EGRA Research Design

Prepared for the USAID workshop
"EGRA Basics and Beyond Workshop"

November 2015

Recap of Where We Are

- **Identification of Research and Sample Design**
- **Development/Adaptation of EGRA Instrument**
- **Procedures for EGRA Administration and Scoring**
- **Establishment of Electronic Data Capture System**
- **Assessor Evaluation and Selection**
- **Pilot and Full Data Collection**
- **Use and Dissemination of EGRA Results**
- **Planning and Managing EGRA Implementation**

Session Objectives

- Review goals of monitoring, evaluation and impact evaluation
- Discuss common research designs that employ EGRA
  - Snapshot
  - Performance Evaluation
  - Experimental Designs
  - Quasi Experimental Designs

Why Assess Early Grade Reading?

The research question focuses the study, determines the methodology, and guides all stages of inquiry, analysis, and reporting.

- *What proportion of children meet a benchmark ORF score?*
- *What challenges do teachers face implementing a reading curriculum?*
- *How does a particular program impact EGRA scores?*

The research question is based on project needs.

Monitoring

Regular collection and reporting of information to track whether a program is implemented as planned

Evaluation

Analytical efforts to answer specific questions about performance of a program/activities.

Impact evaluation

Analytical efforts to relate cause and effect.

"What would have happened in the absence of the intervention"

Donor requirements also often shape the research questions.

- **Standard Indicator 3.2.1-27:** Proportion of students who, after two years of schooling, are able to read and understand grade-level text.

- **The global count:** USAID measures, worldwide, the progress its investments make in achieving the goal one target of “improved reading skills for 100 million children in the primary grades”

Monitoring

Regular collection and reporting of information to track whether a program is implemented as planned

- Periodically collect data on the indicators and compare actual results with targets
- To identify bottlenecks and red flags (time-lags, fund flows)
- Points to what should be further investigated

Evaluation

Analytical efforts to answer specific questions about performance of a program/activities.

- Analyzes why intended results were or were not achieved
- Explores targeting effectiveness and unintended results
- Provides lessons learned and recommendations for improvement
- Does not allow for attribution

Impact evaluation

Analytical efforts to relate cause and effect.

"what would have happened in the absence of the intervention"

- What is effect of program on particular outcomes?
- How would outcomes change under alternative program designs.
- Does the program impact people differently (e.g. females, poor, minorities)
- Is the program cost-effective?
- Changes in particular outcomes can be confidently attributed to the program through a counterfactual.

USAID
FROM THE AMERICAN PEOPLE

Inputs	Activities (What the program does)	Outputs	Outcomes	Long-term Results
• Teachers • Text books • Grants	• Teachers' training • School councils established • Conditional Cash transfers • Salary incentives	• Increased enrollment • Lower teachers' absenteeism • Higher oral reading fluency	• Increased student retention • Reduced drop-out • Improved learning outcomes	• Lower unemployment • Poverty reduction • Better income distribution

This logic model provides guidance on **what** to measure and **when** to measure it

USAID
FROM THE AMERICAN PEOPLE

Inputs	Activities (What the program does)	Outputs	Outcomes	Long-term Results
• Teachers • Text books • Grants	• Teachers' training • School councils established • Conditional Cash transfers • Salary incentives	• Increased enrollment • Lower teachers' absenteeism • Higher oral reading fluency	• Increased student retention • Reduced drop-out • Improved learning outcomes	• Lower unemployment • Poverty reduction • Better income distribution

Indicators related to the implementation of a program – focus of monitoring and evaluation

USAID
FROM THE AMERICAN PEOPLE

Inputs	Activities (What the program does)	Outputs	Outcomes	Long-term Results
• Teachers • Text books • Grants	• Teachers' training • School councils established • Conditional Cash transfers • Salary incentives	• Increased enrollment • Lower teachers' absenteeism • Higher oral reading fluency	• Increased student retention • Reduced drop-out • Improved learning outcomes	• Lower unemployment • Poverty reduction • Better income distribution

