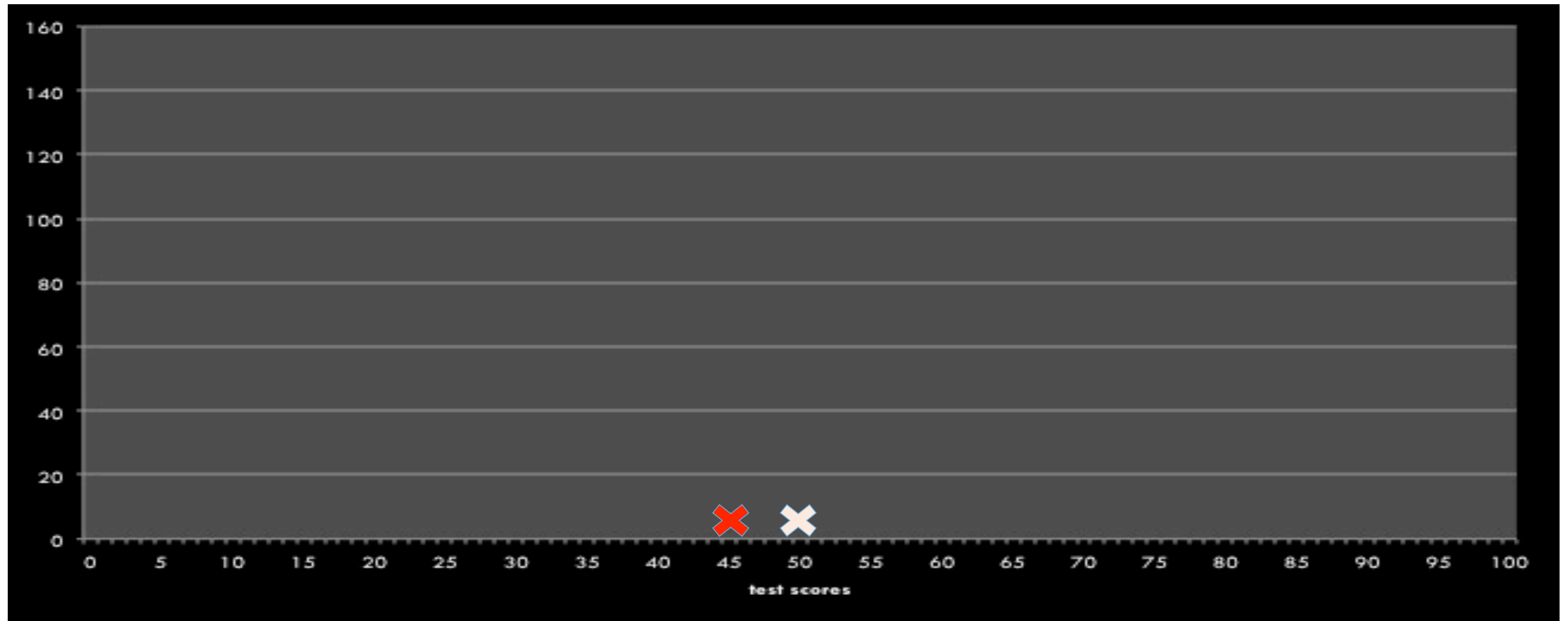


# One experiment: 6 points

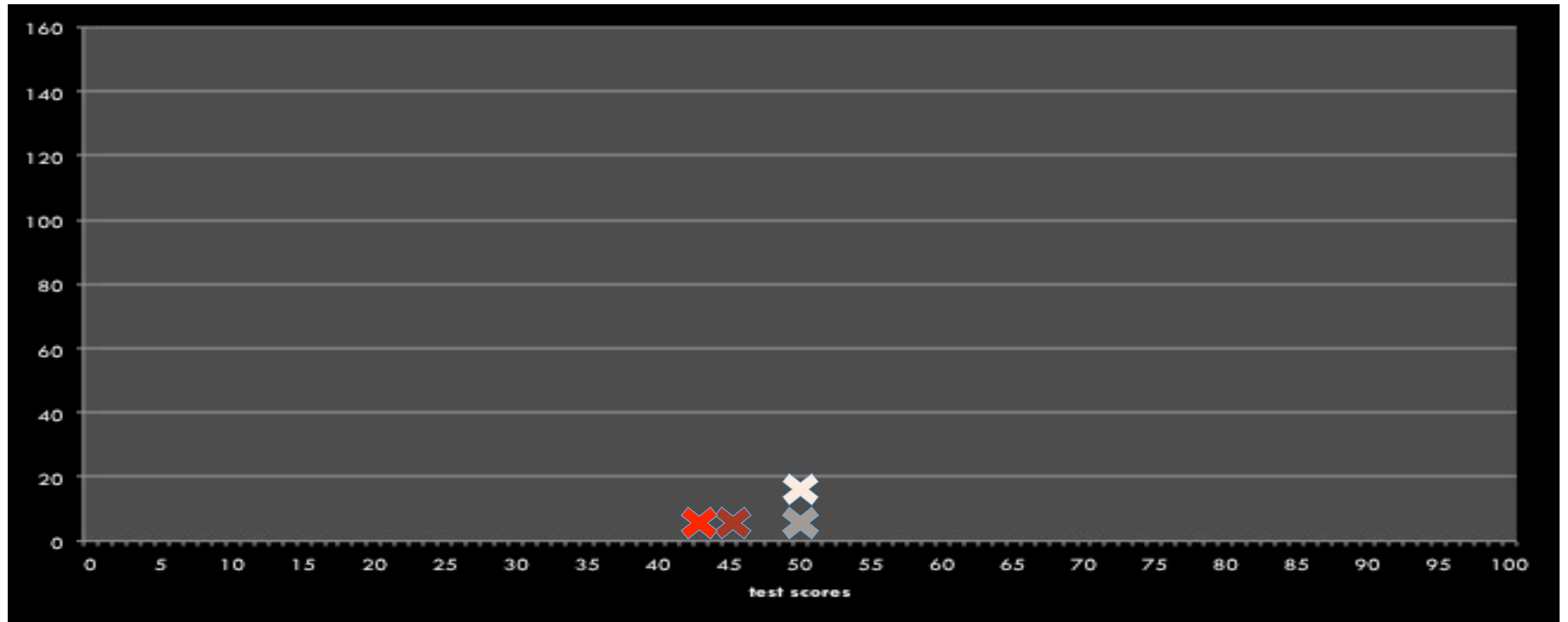




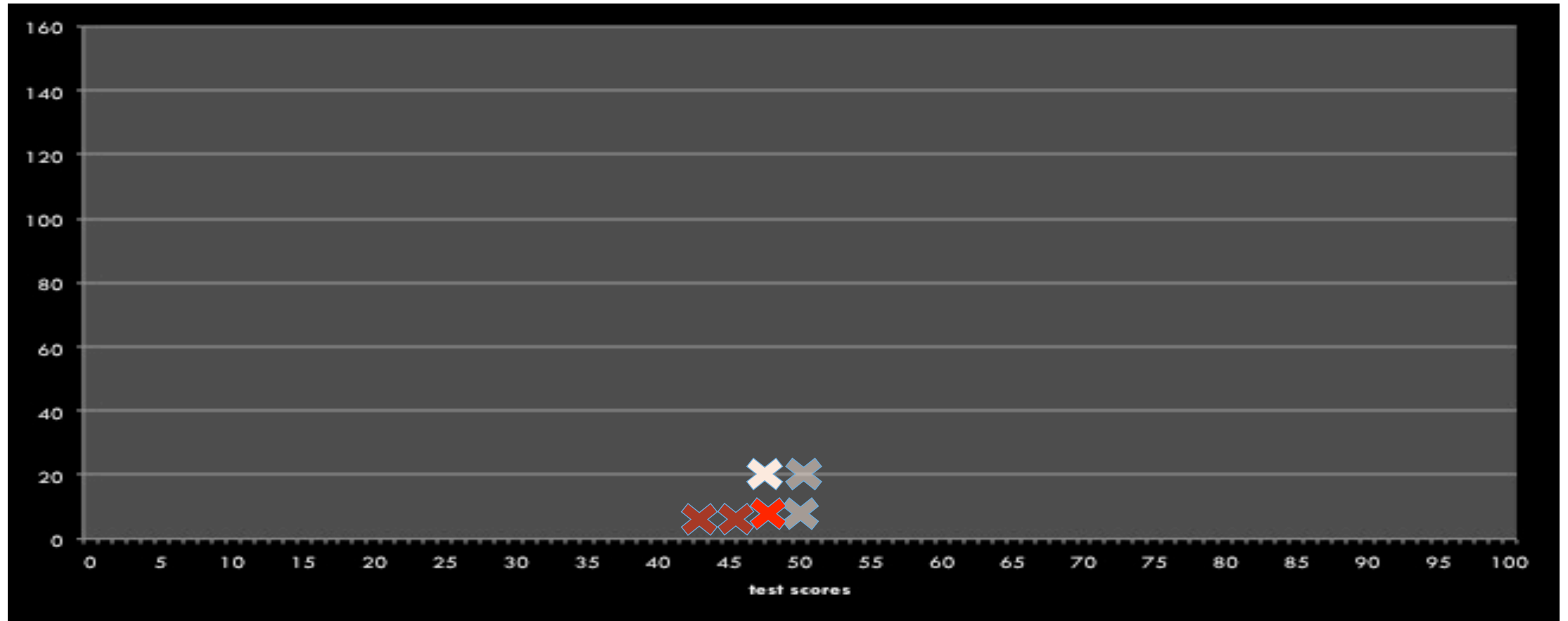
# One experiment



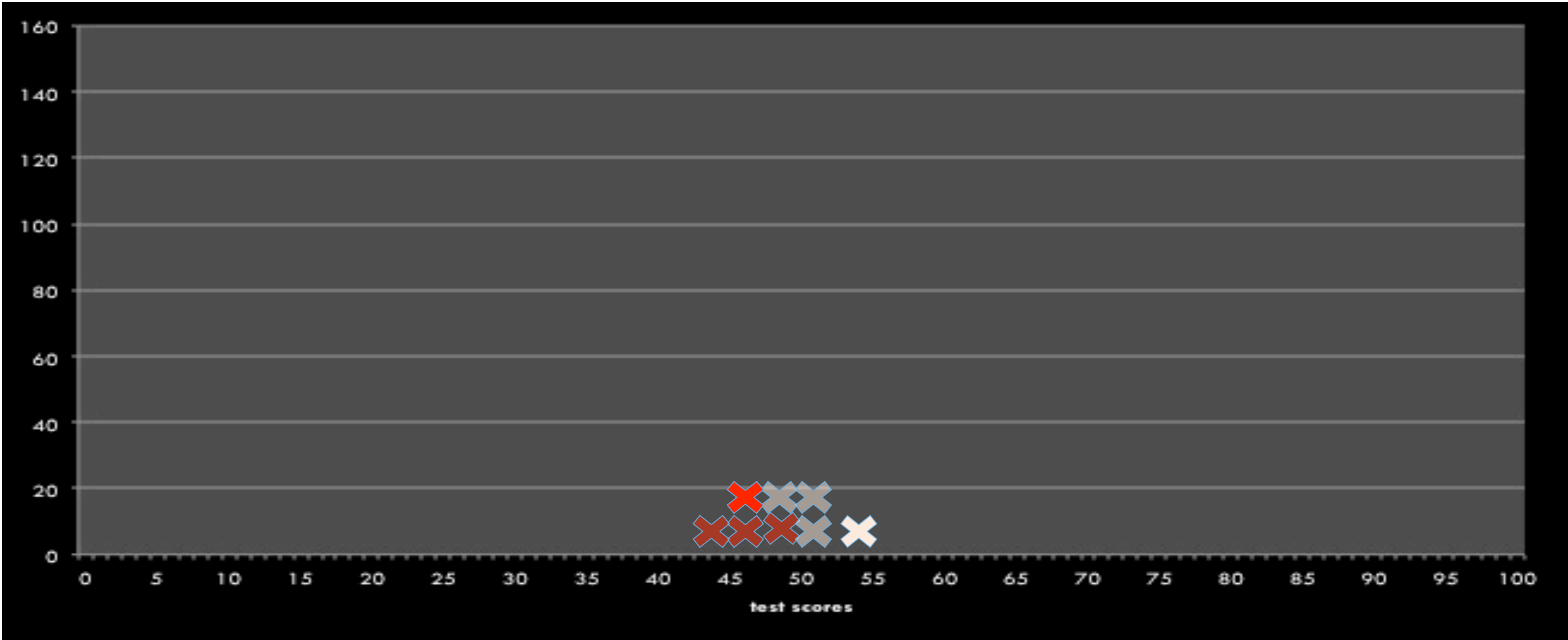
# Two experiments



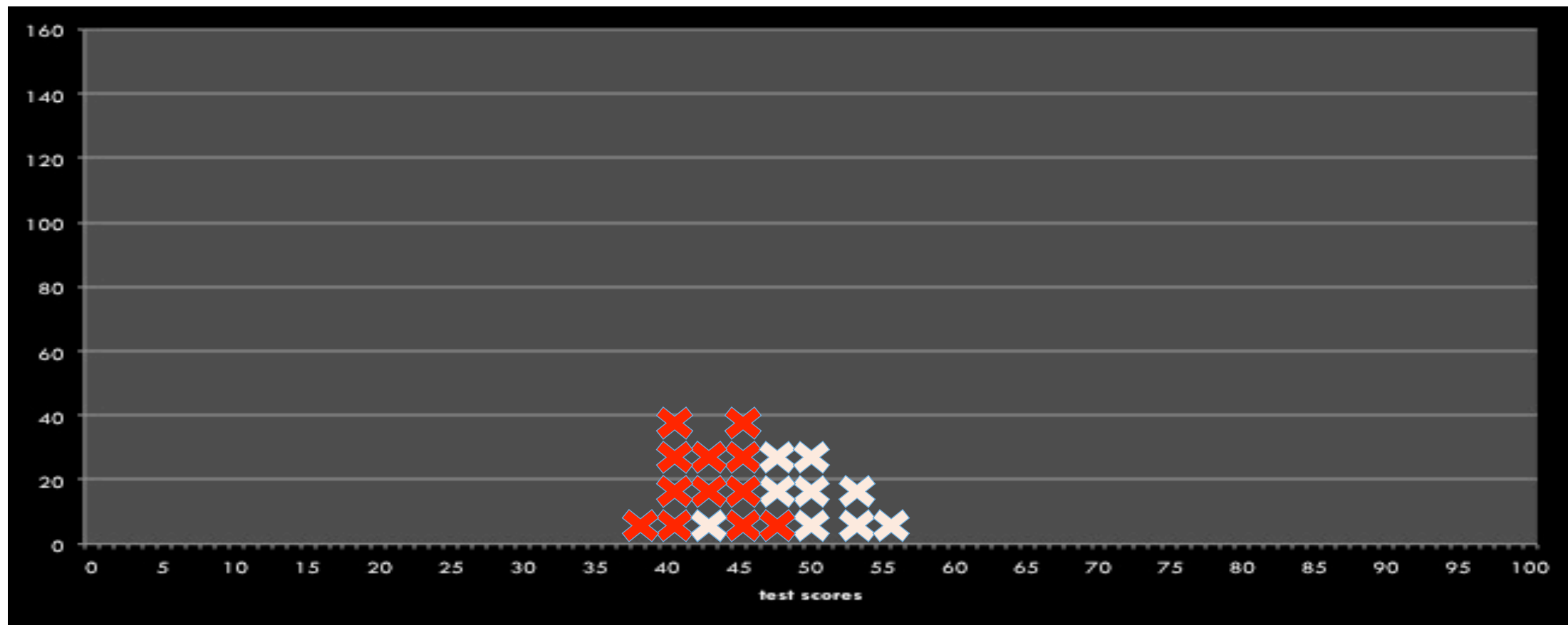
# A few more...



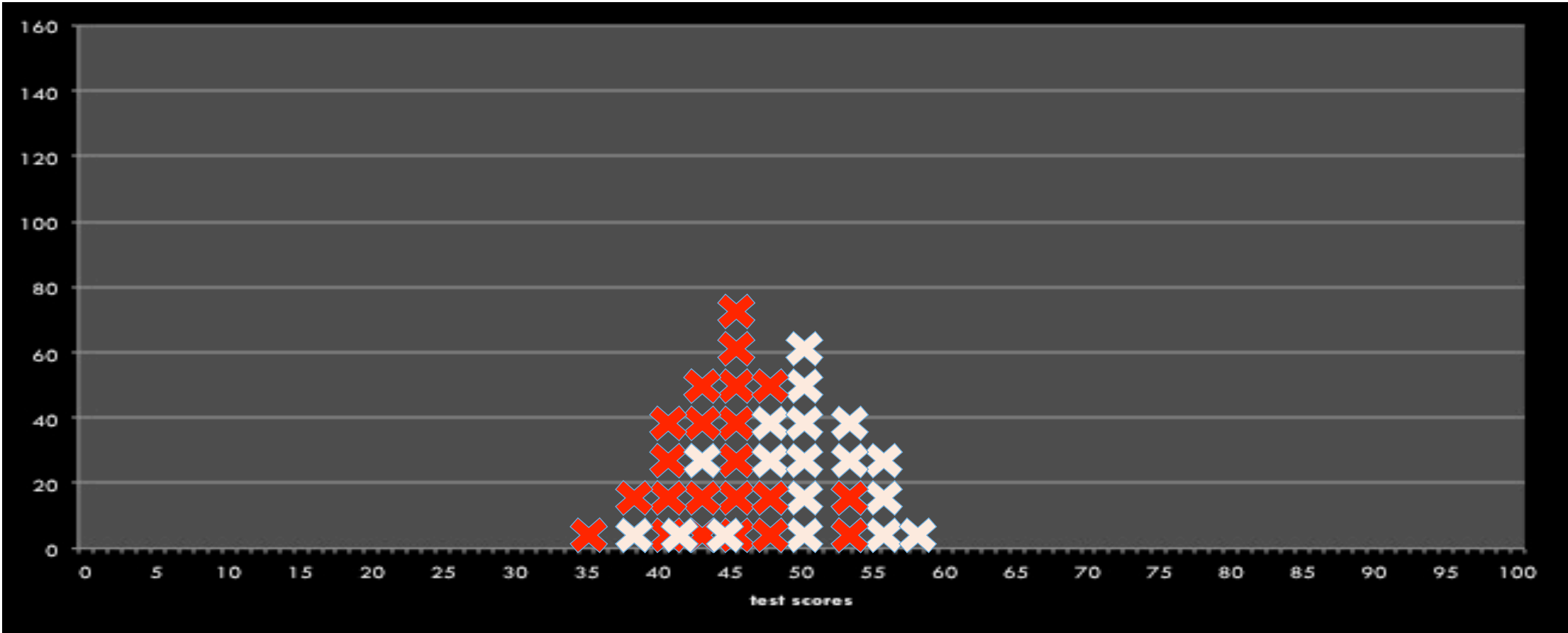
# A few more...



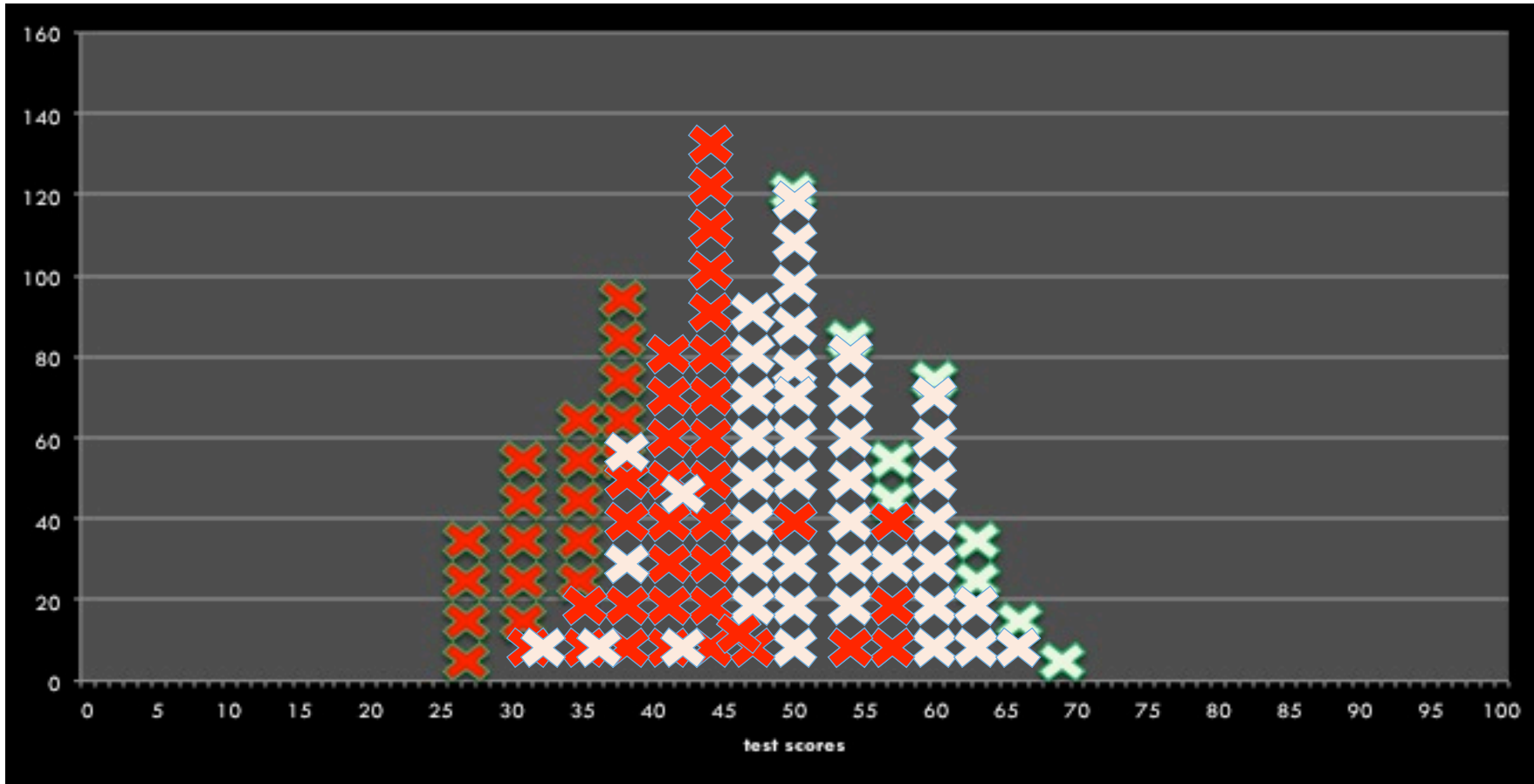
# Many more...



# A whole lot more...

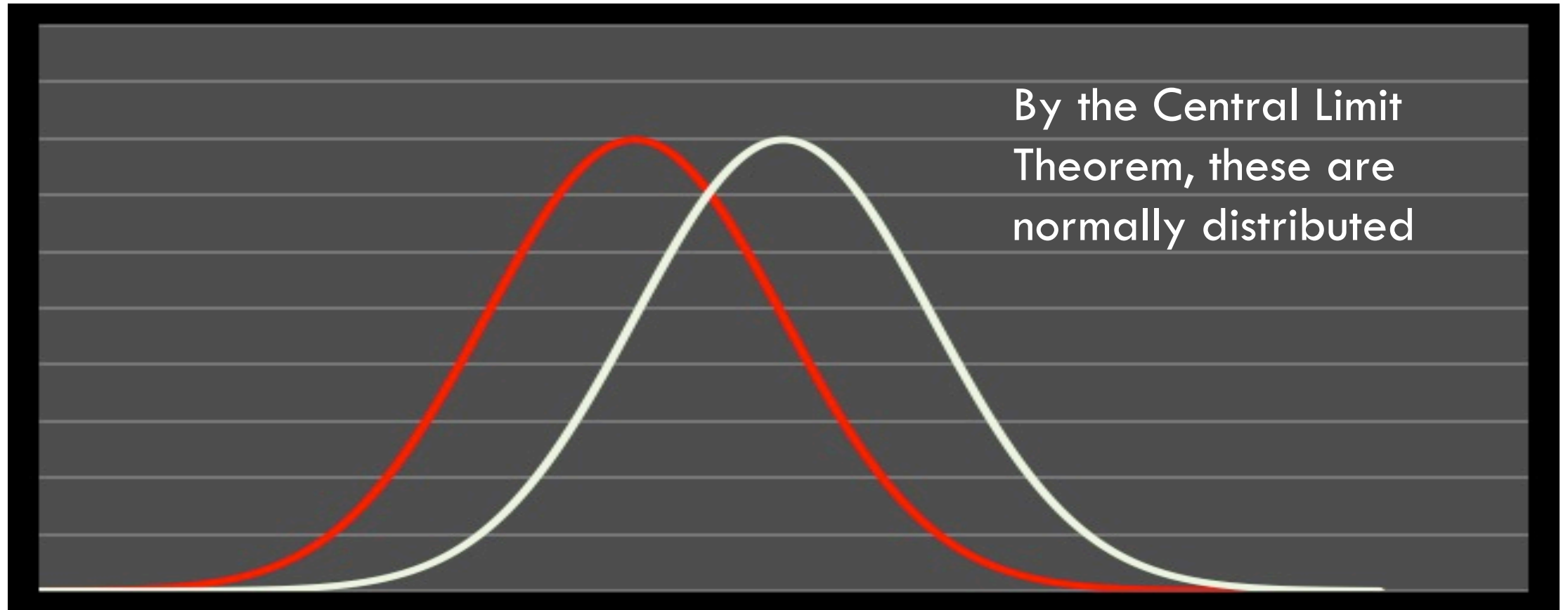


• • •



# Running the experiment thousands of times...

---





# Hypothesis Testing

---

- In criminal law, most institutions follow the rule: “innocent until proven guilty”
- The presumption is that the accused is innocent and the burden is on the prosecutor to show guilt
  - The jury or judge starts with the “null hypothesis” that the accused person is innocent
  - The prosecutor has a hypothesis that the accused person is guilty

# Hypothesis Testing

---

- In program evaluation, instead of “presumption of innocence,” the rule is: “presumption of insignificance”
- The “Null hypothesis” ( $H_0$ ) is that there was no (zero) impact of the program
- The burden of proof is on the evaluator to show a significant effect of the program

# Hypothesis Testing: Conclusions

---





- If it is very unlikely (less than a 5% probability) that the difference is solely due to chance:
  - We “reject our null hypothesis”
- We may now say:
  - “our program has a statistically significant impact”

# What is the significance level?

---

- Type I error: rejecting the null hypothesis even though it is true (false positive)
- Significance level: The probability that we will reject the null hypothesis even though it is true

# Hypothesis testing: 95% confidence

		YOU CONCLUDE	
		<i>Effective</i>	<i>No effect</i>
THE TRUTH	<i>Effective</i>		<i>Type II Error</i> (low power) 
	<i>No effect</i>	<i>Type I Error</i> (5% of the time) 	

# What is Power?

---

- Type II Error: Failing to reject the null hypothesis (concluding there is no difference), when indeed the null hypothesis is false.
- Power: If there is a measureable effect of our intervention (the null hypothesis is false), the probability that we will detect an effect (reject the null hypothesis)

# Hypothesis Testing: Steps

---

1. Determine the (size of the) sampling distribution around the null hypothesis  $H_0$  by calculating the standard error
2. Choose the confidence interval, e.g. 95% (or significance level:  $\alpha$ ) ( $\alpha=5\%$ )
3. Identify the critical value (boundary of the confidence interval)
4. If our observation falls in the critical region we can reject the null hypothesis

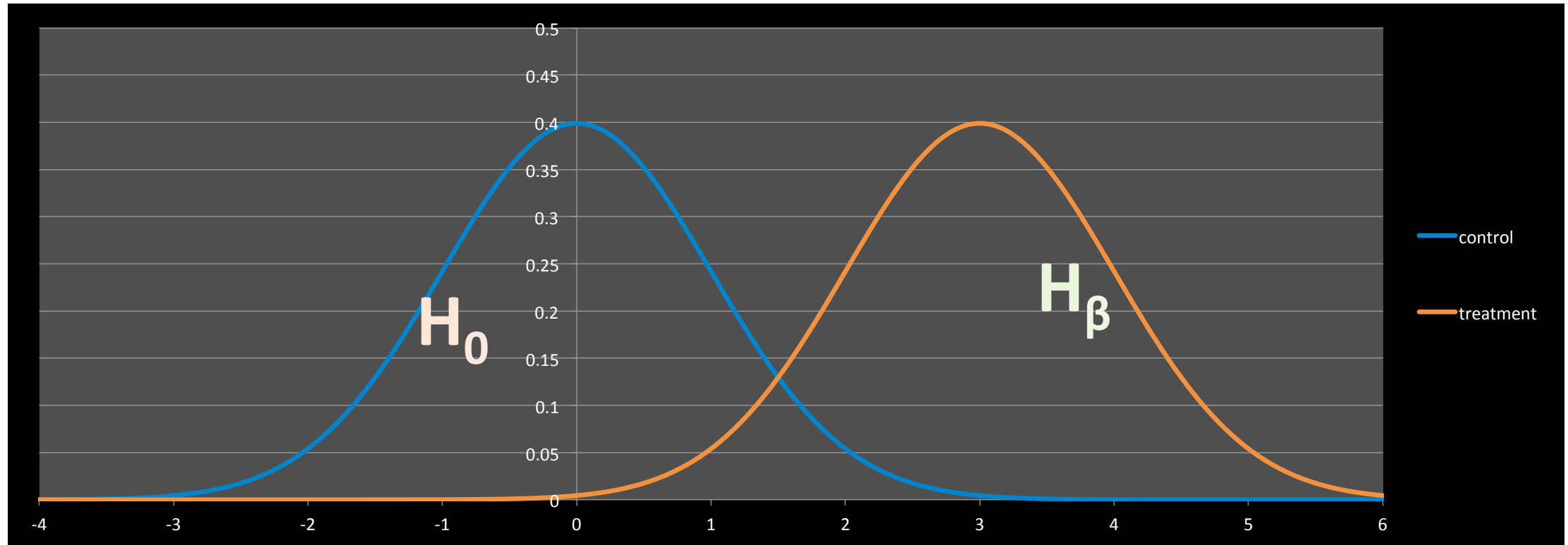
# Determining Power: Steps

---

1. Determine the (size of the) sampling distribution around the null hypothesis  $H_0$  by calculating the standard error
2. Hypothesize an effect size  $H_\beta$
3. Determine the (size of the) sampling distribution around the alternate hypothesis
4. Choose the confidence interval, e.g. 95% (or significance level:  $\alpha$ ) ( $\alpha=5\%$ )
5. Identify the critical value (boundary of the confidence interval)
6. Determine where in the  $H_\beta$  sampling distribution, the critical value lies.
7. Calculate the proportion of the mass under the  $H_\beta$  sampling distribution that lies on the other side of the critical value (away from the null hypothesis)

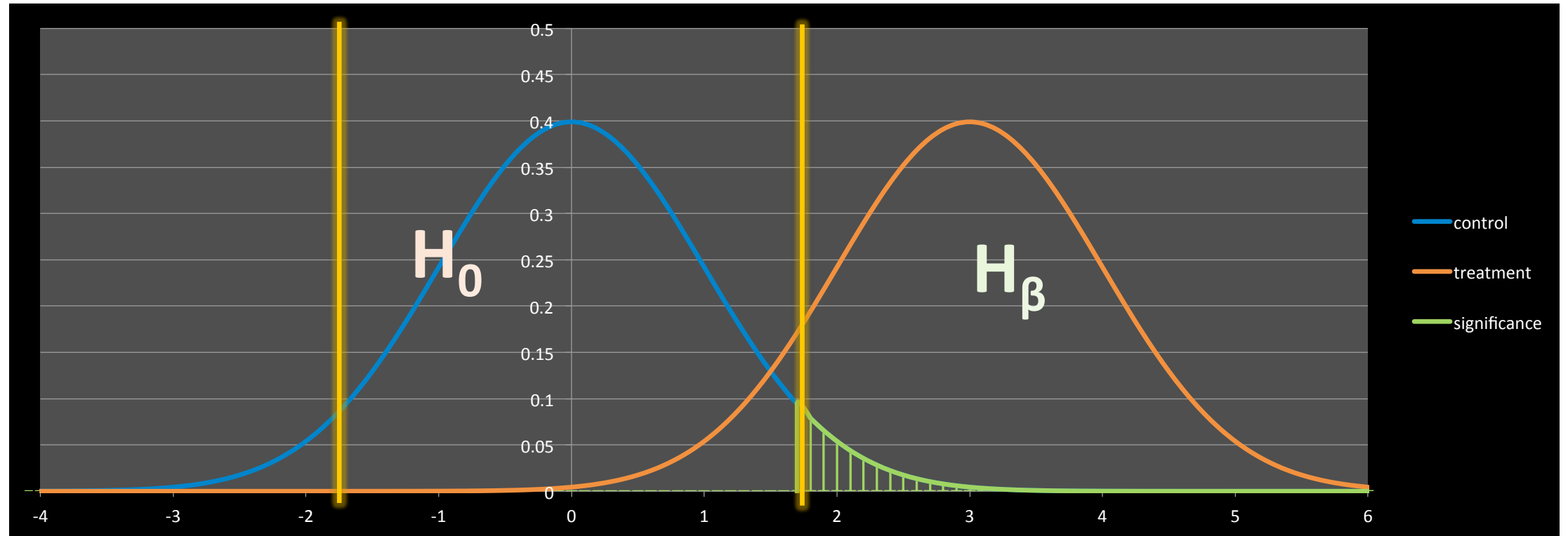


# Before the experiment



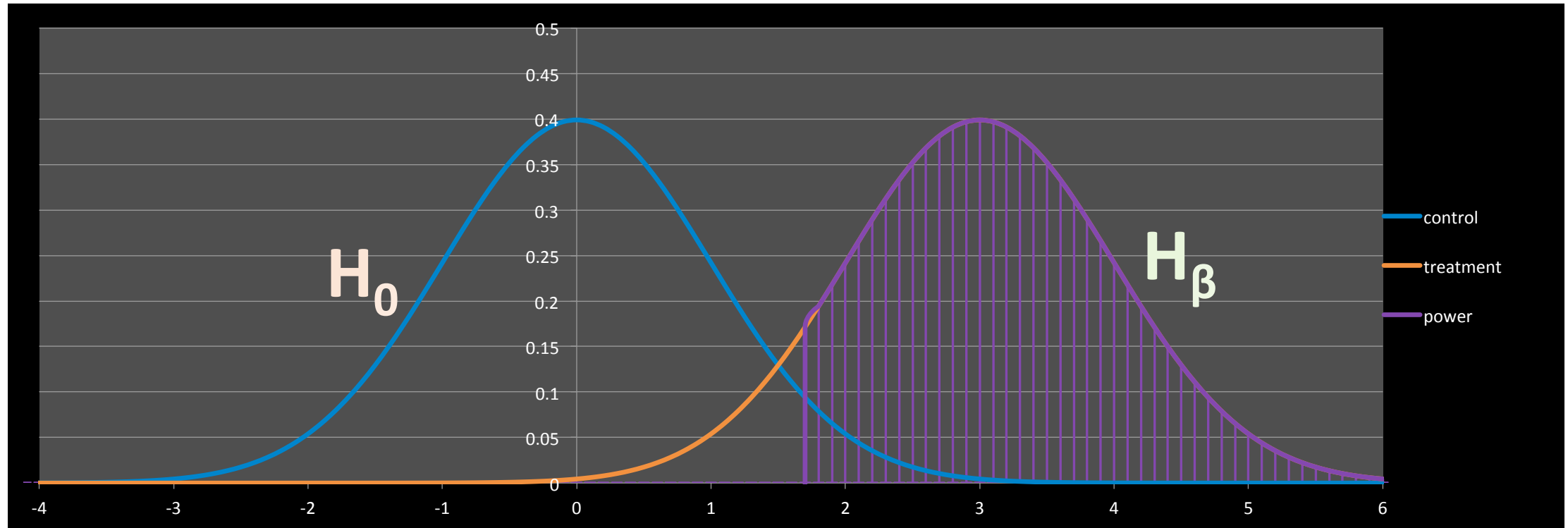
- Assume two effects: no effect and treatment effect  $\beta$

