Personalized Learning Summit 2017

Running Rigorous Evaluations in Personalized Learning

Vincent Quan

Policy Manager (Education Sector Lead)

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Running Rigorous Evaluations in Personalized Learning

Vincent Quan

Policy Manager (Education Sector Lead)

J-PAL North America, MIT

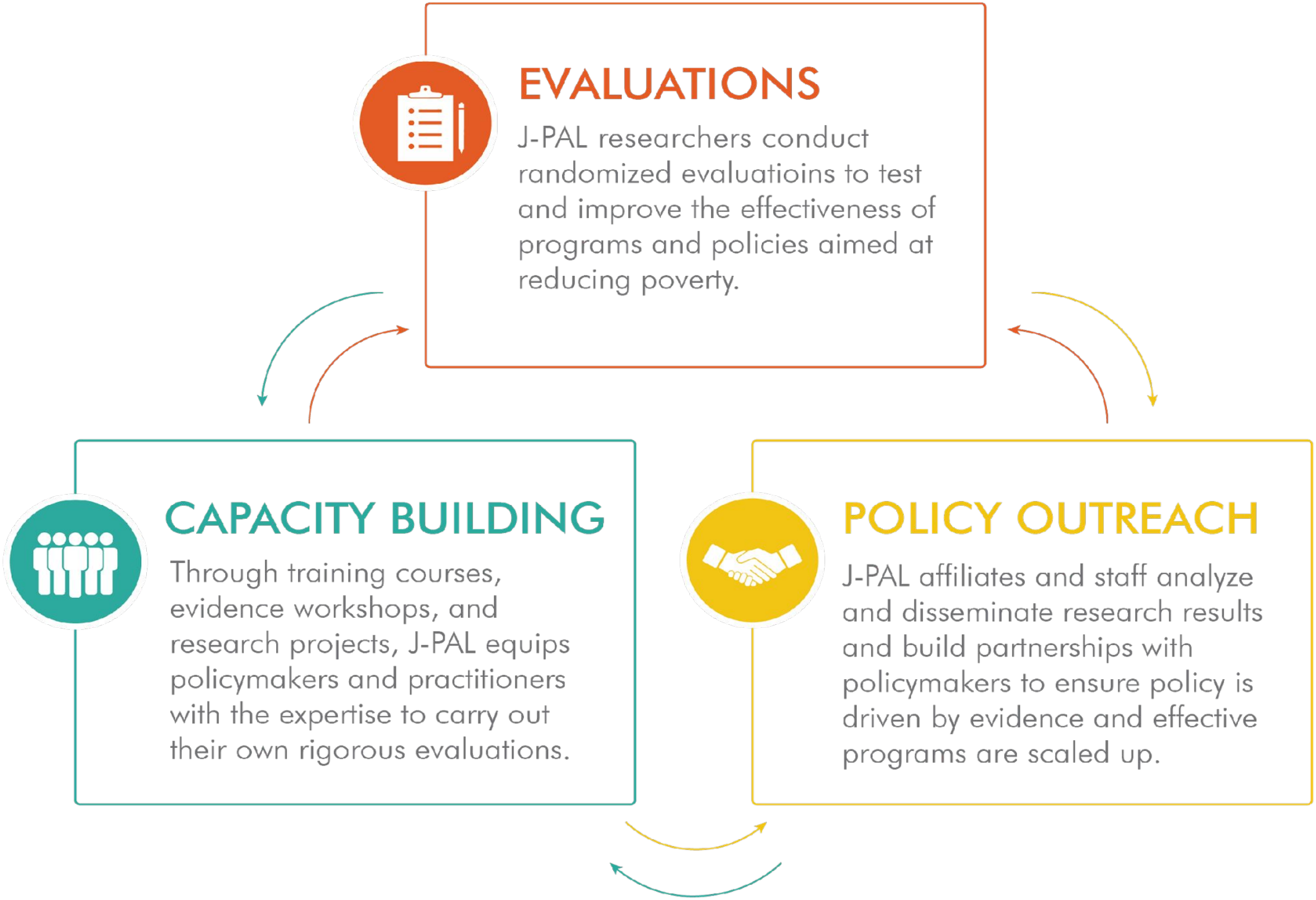
Twitter: @JPAL_NA

@edelements #PLSummit

bit.ly/PLSWorkshopSurvey

J-PAL'S MISSION IS TO ENSURE THAT POLICY IS DRIVEN BY EVIDENCE AND RESEARCH IS TRANSLATED INTO ACTION

www.povertyactionlab.org



EVALUATIONS

J-PAL researchers conduct randomized evaluatioins to test and improve the effectiveness of programs and policies aimed at reducing poverty.