# Impose significance level of 5%



Anything between lines cannot be distinguished from 0

# Can we distinguish $H_\beta$ from $H_0$ ?



Shaded area shows % of time we would find  $H_\beta$  true if it was

# What influences power?

---

- What are the factors that change the proportion of the research hypothesis that is shaded—i.e. the proportion that falls to the right (or left) of the null hypothesis curve?
- Understanding this helps us design more powerful experiments

# Power: main ingredients

---

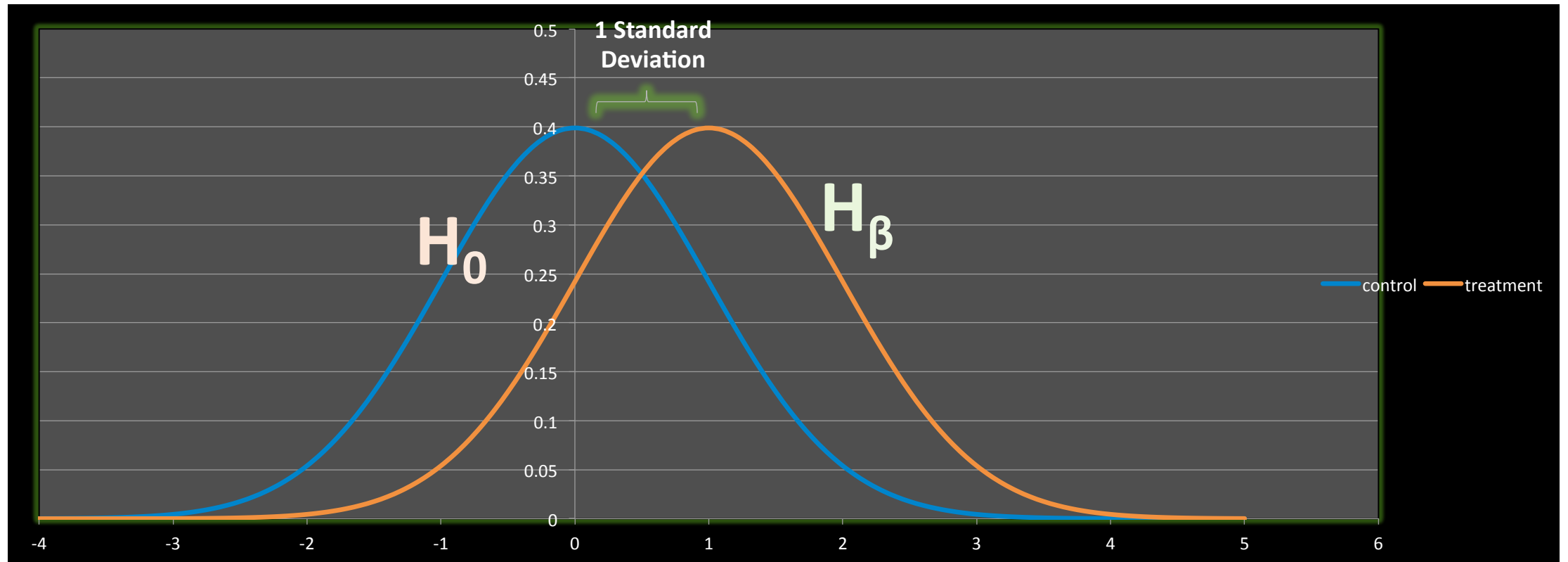
1. Effect Size
2. Sample Size
3. Variance
4. Proportion of sample in T vs. C
5. Clustering

# Power: main ingredients

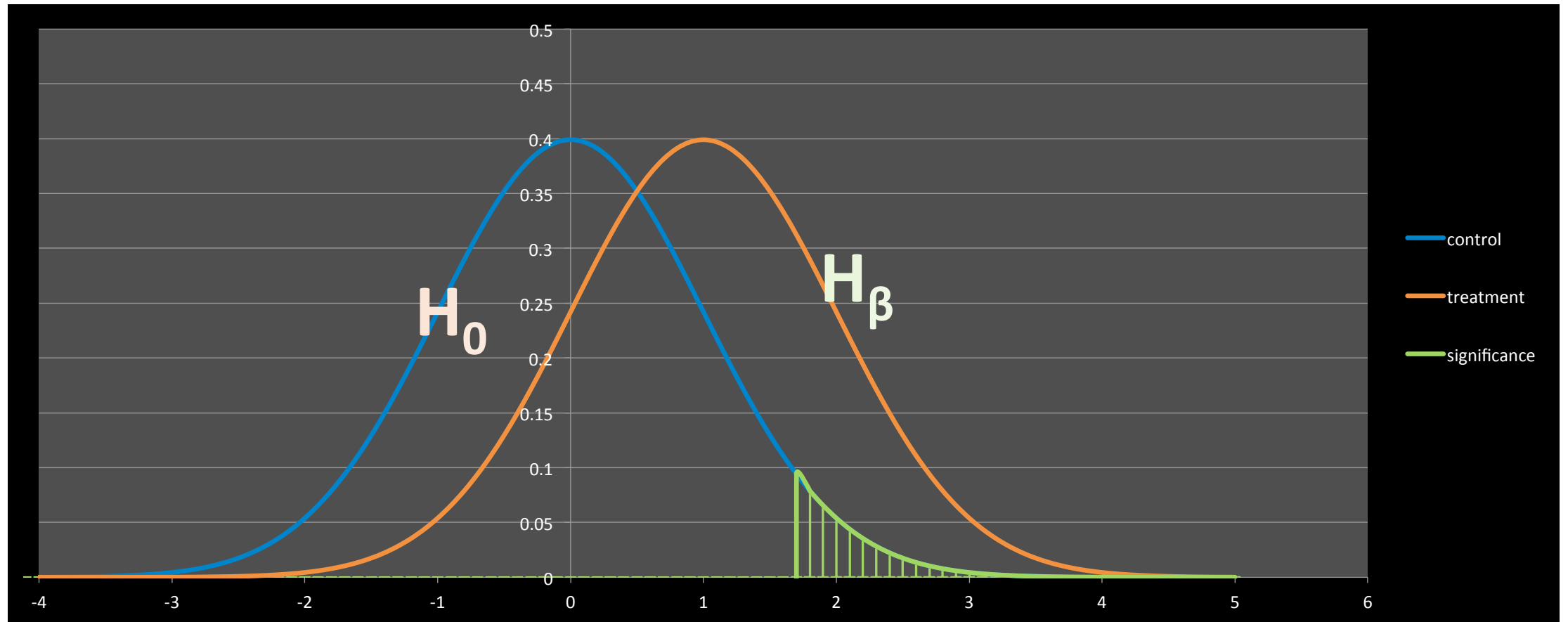
---

- 1. Effect Size to be detected***
2. Sample Size
3. Variance
4. Proportion of sample in T vs. C
5. Clustering

# Effect Size: $1 * SE$

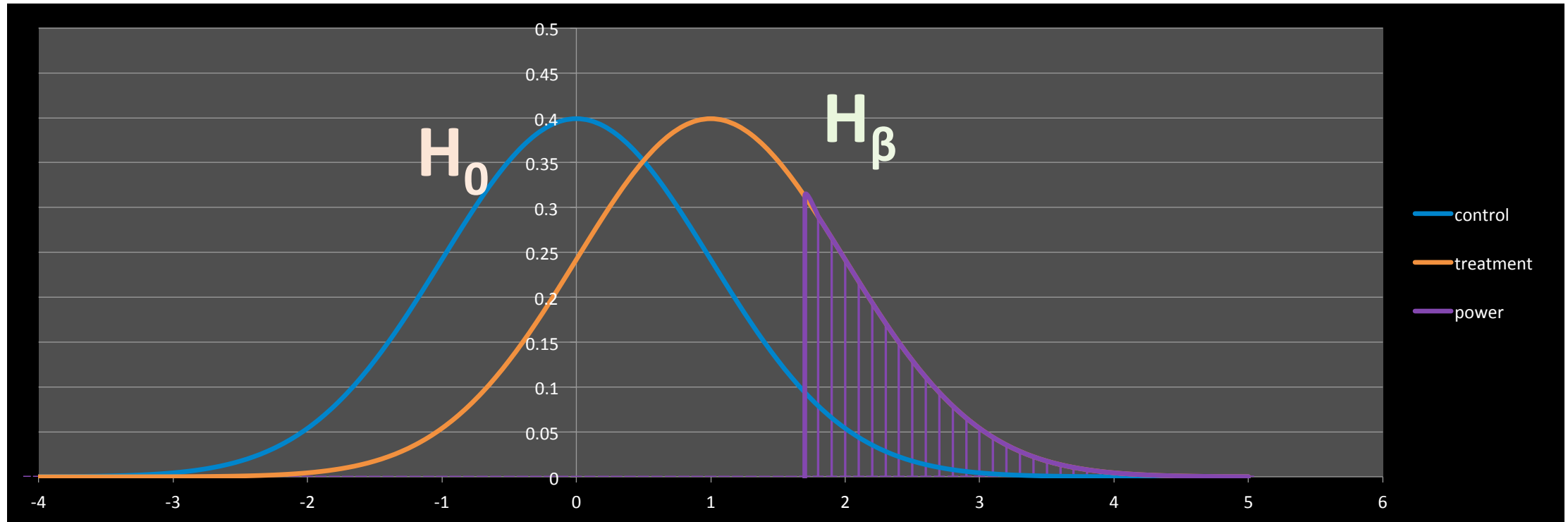


# Effect Size = $1 * SE$



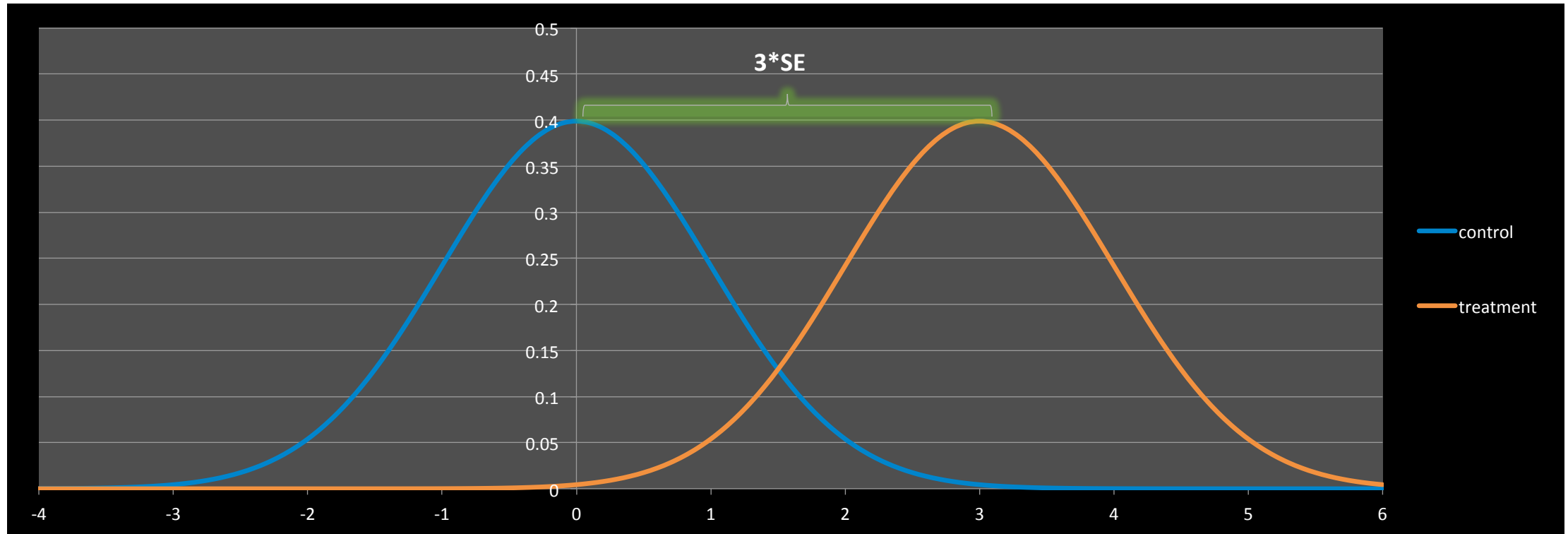


Power: 26%  
If the true impact was  $1 \cdot SE \dots$



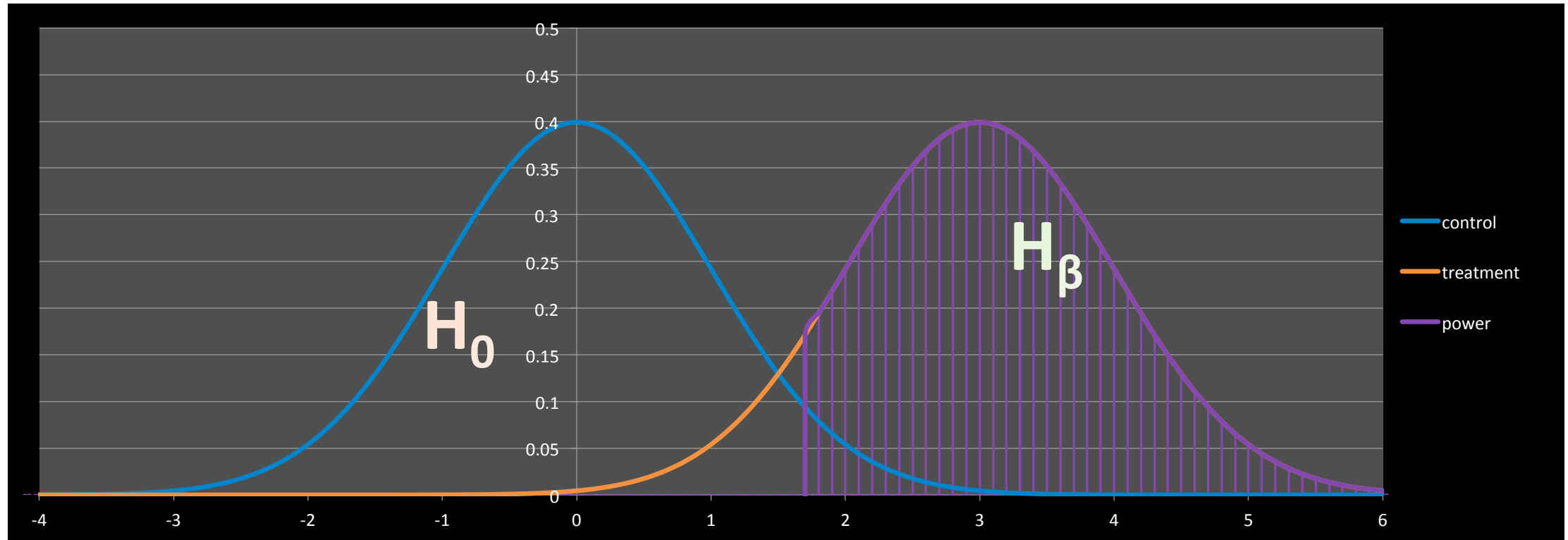
The Null Hypothesis would be rejected only 26% of the time

# Effect Size: $3*SE$



Bigger hypothesized effect size  $\rightarrow$  distributions farther apart

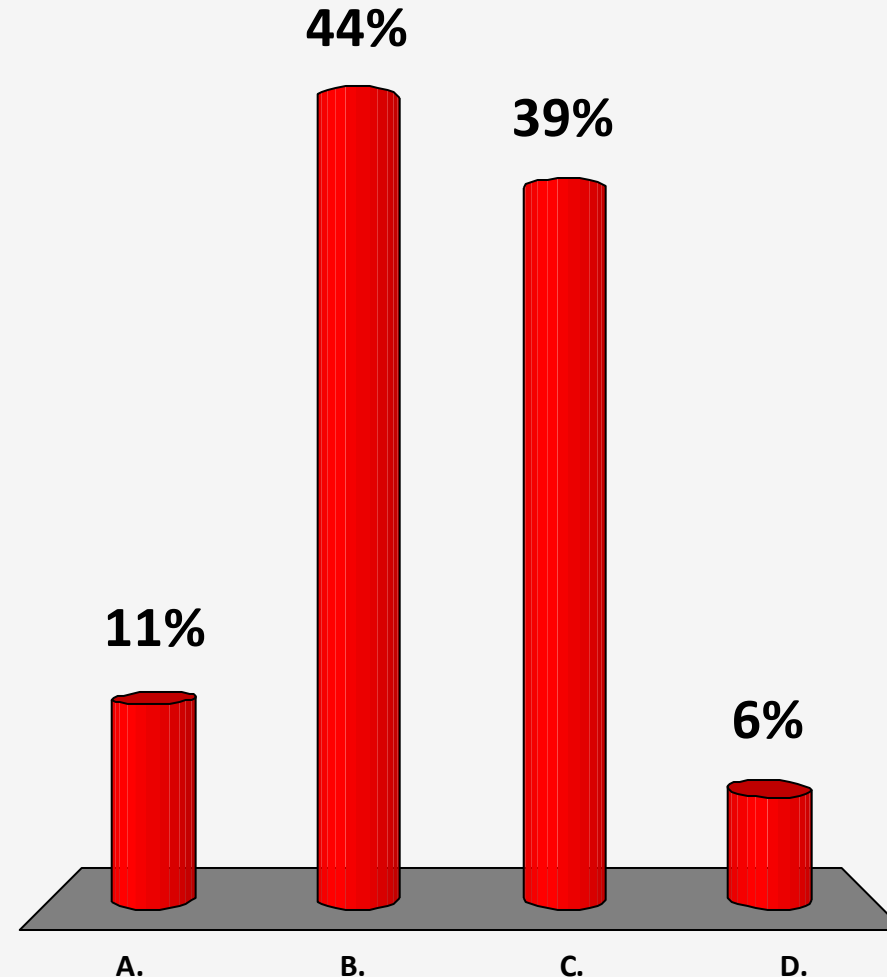
# Effect size $3 * SE$ : Power= 91%



Bigger Effect size means more power

# What effect size should you use when designing your experiment?

- A. Smallest effect size that is still cost effective
- B. Largest effect size you expect your program to produce
- C. Both
- D. Neither

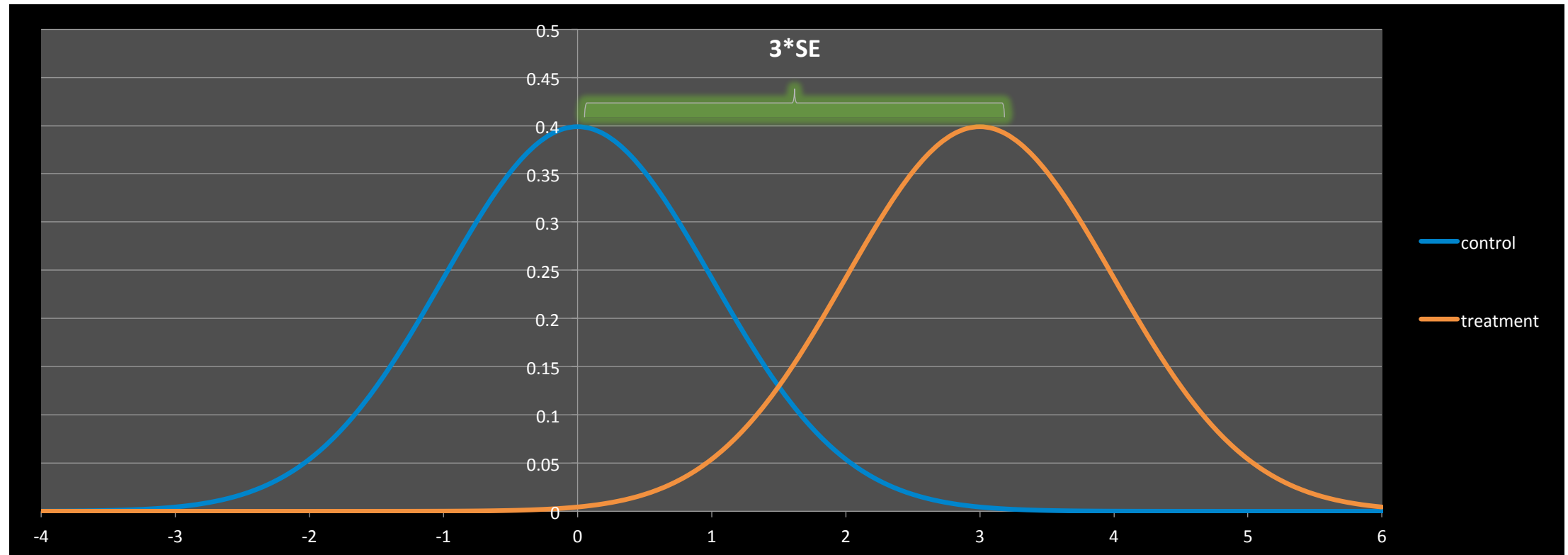


# Effect size and take-up

---

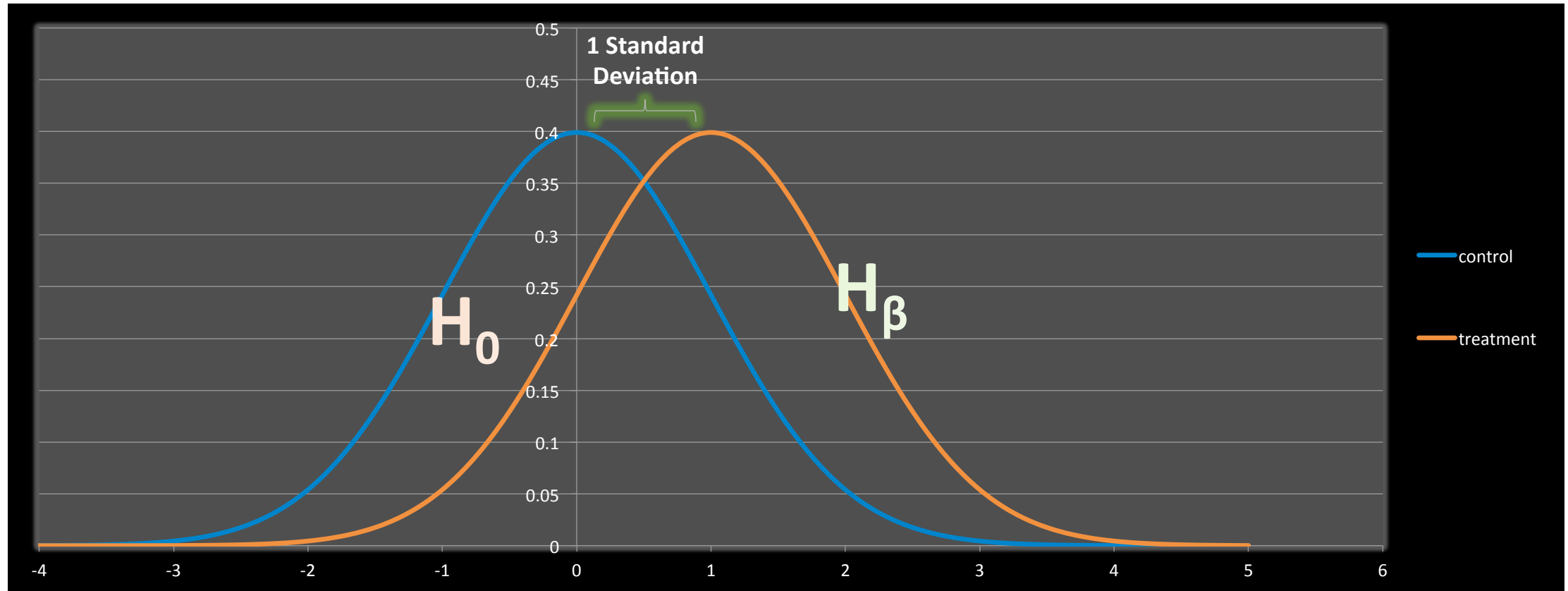
- Let's say we believe the impact on our participants is “3”
- What happens if take up is 1/3?
- Let's show this graphically

# Effect Size: $3*SE$

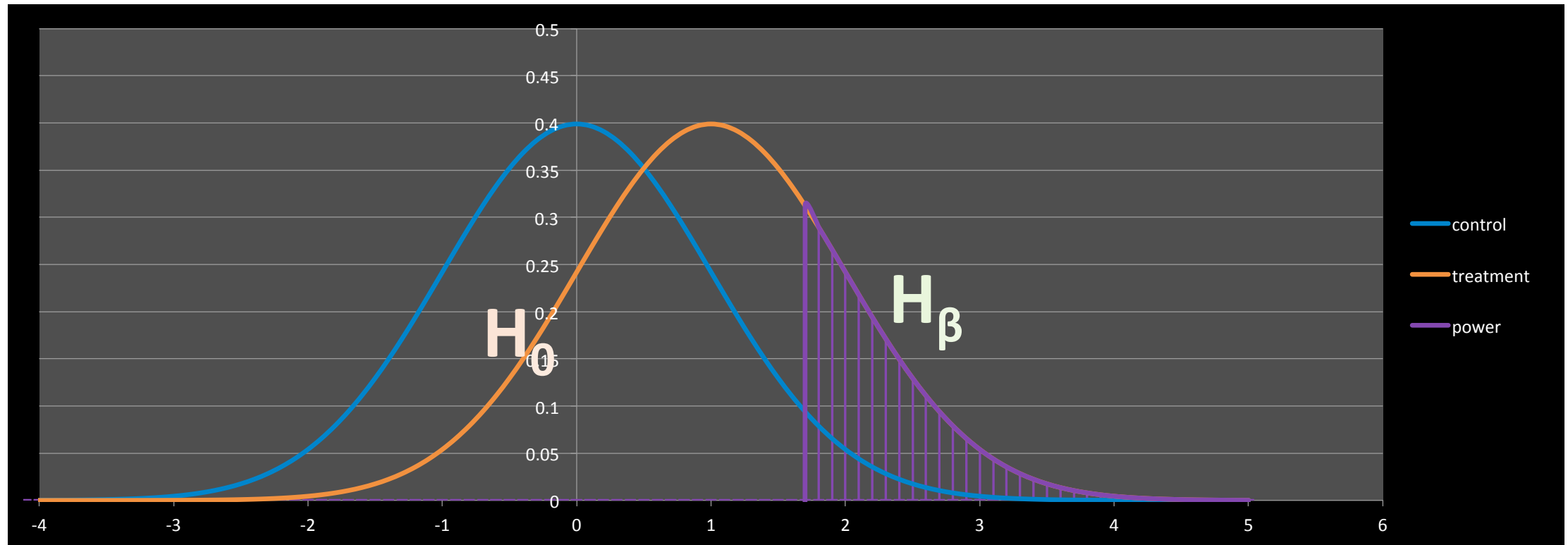


Let's say we believe the impact on our participants is "3"

# Take up is 33%. Effect size is 1/3rd



# Back to: Power = 26%



**Take-up is reflected in the effect size**



# Picking an effect size

---

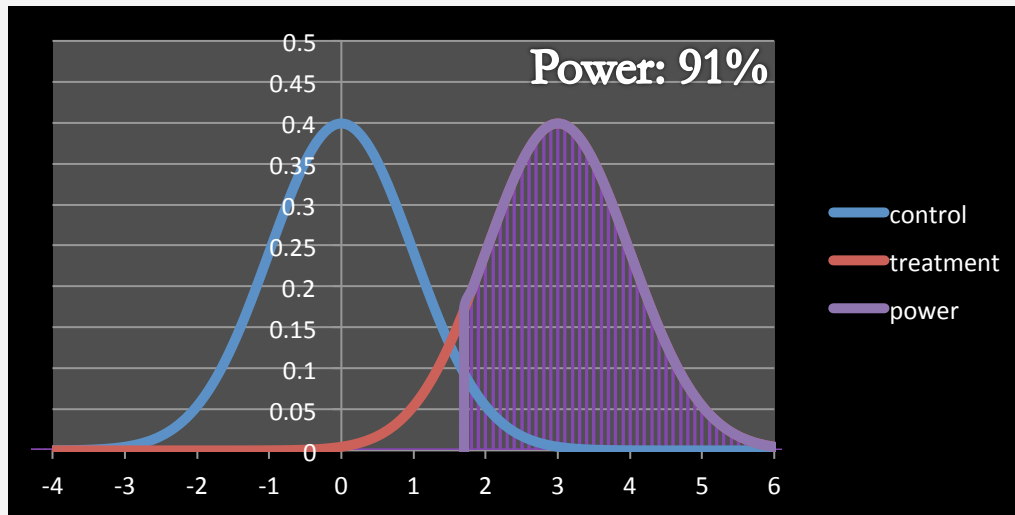
- What is the smallest effect that should justify the program being adopted?
- If the effect is smaller than that, it might as well be zero: we are not interested in proving that a very small effect is different from zero
- In contrast, any effect larger than that would justify adopting this program: we want to be able to distinguish it from zero

# Power: main ingredients

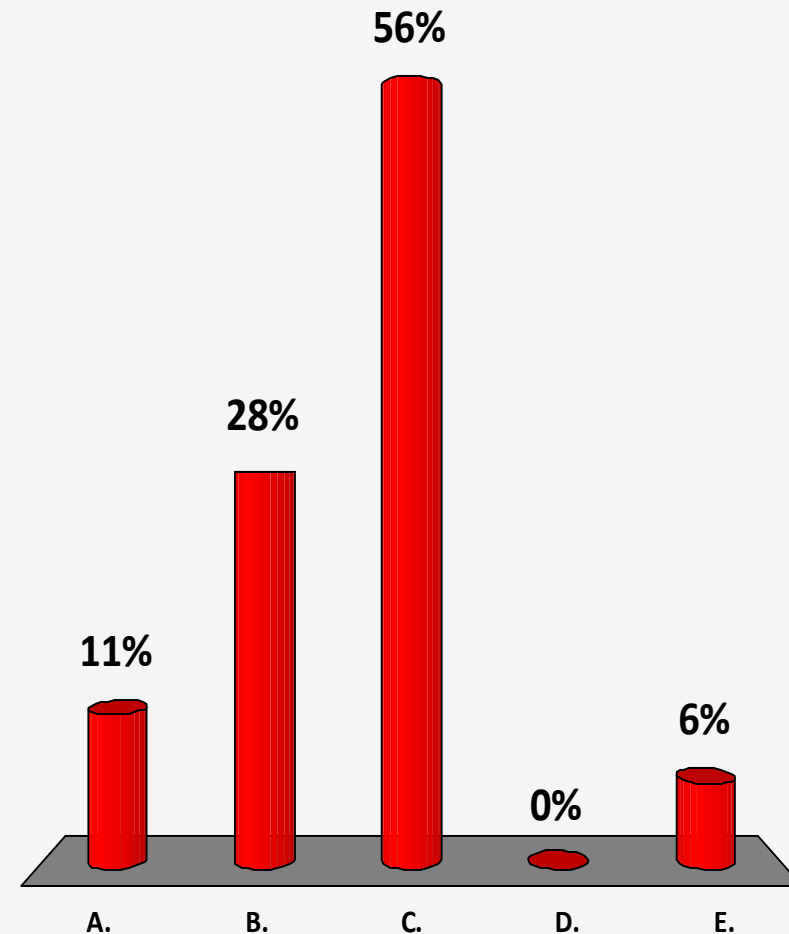
---

1. Effect Size
- 2. *Sample Size***
3. Variance
4. Proportion of sample in T vs. C
5. Clustering

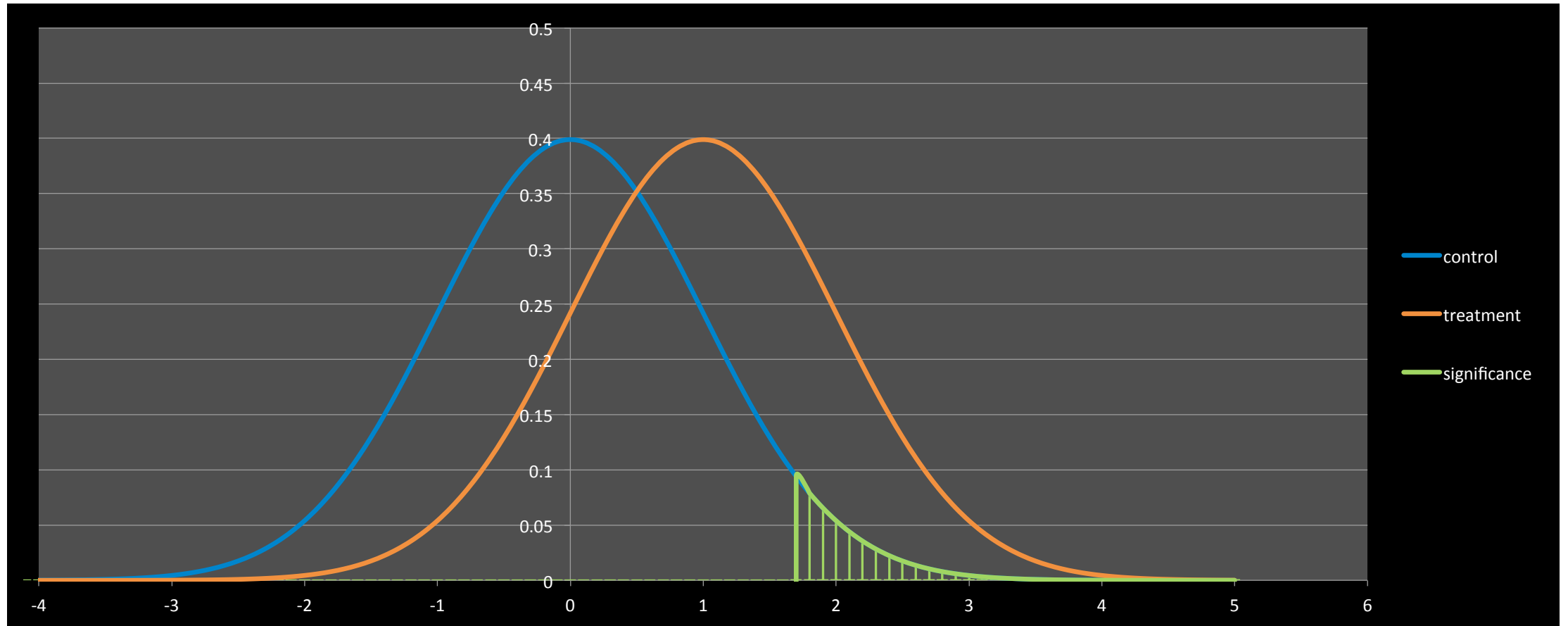
By increasing sample size  
you increase...



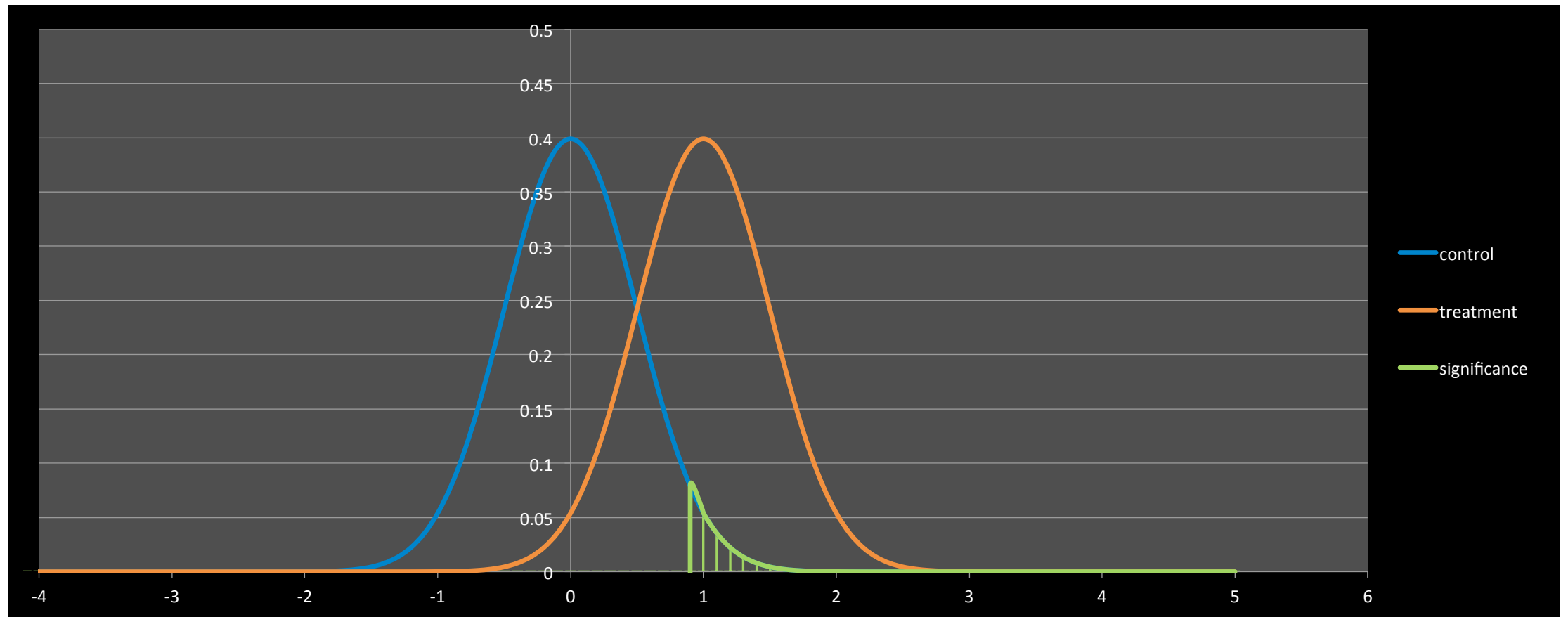
- A. Accuracy
- B. Precision
- C. Both
- D. Neither
- E. Don't know



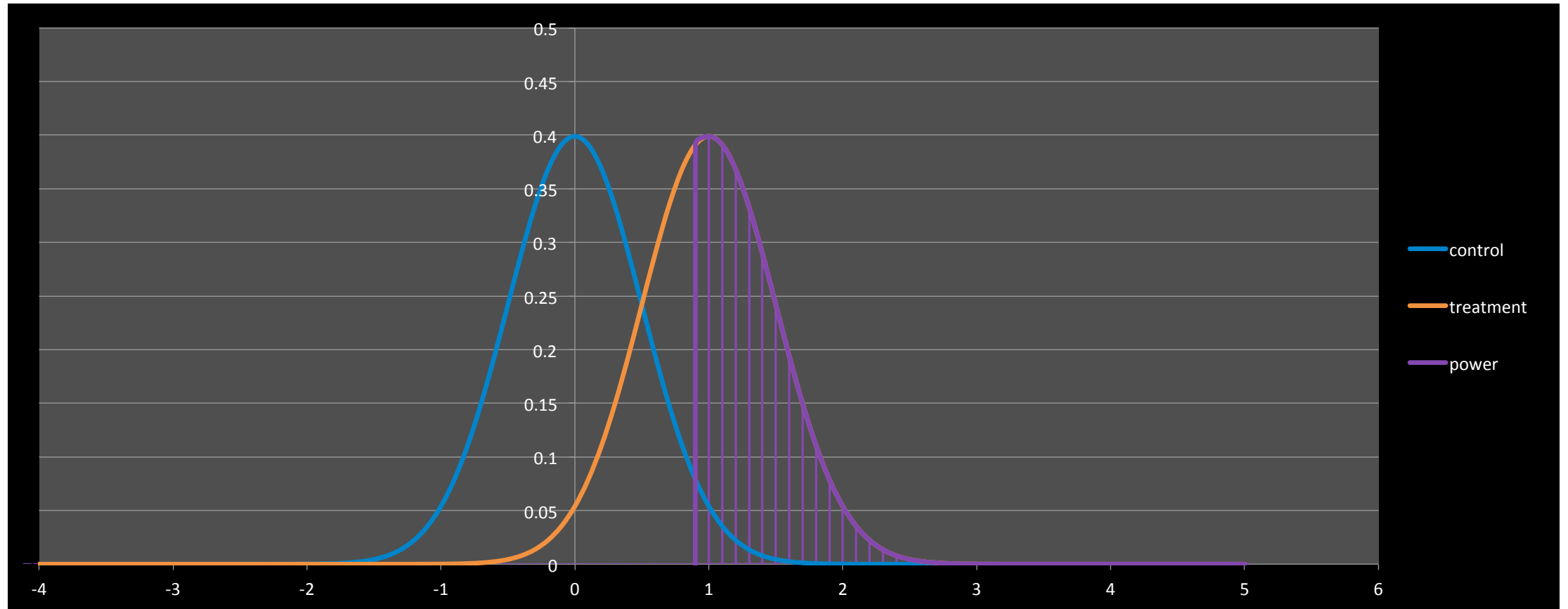
# Power: Effect size = 1SD, Sample size = N



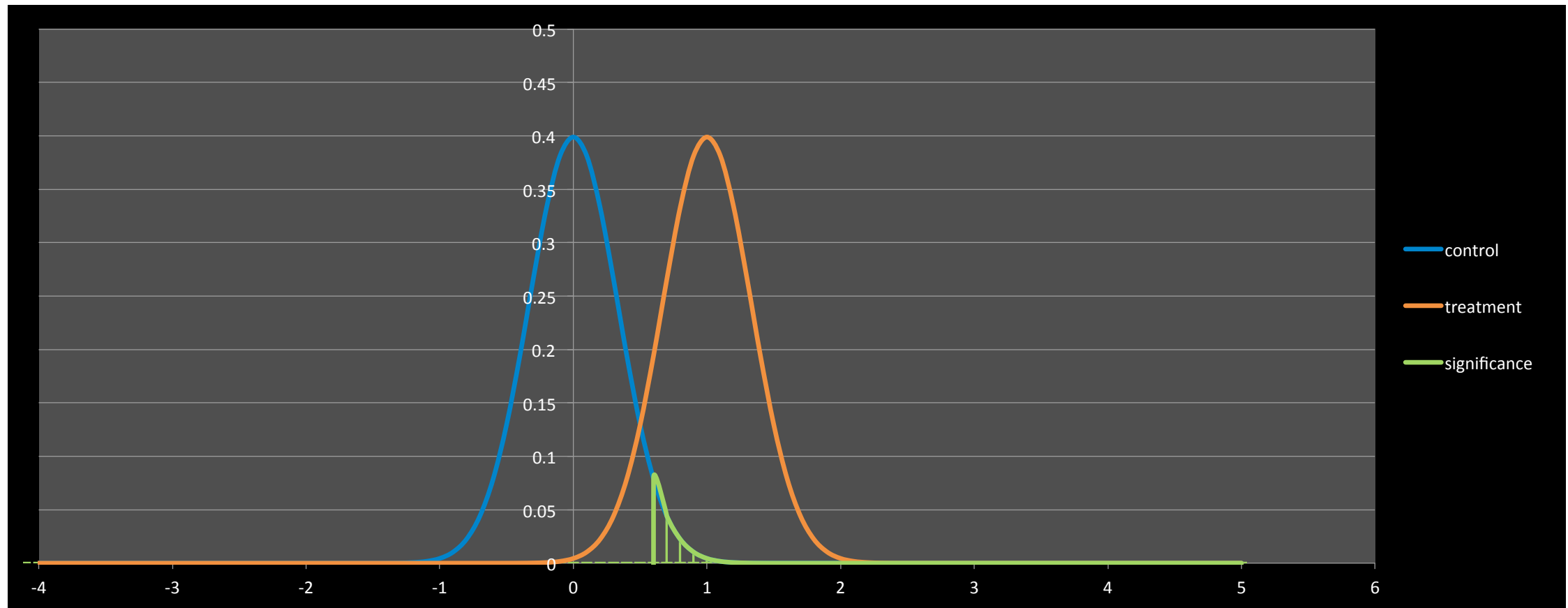
# Power: Sample size = 4N



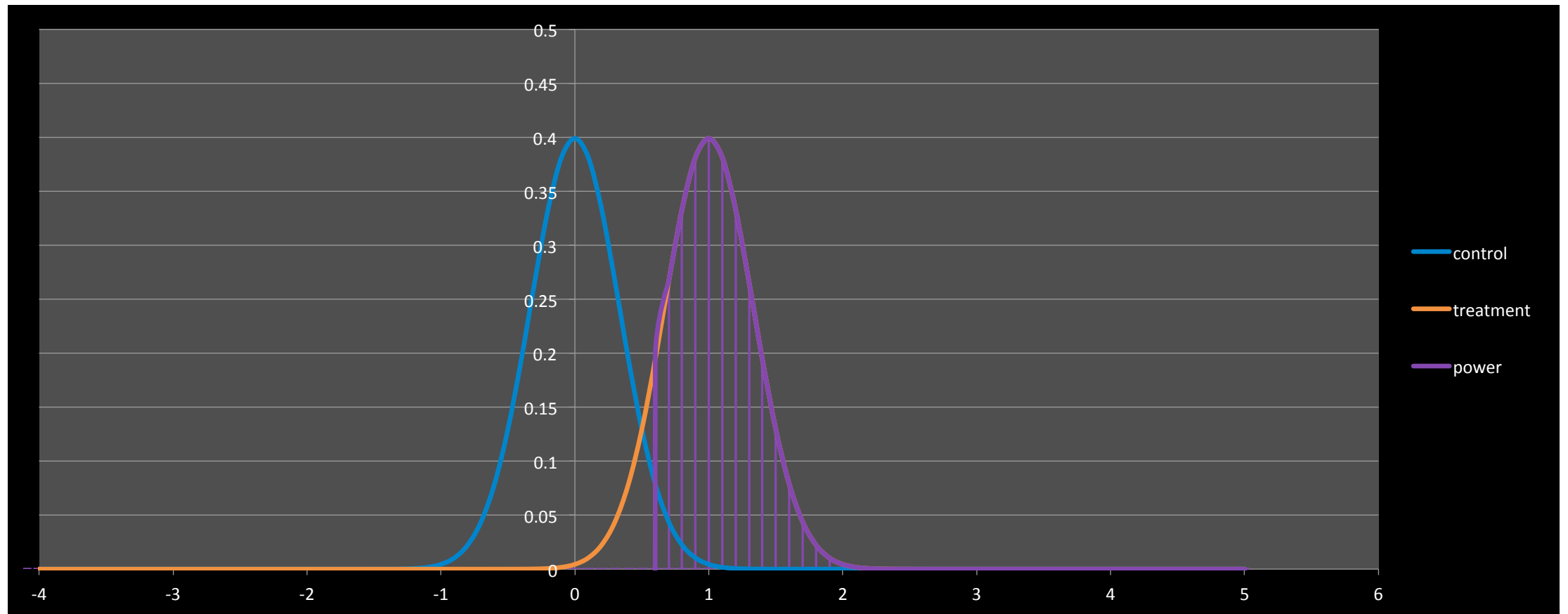
# Power: 64%



# Power: Sample size = 9



# Power: 91%





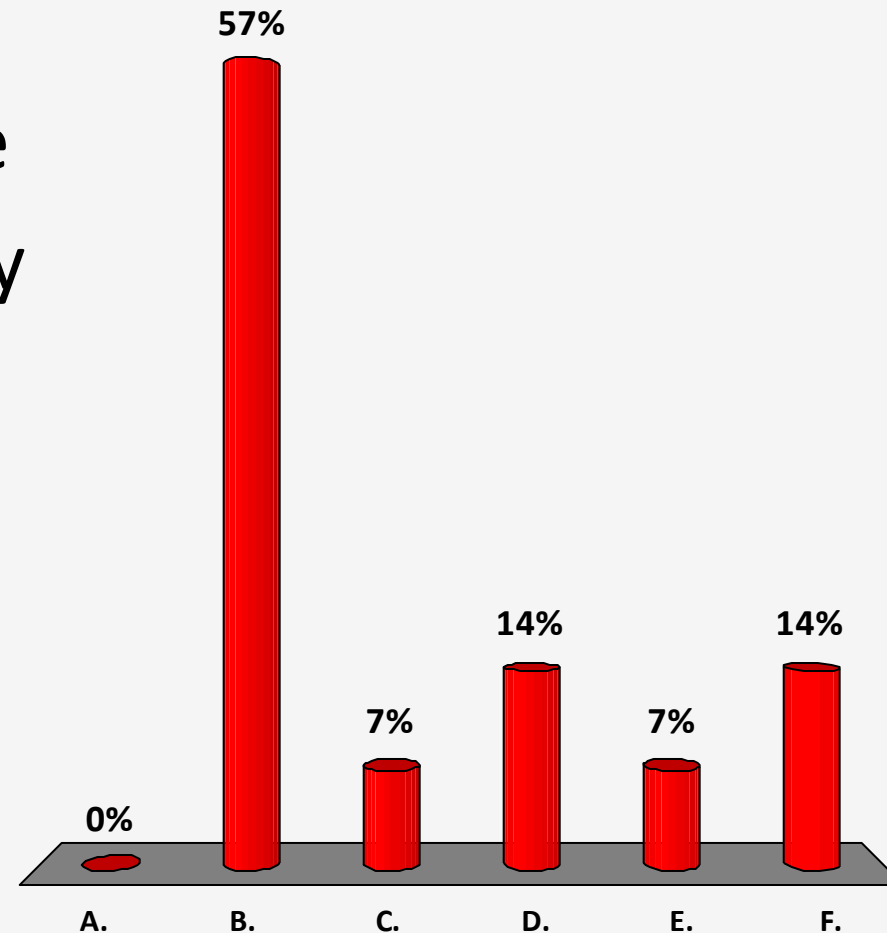
# Power: main ingredients

---

1. Effect Size
2. Sample Size
- 3. Variance**
4. Proportion of sample in T vs. C
5. Clustering

# What are typical ways to reduce the underlying (population) variance?

- A. Include covariates
- B. Increase the sample
- C. Do a baseline survey
- D. All of the above
- E. A and B
- F. A and C



# Variance

---

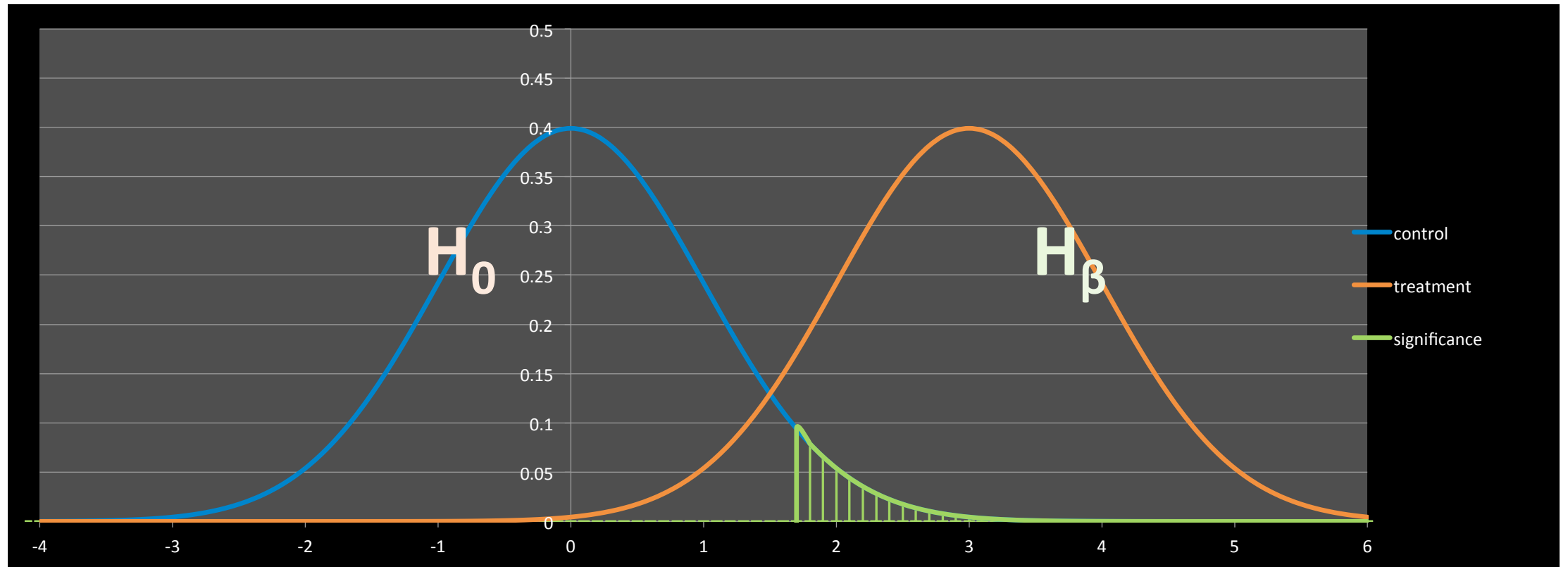
- There is sometimes very little we can do to reduce the noise
- The underlying variance is what it is
- We can try to “absorb” variance:
  - using a baseline
  - controlling for other variables
    - In practice, controlling for other variables (besides the baseline outcome) buys you very little

# Power: main ingredients

---

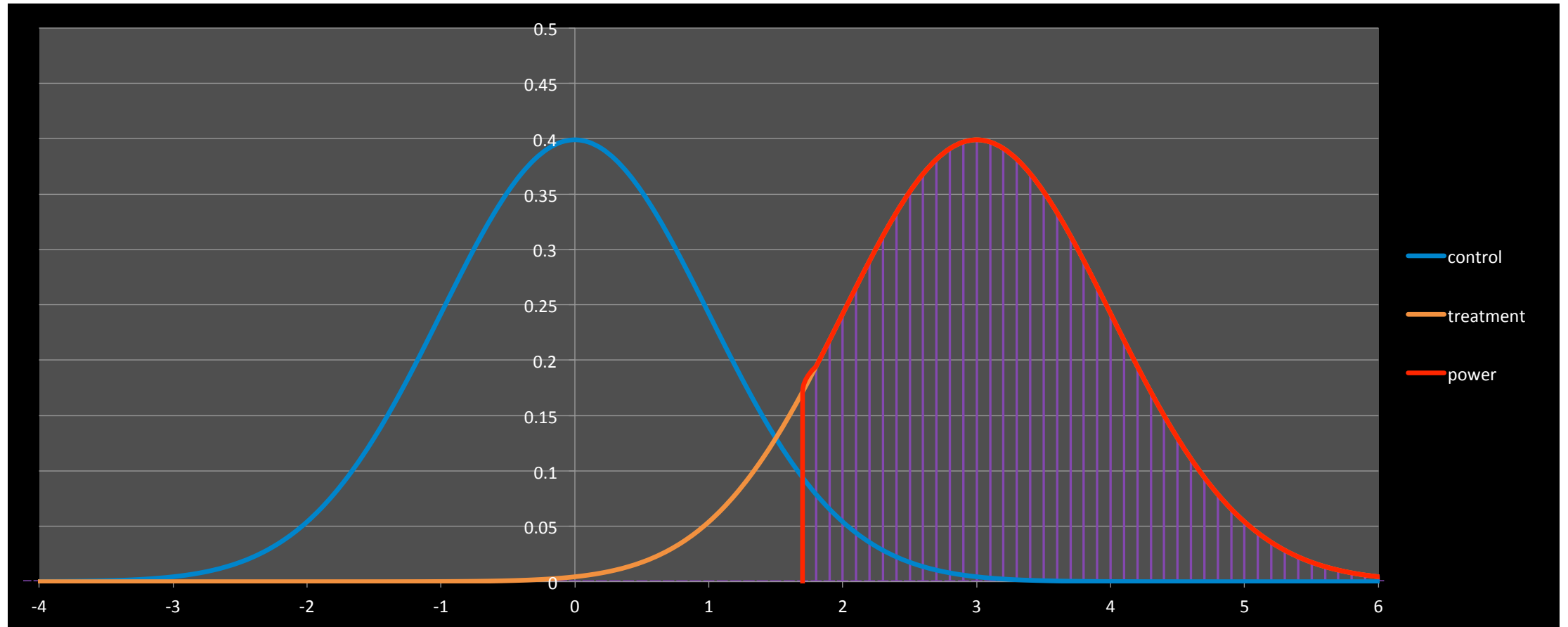
1. Effect Size
2. Sample Size
3. Variance
- 4. Proportion of sample in T vs. C**
5. Clustering

# Sample split: 50% C, 50% T



Equal split gives distributions that are the same “fatness”

# Power: 91% at effect size 3SD

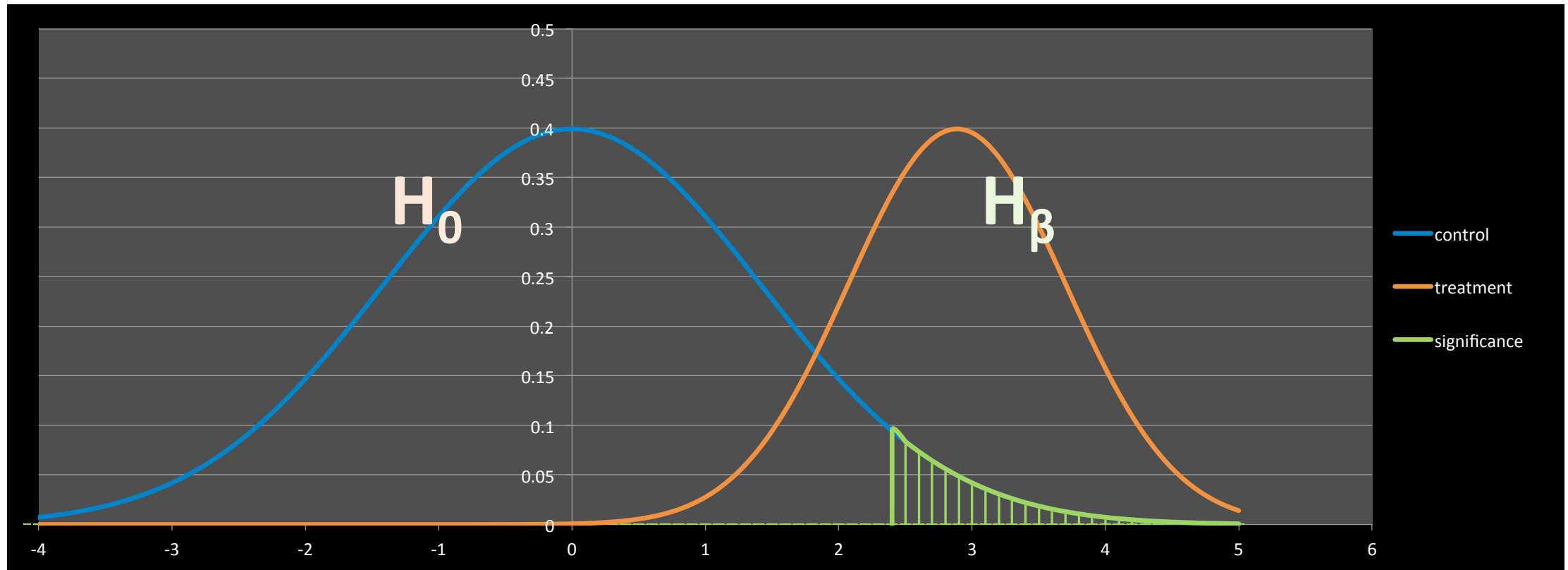


# If it's not 50-50 split?

---

- What happens to the relative fatness if the split is not 50-50.
- Say 25-75?

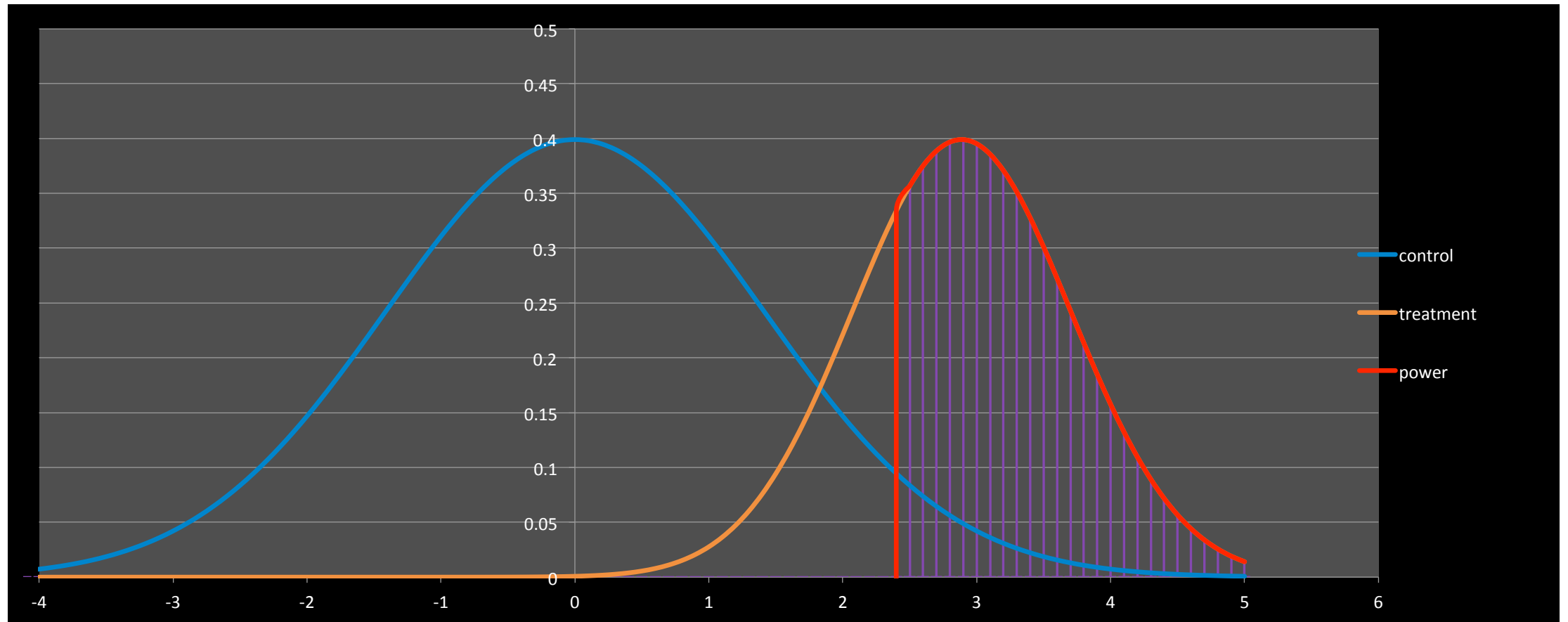
# Sample split: 25% C, 75% T



**Uneven distributions, not efficient, i.e. less power**



# Power: 83%



# Allocation to T v C

---

$$sd(X_1 - X_2) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

$$sd(X_1 - X_2) = \sqrt{\frac{1}{2} + \frac{1}{2}} = \sqrt{\frac{2}{2}} = 1$$

$$sd(X_1 - X_2) = \sqrt{\frac{1}{3} + \frac{1}{1}} = \sqrt{\frac{4}{3}} = 1.15$$

# Power: main ingredients

---

1. Effect Size
2. Sample Size
3. Variance
4. Proportion of sample in T vs. C
- 5. Clustering**

# Clustered design: Definition

---

- In sampling:
  - When clusters of individuals (e.g. schools, communities, etc) are randomly selected from the population, before selecting individuals for observation
- In randomized evaluation:
  - When clusters of individuals are randomly assigned to different treatment groups

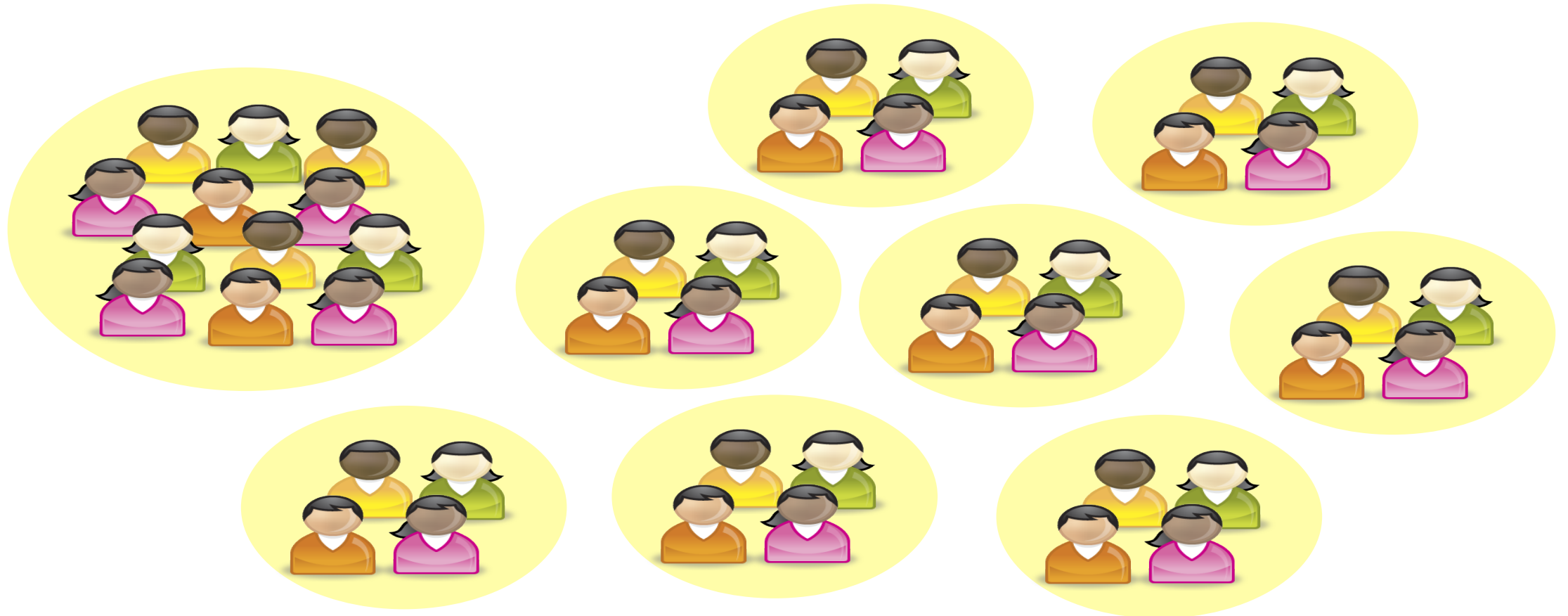
# Clustered design: intuition

---

- You want to know how close the upcoming national elections will be
- Method 1: Randomly select 50 people from entire Indian population
- Method 2: Randomly select 5 families, and ask ten members of each family their opinion

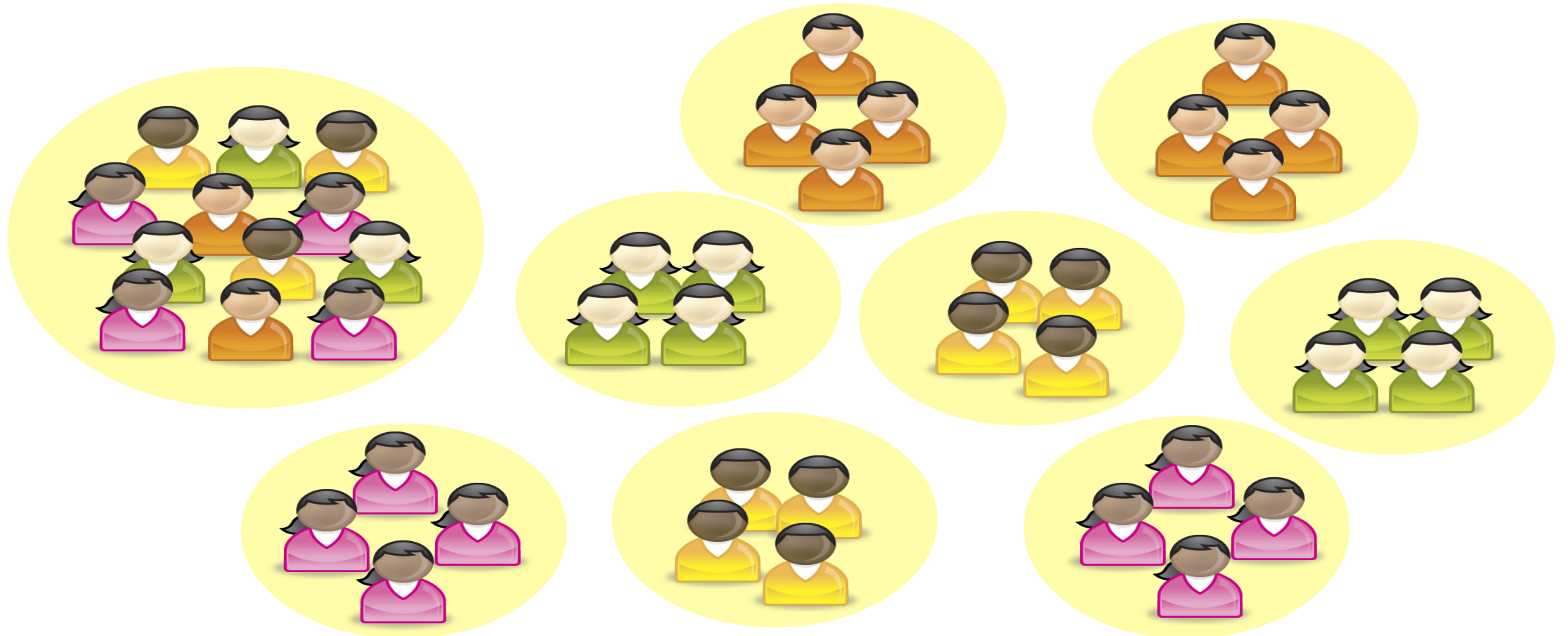
# Low intra-cluster correlation (ICC) aka $\rho$ (rho)

---



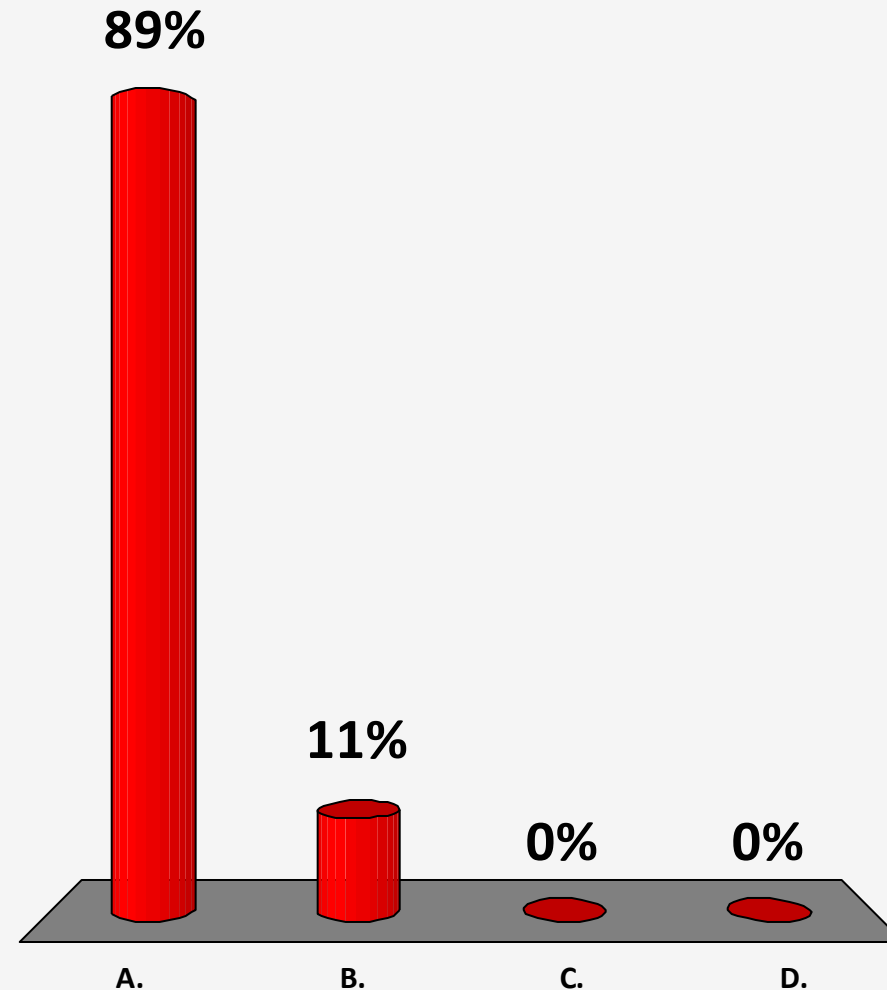
# High intra-cluster correlation (ICC) aka $\rho$ (rho)

---



All uneducated people live in one village. People with only primary education live in another. College grads live in a third, etc. ICC ( $\rho$ ) on education will be..