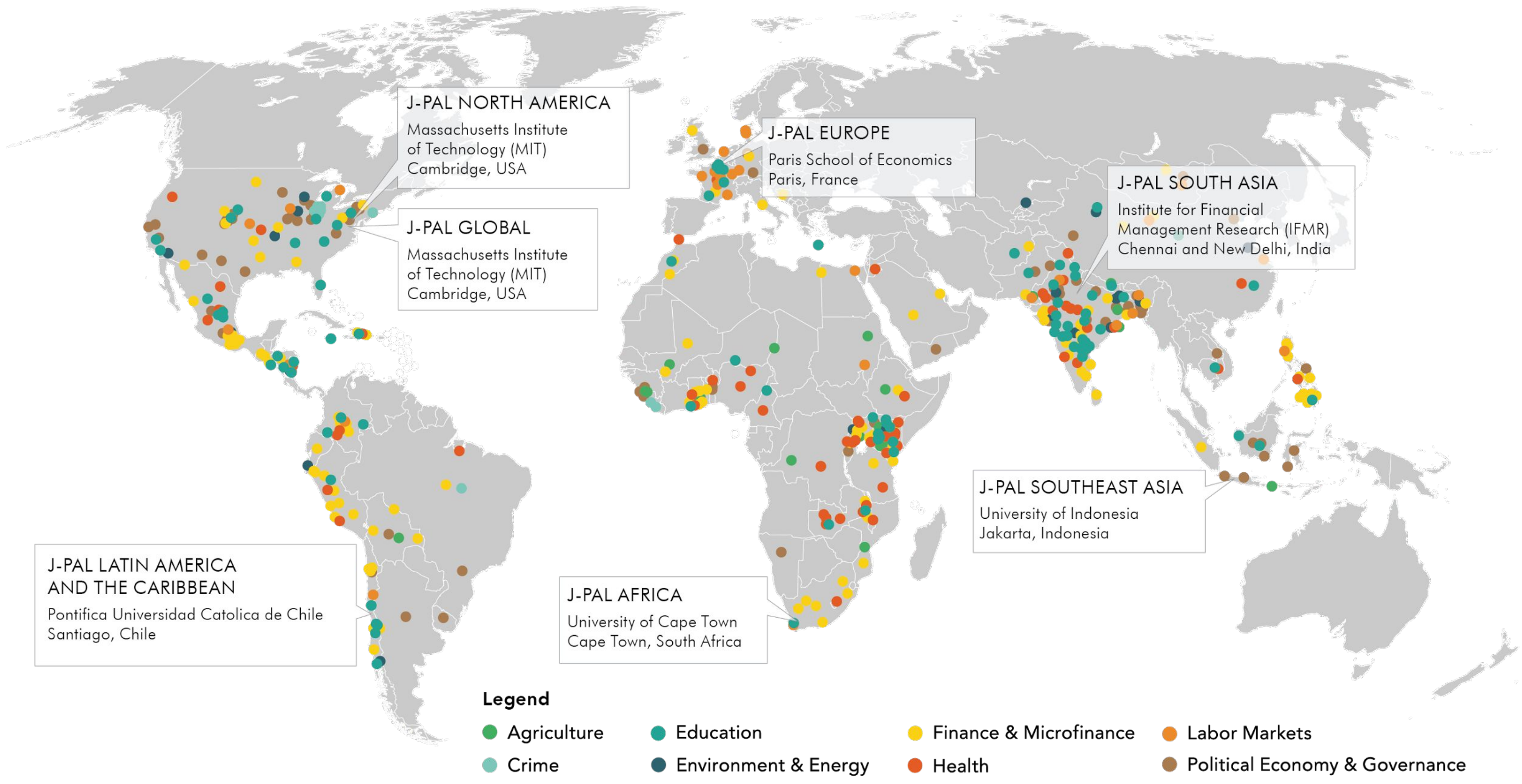
CAPACITY BUILDING

Through training courses, evidence workshops, and research projects, J-PAL equips policymakers and practitioners with the expertise to carry out their own rigorous evaluations.

POLICY OUTREACH

J-PAL affiliates and staff analyze and disseminate research results and build partnerships with policymakers to ensure policy is driven by evidence and effective programs are scaled up.

800+ ongoing and completed projects in 60+ countries
203+ million lives touched by the scale up of proven programs



J-PAL NORTH AMERICA
Massachusetts Institute
of Technology (MIT)
Cambridge, USA

J-PAL GLOBAL
Massachusetts Institute
of Technology (MIT)
Cambridge, USA

J-PAL EUROPE
Paris School of Economics
Paris, France

J-PAL SOUTH ASIA
Institute for Financial
Management Research (IFMR)
Chennai and New Delhi, India

J-PAL LATIN AMERICA
AND THE CARIBBEAN
Pontifica Universidad Catolica de Chile
Santiago, Chile

J-PAL AFRICA
University of Cape Town
Cape Town, South Africa

J-PAL SOUTHEAST ASIA
University of Indonesia
Jakarta, Indonesia

Legend

● Agriculture	● Education	● Finance & Microfinance	● Labor Markets
● Crime	● Environment & Energy	● Health	● Political Economy & Governance

Workshop overview

- Overview of Ed-Tech, Personalized Learning & Research
- What is Evaluation?
- Measuring Impact
- Randomized Evaluations
  - Different ways to randomize
  - Opportunities to randomize
  - Pitfalls to watch for
- Ethics of randomization

In the digital age, technology has the potential to transform learning

- Personalizing learning
- Engaging learners in innovative ways
- Increasing access to education for underserved groups
- Overcome spatial mismatches between learners and educational resources
- Providing teachers with new tools to strengthen their practice

Explosion of Innovation in Personalized Learning and Education Technology

Digital Promise
Accelerating Innovation in Education

ZEARN
Numeracy for All

LEAPinnovations

LEAP learning framework
for personalized learning

How Do We Know What Works?

- While we celebrate the explosion of innovation, we need to recognize that we don’t have all the answers yet.
- Not everything we try works.
- We want to invest scarce money and effort to improve education and lives.
- **Important to get it right**: if you invest in things that do not work rather than those that do, real people’s lives are affected in dramatic ways.

Can We Do Better Than Medieval Doctors?

- Problem with pre-modern medicine: no way of knowing whether the treatment caused the effect because there’s no counterfactual.

- Now we take a more rigorous approach. Through randomized control trials, we can get good data about what works and why.

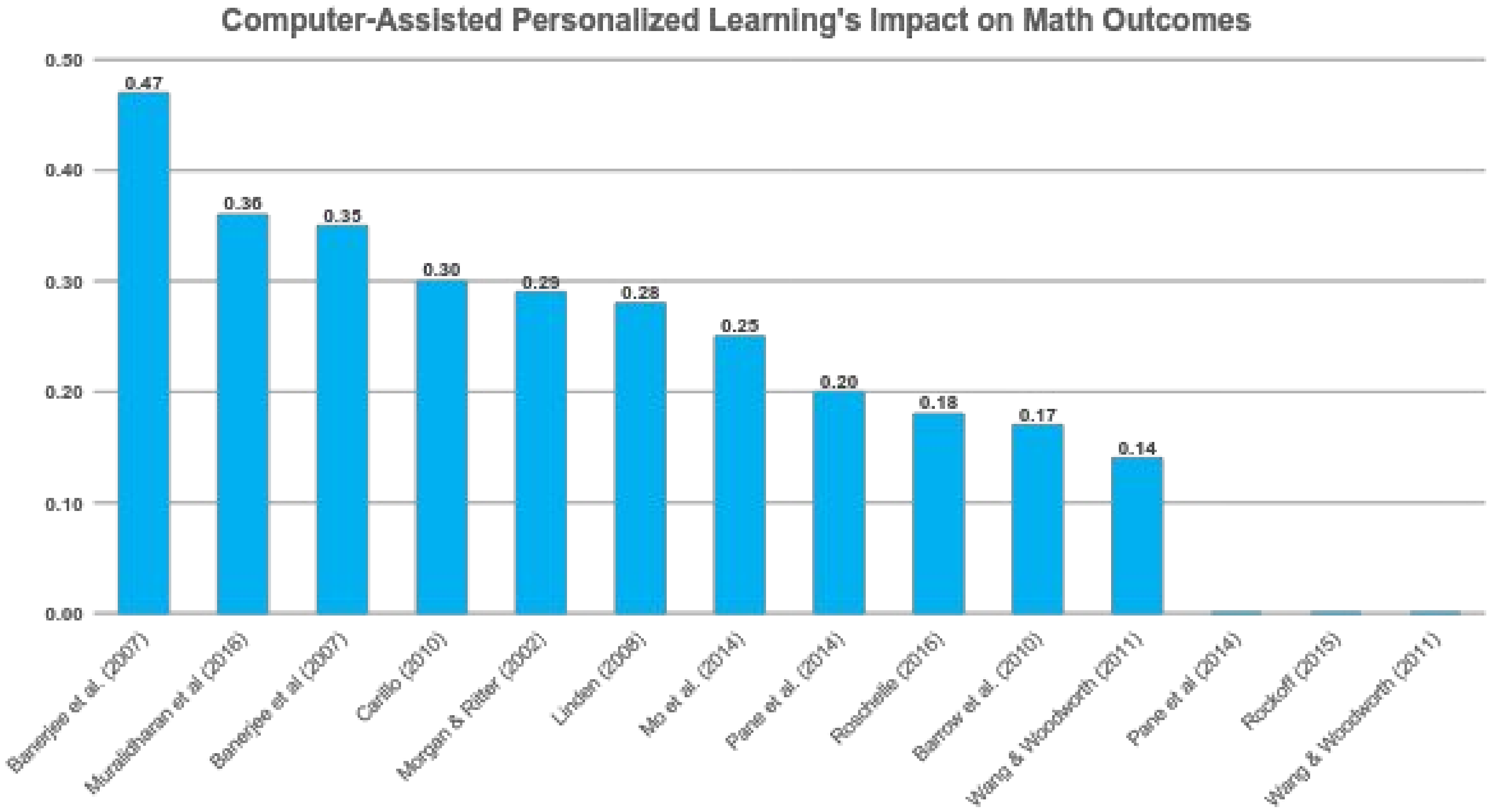
- Silver bullets are rare. Sometimes there are true breakthroughs, but most progress is made by examining particular problems, learning over time.