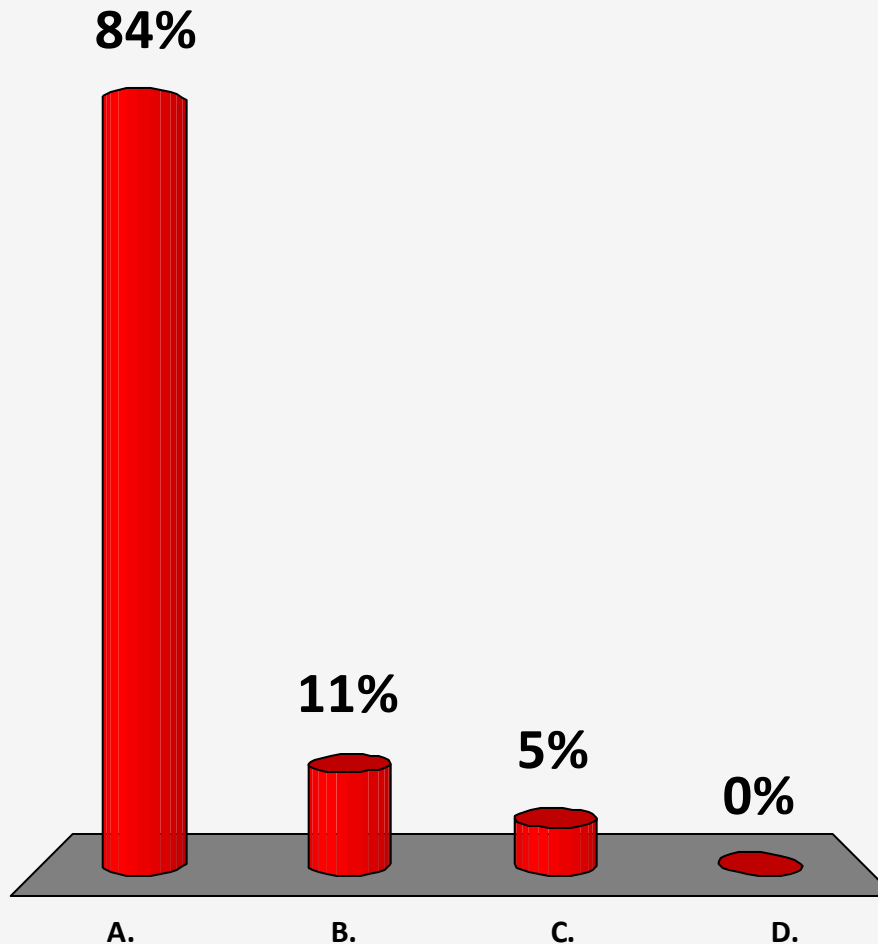
- A. High
- B. Low
- C. No effect on rho
- D. Don't know





If ICC ( $\rho$ ) is high, what is a more efficient way of increasing power?

- A. Include more clusters in the sample
- B. Include more people in clusters
- C. Both
- D. Don't know



# Threats and Analysis

---

Manila, November 2015



ABDUL LATIF JAMEEL  
Poverty Action Lab

J-PAL SOUTHEAST ASIA

# Course Overview

---

1. What is Evaluation?
2. Measuring Impacts
3. Why Randomize?
4. How to Randomize
5. Sampling and Sample Size
6. Threats and Analysis
7. Raskin: RCT Project from Start to Finish
8. Cost Effectiveness Analysis and Scaling Up

# Course Overview

---

1. What is Evaluation?
2. Measuring Impacts
3. Why Randomize?
4. How to Randomize
5. Sampling and Sample Size
- 6. *Threats and Analysis***
7. Raskin: RCT Project from Start to Finish
8. Cost Effectiveness Analysis and Scaling Up

# Lecture Overview

---

- Attrition
- Spillovers
- Partial Compliance and Sample Selection Bias
- Intention to Treat & Treatment on Treated
- Choice of outcomes
- External Validity
- Communication and Implementation
- Conclusion

# Lecture Overview

---

- ***Attrition***
- Spillovers
- Partial Compliance and Sample Selection Bias
- Intention to Treat & Treatment on Treated
- Choice of outcomes
- External Validity
- Communication and Implementation
- Conclusion

# Attrition

---

- Is it a problem if some of the people in the experiment vanish before you collect your data?
  - It is a problem if the type of people who disappear is correlated with the treatment.
- Why is it a problem?
- Why should we expect this to happen?

# Attrition bias: an example

---

- The problem you want to address:
  - Some children don't come to school because they are too weak (undernourished)
- You start a school feeding program and want to do an evaluation
  - You have a treatment and a control group
- Weak, stunted children start going to school more if they live next to a treatment school
- First impact of your program: increased enrollment.
- In addition, you want to measure the impact on child's growth
  - Second outcome of interest: Weight of children
- You go to all the schools (treatment and control) and measure everyone who is in school on a given day
- Will the treatment-control difference in weight be over-stated or understated?



	Before Treatment			After Treatment	
	T	C		T	C
	<b>20</b>	20		<b>22</b>	20
	<b>25</b>	25		<b>27</b>	25
	<b>30</b>	30		<b>32</b>	30
Ave.					
	Difference			Difference	

	Before Treatment			After Treatment	
	T	C		T	C
	<b>20</b>	20		<b>22</b>	20
	<b>25</b>	25		<b>27</b>	25
	<b>30</b>	30		<b>32</b>	30
Ave.	<b>25</b>	<b>25</b>		<b>27</b>	<b>25</b>
	Difference	<b>0</b>		Difference	<b>2</b>

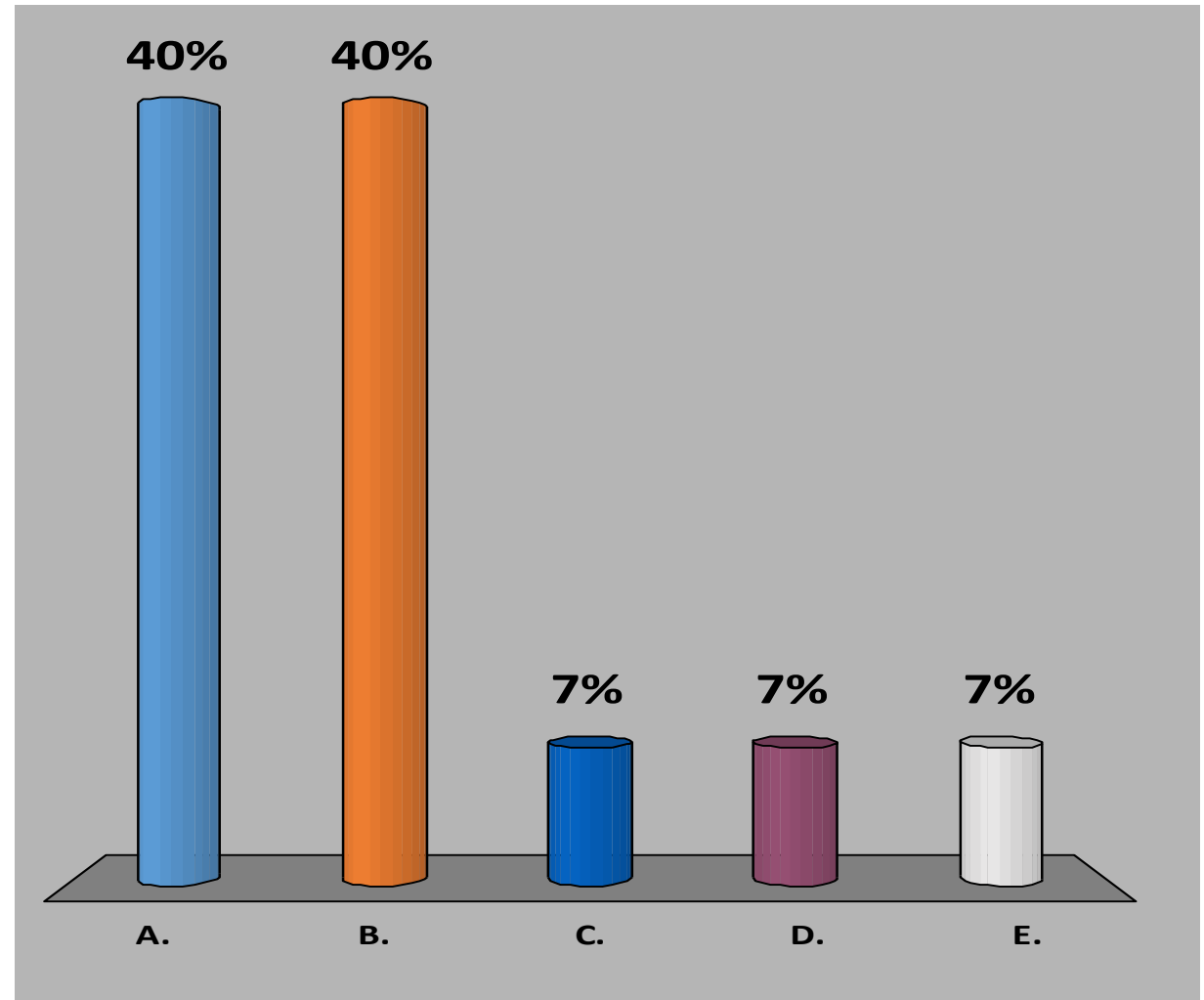
# What if only children $> 21$ Kg come to school?

---

# What if only children > 21 Kg come to school?

Before Treatment		After Treatment	
T	C	T	C
20	20	22	20
25	25	27	25
30	30	32	30

- A. Will you underestimate the impact?
- B. Will you overestimate the impact?
- C. Neither
- D. Ambiguous
- E. Don't know

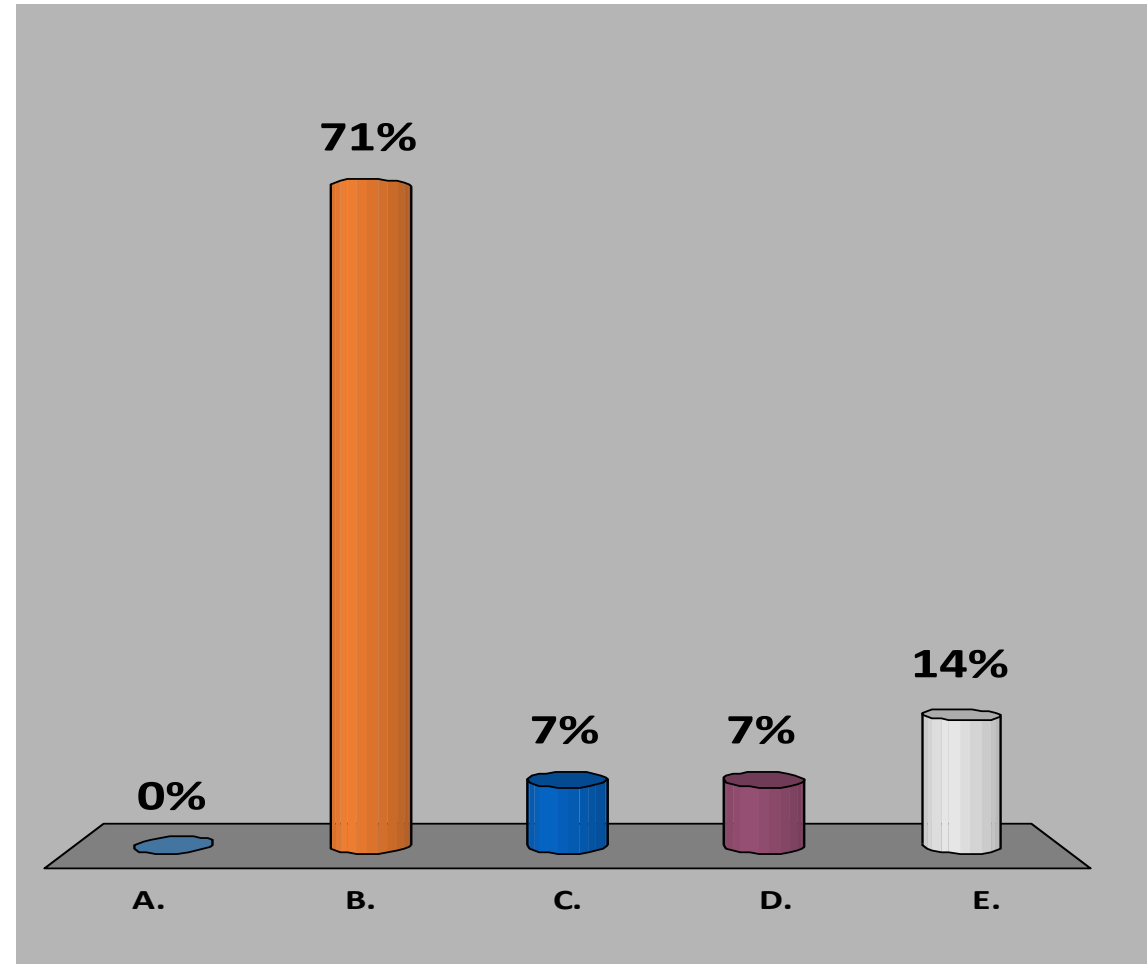


# What if only children > 21 Kg come to school?

	Before Treatment			After Treatment	
	T	C		T	C
	[absent]	[absent]		22	[absent]
	25	25		27	25
	30	30		32	30
Ave.	27.5	27.5		27	27.5
	Difference	0		Difference	-0.5

# When is attrition not a problem?

- A. When it is less than 25% of the original sample
- B. When it happens in the same proportion in both groups
- C. When it is correlated with treatment assignment
- D. All of the above
- E. None of the above



# Attrition Bias

---

- Devote resources to tracking participants after they leave the program
- If there is still attrition, check that it is not different in treatment and control. Is that enough?
- Also check that it is not correlated with observables.
- Try to bind the extent of the bias
  - suppose everyone who dropped out from the treatment got the lowest score that anyone got; suppose everyone who dropped out of control got the highest score that anyone got...
  - Why does this help?

# Lecture Overview

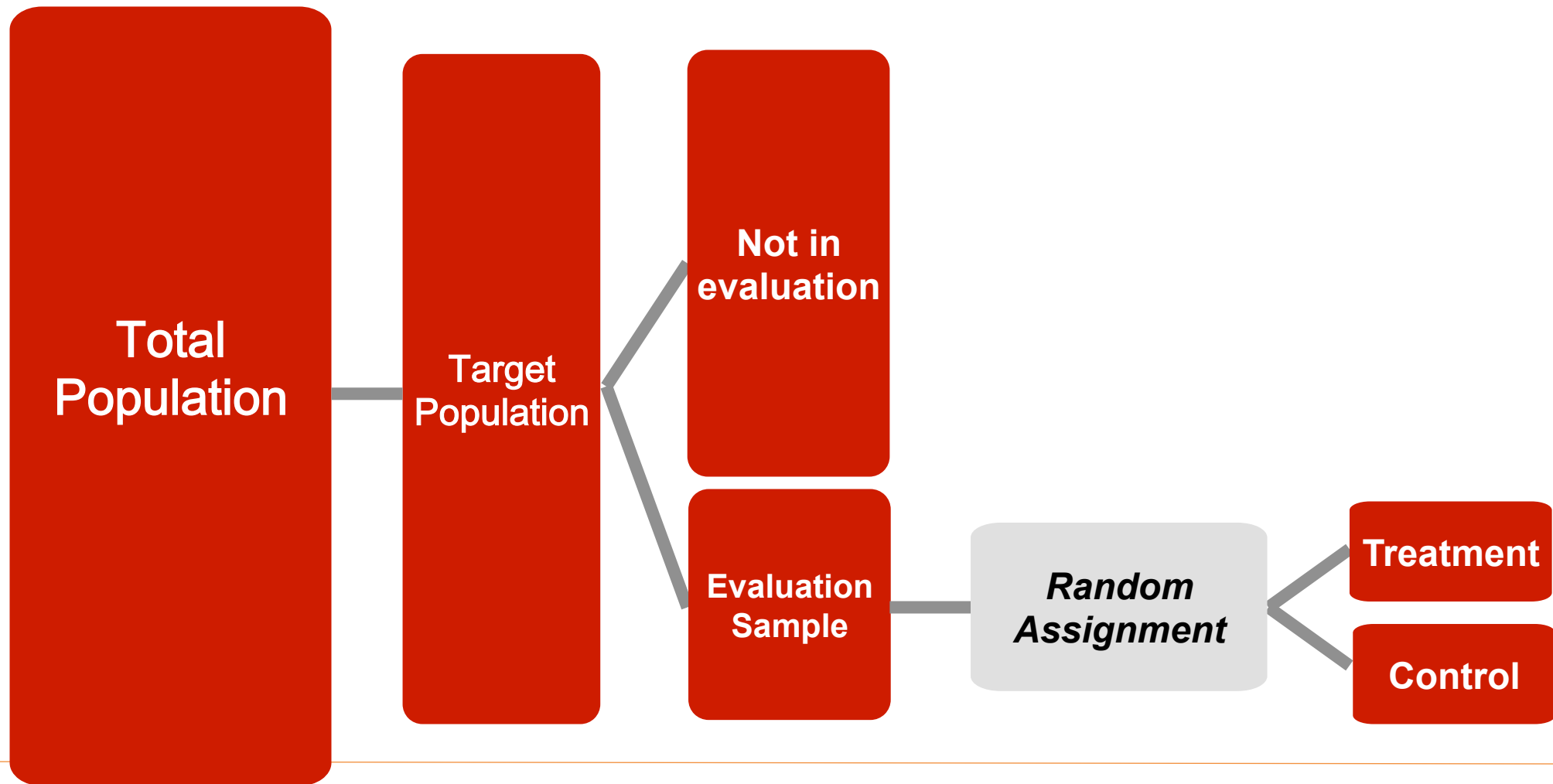
---

- Attrition
- *Spillovers*
- Partial Compliance and Sample Selection Bias
- Intention to Treat & Treatment on Treated
- Choice of outcomes
- External Validity
- Data Quality Assurance
- Communication and Implementation
- Conclusion



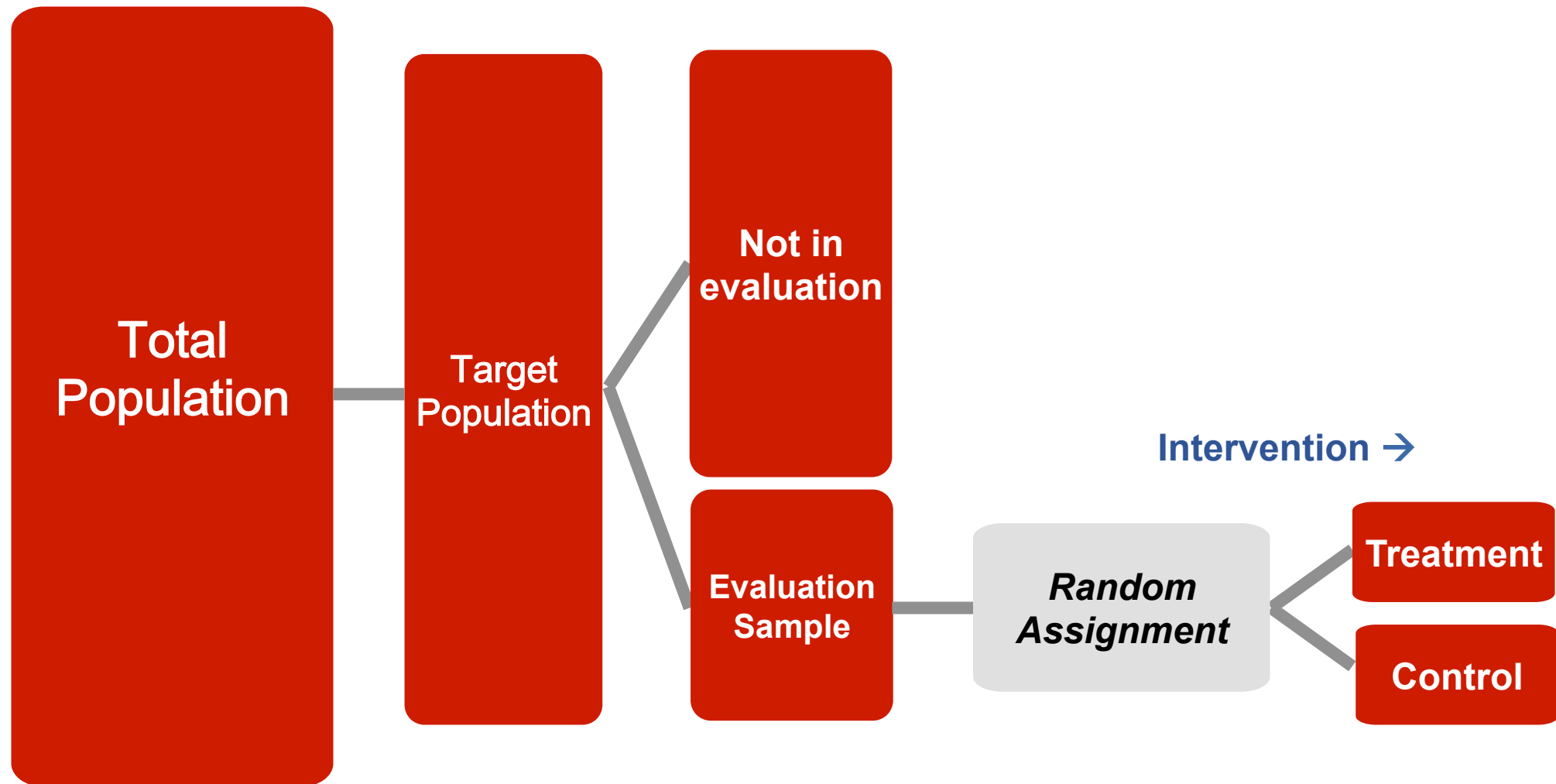
# What else could go wrong?

---



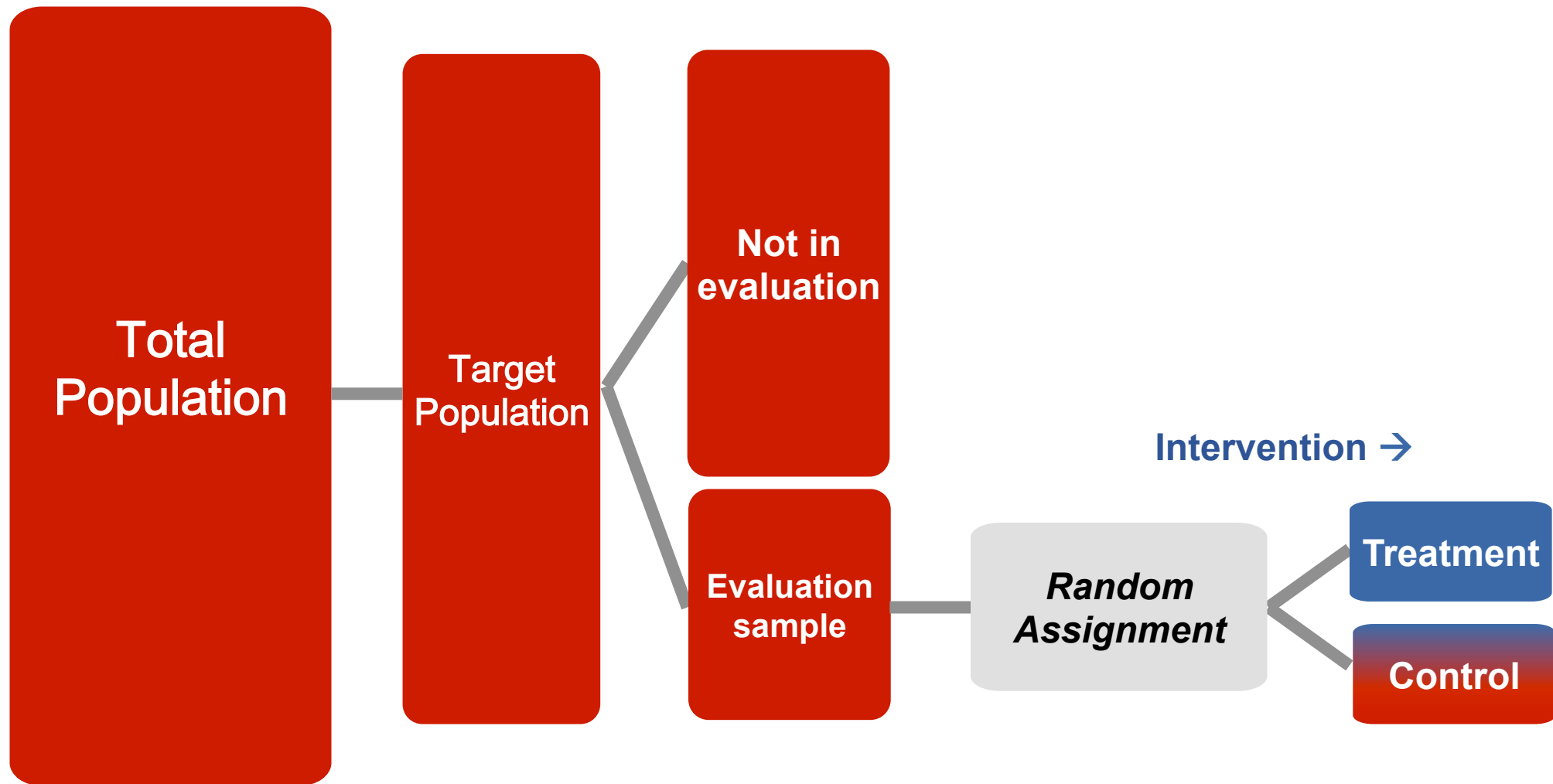
# What else could go wrong?

---



# Spillovers, contamination

---

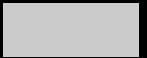

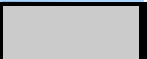





# Example: Vaccination for chicken pox

---

- Suppose you randomize chicken pox vaccinations within schools
  - Suppose that prevents the transmission of disease, what problems does this create for evaluation?
  - Suppose externalities are local? How can we measure total impact?

## Externalities Within School

Without Externalities				
School A	Treated?	Outcome		
<b>Pupil 1</b>	Yes	no chicken pox	Total in Treatment with chicken pox	
Pupil 2	No	chicken pox	Total in Control with chicken pox	
<b>Pupil 3</b>	Yes	no chicken pox	<b>Treatment Effect</b>	
Pupil 4	No	chicken pox		
<b>Pupil 5</b>	Yes	no chicken pox		
Pupil 6	No	chicken pox		

With Externalities				
School A	Treated?	Outcome		
<b>Pupil 1</b>	Yes	no chicken pox	Total in Treatment with chicken pox	
Pupil 2	No	no chicken pox	Total in Control with chicken pox	
<b>Pupil 3</b>	Yes	no chicken pox	<b>Treatment Effect</b>	
Pupil 4	No	chicken pox		
<b>Pupil 5</b>	Yes	no chicken pox		
Pupil 6	No	chicken pox		

## Externalities Within School

Without Externalities				
School A	Treated?	Outcome		
<b>Pupil 1</b>	Yes	no chicken pox	Total in Treatment with chicken pox	0%
Pupil 2	No	chicken pox	Total in Control with chicken pox	100%
<b>Pupil 3</b>	Yes	no chicken pox	<b>Treatment Effect</b>	-100%
Pupil 4	No	chicken pox		
<b>Pupil 5</b>	Yes	no chicken pox		
Pupil 6	No	chicken pox		

With Externalities				
School A	Treated?	Outcome		
Suppose, because prevalence is lower, some children are not re-infected with chicken pox				
<b>Pupil 1</b>	Yes	no chicken pox	Total in Treatment with chicken pox	0%
Pupil 2	No	no chicken pox	Total in Control with chicken pox	67%
<b>Pupil 3</b>	Yes	no chicken pox	<b>Treatment Effect</b>	-67%
Pupil 4	No	chicken pox		
<b>Pupil 5</b>	Yes	no chicken pox		
Pupil 6	No	chicken pox		

# How to measure program impact in the presence of spillovers?

---

- Design the unit of randomization so that it encompasses the spillovers
- If we expect externalities that are all within school:
  - Randomization at the level of the school allows for estimation of the overall effect

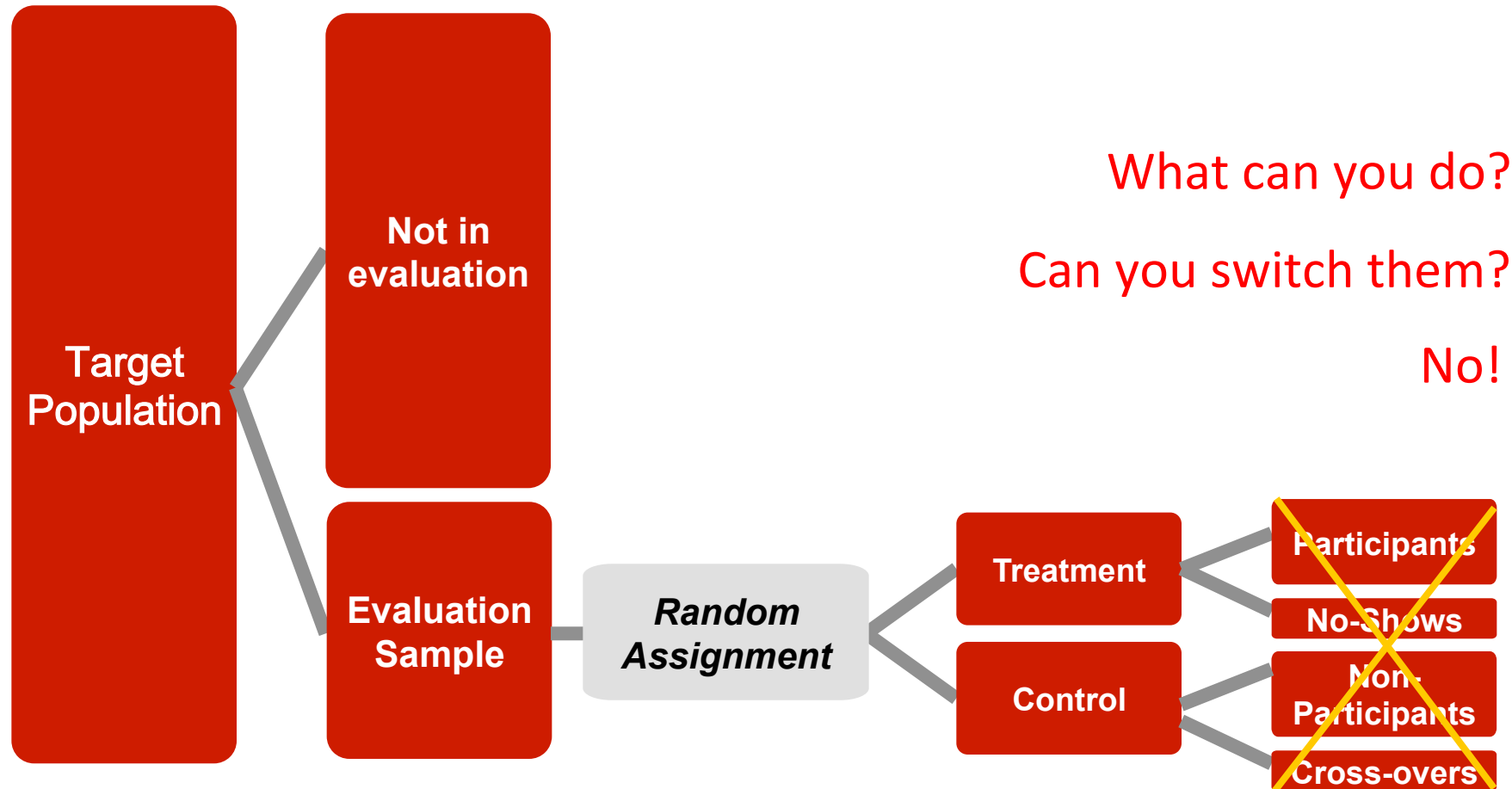
# Lecture Overview

---

- Attrition
- Spillovers
- ***Partial Compliance and Sample Selection Bias***
- Intention to Treat & Treatment on Treated
- Choice of outcomes
- External Validity
- Communication and Implementation
- Conclusion

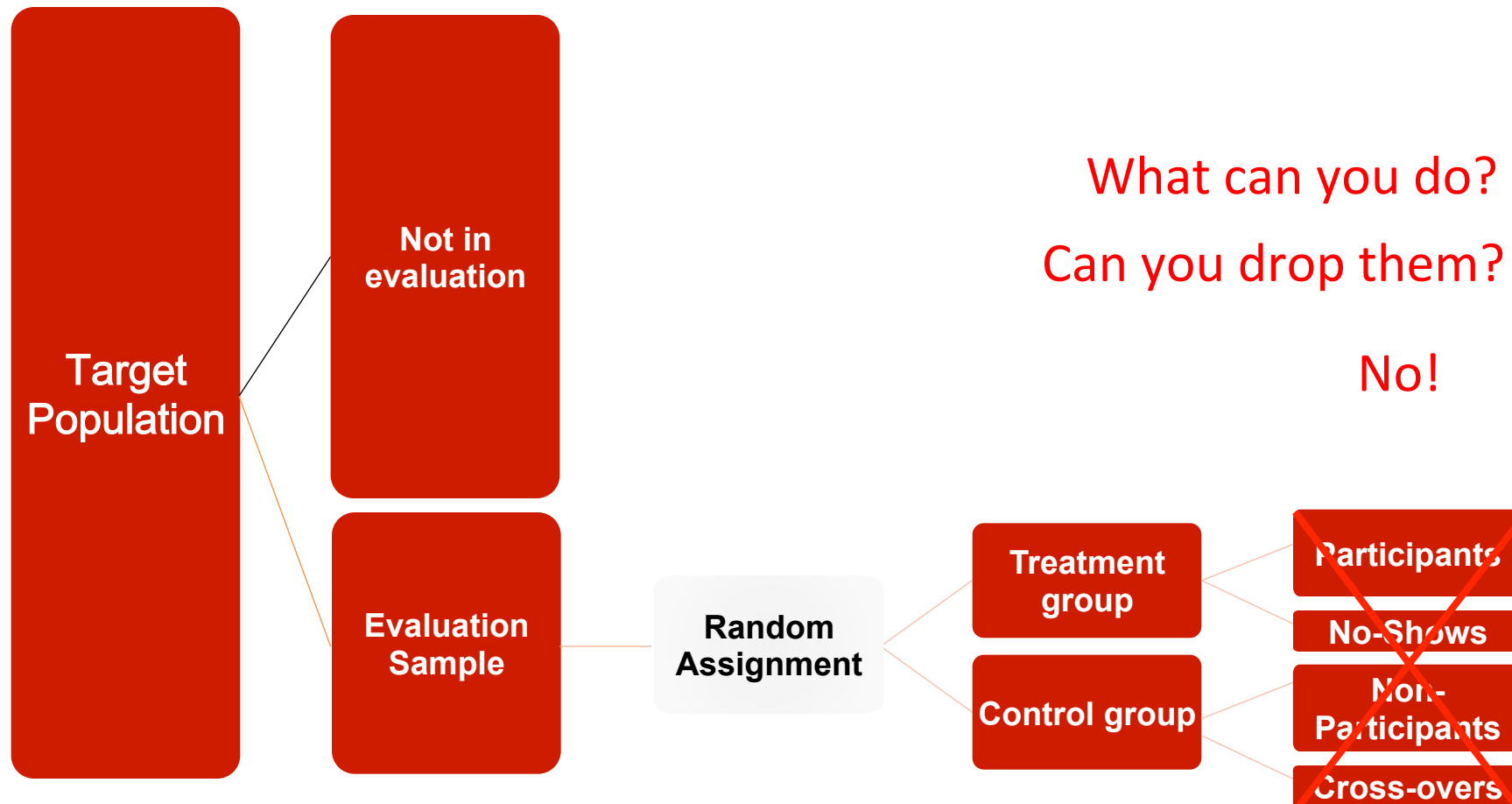


# Non compliers



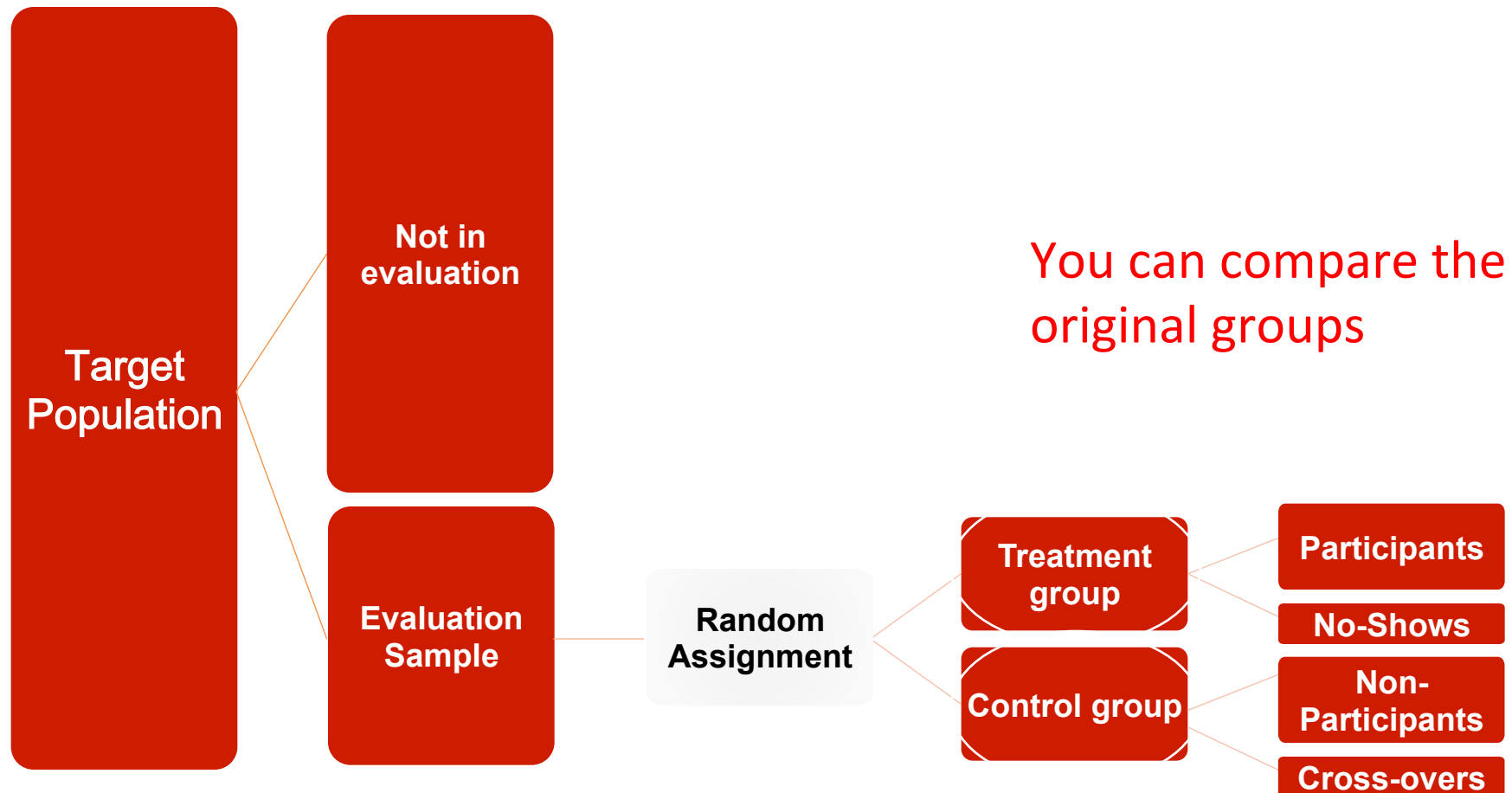
# Non compliers

---



# *Non compliers*

---



# *Sample selection bias*

---

- Sample selection bias could arise if factors other than random assignment influenced program allocation
  - Even if intended allocation of program was random, the actual allocation may not be

# *Sample selection bias*

---

- Individuals assigned to comparison group could attempt to move into treatment group
  - School feeding program: parents could attempt to move their children from comparison school to treatment school
- Alternatively, individuals allocated to treatment group may not receive treatment
  - School feeding program: some students assigned to treatment schools bring and eat their own lunch anyway, or choose not to eat at all.

# What could we do to the *non-compliers*?

---

- Intent to treat (ITT) estimation: compare treatment and control group (as originally randomly assigned)
- Treatment on the Treated (TOT) estimation: calculate the intervention effect for those who participate in the program (compliers)
  - Divide “ITT Estimation” with “proportion of program participant who really complies with the intervention”
- What conclusion can be drawn? What is the caution?

# Lecture Overview

---

- Attrition
- Spillovers
- Partial Compliance and Sample Selection Bias
- ***Intention to Treat & Treatment on Treated***
- Choice of outcomes
- External Validity
- Communication and Implementation
- Conclusion

# ITT and ToT

---

- Vaccination campaign in villages
- Some people in treatment villages not treated
  - 78% of people assigned to receive treatment received some treatment
- What do you do?
  - Compare the beneficiaries and non-beneficiaries?
  - Why not?



# Which groups can be compared ?

---

Treatment group: vaccination	Control group
<b>TREATED</b>	
<b>NON-TREATED</b>	<b>NON-TREATED</b>

# What is the difference between the 2 random groups?

---

Treatment Group	Control Group
1: treated – not infected 2: treated – not infected 3: treated – infected	5: non-treated – infected 6: non-treated – not infected 7: non-treated – infected 8: non-treated – infected
4: non-treated – infected	

# Intention to Treat - ITT

---

Treatment Group: 50% infected

Control Group: 75% infected

- $Y(T)$  = Average Outcome in Treatment Group
- $Y(C)$  = Average Outcome in Control Group

$$\text{ITT} = Y(T) - Y(C)$$

- $\text{ITT} = 50\% - 75\% = -25$  percentage points

# Intention to Treat (ITT)

---

- What does “intention to treat” measure?  
*“What happened to the average child who is in a treated school in this population?”*
- Is this difference the causal effect of the intervention?

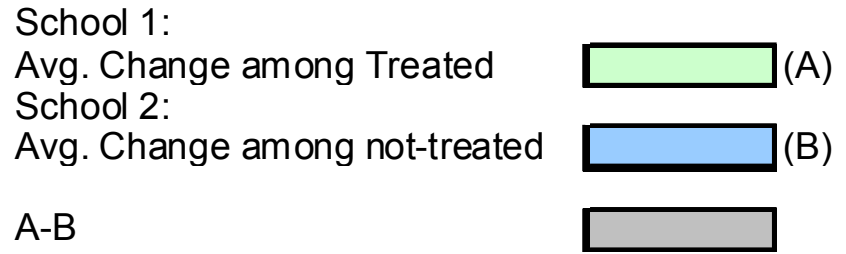
# When is ITT useful?

---

- May relate more to actual programs
- For example, we may not be interested in the medical effect of deworming treatment, but what would happen under an actual deworming program?
- If students often miss school and therefore don't get the deworming medicine, the intention to treat estimate may actually be most relevant.

School 1	Intention to Treat ?	Treated?	Observed Change in weight
Pupil 1	yes	yes	4
Pupil 2	yes	yes	4
Pupil 3	yes	yes	4
Pupil 4	yes	no	0
Pupil 5	yes	yes	4
Pupil 6	yes	no	2
Pupil 7	yes	no	0
Pupil 8	yes	yes	6
Pupil 9	yes	yes	6
Pupil 10	yes	no	0
<b>Avg. Change among Treated A=</b>			<b>4</b>

School 2	Intention to Treat ?	Treated?	Observed Change in weight
Pupil 1	no	no	2
Pupil 2	no	no	1
Pupil 3	no	yes	3
Pupil 4	no	no	0
Pupil 5	no	no	0
Pupil 6	no	yes	3
Pupil 7	no	no	0
Pupil 8	no	no	0
Pupil 9	no	no	0
Pupil 10	no	no	0
<b>Avg. Change among Not-Treated B=</b>			<b>0</b>



School 1	Intention to Treat ?	Treated?	Observed Change in weight
Pupil 1	yes	yes	4
Pupil 2	yes	yes	4
Pupil 3	yes	yes	4
Pupil 4	yes	no	0
Pupil 5	yes	yes	4
Pupil 6	yes	no	2
Pupil 7	yes	no	0
Pupil 8	yes	yes	6
Pupil 9	yes	yes	6
Pupil 10	yes	no	0
<b>Avg. Change among Treated A=</b>			<b>3</b>

School 2	Intention to Treat ?	Treated?	Observed Change in weight
Pupil 1	no	no	2
Pupil 2	no	no	1
Pupil 3	no	yes	3
Pupil 4	no	no	0
Pupil 5	no	no	0
Pupil 6	no	yes	3
Pupil 7	no	no	0
Pupil 8	no	no	0
Pupil 9	no	no	0
Pupil 10	no	no	0
<b>Avg. Change among Not-Treated B=</b>			<b>0.9</b>

School 1:  
 Avg. Change among Treated  (A)  
 School 2:  
 Avg. Change among not-treated  (B)  
 A-B

# From ITT to effect of treatment on the treated (TOT)

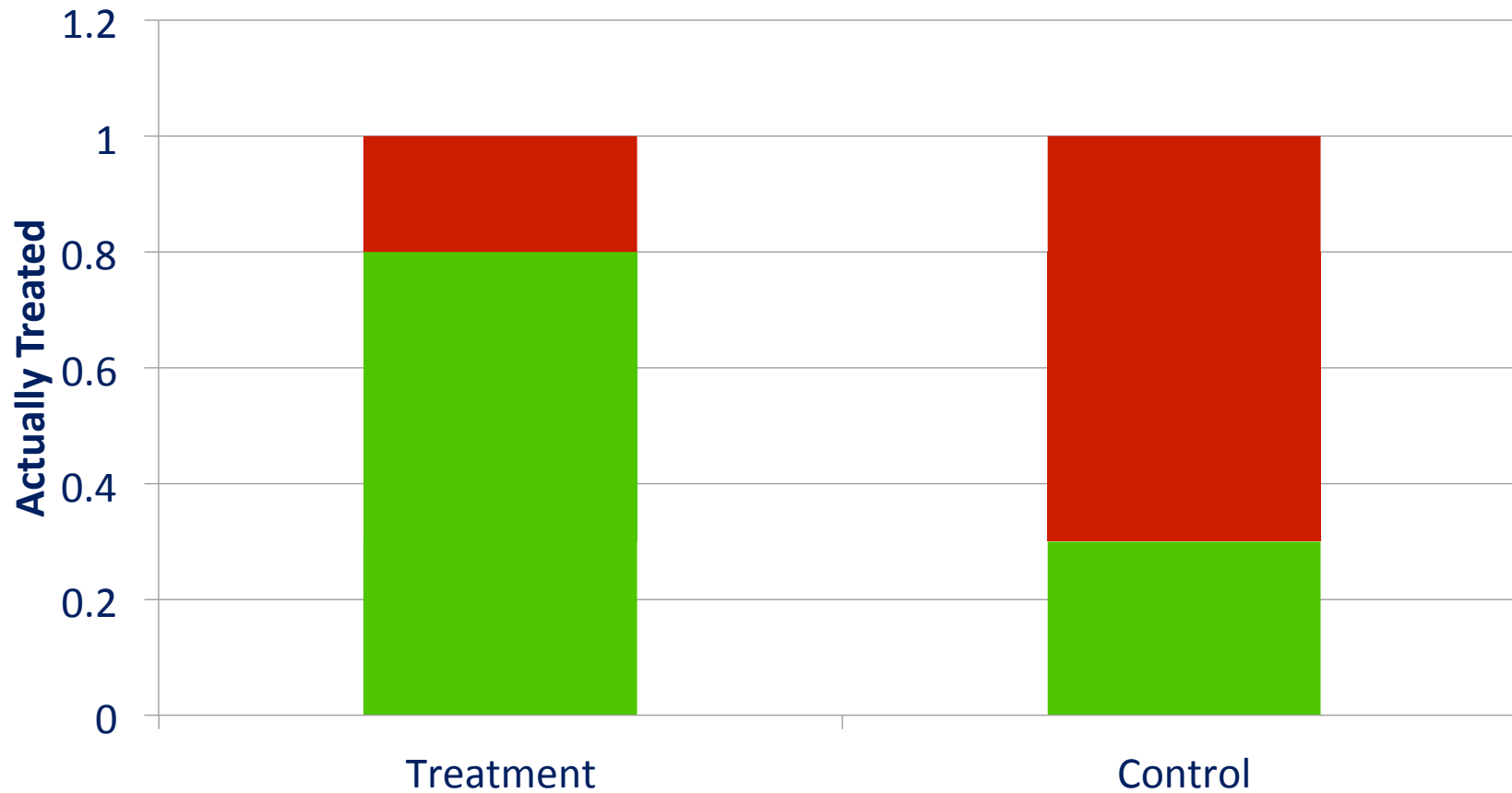
---

- The point is that if there is leakage across the groups, the comparison between those originally assigned to treatment and those originally assigned to control is smaller
- But the difference in the probability of getting treated is also smaller
- Formally this is done by “instrumenting” the probability of treatment by the original assignment



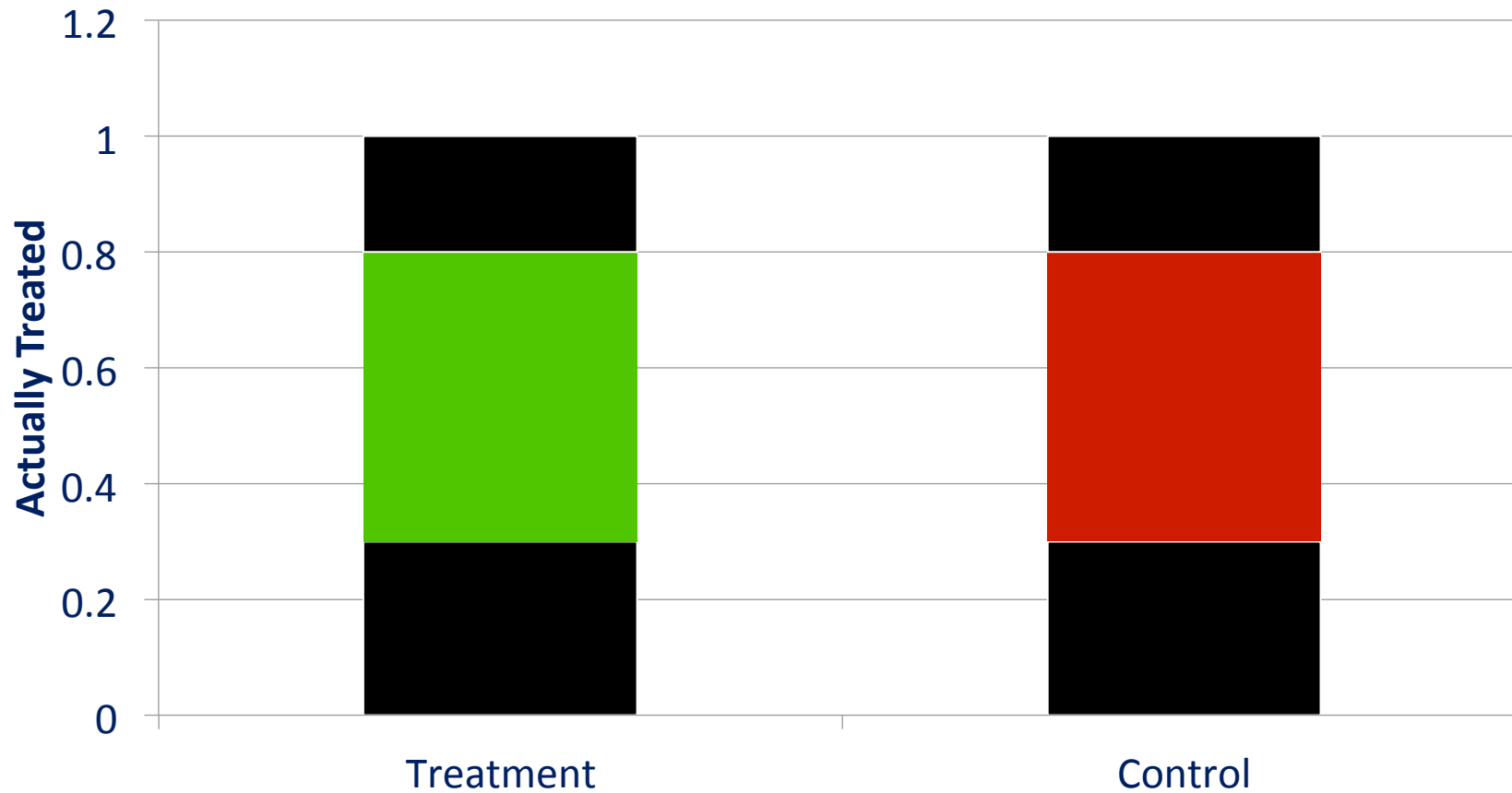
# Estimating ToT from ITT: Wald

---



# Interpreting ToT from ITT: Wald

---



# Estimating TOT

---

- What values do we need?
- $Y(T)$
- $Y(C)$
  
- $\text{Prob}[\text{treated} | T]$
- $\text{Prob}[\text{treated} | C]$

# Treatment on the treated (TOT)

---

- Starting from a simple regression model:
- $Y_{it} = a + B * S_{it} + e_{it}$
- [Angrist and Pischke, p. 67 show]:

$$B = \frac{E(Y_{it} | z_{it} = 1) - E(Y_{it} | z_{it} = 0)}{E(S_{it} | z_{it} = 1) - E(S_{it} | z_{it} = 0)}$$

# Treatment on the treated (TOT)

---

$$B = EY_{i,t} | z_{i,t} = 1 - EY_{i,t} | z_{i,t} = 0 / E s_{i,t} | z_{i,t} = 1 - E s_{i,t} | z_{i,t} = 0$$

$$Y(T) - Y(C) / \text{Prob}[\text{treated} | T] - \text{Prob}[\text{treated} | C]$$

# TOT estimator

School 1	Intention to Treat ?	Treated?	Observed Change in weight
Pupil 1	yes	yes	4
Pupil 2	yes	yes	4
Pupil 3	yes	yes	4
Pupil 4	yes	no	0
Pupil 5	yes	yes	4
Pupil 6	yes	no	2
Pupil 7	yes	no	0
Pupil 8	yes	yes	6
Pupil 9	yes	yes	6
Pupil 10	yes	no	0
<b>Avg. Change Y(T)=</b>			<input type="text"/>

A = Gain if Treated  
B = Gain if not Treated

ToT Estimator: A-B

$$A-B = \frac{Y(T)-Y(C)}{\text{Prob(Treated|T)}-\text{Prob(Treated|C)}}$$

School 2	Intention to Treat ?	Treated?	Observed Change in weight
Pupil 1	no	no	2
Pupil 2	no	no	1
Pupil 3	no	yes	3
Pupil 4	no	no	0
Pupil 5	no	no	0
Pupil 6	no	yes	3
Pupil 7	no	no	0
Pupil 8	no	no	0
Pupil 9	no	no	0
Pupil 10	no	no	0
<b>Avg. Change Y(C) =</b>			<input type="text"/>

Y(T)

Y(C)

Prob(Treated|T)

Prob(Treated|C)

Y(T)-Y(C)

Prob(Treated|T)-Prob(Treated|C)

**A-B**

# TOT estimator

School 1	Intention to Treat ?	Treated?	Observed Change in weight
Pupil 1	yes	yes	4
Pupil 2	yes	yes	4
Pupil 3	yes	yes	4
Pupil 4	yes	no	0
Pupil 5	yes	yes	4
Pupil 6	yes	no	2
Pupil 7	yes	no	0
Pupil 8	yes	yes	6
Pupil 9	yes	yes	6
Pupil 10	yes	no	0
<b>Avg. Change Y(T)=</b>			<b>3</b>

A = Gain if Treated  
B = Gain if not Treated

ToT Estimator: A-B

$$A-B = \frac{Y(T)-Y(C)}{\text{Prob(Treated|T)}-\text{Prob(Treated|C)}}$$

School 2	Intention to Treat ?	Treated?	Observed Change in weight
Pupil 1	no	no	2
Pupil 2	no	no	1
Pupil 3	no	yes	3
Pupil 4	no	no	0
Pupil 5	no	no	0
Pupil 6	no	yes	3
Pupil 7	no	no	0
Pupil 8	no	no	0
Pupil 9	no	no	0
Pupil 10	no	no	0
<b>Avg. Change Y(C) =</b>			<b>0.9</b>

Y(T)	3
Y(C)	0.9
Prob(Treated T)	60%
Prob(Treated C)	20%

Y(T)-Y(C)	2.1
Prob(Treated T)-Prob(Treated C)	40%

<b>A-B</b>	<b>5.25</b>
------------	-------------

# Generalizing the ToT Approach: Instrumental Variables

---

1. First stage regression:

$$T \downarrow \text{Actual} \downarrow i = \alpha \downarrow 0 \downarrow i + \alpha \downarrow 1 \downarrow i T \downarrow 1 \downarrow i + \alpha \downarrow i \downarrow X \downarrow i + e$$

2. Predict treatment status using estimated coefficients

$$T \downarrow \text{predicted} \downarrow i = a \downarrow 0 \downarrow i + a \downarrow 1 \downarrow i T \downarrow 1 \downarrow i + a \downarrow i \downarrow X \downarrow i$$

3. Regress outcome variable on predicted treatment status

$$Y \downarrow i = \beta \downarrow 0 \downarrow i + \beta \downarrow 1 \downarrow i T \downarrow \text{predicted} \downarrow i + \beta \downarrow X \downarrow i + \varepsilon$$

4.  $\beta \downarrow 1 \downarrow i$  gives treatment effect



# Requirements for Instrumental Variables

---

- First stage
  - Your experiment (or instrument) meaningfully affects probability of treatment
- Exclusion restriction
  - Your experiment (or instrument) does not affect outcomes through another channel

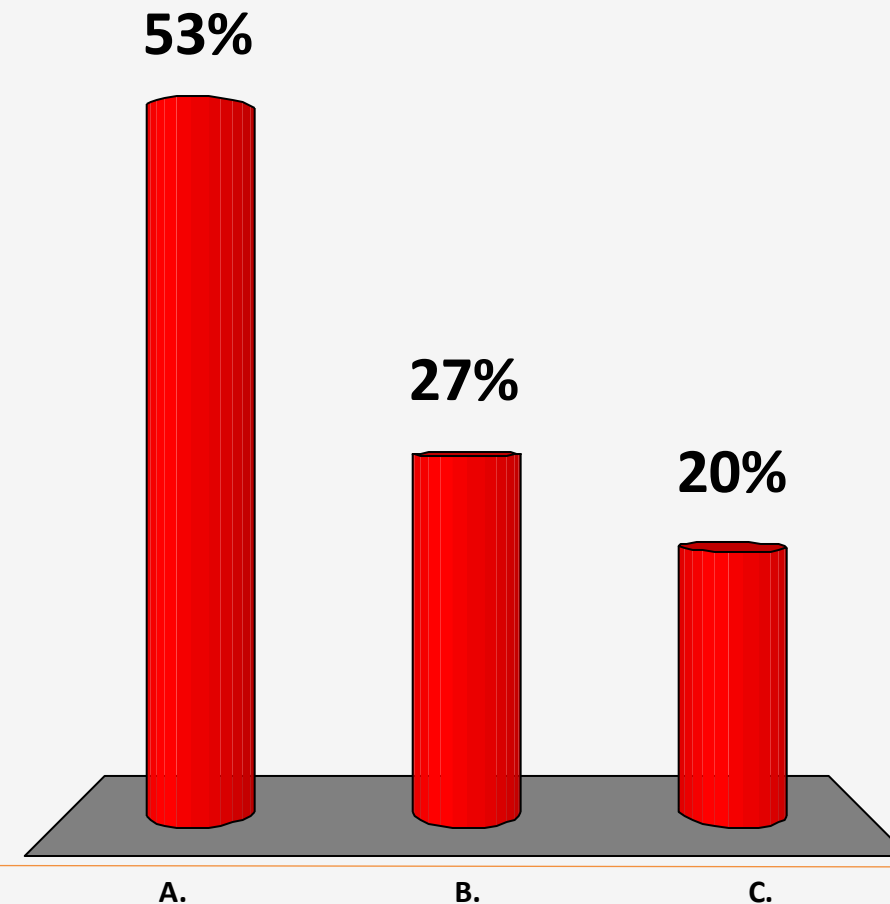
The ITT estimate will always be smaller (e.g., closer to zero) than the ToT estimate

---

A.True

B.False

C.Don't Know



# TOT is not always appropriate...

---

- Example: send 50% of MIT staff a letter warning of flu season, encourage them to get vaccines
- Suppose 50% in treatment, 0% in control get vaccines
- Suppose incidence of flu in treated group drops 35% relative to control group
- Is  $(.35) / (.5 - 0) = 70\%$  the correct estimate?
- What effect might letter alone have?

# Lecture Overview

---

- Attrition
- Spillovers
- Partial Compliance and Sample Selection Bias
- Intention to Treat & Treatment on Treated
- ***Choice of outcomes***
- External Validity
- Communication and Implementation
- Conclusion

# Multiple outcomes

---

- Can we look at various outcomes?
- The more outcomes you look at, the higher the chance you find at least one significantly affected by the program
  - Pre-specify outcomes of interest
  - Report results on all measured outcomes, even null results
  - Correct statistical tests (Bonferroni)

# *Covariates*

---

- Why include covariates?
  - May explain variation, improve statistical power
- Why not include covariates?
  - Appearances of “specification searching”
- What to control for?
  - If stratified randomization: add strata fixed effects
  - Other covariates

Rule: Report both “raw” differences and regression-adjusted results

# Lecture Overview

---

- Attrition
- Spillovers
- Partial Compliance and Sample Selection Bias
- Intention to Treat & Treatment on Treated
- Choice of outcomes
- ***External Validity***
- Communication and Implementation
- Conclusion

# Threats to external validity

---

- Behavioral responses to evaluations
- Generalizability of results



# Threat to external validity: Behavioral responses to evaluations

---

- One limitation of evaluations is that the evaluation itself may cause the treatment or comparison group to change its behavior
  - Treatment group behavior changes: Hawthorne effect
  - Comparison group behavior changes: John Henry effect
- Minimize salience of evaluation as much as possible
- Consider including controls who are measured at end-line only

# Generalizability of results

---

- Depend on three factors:
  - Program Implementation: can it be replicated at a large (national) scale?
  - Study Sample: is it representative?
  - Sensitivity of results: would a similar, but slightly different program, have same impact?

# Presentation outline

---

- Attrition
- Spillovers
- Partial Compliance and Sample Selection Bias
- Intention to Treat & Treatment on Treated
- Choice of outcomes
- External Validity
- ***Communication and Implementation***
- Conclusion

# Communication and Implementation

---

- Unlike other evaluation techniques, randomization will affect how certain program is designed and implemented – not only how the program is evaluated
- This may affect non-traditional evaluations (other than baseline / endline)
  - Program design - who, what, where
  - Program implementation
  - Evaluation strategy

# Communication and Implementation

---

- This means that further communication between program and evaluation is necessary
  - Can the randomization be done? If yes, how? Will it change the program implementation?
  - Is the randomization successfully completed?
  - Is there any change in implementation after program takes place?
  - Is there any drop-out? Imperfect compliance?
- Evaluator should provide feedback and data timely

# Lecture Overview

---

- Attrition
- Spillovers
- Partial Compliance and Sample Selection Bias
- Intention to Treat & Treatment on Treated
- Choice of outcomes
- External Validity
- Data Quality Assurance
- Communication and Implementation
- ***Conclusion***

# Conclusion

---

- There are many threats to the internal and external validity of randomized evaluations...
- ...as are there for every other type of study
- Randomized trials:
  - Facilitate simple and transparent analysis
    - Provide few “degrees of freedom” in data analysis (this is a good thing)
  - Allow clear tests of validity of experiment

# Further resources

---

- Using Randomization in Development Economics Research: A Toolkit (Duflo, Glennerster, Kremer)
- Mostly Harmless Econometrics (Angrist and Pischke)
- Identification and Estimation of Local Average Treatment Effects (Imbens and Angrist, Econometrica, 1994).



# RCT Evaluation from Start to Finish: Raskin project

---

Manila, November 2015



ABDUL LATIF JAMEEL  
Poverty Action Lab

J-PAL SOUTHEAST ASIA

# Course Overview

---

1. What is Evaluation?
2. Measuring Impacts
3. Why Randomize?
4. How to Randomize
5. Sampling and Sample Size
6. Threats and Analysis
7. Raskin: RCT Project from Start to Finish
8. Cost Effectiveness Analysis and Scaling Up

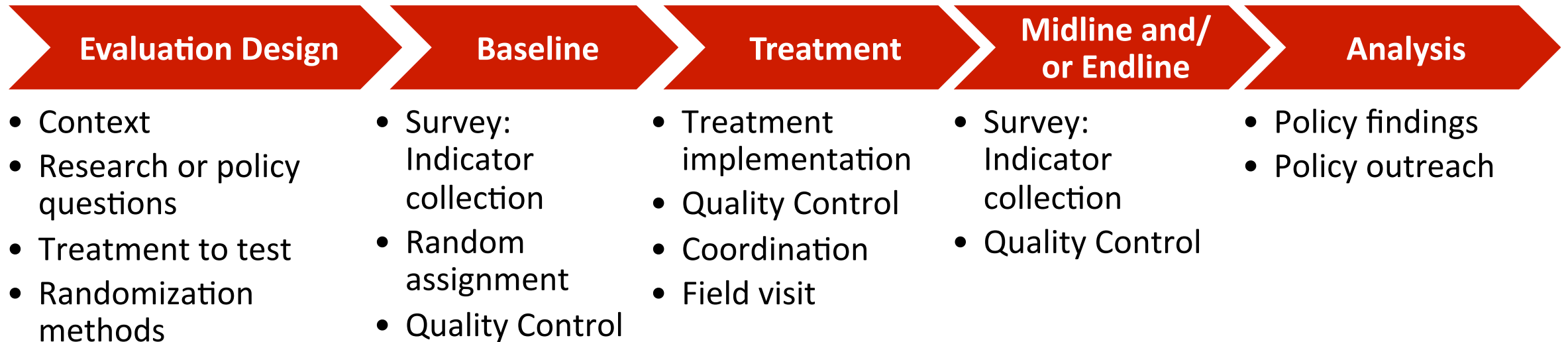
# Course Overview

---

1. What is Evaluation?
2. Measuring Impacts
3. Why Randomize?
4. How to Randomize
5. Sampling and Sample Size
6. Threats and Analysis
- 7. *Raskin: RCT Project from Start to Finish***
8. Cost Effectiveness Analysis and Scaling Up

# Evaluation flow in general

---



# Evaluation from Start to Finish

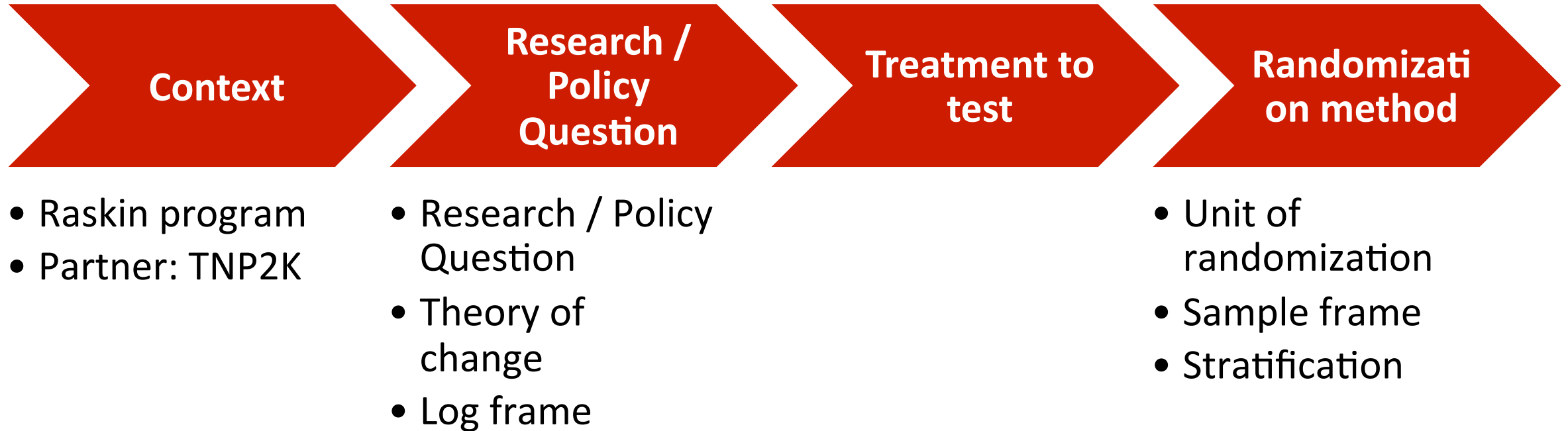
---

1. Background and Evaluation Design
2. Implementation: Treatment
3. Implementation: Data Collection
4. Analysis and Scaling up

# Evaluation Background

# Design and Background

---



# Design and Background

---





# Raskin program

---

- One of the largest social protection program in Indonesia
  - US \$1.5 billion annually
  - 53% of government spending for social aid (World Bank 2012)
- Providing subsidized rice to poor households
  - 15 kg of rice at Rp. 1,600 per kg
  - Target: 30% poorest RT (PPLS'11)
- Realization: Targeted beneficiary (poor households) only received 32% of total subsidy provided (Raskin HH Survey)

# Partner: TNP2K

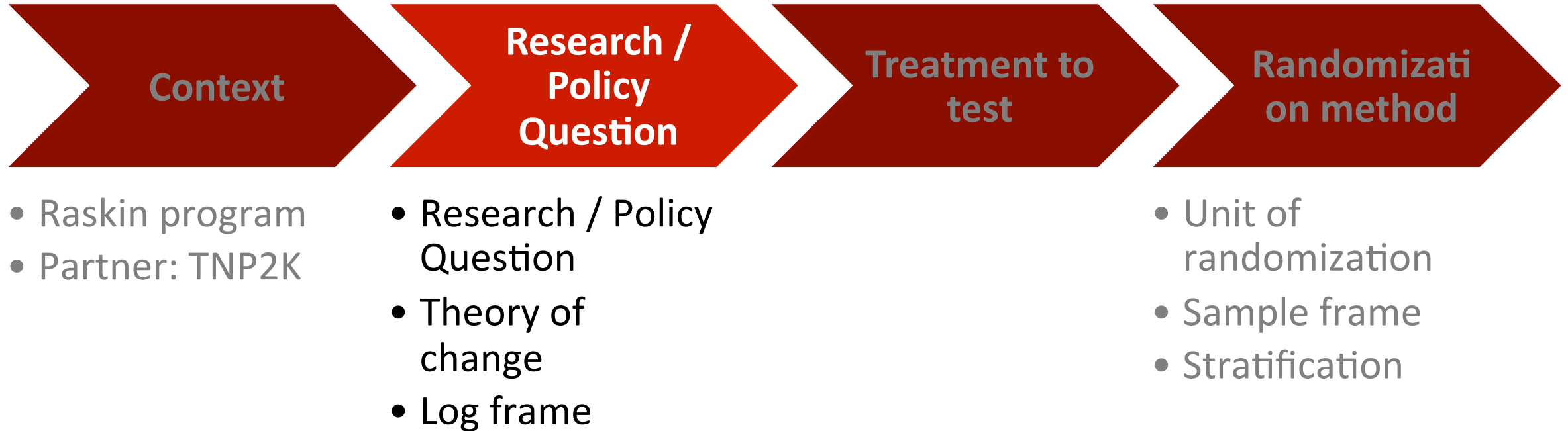
---



- Tim Nasional Percepatan Penanggulangan Kemiskinan (TNP2K; National team for the Acceleration of Poverty Reduction)
- Established by and reporting to then-Vice President Boediono
- TNP2K responsibilities:
  - To formulate evidence-based policy for increasing the effectiveness of social aid program
  - Mengordinasikan aktor–aktor pemerintahan untuk tujuan ini
- One of TNP2K priorities is to increase the quality of targeting and service delivery. Raskin belongs to Cluster 1 TNP2K.