Photo Credit: Triin Erg

What do we know from the RCT evidence on Personalized Learning?

- **Promising evidence of effectiveness on learning**
  - Computer-assisted personalized learning leads to consistently positive impacts especially when used as a complement
  - One study finds a **0.57 SD decrease** when the program is used as a substitute, but a **0.28 SD increase** when used as a complement (Linden 2008)
- **Math interventions seem especially successful**
  - 11 studies showing positive effect and only 2 studies showing no effects
- **Evidence for language is more mixed**
  - 4 studies showing positive effect and 4 studies showing no effects

Positive Impacts on Math



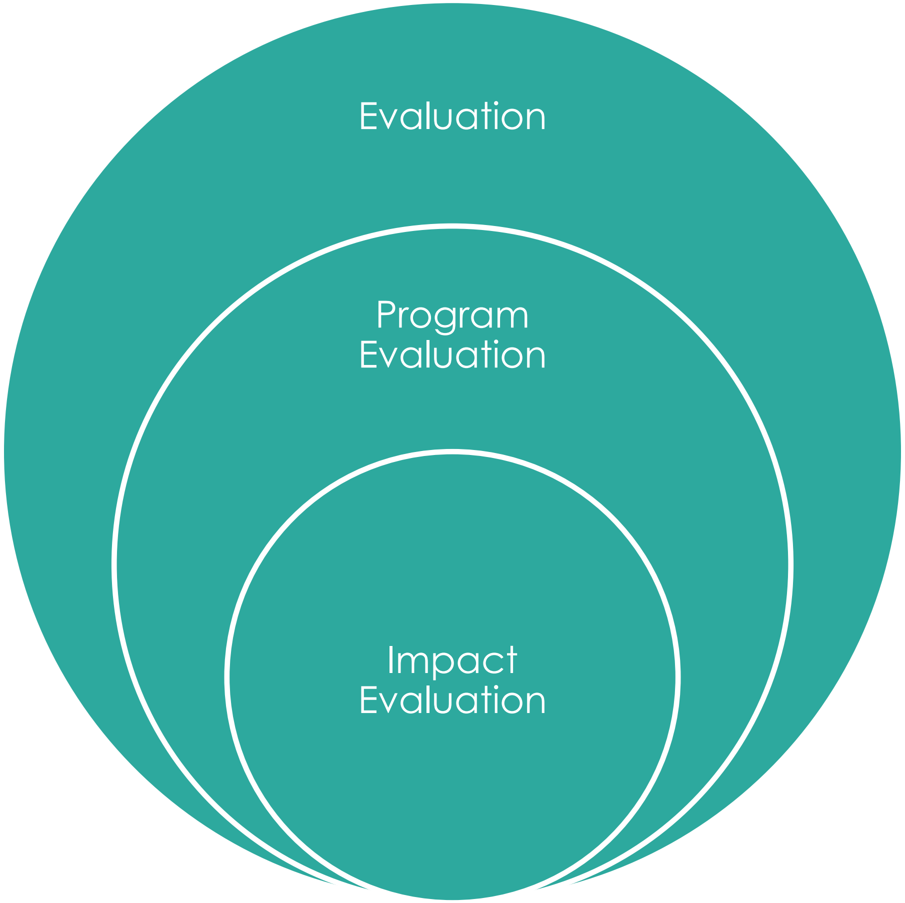
The need for more rigorous evaluation

- **Potentially ineffective untested programs and approaches**
  - Reliance on educational software that lack evidence
  - Technology-based tools designed without sufficient grounding in the needs of parents, students, and teachers
- **A need to understand mechanisms, context, and generalizability**
  - Rollout and implementation
  - Quality of substitutes (e.g., the quality of instruction that a software module is replacing)
- **Relatively low costs and high potential benefits for ed-tech and personalized learning evaluations**
  - Once a platform is established, costs of scale-up frequently approach zero
  - Ed-Tech platforms often support built-in data collection

Lecture overview

- Overview of Ed-Tech, Personalized Learning & Research
- What is Evaluation?
- Measuring Impact
- Randomized Evaluations
  - Different ways to randomize
  - Opportunities to randomize
  - Pitfalls to watch for
- Ethics of randomization

What is evaluation?