# Design and Background

---



# Evaluation Objective

---

- To collect evidence in order to learn whether the distribution of Raskin card as proposed by TNP2K will improve the effectiveness of Raskin program

## **The pilot aims to answer three key questions:**

1. Will the Raskin card increase the quality of targeting and take-up rate of Raskin program as well as decrease the Raskin price for eligible household?
2. Will the Raskin card be socially acceptable?
3. What is the most effective way in implementing the distribution of Raskin card?

# Theory of Change

**DISTRIBUTING RASKIN CARD TO ELIGIBLE HOUSEHOLDS (HH)**

**Assumption:** The distribution of Raskin card to household is successful, there's no distribution challenge/constraint

**HH RECEIVES RASKIN CARD**

**Assumption:** HH understand the information provided in the card, use the card, and it is less likely they use it interchangeably with previous version of Raskin card/version

**HH RECEIVES MORE SUBSIDIZED RICE**

**Assumption:** Eligible households ask for lower price of Raskin, village officer is following up and able to create any difference

**TRANSPARENCY AND EFFECTIVENESS OF SOCIAL PROTECTION PROGRAM INCREASES**

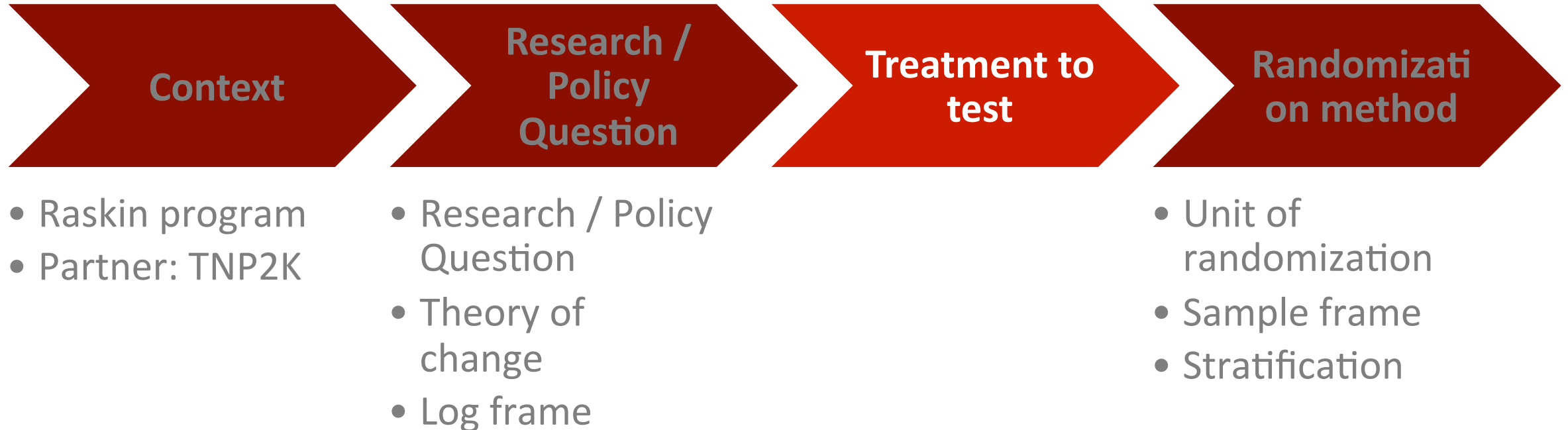
**Assumption:** The effectiveness of social aid program is due to lack of transparency and the change is sustainable moving forward

# Log Frame

	Objectives Hierarchy	Indicators	Source of Verification	Assumptions/Threats
<b>Impact (Goal/ Overall Objective)</b>	Increased transparency and effectiveness of social distribution program	Quantity and price of the Raskin	Household survey	The ineffectiveness is due to lack of transparency, better implementation continues.
<b>Outcome (Project Objective)</b>	Beneficiaries may receive more subsidized Raskin	Quantity and price of the Raskin	Household survey	Eligible households ask for lower price of Raskin, village officer is following up and able to create any difference
<b>Outputs</b>	Eligible households receive Raskin card	Do eligible households receive the card?	Household survey	HH understand the information provided in the card, use the card, and it is less likely they use it interchangeably with previous version of Raskin card/version
<b>Inputs (Activities)</b>	Distributing card to eligible households	Is the card successfully delivered to eligible HH?	Household survey, administrative data from PT Pos	The distribution of Raskin card to household is successful, there's no distribution challenge/constraint

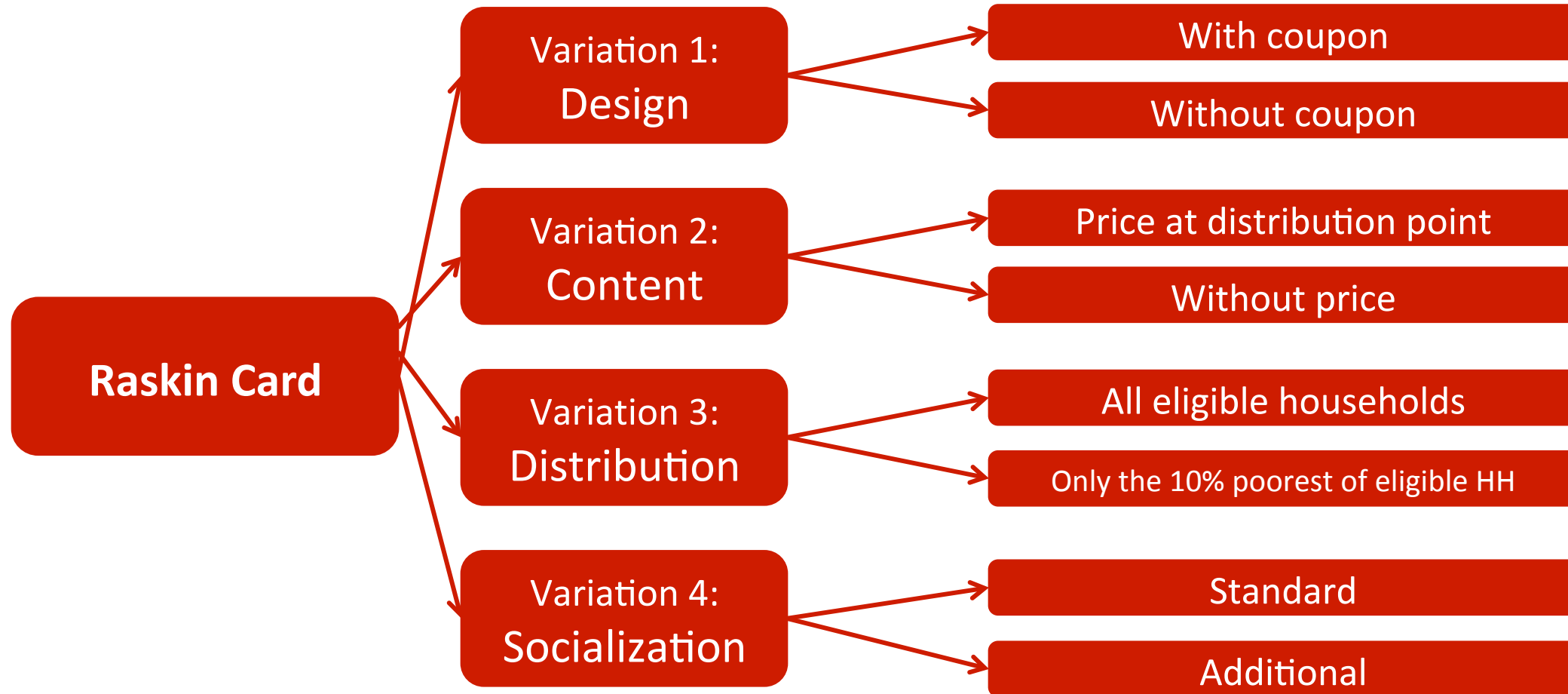
# Design and Background

---



# Intervention Summary

---





# Raskin Card



Raskin card without coupon, with price



Raskin card with both coupon and price

# Additional Socialization Poster

**1**

**MAU BELI RASKIN?**  
**GUNAKAN KARTU RASKIN ANDA!**




**PENGUMUMAN:**

1. Rumah Tangga yang berhak membeli Raskin tercatat di Daftar Penerima Manfaat (DPM)
2. Rumah tangga tersebut akan memperoleh Kartu Raskin
3. Kartu Raskin harus dibawa saat membeli Raskin




**4**

**MAU BELI RASKIN?**  
**GUNAKAN KARTU RASKIN ANDA!**



**PENGUMUMAN:**

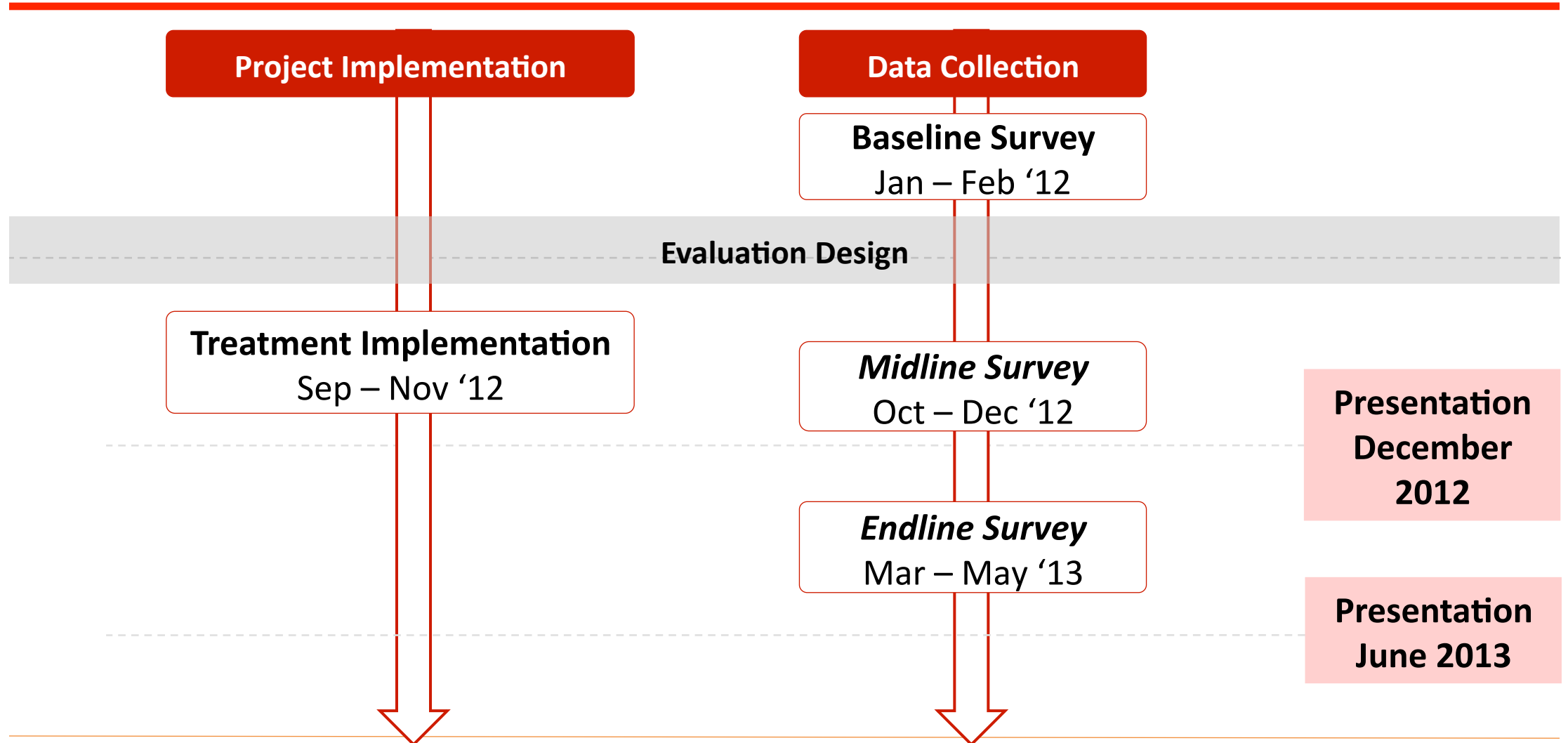
1. Rumah Tangga yang berhak membeli Raskin tercatat di Daftar Penerima Manfaat (DPM)
2. Kelompok rumah tangga paling miskin memperoleh Kartu Raskin
3. Penerima kartu harus membawa Kartu Raskin saat membeli Raskin



# Treatment Variation

Card variation			Standard socialization	Additional Socialization
All eligible households	With price	With coupon	Group 1	Group 2
		Without coupon	Group 3	Group 4
	Without price	Coupon	Group 5	Group 6
		Without coupon	Group 7	Group 8
Only the 10% poorest of eligible households	With price	Coupon	Group 9	Group 10
		Without coupon	Group 11	Group 12
	Without price	Coupon	Group 13	Group 14
		Without coupon	Group 15	Group 16
			Control (Without card and socialization at all)	

# Evaluation and Its Impact



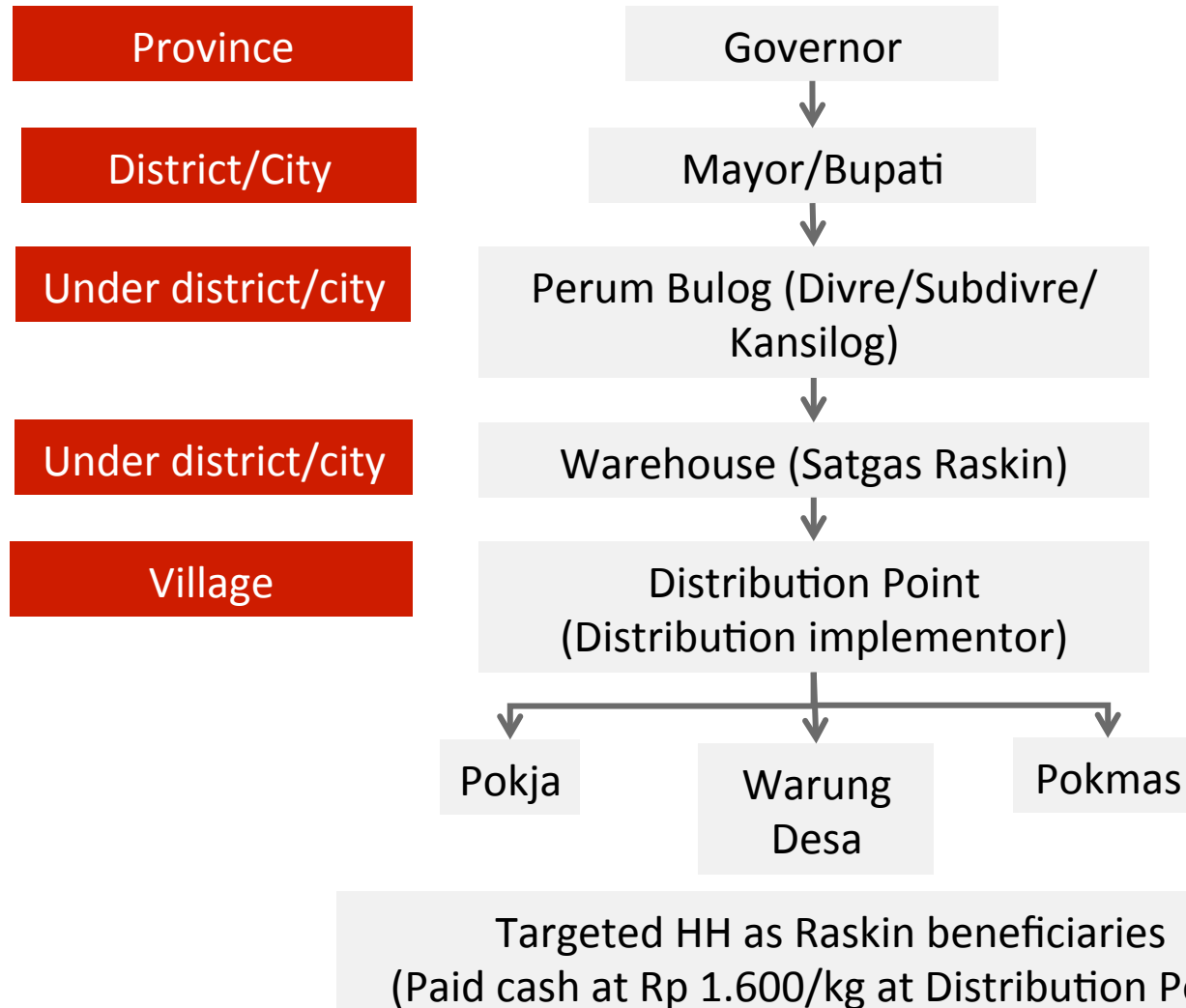
# Design and Background

---



# Identifying Unit of Randomization

Head of National Raskin Team (Ministry of Social Affairs)



- What is the smallest administrative unit where Raskin is distributed?
- Sub-district? Bulog warehouse? Village? Sub-village?

# Sample Frame

---

- Raskin sample is identical to previous project (Targeting II)
- 600 villages (including *control village*)
  - 28 is excluded from sample due to its high risk and remoteness
- 572 villages di 6 districts/cities
  - Pemalang and Wonogiri (Central Java),
  - Palembang and Ogan Komering Ilir (South Sumatera),
  - Bandar Lampung and Central Lampung (Lampung)



# Stratification

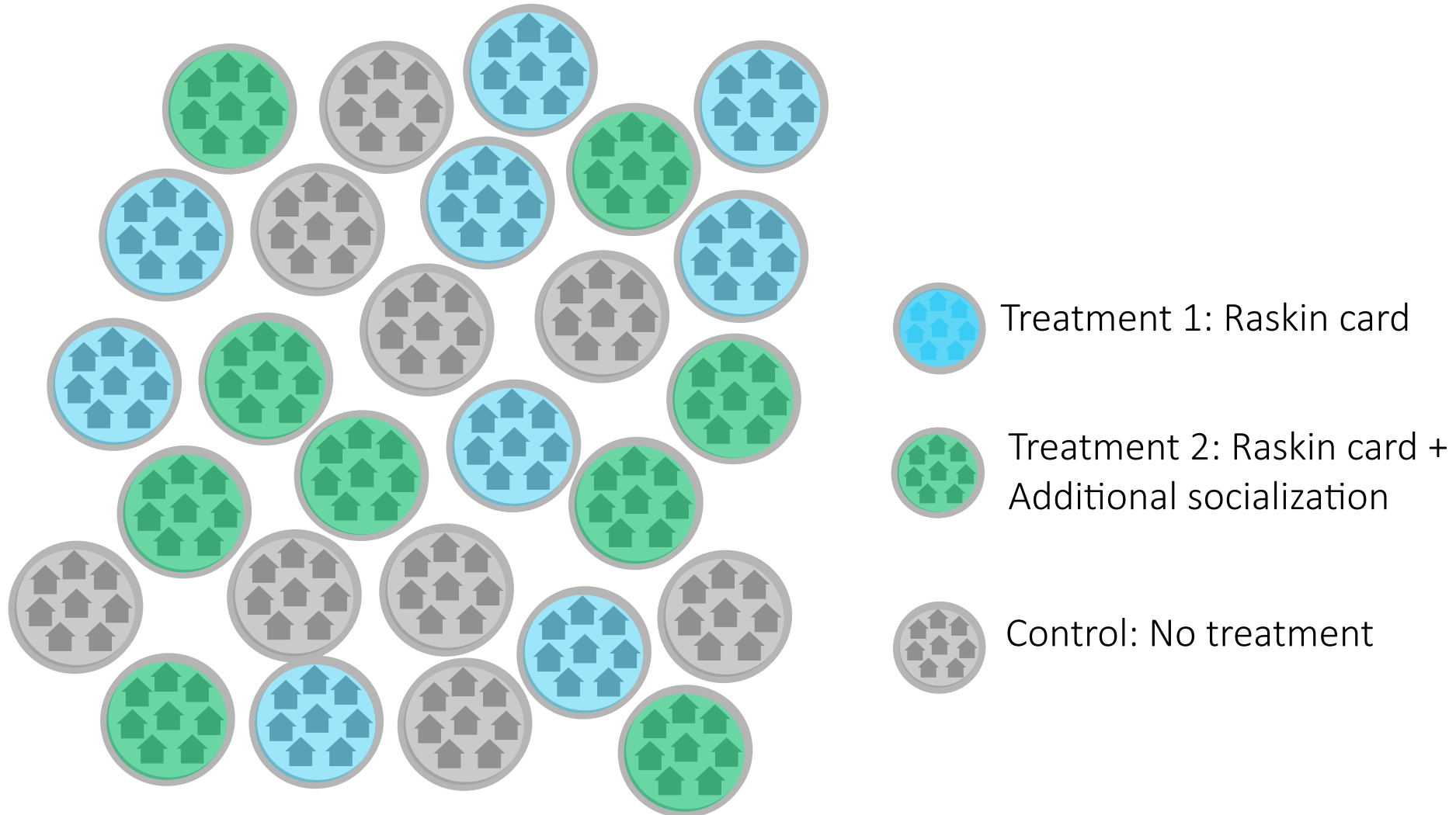
---

- The treatment is stratified based on:
  - District
  - Treatment group in *Targeting II* project
  - Sub-district
  - Urban to rural ratio must be at 2:3



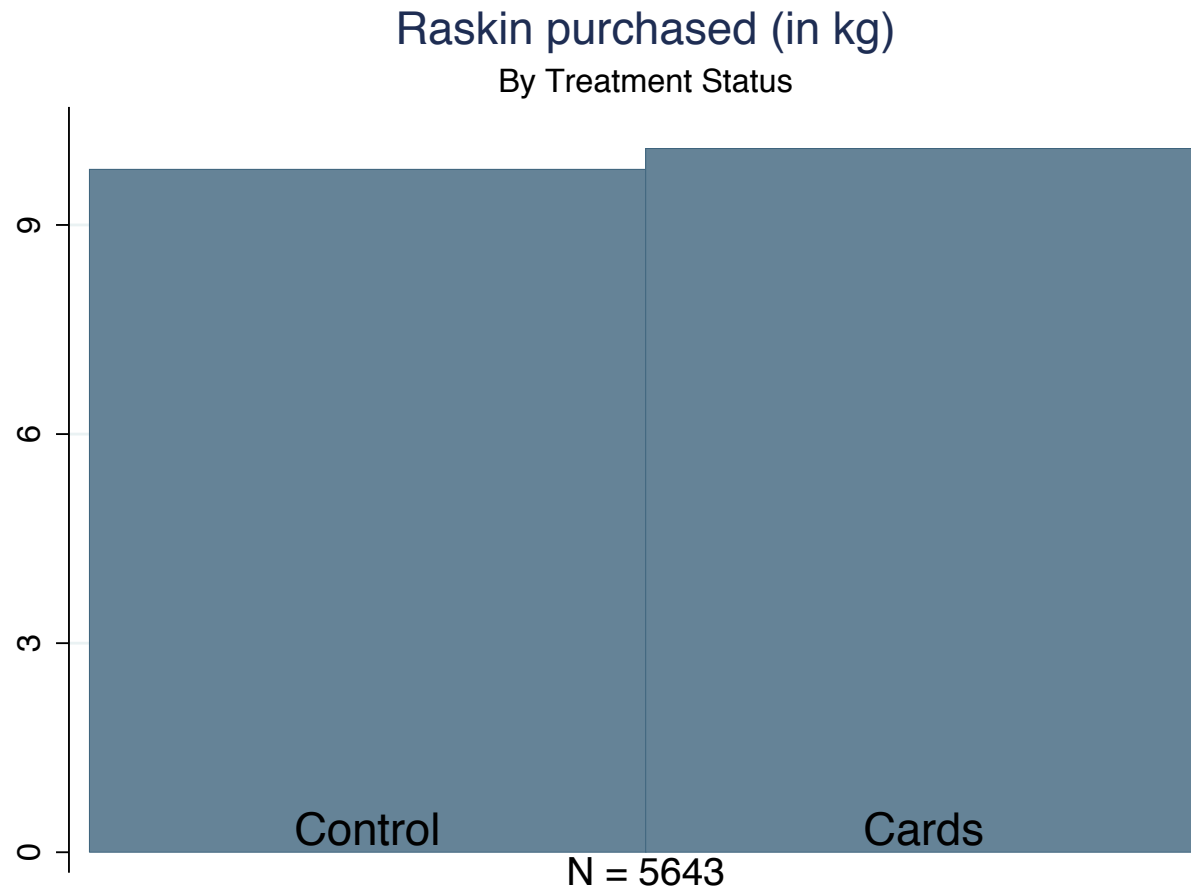
# Illustration of Randomization

---



# Statistically, villages who receive treatment and the control group are identical prior to pilot

---



# Implementation: Treatment

# Implementation: Treatment

---

**Treatment  
operational**

***Set quality  
control checks***

**Treatment  
preparation and  
launching**

**Coordination  
with government**

**Field visit**

# Implementation Plan

---

- Identifying how to implement treatment in the field
  - Developing detailed implementation plan with facilitation specialist and government
  - Use vendor chosen by government
- Set strong internal quality management
  - *Random checks* during the printing of cards
  - Standards for facilitator recruitment
  - Documenting treatment implementation with standardized forms
  - Clear reporting procedure

# Implementation Plan

---

- Coordination with government
  - Workshop in Jakarta for representatives from province, district/city.
  - Pre-field work coordination with head of district/city, sub-district, and village
- Treatment preparation and launching
  - Scheduling, training, and logistic coordination
- Field visit: to observe response from treatment given

# Treatment variation

Card variation			Standard socialization	Additional Socialization
All eligible households	With price	With coupon	Group 1	Group 2
		Without coupon	Group 3	Group 4
	Without price	Coupon	Group 5	Group 6
		Without coupon	Group 7	Group 8
Only the 10% poorest of eligible households	With price	Coupon	Group 9	Group 10
		Without coupon	Group 11	Group 12
	Without price	Coupon	Group 13	Group 14
		Without coupon	Group 15	Group 16
			Control (Without card and socialization at all)	

# Recap: Raskin card variation

---

- Randomly, card is customized based on:
  - *Design*: with or without coupon
  - *Content*: with or without price
  - *Distribution*: to all eligible households or only 10% poorest
- **Distribution of Raskin card to eligible households**

Sept. to mid of Oct. '12

  - 378 villages received card;
  - 194 control villages (did not receive card)



# Recap: Raskin card variation

---

Randomly, village with Raskin card receive:

- *Standard socialization*: Letter, DPM; or
- *Additional Socialization*:
  - + 3 DPM per sub-village
  - + 3 information poster per sub-village
  - Socialization to tokoh masyarakat
  - Announcement through mosque

# Recap: Raskin card variation

---

- End of Sept. to mid-Nov. '12
- 378 villages who receive card and socialization
  - 186 villages: Standard socialization
  - 192 villages: Additional socialization
- 194 control villages did not receive socialization

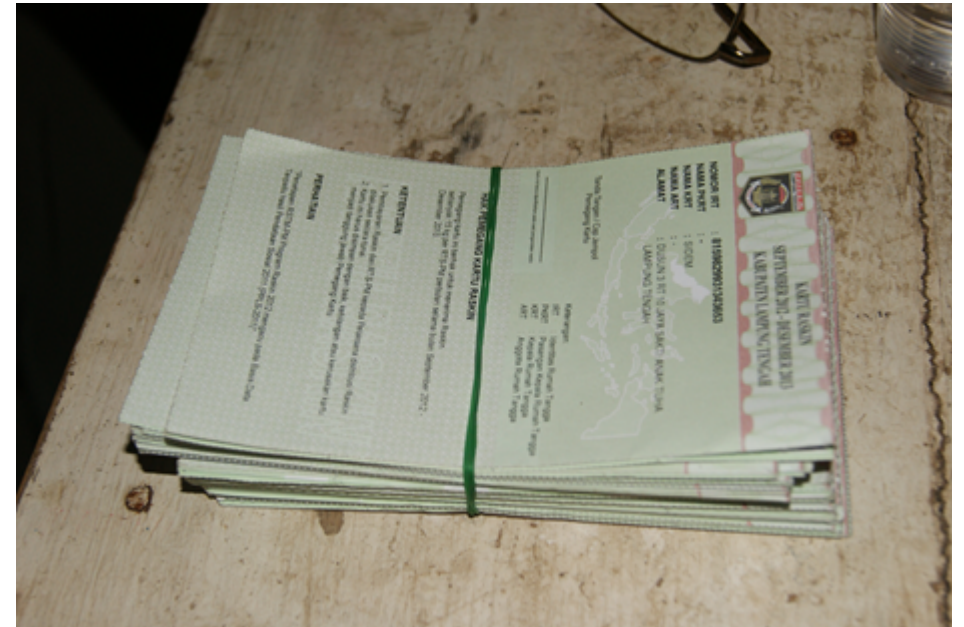


*Seorang fasilitator menjelaskan mengenai kartu Raskin kepada pemimpin kampung di Lampung Tengah*

# Challenges in Implementing Treatment

---

- Some treatment villages are not safe and easily accessible
- There's a lot of card who are not distributed to beneficiaries



# Challenges in Implementing Treatment

---

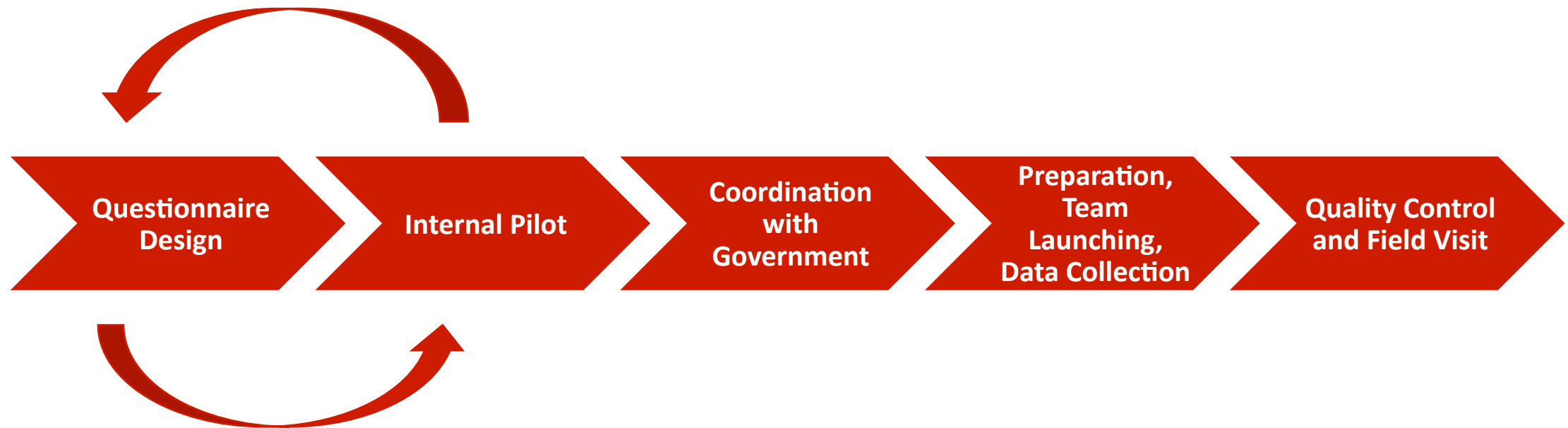
- The availability of poster explaining the beneficiaries list was not sufficient for facilitators
- During the meeting, participants were complaining on other aspect of the Raskin/ government



# Implementation: Data Collection

# Implementation: Data Collection

---



# Data Collection Plan

---

- Survey instrument: HH survey and people
- Baseline survey—use *endline survey* from previous project
  - Ensure that both treatment and control group are statistically indifferent
- Respondent identification
  - Respondent:
    - Eligible households (poor)
    - Eligible households (extremely poor/10% poorest)
    - Ineligible households
  - *Listing*, to identify ineligible households
  - Use PPLS'10 data, to identify eligible households

# Challenges in Data Collection

---

- To match administrative data and field data
  - *Human error*, change of poverty status/address
- To accommodate change in administrative area i.e. new province/district
- Time constraint
- To collect enough human resource to collect data
- Other challenges: how can respondent recall the memory? How to explain “Distribution Point”?



# Data Collection

---

Survey	Sumber	Responden	Data yang dikumpulkan
<b>Baseline</b> 2011	Endline from Targeting II, previous project	PKH recipient, non-poor	Main objective: to ensure control and treatment group are statistically indifferent
<b>Midline</b> Oct-Dec'12	5,148 HHs, through HH survey and community survey (target: Head of village)	Non-poor, eligible HHs (poor and non-poor)	Quantity and price of Raskin purchased, knowledge about Raskin program, satisfaction rate, HH consumption, relative wealth level, etc.
<b>Endline</b> Mar-May '13	6,292 HHs, through HH and community survey	<i>Ibid</i>	<i>Ibid</i>

# Challenges in Program Evaluation

---

- **Attrition:** when evaluator failed to collect data from selected individuals as part of original sample
  - *Midline:* 9% was replaced (418/4,572), *Endline:* 9.8% (561/5,706)
  - Respondent change was integrated within data collection process
  - Excluded 28 villages from evaluation

DAFTAR SAMPLE - SUMSEL

[A-D]-6 C-3

Sample A	Area	Typ	KKID	u1	u2	u3	Nama KET	Nama PERT	Alamat	Informasi / Tolong	Keterangan Lokasi	Nomor Telpon	# Sampel program
	12345	Sampled B1	123456789	Dxxxx B3	RW B4	RT B2	XXXXXXXX	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	
	12345	Sampled B2	123456789	Dxxxx B3	RW B4	RT B2	XXXXXXXX	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	
	12345	Sampled B3	123456789	Dxxxx B3	RW B4	RT B2	XXXXXXXX	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	
	12345	Replacement B1	123456789	Dxxxx B3	RW B4	RT B2	X	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	

Sample B	Area	Typ	KKID	u1	u2	u3	Nama KET	Nama PERT	Alamat	Informasi / Tolong	Keterangan Lokasi	Nomor Telpon	Instruksi	# Sampel program
	12345	Sampled B1	123456789	Dxxxx B3	RW B4		XXXXXXXX	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	Periksa SL53!	
	12345	Sampled B2	123456789	Dxxxx B3	RW B4		XXXXXXXX	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	Periksa SL53!	
	12345	Sampled B3	123456789	Dxxxx B3	RW B4		XXXXXXXX	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	Periksa SL53!	
	12345	Replacement B1	123456789	Dxxxx B3	RW B4	RT B2	X	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	Periksa SL53!	
	12345	Replacement B2	123456789	Dxxxx B3	RW B4	RT B3	X	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	Periksa SL53!	
	12345	Replacement B3	123456789	Dxxxx B3	RW B4		XXXXXXXX	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	Periksa SL53!	
	12345	Replacement B4	123456790	Dxxxx B4	RW B5		XXXXXXXX	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456790	Periksa SL53!	
	12345	Replacement B5	123456791	Dxxxx B5	RW B6		XXXXXXXX	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456791	Periksa SL53!	
	12345	Replacement B6	123456792	Dxxxx B6	RW B7		XXXXXXXX	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456792	Periksa SL53!	
	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Sample C	Area	Typ	KKID	u1	u2		Nama KET	Nama PERT	Alamat	Informasi / Tolong	Keterangan Lokasi	Nomor Telpon	# Sampel program
	12345	Sampled B1	123456789	Dxxxx B3	RW B4		XXXXXXXX	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	
	12345	Sampled B2	123456789	Dxxxx B3	RW B4		XXXXXXXX	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	
	12345	Sampled B3	123456789	Dxxxx B3	RW B4		XXXXXXXX	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	
	12345	Replacement B1	123456789	Dxxxx B3	RW B4		X	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	
	12345	Replacement B2	123456789	Dxxxx B3	RW B4		X	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	
	12345	Replacement B3	123456789	Dxxxx B3	RW B4		X	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	
	12345	Replacement B4	123456789	Dxxxx B3	RW B4		X	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	
	12345	Replacement B5	123456789	Dxxxx B3	RW B4		X	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	
	12345	Replacement B6	123456789	Dxxxx B3	RW B4		X	YYYYYYY	222222	AAAAAAAA	DDDDDDDD	123456789	
	.	.	.	.	.	.	.	.	.	.	.	.	.

# Analysis

# Analysis



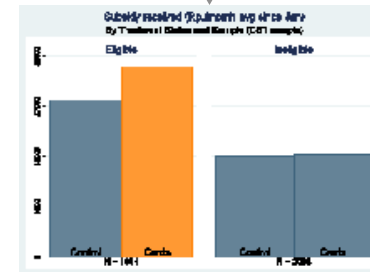
1. Formulate the analysis plan



2. Write the *STATA do.file*



3. Process the data using program



4. Produce the charts

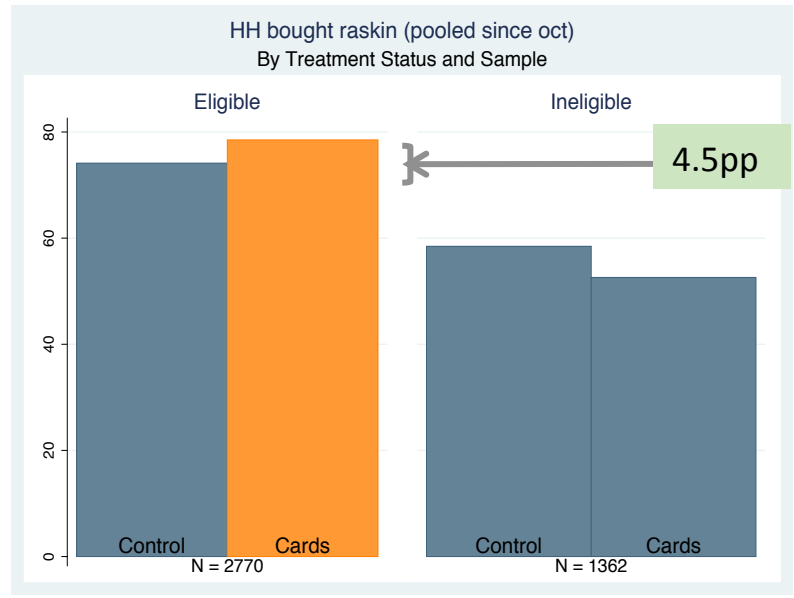


5. Analysis: Relate the qualitative field observation

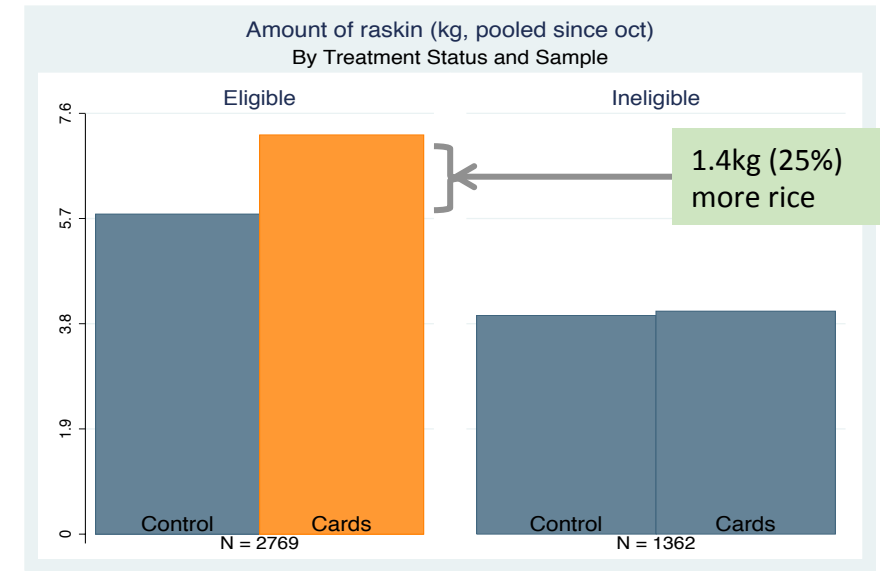


6. Disseminate the findings

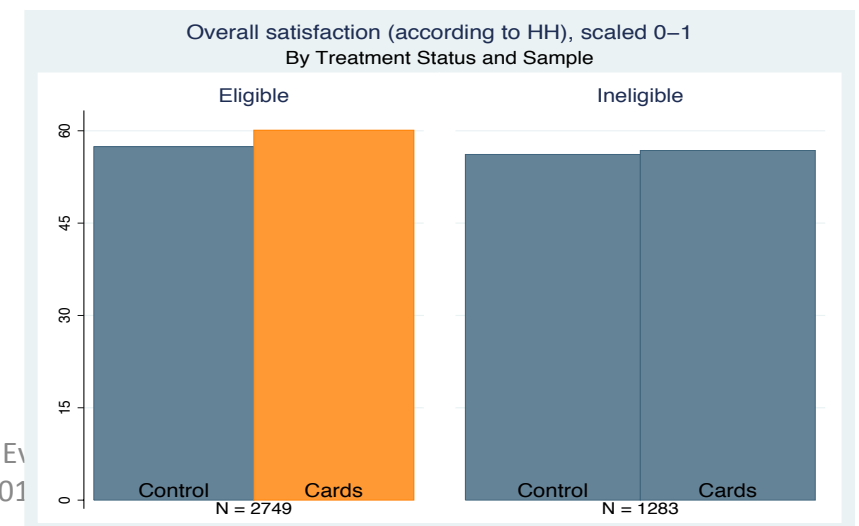
### Increased Raskin purchase by those who are eligible vs. Control



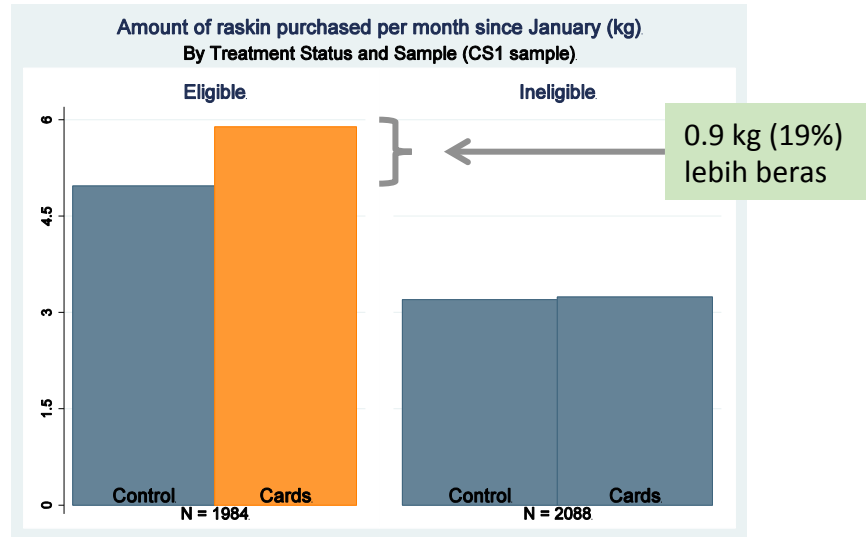
### Eligible HH in treatment group purchased more Raskin vs. Control



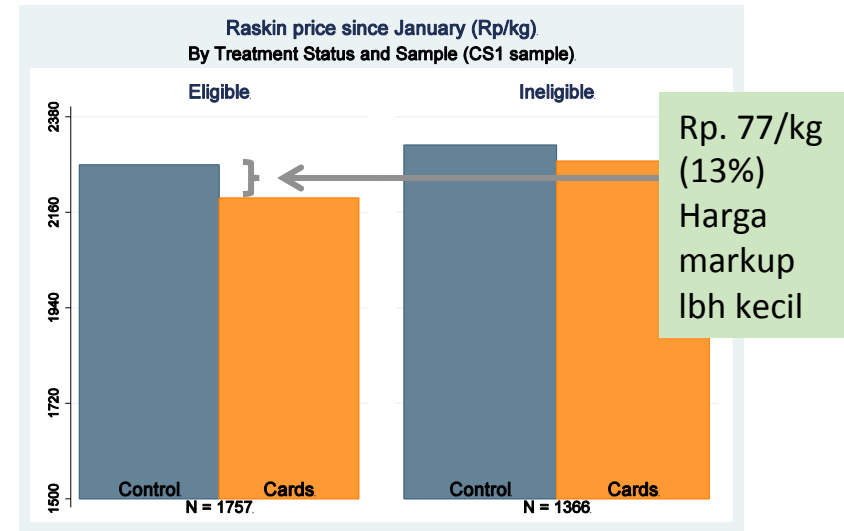
### Eligible HH in villages with card has higher level of satisfaction upon Raskin Relative vs. Control



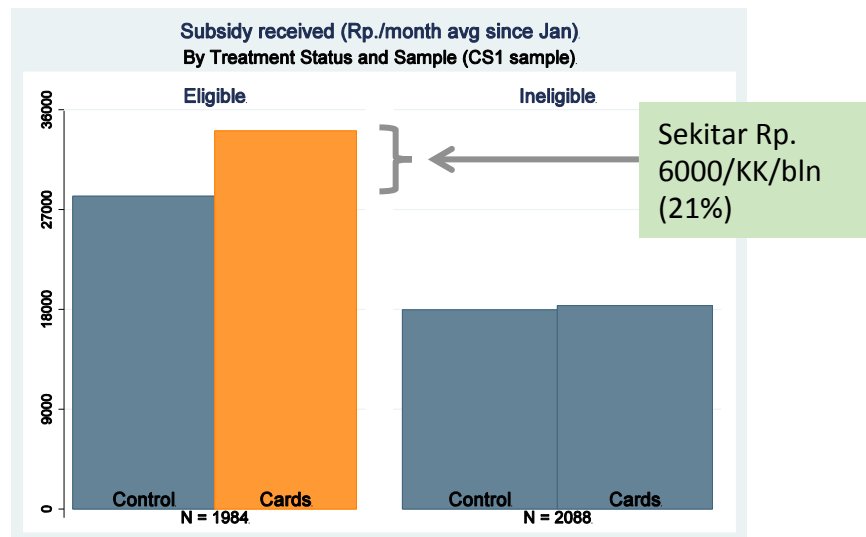
### Eligible HH in treatment village purchased more Raskin



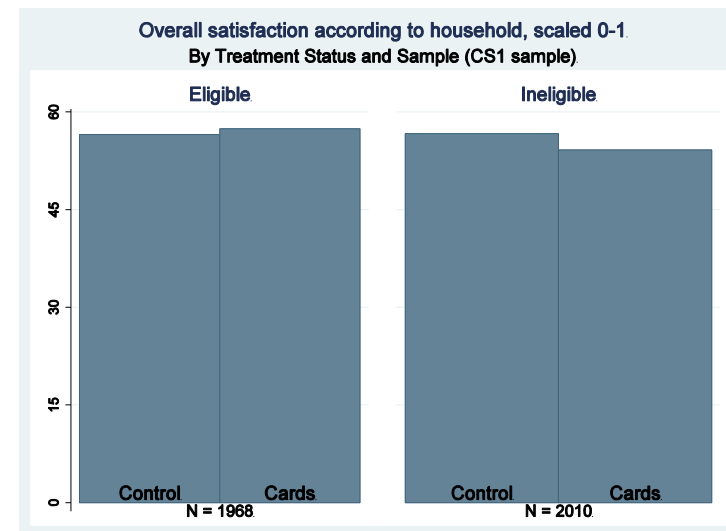
### Eligible HH in treatment village paid with lower markup



### Raskin subsidy received by eligible HH in treatment village increased



### There's no difference on household satisfaction rate between Treatment vs. Control



# Result Development

Treatment	Midline	Endline
<b>Card</b> >> <i>take-up</i> , purchase, and subsidy	(+)	(+)
<b>Card</b> >> HH satisfaction rate	(+) <i>eligible</i>	No effect
<b>Additional Socialization</b> >> purchase, subsidy	(+) <i>eligible</i>	(+) <i>eligible</i>
<b>Additional Socialization</b> >> satisfaction	(+) households (-) leader	(+) households No effect from leader
<b>Card with Raskin Price</b> >> <i>price mark-up</i>	(-)	(-), only with additional socialization
<b>Coupon</b> >> subsidy	(-) <i>ineligible</i>	(+) <i>eligible</i>
<b>Distribution to 10% poorest</b> >> Rice purchase	No effect	(+), due to additional socialization



# Input for Policy

---

- **The distribution of Raskin card improved the program implementation**
  - Eligible HHs who purchase and the amount of purchase increased
  - Decreased number of incident where there is a prie mark-up
  - Net, it is a Rp 6,000 subsidy adjustment for eligible HH, and no change for those who are ineligible
- **Additional socialization increased subsidy and the satisfaction level of beneficiary**
  - Achieved with only 2-3 person-days of external facilitation and 3 posters per dusun.

# Input for Policy (cont.)

---

- **Adding Raskin price in card increase the effectiveness of the program**
  - When combined with additional socialization, the short term effect remains until medium-term
- **The distribution of card to only 10% poorest HH may become the effective way to improve targeting**
  - In the beginning, 10% poorest HH reported lower level of satisfaction and no significant difference is observed for take-up level
  - However, in the last survey, they reported higher amount of subsidy and higher level of satisfaction rate in the village with additional socialization

# Input for Policy (cont.)

---

- Using coupon is potentially effective if also combined with additional socialization
- In the medium-term, the card variation effect depends on the additional socialization of the program.
  - This applies to card with coupon price and targeting to 10% poorest HHs.

# Policy Scale-up

---



# Kartu Perlindungan Sosial (KPS; Social Protection Card)

- TNP2K upgraded Raskin card to be Kartu Perlindungan Sosial (KPS).
- As per June 2013, the KPS has been distributed nationally to 15.5 million households (65.6 million people).
- The card can be used as kartu Bantuan Langsung Sementara Masyarakat (BLSM); kartu Bantuan Siswa Miskin (BSM); and Raskin.



*Tampak depan*

*Tampak belakang*

# Cost-effectiveness analysis and Scaling up

---

Manila, November 2015



ABDUL LATIF JAMEEL  
Poverty Action Lab

J-PAL SOUTHEAST ASIA

# Cost-effectiveness analysis

---

Manila, November 2015



ABDUL LATIF JAMEEL  
Poverty Action Lab

---

J-PAL SOUTHEAST ASIA

# Course Overview

---

1. What is Evaluation?
2. Measuring Impacts
3. Why Randomize?
4. How to Randomize
5. Sampling and Sample Size
6. Threats and Analysis
7. Raskin: RCT Project from Start to Finish
8. Cost Effectiveness Analysis and Scaling Up



# Course Overview

---

1. What is Evaluation?
2. Measuring Impacts
3. Why Randomize?
4. How to Randomize
5. Sampling and Sample Size
6. Threats and Analysis
7. Raskin: RCT Project from Start to Finish
- 8. *Cost Effectiveness Analysis and Scaling Up***

# Course Overview

---

1. What is Evaluation?
2. Measuring Impacts
3. Why Randomize?
4. How to Randomize
5. Sampling and Sample Size
6. Threats and Analysis
7. Raskin: RCT Project from Start to Finish
- 8. *Cost Effectiveness Analysis***

# Outline

---

- 1. Example: From impact to cost-effectiveness analysis***
2. What is CEA? (vs. CBA)
3. Common uses of CEA
4. Key challenges in doing CEA



# Evaluating Immunization Camps and Incentives in Udaipur, India – Supply Side

---

- Immunization is really low in Rajasthan (less than 5% in Udaipur)
- One possibility is that the supply channel is the problem:
  - Hilly, tribal region with low attendance by city based health staff to local health clinics (45% absenteeism)
  - Conducted monthly immunization camps in 60 villages: regular camps held rain or shine from 11a-2p (95% held)
  - Camera Monitoring



# The Demand Side of Immunization

---

- Second possibility: There is a problem of demand:
  - People not interested in immunization, scared?
  - Opportunity cost of going for 5 rounds of vaccination
  - Can demand be affected?

# Incentivizing Demand

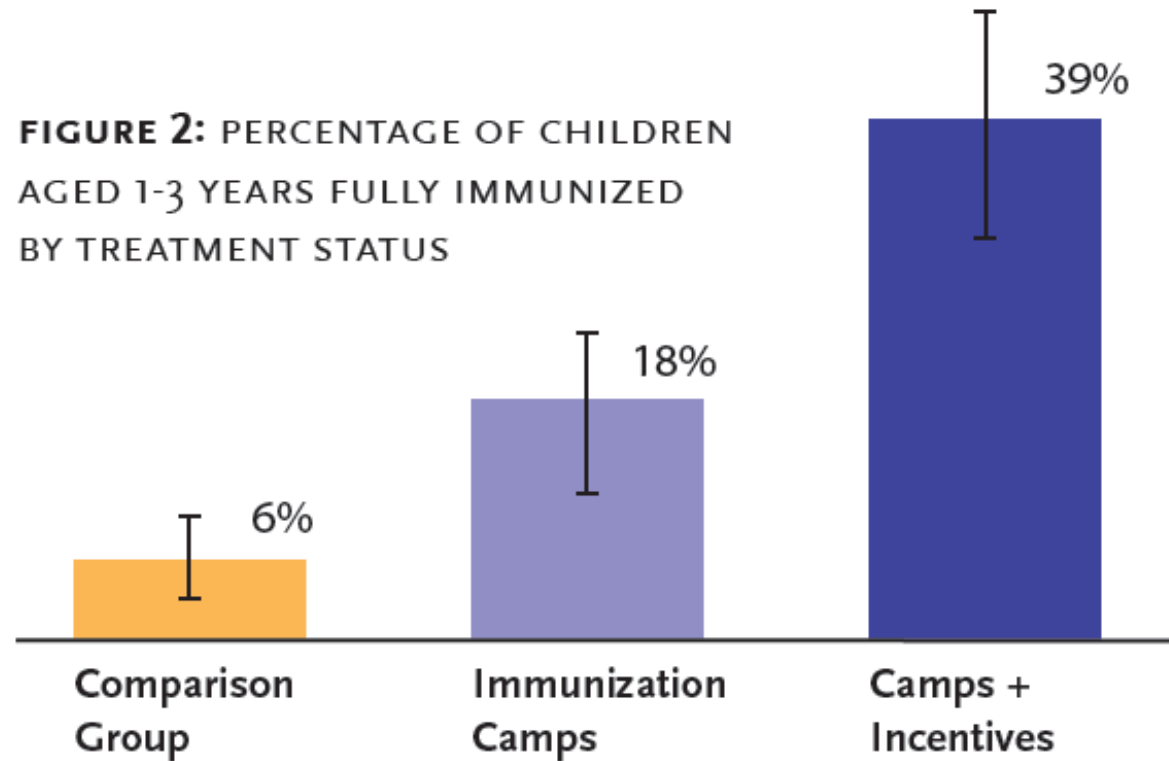
---

- Extra incentive: provided one kilogram of lentils for each immunization (Rs. 40 – one day's wage) plus thali set for full course
- Treatment 1: Reliable camps
  - 30 villages
- Treatment 2: Reliable camps + incentives
  - 30 villages
- Control group
  - 60 villages
- Collected data on immunization rates



# Regular Supply Increased Immunization, Incentives Helped it Even More

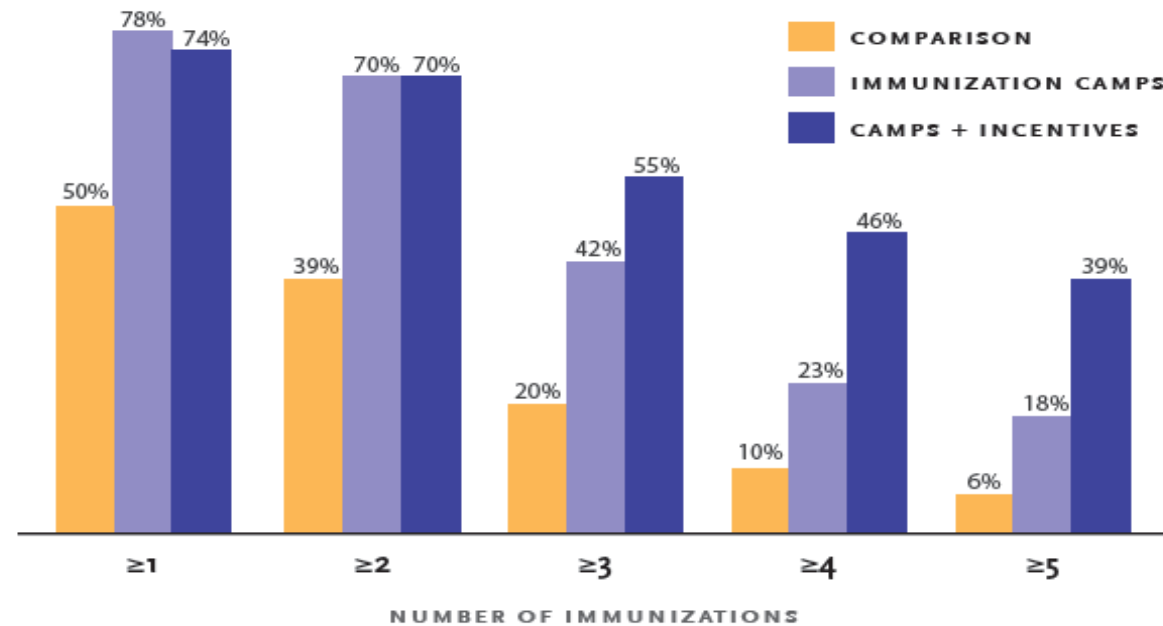
---



# Regular Supply Increased Immunization, Incentives Helped it Even More

---

**FIGURE 1: NUMBER OF IMMUNIZATIONS RECEIVED BY CHILDREN AGED 1-3 YEARS**

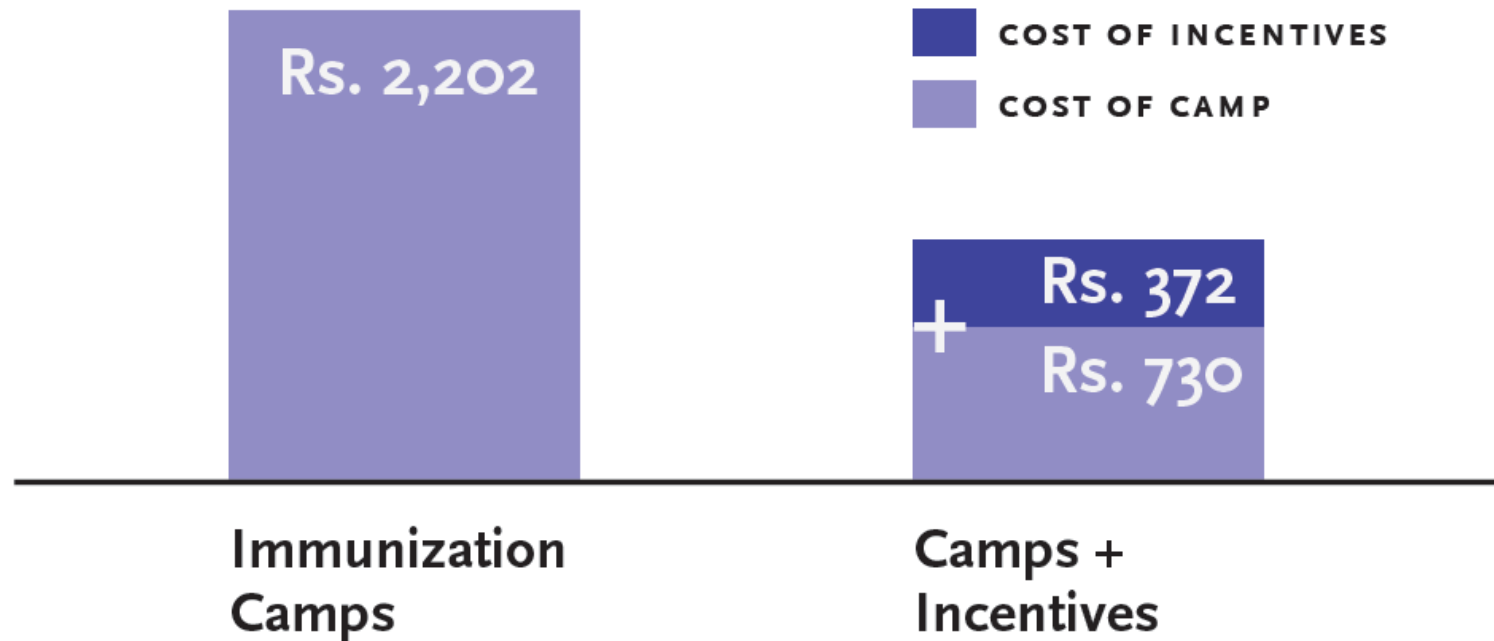




# Giving incentives was twice as cost-effective

---

**FIGURE 3: COSTS PER FULLY IMMUNIZED CHILD**



# Outline

---

1. Example: From impact to cost-effectiveness analysis
- 2. *What is CEA? (vs. CBA)***
3. Common uses of CEA
4. Key challenges in doing CEA



# Cost-effectiveness Analysis (CEA)

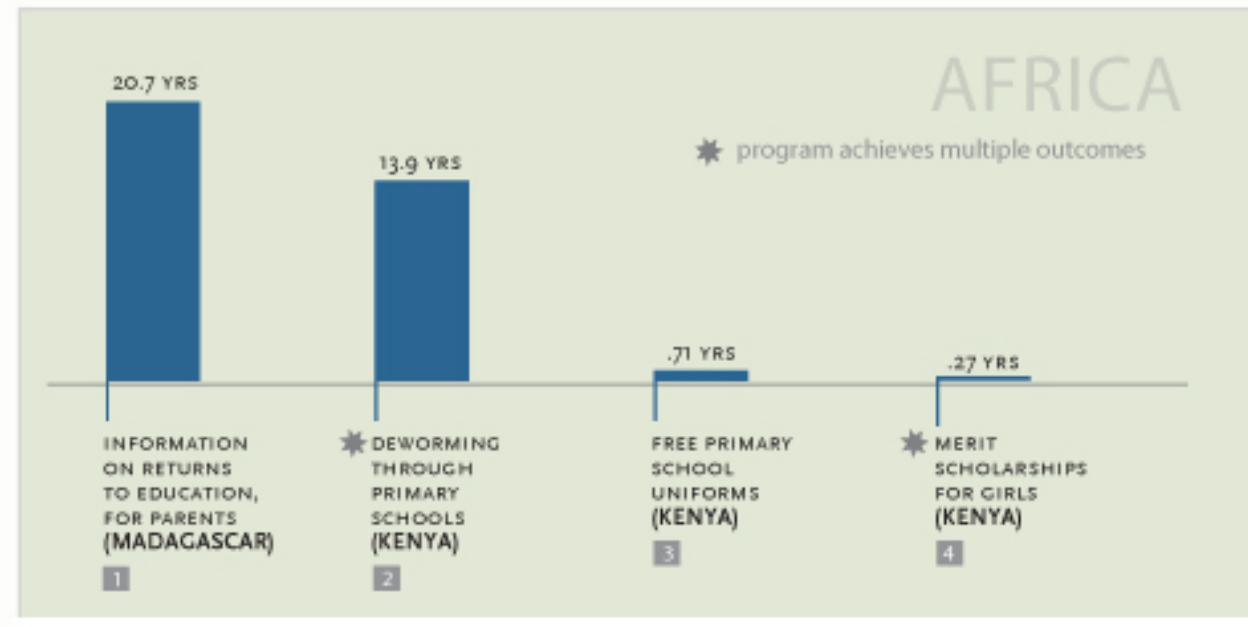
---

- Cost-effectiveness analysis measures the ratio of the costs of a program to the effects it has on one outcome
  - Measure the cost for a given level of effectiveness: e.g. cost to increase school attendance by 1 year
  - Or, measure the level of effectiveness for a given cost: years of additional attendance induced by spending \$100

# Comparative Cost-effectiveness Analysis

- **Comparative** cost-effectiveness then compares this cost-effectiveness ratio for multiple programs

Example: Years of schooling gained per \$100 spent



# Cost-effectiveness Analysis (CEA) and Comparative CEA

---

- **Comparative** cost-effectiveness then compares this cost-effectiveness ratio for multiple programs
  - Must compute costs and benefits using similar methodology for all programs
- Good way to help policymakers synthesize information from many evaluations
  - Provides a summary of a single program in terms of its costs and effects on one outcome
  - Can be used to compare many programs, find the most cost-effective option (comparative analysis)

# Cost-Effectiveness (CEA) vs. Cost-Benefit Analysis (CBA)

---

- CEA: Ratio of costs to effect on one outcome
- CBA: Ratio of costs to monetary value of effects on all outcomes
  - Can deliver absolute judgment on whether a program is worth the investment.
  - But, also requires assumptions about the monetary value of all the different benefits. (cost of life, disability, lower crime among school kids)
- Advantage of CEA is its simplicity:
  - Allows user to choose an objective outcome measure (e.g. cost to induce an additional day of schooling) – no need for making judgments on monetary value of that schooling
  - Easier for policymakers to compare programs when they are primarily concerned about one outcome of interest (e.g. increasing school attendance, not child health)

# When is cost-effectiveness analysis useful?

---

- You have a specific outcome measure you want to affect
  - There are many possible interventions to address this goal, and you are unsure which will get the most impact at the least cost
- You want to convince a decision maker that a non-obvious program is a good idea (Example: Deworming)
- You want to understand how the CE of a program could vary with contextual and implementation factors

# What info is needed?

---

- Take impact measures from rigorous impact evaluations
  - Need some other info, like number of beneficiaries, when impacts were measured
- Take cost data from...?
  - Most projects don't record their implementation costs
  - Need fairly disaggregated specific data on exactly what items were purchased, how much staff time was spent (on what), transportation costs, etc.