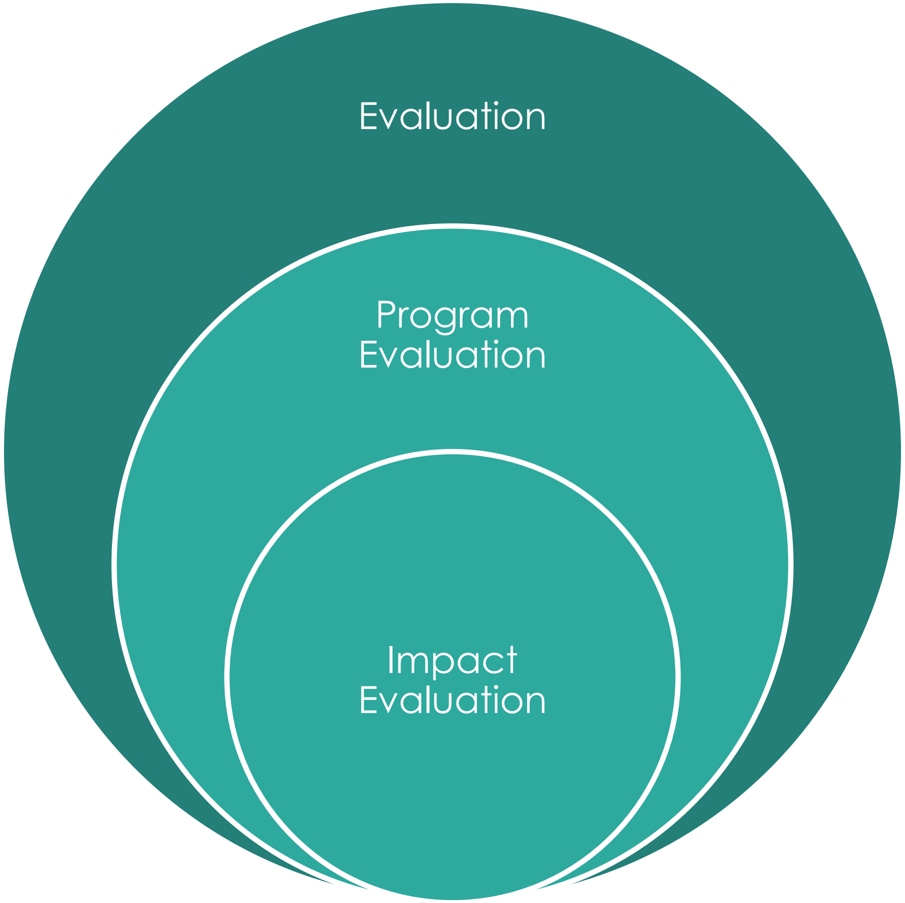


Evaluation

Program Evaluation

Impact Evaluation

Program evaluation



Evaluation

Program Evaluation

Impact Evaluation

Components of program evaluation

Needs Assessment

Theory of Change

Process Evaluation

Impact Evaluation

Cost-Effectiveness Analysis

Other quasi-experimental methods

Methodology	
Pre-Post (Before-and-after)	Measure how the same program participants improved (or changed) over time
Simple Difference	Measure the difference between program participants and non-participants after the program is completed.
Difference in Differences	Measure the before-and-after change in outcomes for the program participants, then subtract the before-and-after change in outcomes of the non-participants
Multiple Linear Regression	Compare participants to non-participants, and estimate the effects of the program by controlling for observed characteristics
Statistical Matching	Individuals who received a program are compared to similar individuals who did not receive it.
Regression Discontinuity Design	Compare similar individuals right above and right below a cutoff (e.g. SAT score of 600, GPA of 3.3)
Instrumental Variables	Individuals who, because of this "instrumental" factor, are predicted not to participate and (possibly as a result) did not participate.
Randomized Evaluation	Random assignment (e.g. a coin toss or random number generator) determines who may participate in the program so that those assigned to participate in the program are, on average, the same as those who are not

Lecture overview

- Overview of Ed-Tech, Personalized Learning & Research
- What is Evaluation?
- Measuring Impact
- Randomized Evaluations
  - Different ways to randomize
  - Opportunities to randomize
  - Pitfalls to watch for
- Ethics of randomization

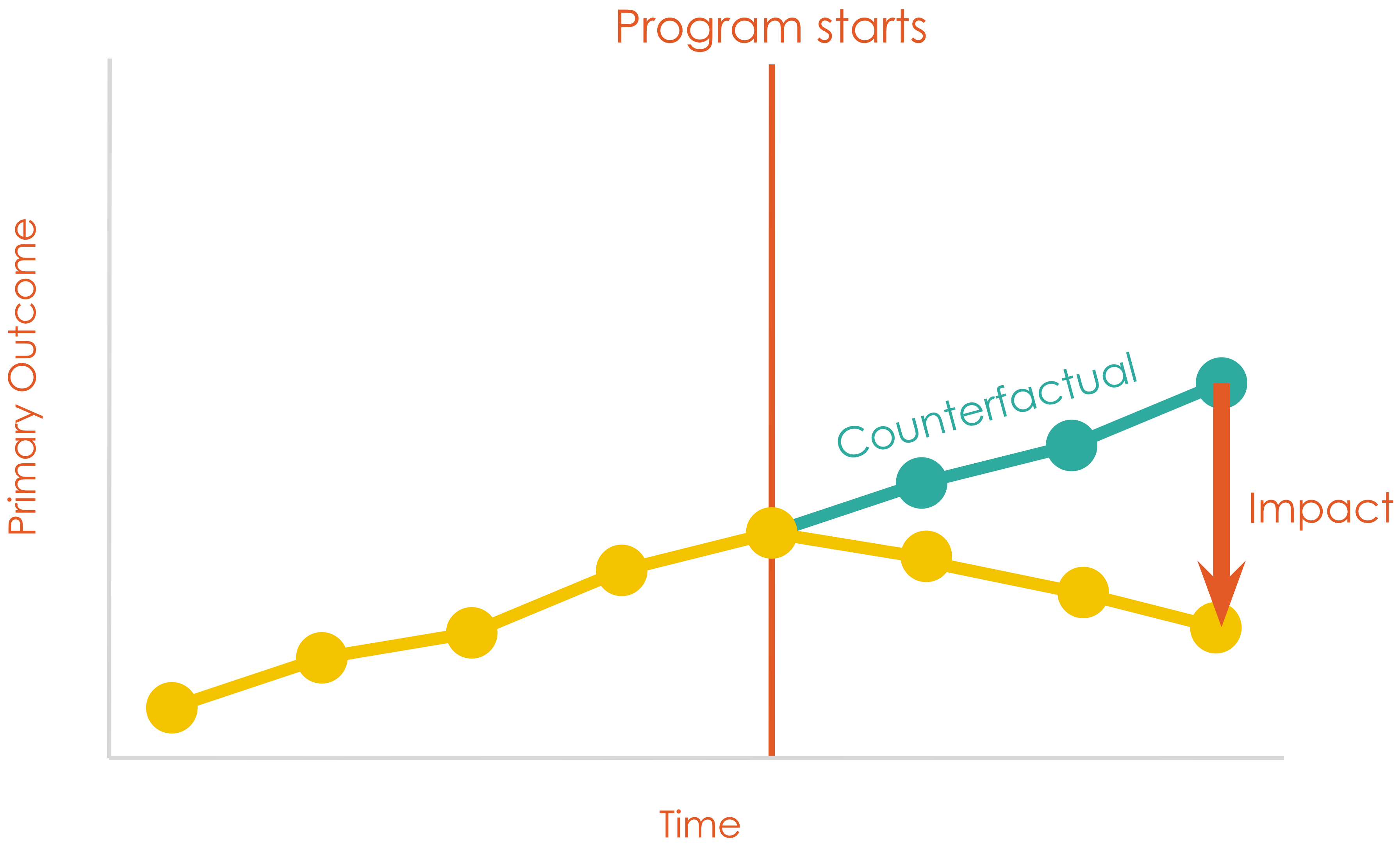
RCTs and Measuring Impact

Impact is defined as a comparison between:

What actually happened and

What would have happened, had the program not been introduced (i.e., the "counterfactual")

What is the impact of this program?



Counterfactual

The counterfactual represents what would have happened to program participants in the absence of the program

Problem: Counterfactual cannot be observed

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

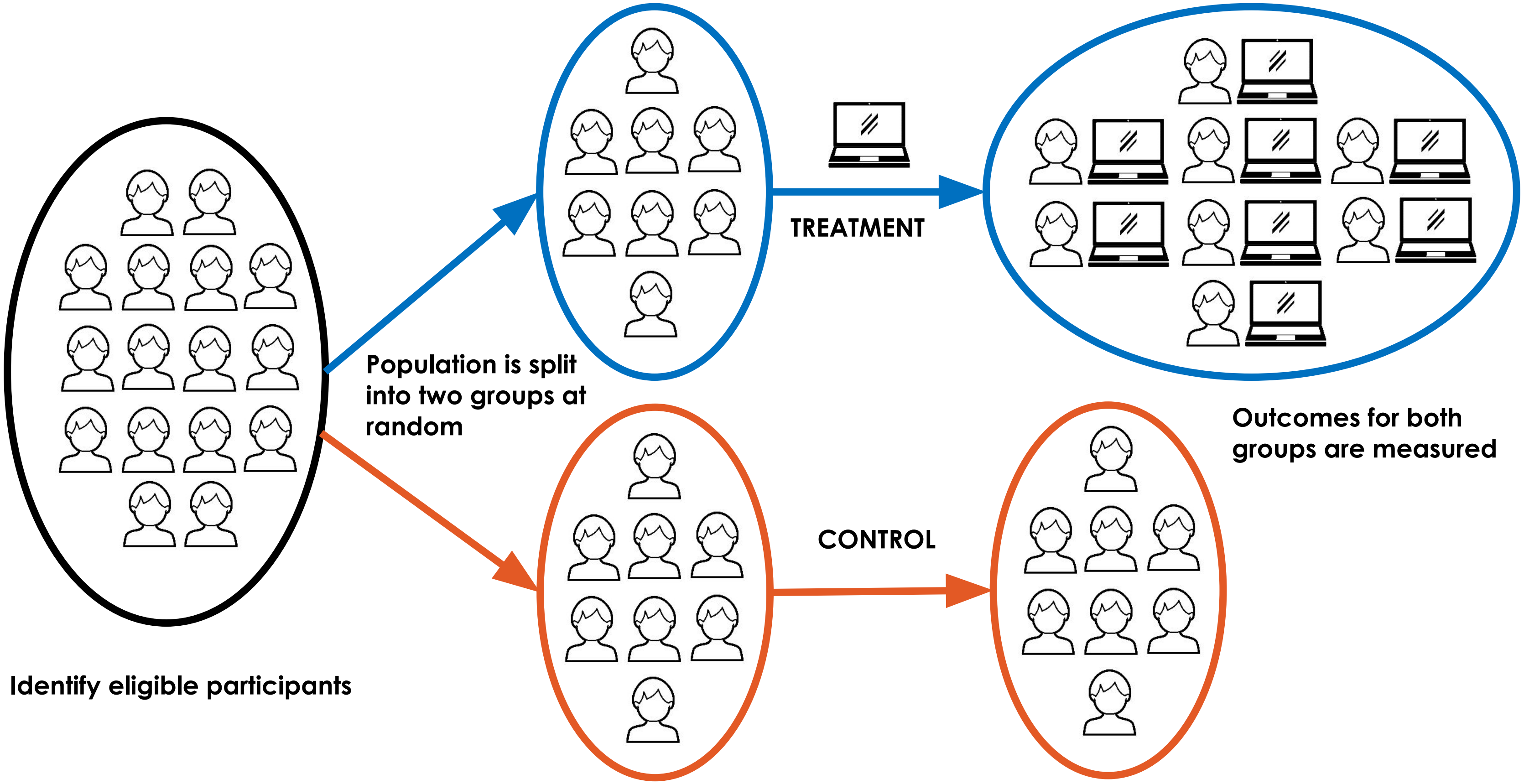
The critical objective of impact evaluation is to establish a credible comparison group.

Randomized Control Trials (RCTs) work by mimicking a comparison group as close to the counterfactual as possible through randomization.

Lecture overview

- Overview of EdTech & Research
- What is Evaluation?
- Measuring Impact
- Randomized Evaluations
  - Different ways to randomize
  - Opportunities to randomize
  - Pitfalls to watch for
- Ethics of randomization

Randomized Evaluations (RCTs)



Population is split into two groups at random

Outcomes for both groups are measured

Identify eligible participants

The RCT Game: The Candy Experiment

- Theory of change
- Generating the list
- Consent (asking first!)
- Baseline (optional)
- Randomization
- Treatment
- Process evaluation
- Endline

Selecting the comparison group

Idea: Select a group that is exactly like the group of participants in all ways except one—their exposure to the program being evaluated

Goal: To be able to attribute differences in outcomes to the program (and not to other factors)

Different ways to randomize

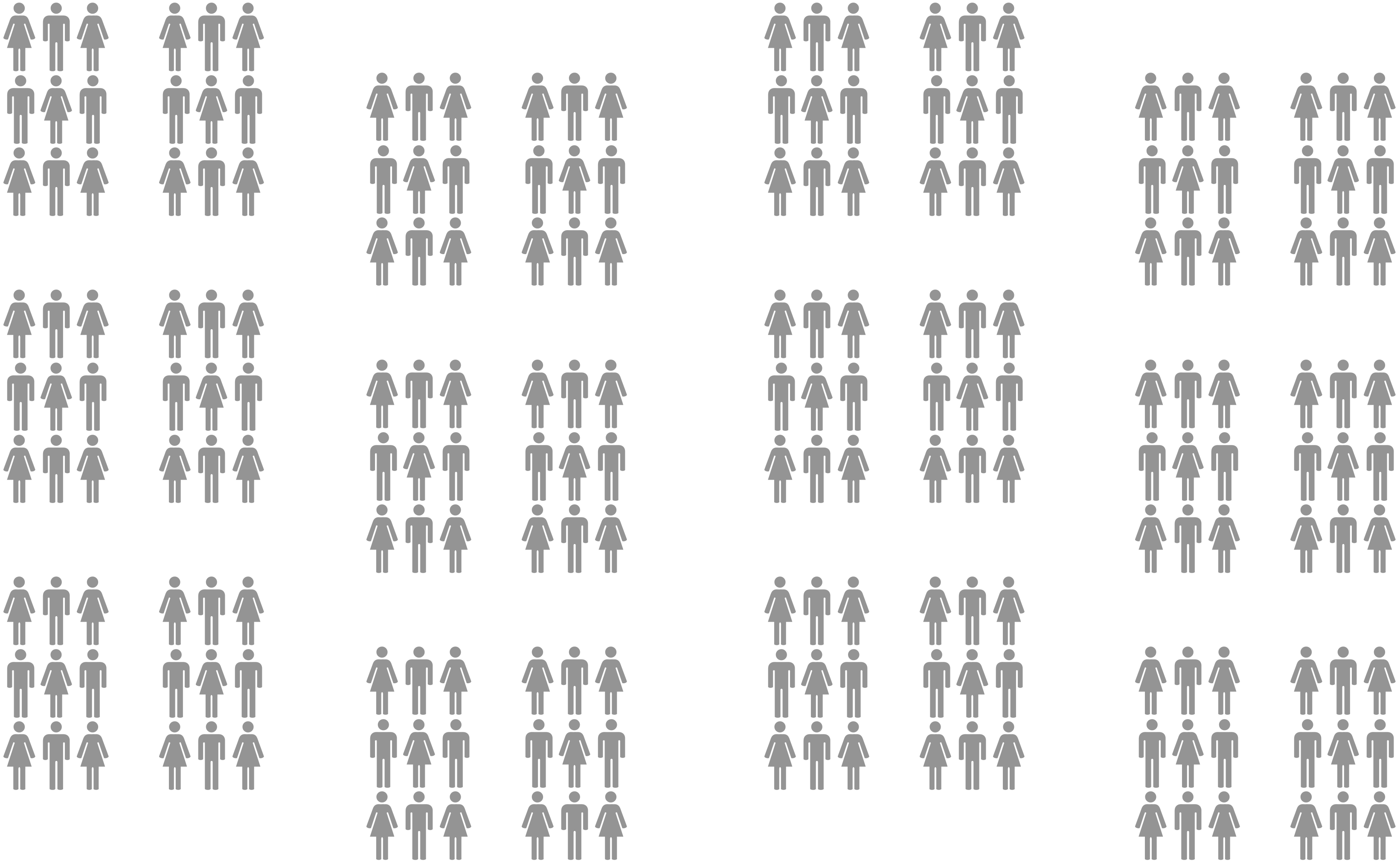
- Different **units of randomization**: individual students vs. "cluster" (e.g. classrooms, schools, districts)