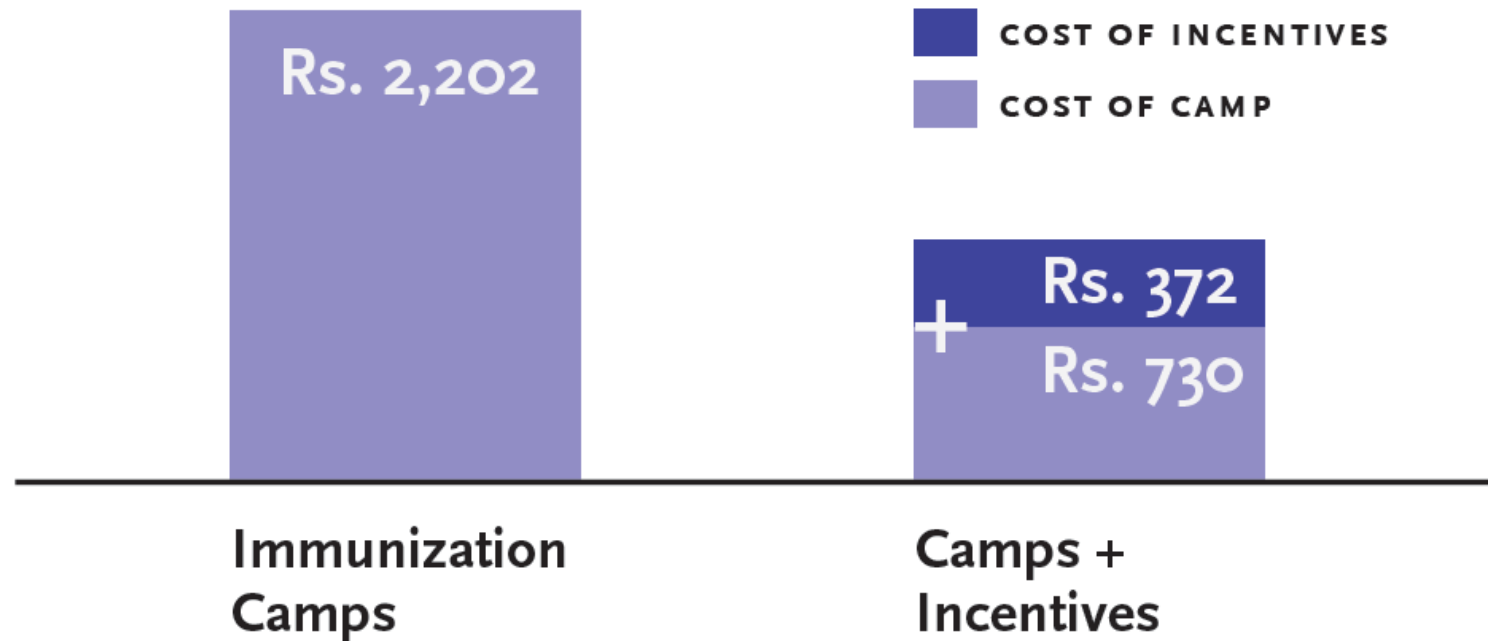
# Tally the full Costs of the Program – Ingredients Method

Cost Components	Details	Camps with Incentives	% of Total	Camps without Incentives	% of Total
Salary	Team of 4 GNMs and 4 GNM Assistants + Coordinators Salary	558,500	29%	558,500	46%
Travel	Staff and Incentive transport to camps	171,460	9%	63,460	5%
<i>Honourarium</i>	<i>USD 0.26 per child under 2 yrs per shot , given to village workers.</i>	<i>119,580</i>	<i>6%</i>	<i>62,370</i>	<i>5%</i>
Daily allowance	USD 1.10 for attending bi monthly meetings, given to village workers.	19,500	1%	19,500	2%
Consultancy fees	Paid for training of nurses and assistants.	2,200	0%	2,200	0%
Lodging & boarding	Expenses incurred during trainings.	7,333	0%	7,333	1%
Travel	For village worker's transport to trainings	4,645	0%	4,645	0%
Training Material	Office supplies disbursed during trainings.	1,500	0%	1,500	0%
Medicines	Includes paracetamol, syringes and needles, needle cutters, blood pressure instruments, and stethoscopes.	43,925	2%	15,320	1%
Refrigerators	Four for vaccine storage.	25,178	1%	25,178	2%
Cost of Monitoring	Includes cameras, film, and manpower required for monitoring camps, entering, and analyzing data.	446,480	23%	446,480	37%
<i>Incentive</i>	<i>Utensils and lentils (includes storage boxes)</i>	<i>550,164</i>	<i>28%</i>	<i>-</i>	<i>0%</i>
<b>Total</b>		<b>1,950,465</b>	<b>100%</b>	<b>1,206,486</b>	<b>100%</b>

# Giving incentives was twice as cost-effective

---

**FIGURE 3: COSTS PER FULLY IMMUNIZED CHILD**



# Outline

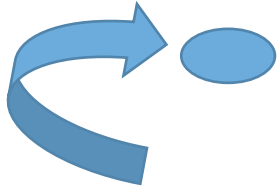
---

1. Example: From impact to cost-effectiveness analysis
2. What is CEA? (vs. CBA)
- 3. Common uses of CEA***
4. Key challenges in doing CEA

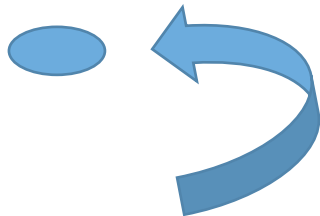


# Common CEA Uses

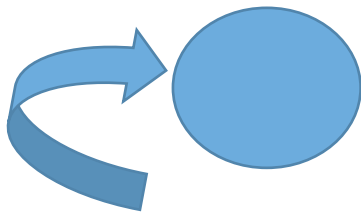
---



- Prospective analysis of pilot programs
  - “Roughly how cost-effective could this proposed pilot be?”
  - “How big an impact must this achieve to meet our threshold?”

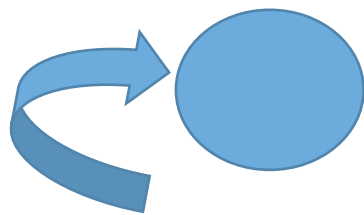
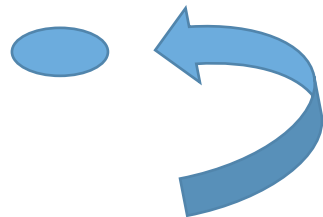
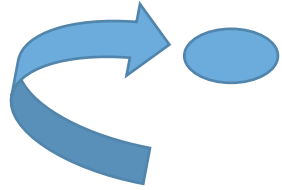


- Retrospective analysis of pilot programs
  - “Exactly how cost-effective was that pilot program?”



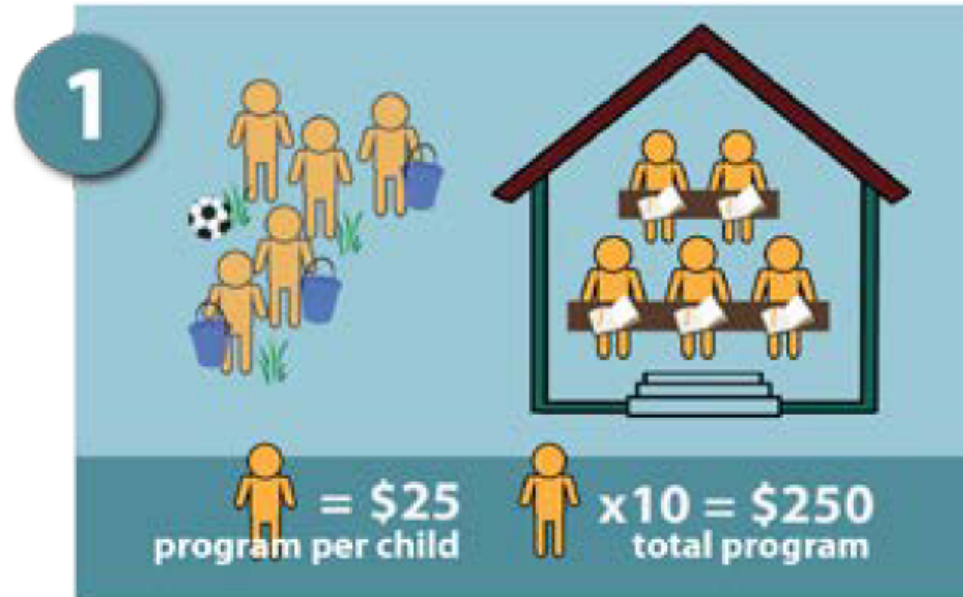
- Prospective analysis of programs at scale
  - “Roughly how cost-effective might this proposed national program be?”

# Common CEA Uses



	Necessary Data	Strengths	Weaknesses
Prospective Analysis of Pilot Programs	<ul style="list-style-type: none"> <li>Projected costs</li> <li>Impact estimates from a similar program</li> </ul>	Even rough calculations can help rule out programs that can't be cost-effective	Cost projections and impact estimates from similar programs may not be accurate
Retrospective Analysis of Pilot Programs	<ul style="list-style-type: none"> <li>Cost data from exact program that was evaluated</li> <li>Rigorous impact estimates</li> </ul>	Gives precise estimates of how cost-effective a program was in that context	Still suffers from external validity problem for cost and impact estimates
Prospective Analysis of Programs at Scale	<ul style="list-style-type: none"> <li>Projected cost data for program at scale</li> <li>Rigorous impact estimates from pilot evaluation</li> </ul>	Producing customized prospective estimates are a powerful tool when speaking with country governments	Impacts from small-scale pilots may not generalize to at-scale programs

# Using thresholds to assess cost-effectiveness

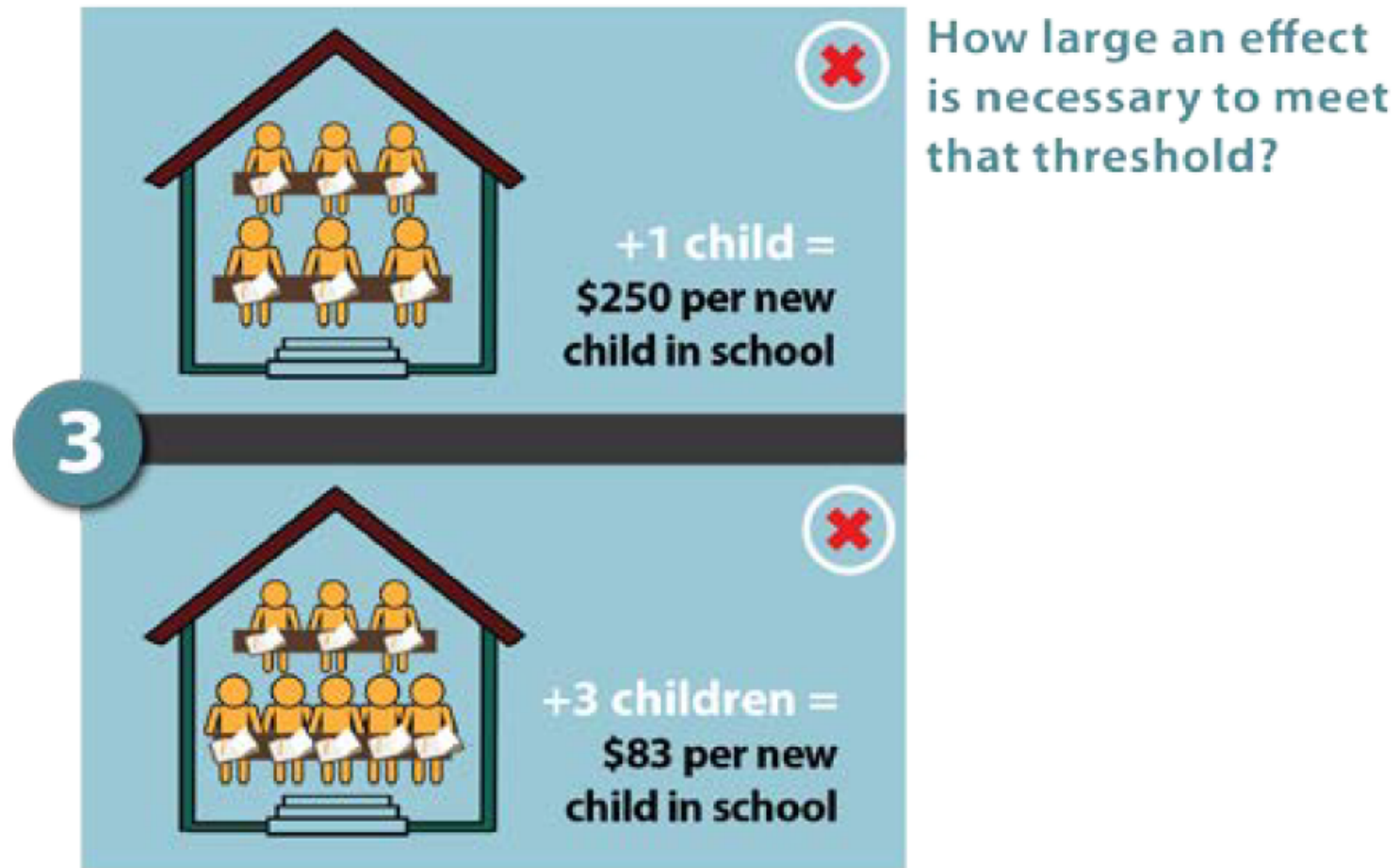


How much will  
the program cost?

**2** **threshold: no more than \$50  
per additional child in school**

What is threshold for  
cost-effectiveness?

# Using thresholds to assess cost-effectiveness



# Using thresholds to assess cost-effectiveness

---

4



Is that effect size likely?

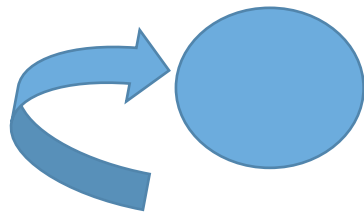
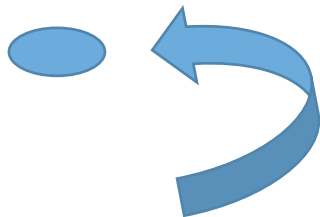
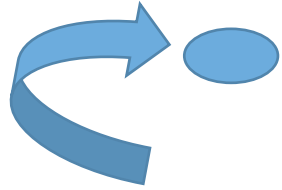
+5 children = \$50 per new child in school

100% increase in school attendance is only way to reach goal → is this attainable?

The diagram shows a schoolhouse icon with a brown roof and a green outline. Inside the schoolhouse, there are several stylized human figures representing children and an adult. The figures are arranged in a way that suggests a classroom setting. To the right of the schoolhouse, there is a green checkmark icon inside a white circle. Below the schoolhouse, there is a dark blue rectangular box containing white text.



# Common CEA Uses



	Necessary Data	Strengths	Weaknesses
Prospective Analysis of Pilot Programs	<ul style="list-style-type: none"> <li>Projected costs</li> <li>Impact estimates from a similar program</li> </ul>	Even rough calculations can help rule out programs that can't be cost-effective	Cost projections and impact estimates from similar programs may not be accurate
Retrospective Analysis of Pilot Programs	<ul style="list-style-type: none"> <li>Cost data from exact program that was evaluated</li> <li>Rigorous impact estimates</li> </ul>	Gives precise estimates of how cost-effective a program was in that context	Still suffers from external validity problem for cost and impact estimates
Prospective Analysis of Programs at Scale	<ul style="list-style-type: none"> <li>Projected cost data for program at scale</li> <li>Rigorous impact estimates from pilot evaluation</li> </ul>	Producing customized prospective estimates are a powerful tool when speaking with country governments	Impacts from small-scale pilots may not generalize to at-scale programs

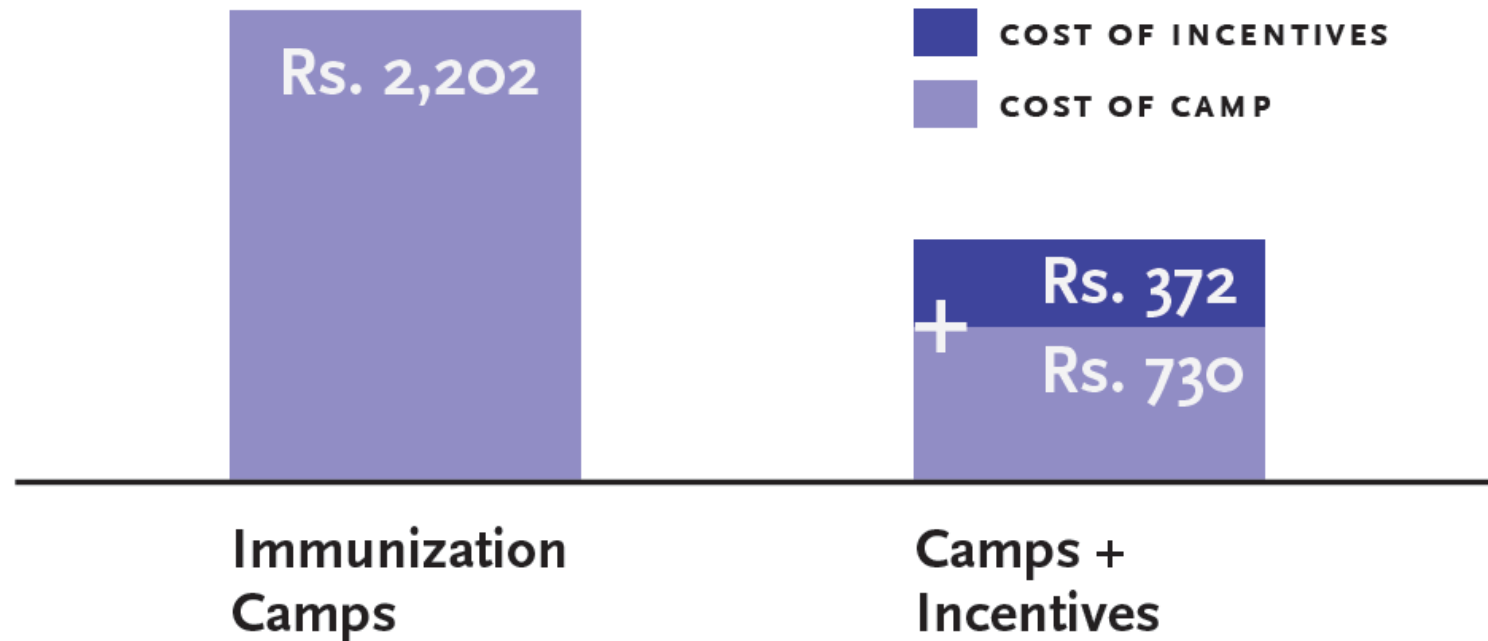
# Tally the full Costs of the Program – Ingredients Method

Cost Components	Details	Camps with Incentives	% of Total	Camps without Incentives	% of Total
Salary	Team of 4 GNMs and 4 GNM Assistants + Coordinators Salary	558,500	29%	558,500	46%
Travel	Staff and Incentive transport to camps	171,460	9%	63,460	5%
<i>Honourarium</i>	<i>USD 0.26 per child under 2 yrs per shot , given to village workers.</i>	<i>119,580</i>	<i>6%</i>	<i>62,370</i>	<i>5%</i>
Daily allowance	USD 1.10 for attending bi monthly meetings, given to village workers.	19,500	1%	19,500	2%
Consultancy fees	Paid for training of nurses and assistants.	2,200	0%	2,200	0%
Lodging & boarding	Expenses incurred during trainings.	7,333	0%	7,333	1%
Travel	For village worker's transport to trainings	4,645	0%	4,645	0%
Training Material	Office supplies disbursed during trainings.	1,500	0%	1,500	0%
Medicines	Includes paracetamol, syringes and needles, needle cutters, blood pressure instruments, and stethoscopes.	43,925	2%	15,320	1%
Refrigerators	Four for vaccine storage.	25,178	1%	25,178	2%
Cost of Monitoring	Includes cameras, film, and manpower required for monitoring camps, entering, and analyzing data.	446,480	23%	446,480	37%
<i>Incentive</i>	<i>Utensils and lentils (includes storage boxes)</i>	<i>550,164</i>	<i>28%</i>	<i>-</i>	<i>0%</i>
<b>Total</b>		<b>1,950,465</b>	<b>100%</b>	<b>1,206,486</b>	<b>100%</b>

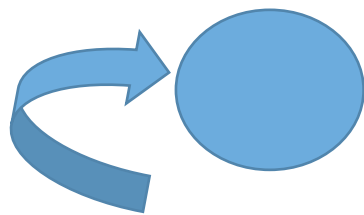
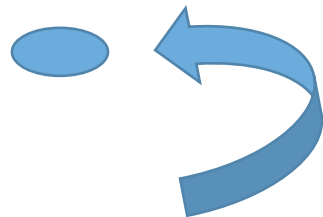
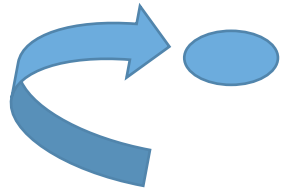
# Giving incentives was twice as cost-effective

---

**FIGURE 3: COSTS PER FULLY IMMUNIZED CHILD**



# Common CEA Uses



	Necessary Data	Strengths	Weaknesses
Prospective Analysis of Pilot Programs	<ul style="list-style-type: none"> <li>Projected costs</li> <li>Impact estimates from a similar program</li> </ul>	Even rough calculations can help rule out programs that can't be cost-effective	Cost projections and impact estimates from similar programs may not be accurate
Retrospective Analysis of Pilot Programs	<ul style="list-style-type: none"> <li>Cost data from exact program that was evaluated</li> <li>Rigorous impact estimates</li> </ul>	Gives precise estimates of how cost-effective a program was in that context	Still suffers from external validity problem for cost and impact estimates
Prospective Analysis of Programs at Scale	<ul style="list-style-type: none"> <li>Projected cost data for program at scale</li> <li>Rigorous impact estimates from pilot evaluation</li> </ul>	Producing customized prospective estimates are a powerful tool when speaking with country governments	Impacts from small-scale pilots may not generalize to at-scale programs

# Cost Effectiveness Analysis - Raskin

---

- Impact data: RCT evaluation of the effectiveness of the Raskin ID cards
- Cost data: TNP2K costs for the national KPS distribution

# Benefits of Raskin ID Card Distribution and Socialization

---

	Increase in Raskin purchased (kg/hh/month)	Increase in Raskin subsidy received (IDR/hh/month)
Raskin ID cards with <u>Regular</u> Socialization	1,12 kg	Rp 6.659 (23%)
Raskin ID Cards with <u>Enhanced</u> Socialization	1,9 kg	Rp 11.098 (38%)

# Cost Calculation Standard Socialization

---

Calculation utilizes TNP2K's actual cost estimates for card printing and distribution of the KPS card:

## **Cards + Standard Socialization: 190.6 billion IDR**

- Card Printing and Distribution: 12,200 IDR/household
  - 3,200 IDR/household to print the card and attached letter
  - 9,000 IDR/household to distribute the card and letter through PT Pos
  - 12,200 IDR \* 15.5 million households = 189.1 billion IDR
- Guideline Book Printing and Distribution: 20,000 IDR/village
  - 20,000 IDR \* 75,000 villages = 1.5 billion IDR

# Cost Calculation Enhanced Socialization

---

Calculation utilizes TNP2K's actual cost estimates for card printing and distribution of the KPS card and RCT cost of enhanced socialization

## **Cards + Enhanced Socialization: 715.6 billion IDR**

- Card Printing and Distribution: 12,200 IDR/household
  - $12,200 \text{ IDR} * 15.5 \text{ million households} = 189.1 \text{ billion IDR}$
- Guideline Book Printing and Distribution: 20,000 IDR/village
  - $20,000 \text{ IDR} * 75,000 \text{ villages} = 1.5 \text{ billion IDR}$
- Enhanced Socialization: 7 million IDR/village
  - $7,000,000 \text{ IDR} * 75,000 \text{ villages} = 525 \text{ billion IDR.}$

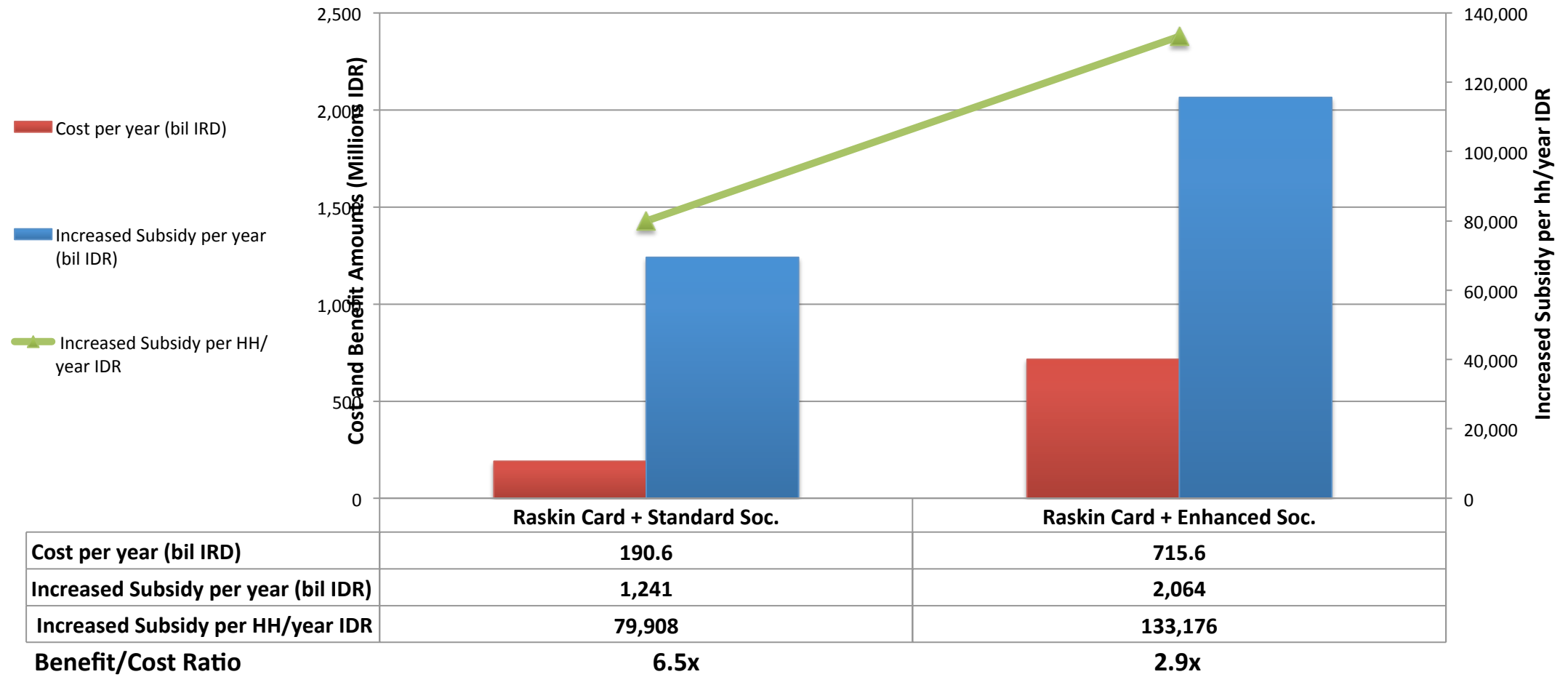


# Benefit to Cost Ratio of Raskin Card

---

Distribution & socialization Scenarios	Total cost per year (bn IDR)	Total Increased subsidy received per year (bn IDR)	Benefit/cost Ratio	Subsidy increase received per household per year
Raskin Card + Standard Soc.	190.6	1,240.6	6.5x	79,908 IDR
Raskin Card + Enhanced Soc.	715.60	2,064.2	2.9 x	133,176 IDR

# Cost Effectiveness Analysis - Raskin



# Outline

---

1. Example: From impact to cost-effectiveness analysis
2. What is CEA? (vs. CBA)
3. Common uses of CEA
- 4. Key challenges in doing CEA***



# Three Key Challenges in Doing CEA

---

## I. Absence of incentives to do CEA:

- What if the program was effective but not really cost-effective?
- No editorial requirement to show CEA in most social-science journals

## II. Not straightforward:

- Number of assumptions are needed to complete the analysis (e.g. multiple outcomes, transfers, spillover effects, exchange rates, inflation etc.)
- No one “right” way, but consistency is important!

# Three Key Challenges in Doing CEAs

---

I. Absence of incentives to do CEA

II. Not straightforward

**III. Costs are hard to gather:**

- Collecting cost data not seen as key part of evaluation unlike impact measures
- Cost data is surprisingly hard to collect from implementers (budgets different from implementation costs; hard to divvy up overhead and existing costs to project)
- Hard to get cost data from other authors for a *comparative* CEA
- Impact measures and cost collection often not harmonized

# Gathering Cost Data - Retrospectively

---

- Retrospectively:
  - J-PAL mostly uses “ingredients” method (Levin and McEwan 2001)
- Gather cost data from multiple sources:
  - Academic paper for description of program structure, ingredients and local conditions like wages
  - Interview researchers for additional ingredients, their costs, additional documents like budgets
  - Program staff and field research staff for unit cost data
  - Supplement with public sources (e.g. local wages, transportation costs etc.)

# Retrospective vs. Prospective Cost Gathering

---

- Challenges with retrospective approach:
  - Data not originally collected by implementer or evaluator and key field staff are hard to locate or do not respond
  - Many important costs are forgotten, or hard to estimate after long lag
  - Program as implemented may be very different from how it was budgeted
  - Aggregate cost data is much less useful for sensitivity analysis or scale-up
- Prospectively:
  - Overcomes challenges of retrospective cost gathering
  - J-PAL Initiatives provide standard templates to assist in data collection
  - Harmonization makes it easier to do *comparative* CEA

# Assumptions for CEA

---

- What are you calculating the cost-effectiveness of?
  - The program, during pilot phase
  - The program, if it was scaled up
  - Some component of the program
- How will you deal with...
  - Exchange, inflation, discounting
  - Spillover effects
  - Multiple outcomes
  - Costs shared with a partner organization
  - Fuzzy costs: administration, overhead, and management



# Issues to Consider in Cost Effectiveness Analysis – *there is no one right way*

---

- *Present Value*: Real discount rate of 10% is used to discount costs and benefits to control for time value of money
- *Inflation*: Adjust costs to today's prices
- *Across Countries*: Standard exchange rates are used to adjust to US\$
- *Multiple Outcome*: Can only examine one type of benefit at a time, which is how many policies are framed anyway



# Some Resources for CEA

---

- *Total vs. Sunk Costs*: Only consider incremental cost to the existing infrastructure (material, personnel, oversight)
- *Proximal Success vs. Final Impact of Programs*: Use global measures to translate proximal outcomes into final outcomes

**There is no one right way of doing a CEA. But we need to make choices (be transparent about assumptions) and apply the same standard across all studies in an analysis.**

# Issues to Consider in Cost Effectiveness Analysis – *there is no one right way*

---

- J-PAL paper on CE methodology:
  - Why CEA is valuable
  - What assumptions are necessary to perform CEA
  - Common problems or mistakes in calculating CEA

**[www.povertyactionlab.org/publication/cost-effectiveness](http://www.povertyactionlab.org/publication/cost-effectiveness)**

- Also includes some very basic templates for cost-gathering and doing CEA

# Conclusion

---

- CEA is a useful first step in comparing alternate programs that are aimed at the same outcome
- Simplicity allows for greater use of evidence in policymaking but need to make user aware of assumptions
- Sensitivity Analysis around CEAs allow policy makers to see the effect of modifying assumptions and local conditions
- Cost Collection process is far more accurate and easier when done prospectively rather than retrospectively

# Outline

---

1. Example: From impact to cost-effectiveness analysis
2. What is CEA? (vs. CBA)
3. Common uses of CEA
4. Key challenges in doing CEA
- 5. Scale ups***



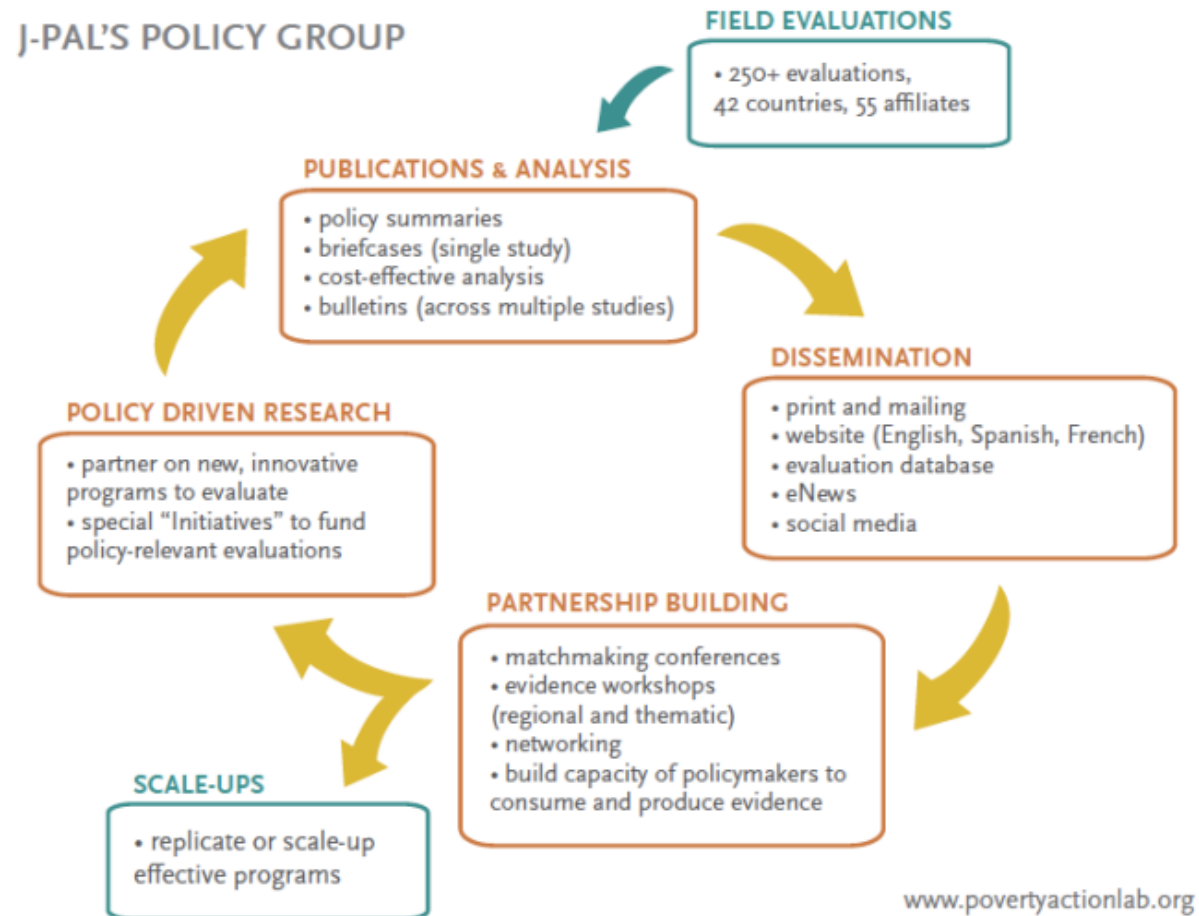
# There are Different Paths from Impact Evaluations to Scale-Ups

1. Government evaluate their pilot programs to demonstrate usefulness to public, gather support for their expansion and learn lessons to make it more effective (e.g. Progresa)
2. Leveraging evidence by implementing organization to expand exiting programs and get more funding (e.g. Pratham)
3. Independent organization can use evidence to replicate or scale-up programs found to be highly cost-effective, and/or simple to implement (e.g. Deworm the World)

# There are Different Paths from Impact Evaluations to Scale-Ups

4. If an evaluation helps provide evidence on a very policy relevant and salient topic, it gets a huge amount of traction very easily (e.g. Pricing)
5. Careful study of the new context, collaboration with original evaluator and implementer and a pilot replication (e.g. TCAI: remedial education in India and Ghana)

# There are Different Paths from Impact Evaluations to Scale-Ups – Here is one





# Final Issues to Consider in Scale Ups – *there are no easy answers*

- *Spillover Effects*: Spillovers may be different in a pilot vs. scaled program
- *Partial vs. General Equilibrium*: Very hard to measure precise nature or directions of such effects
- *Experimental vs. Scalable Mode*: Cost of inputs may become endogenous to the scale up
- *Hard to Control Contextual Differences*: Quality of infrastructure, motivation of local partners and beneficiaries, price differences, cultural differences, local parameters

# Conclusion

---

- CEA is a useful first step in comparing alternate programs that are aimed at the same outcome
- Simplicity allows for greater use of evidence in policymaking but need to make user aware of assumptions
- Sensitivity Analysis around CEAs allow policy makers to see the effect of modifying assumptions and local conditions
- Cost Collection process is far more accurate and easier when done prospectively rather than retrospectively
- The journey from impact evaluation to scale-ups is neither automatic, nor easy
- But we are learning more about the process and there are more and more success stories