- Randomizing **access**: we can choose which people are offered access to a program
  - Simple lottery
  - Randomizing "in the bubble"

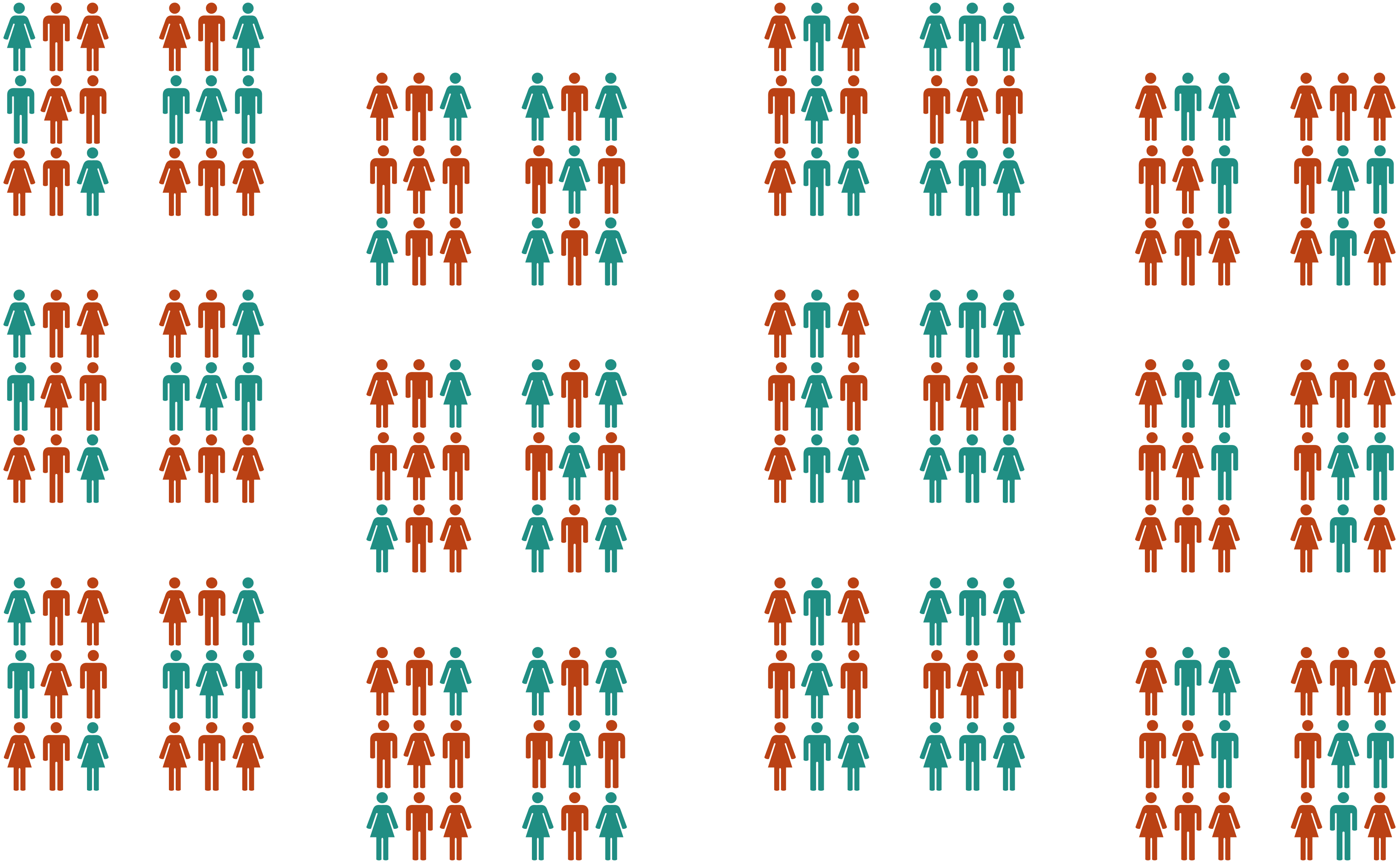
- Randomizing **timing**: we can choose when people are offered access
  - Phase-in design

- Randomizing **encouragement**: we can choose which people are encouraged to participate in a program
  - Encouragement design

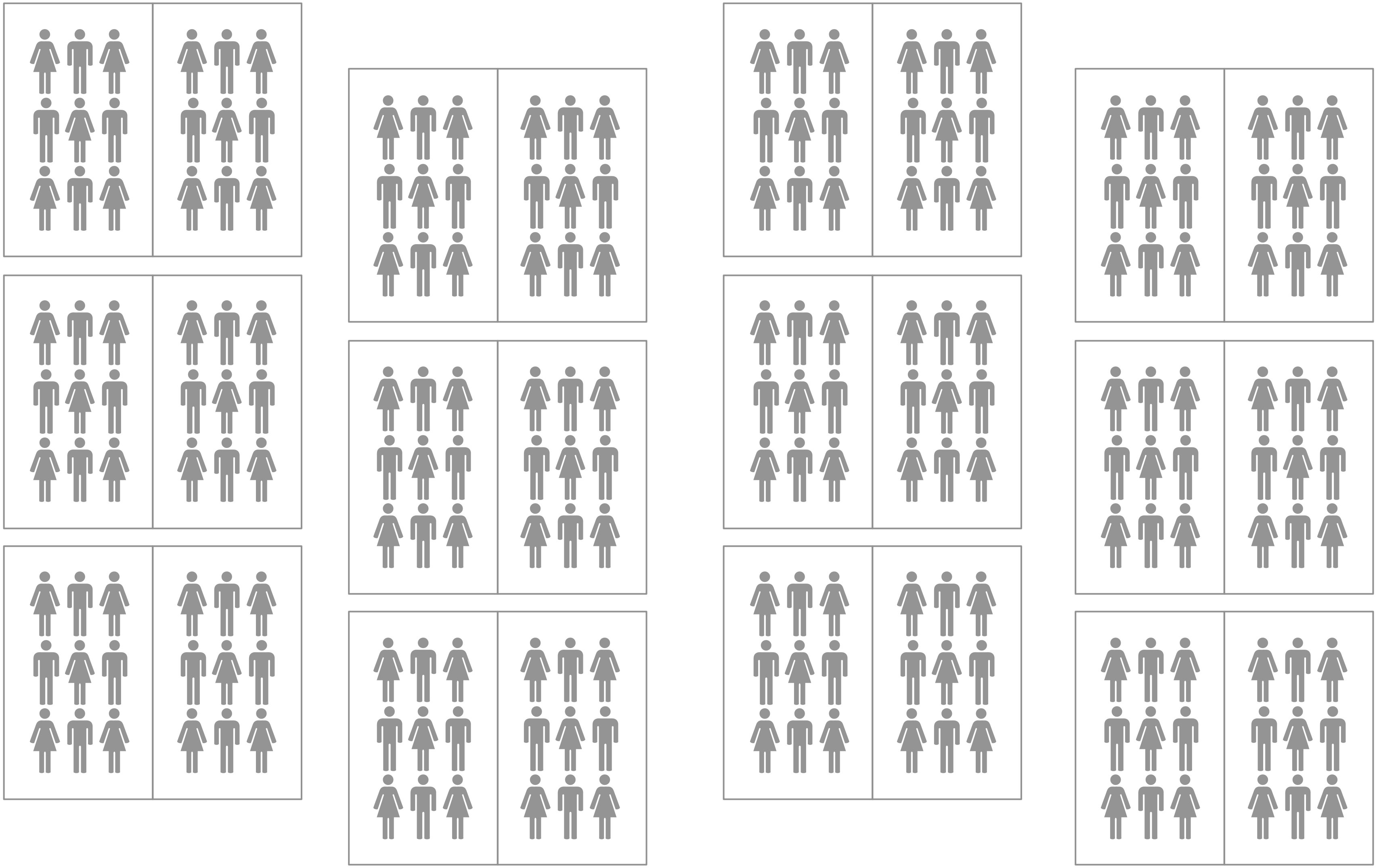
Unit of randomization: Individual Students



Unit of Randomization: Individual Students



Unit of Randomization: Classroom



Unit of randomization: Classroom

How to choose the unit of randomization

- Nature of the intervention
  - Generally, best to randomize at the level at which the program is administered (e.g. individual students, entire classrooms, entire schools etc).
- How wide is the potential impact?
- What level of data is available?
- Sample size and power requirements

Randomizing access: Simple lottery

- Individuals, communities, schools, etc. (units) are randomized to receive access (or not) to the program

- Optimal when:
  - Program is being piloted
  - The program is oversubscribed, there are limited resources

- Advantages:
  - Simple to administer and explain

- Disadvantages:
  - The control group never gets the program
  - Hard to evaluate entitlement programs where everyone who is eligible is entitled to access by law

Randomizing access: "In the bubble"

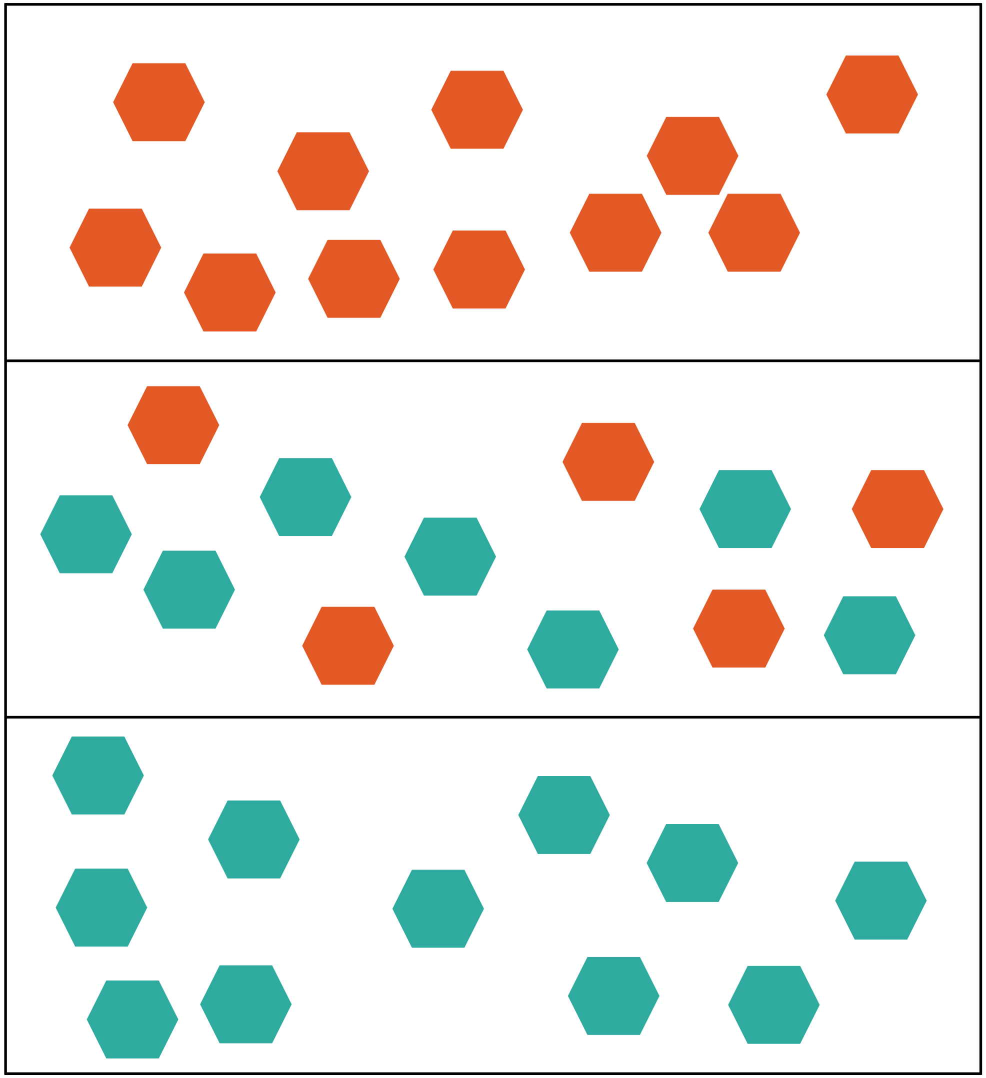
- Individuals or groups are scored on some eligibility criteria
  - High scores all admitted, low scorers not admitted
  - Those with intermediate scores randomized into or out of program
- Optimal when:
  - Clear eligibility criteria
  - The program is oversubscribed, there are limited resources
- Advantages:
  - Program keeps lot of control over who is admitted
  - Answers the question: "should we expand this program?"
- Disadvantages:
  - Does not measure the impact of program on the average participant
  - Some less eligible people admitted instead of those more eligible

Randomization "in the bubble"
SAT Scores Example

Participants
(scores > 700)

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

Non-participants
(scores <500)



Randomizing timing: Phase-in

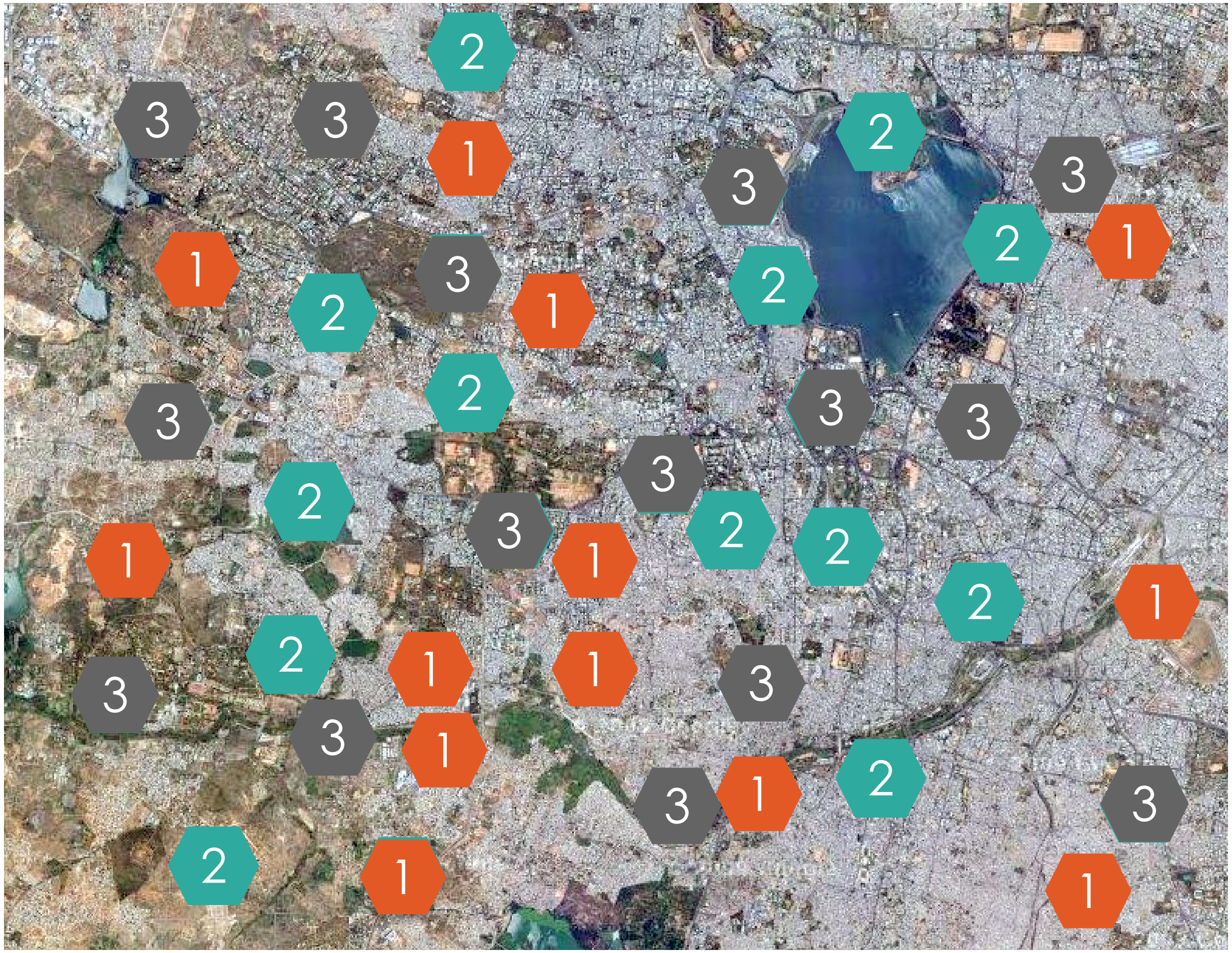
- Individuals or groups are randomly phased into program over time
- Optimal when:
  - Capacity constraints mean cannot roll out everywhere at once
- Advantage:
  - Everyone receives the program eventually
- Disadvantage:
  - Only in special situations can you measure long run effects, as control group disappears in long run

Phase-in design:
A personalized learning program

Round 1
Treatment: 1
Control: 2 & 3

Round 2
Treatment: 1 & 2
Control: 3

Round 3
Treatment: 1 & 2 & 3
Control: None



Some opportunities to look for

New program
New service
New people
New location
Oversubscription
Undersubscription
Admissions cutoff

Lecture overview

- Overview of Ed-Tech, Personalized Learning & Research
- What is Evaluation?
- Measuring Impact
- Randomized Evaluations
  - Different ways to randomize
  - Opportunities to randomize
  - Pitfalls to watch for
- Ethics of randomization

Sample size and statistical power

- An experiment must be sensitive enough to detect outcome differences between the treatment and the comparison groups
- The sensitivity of an experiment is measured by statistical power, which, among other factors, depends on the sample size
- The intervention should be operating on a big enough scale to be able to generate a sample size that will provide enough power for the experiment

Spillovers/crossovers

- Spillovers
  - The intervention unintentionally impacts the control group (either positively or negatively)

- Crossovers
  - Control group members get treated
  - Treatment group members don't get treated

- If control group is different from the counterfactual (what would have happened in the absence of the intervention), our results can be biased

What if you don’t have a pre-existing list?

- To randomize, we generally need to start with a list (of individuals, households, classrooms, etc.)

- If we don’t have a list beforehand, you can randomize “on the spot” like we did with the candy game

Lecture overview

- Overview of Ed-Tech, Personalized Learning & Research
- What is Evaluation?
- Measuring Impact
- Randomized Evaluation
  - Different ways to randomize
  - Opportunities to randomize
  - Pitfalls to watch for
- Ethics of randomization

Ethics of randomization

- Are fewer people being given access to the program? Or is the evaluation just changing who gets access?

- Is the evaluation changing when people have access to the program?

- How much evidence is there that the program will be a benefit?

Interactive Activity: How to Run an RCT

AdaptiveReading: Evaluating a Web-Based Personalized Tutoring Program

- Suppose you are a city Department of Education administrator who has just purchased an initial limited subscription of *AdaptiveReading,* a new popular web-based intelligent tutoring system that is designed to improve reading comprehension for fifth graders. *AdaptiveReading* is meant to be used in the classroom once a week as a supplementary tool and cannot be used at home. While skeptical of failed digital-learning platforms, 200 elementary schools are already on board with trying out *AdaptiveReading,* but want to know whether the program actually helps students read better. As the city administrator, you want to know whether *AdaptiveReading* is effective, which will help you decide whether you should renew the city's subscription.

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Questions?

Contact Vincent Quan at quanv@mit.edu

Please fill out the survey at:
bit.ly/PLSWorkshopSurvey