Indicators related to the results of a program – focus of impact evaluation

Research Question(s) Guide Evaluation Design

- Monitoring and Evaluation
  - Snapshot assessment
  - Performance evaluation
- Impact Evaluation
  - Experimental methods
  - Quasi-experimental methods
- Measuring towards USAID standard indicators and for the global count

Non-experimental Designs

- Focus on descriptive questions
- A snapshot seeks to provide a view of a particular indicator, such as oral reading fluency, at one point in time or over time.
  - How are targeted students performing in relation to nationwide benchmarks?
- A performance evaluation attempts to understand how a program may contribute to outcomes
  - Which EGRA pre-reading task is most highly correlated with oral reading fluency?
  - Was the program implemented as intended?

Impact Evaluation Designs: Experimental

- Focus on questions about cause and effect
- Experimental Designs (RCTs)
  - At least two groups (in the simplest case, an experimental and a comparison group)
  - Random assignment to the groups
  - Change in the independent variable (program activity) before assessment of impact
- By design, treatment and comparison have the same characteristics (observed and unobserved), on average
- Yields unbiased impact estimates

Opportunities for Randomization

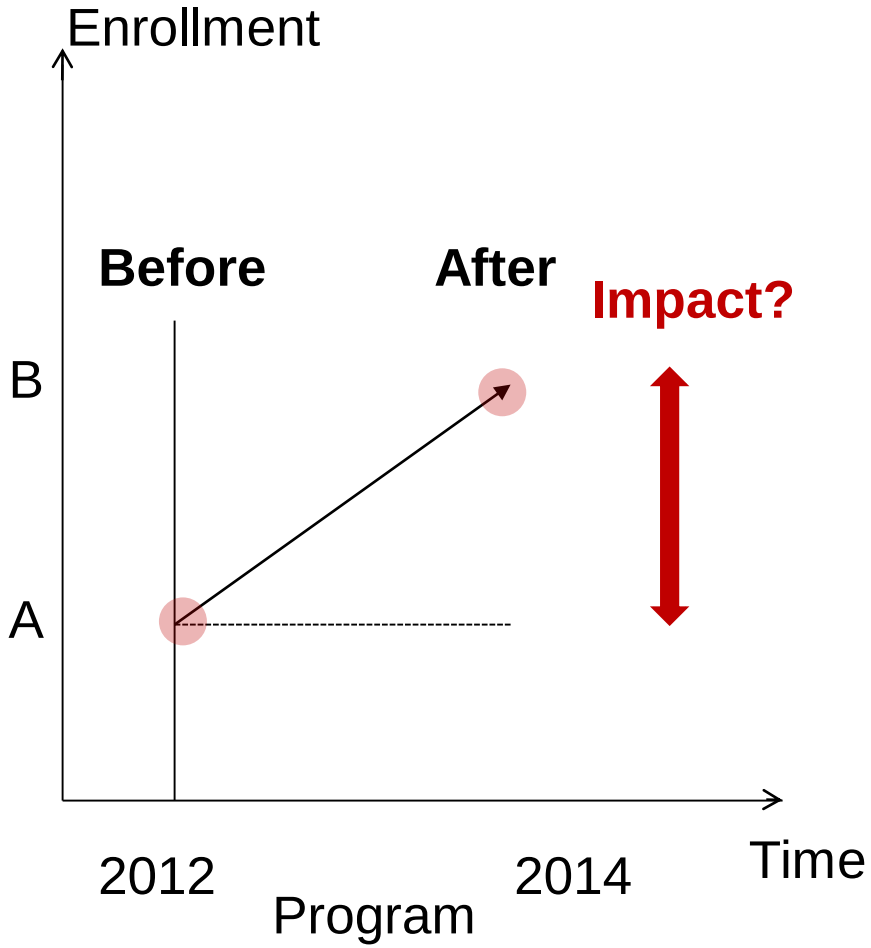
- Budget constraint prevents full coverage
  - Random assignment (lottery) is fair and transparent
- Limited implementation capacity
  - Random phase-in (everyone receives eventually)
- No evidence on which alternative is best
  - Variation in treatment (everyone receives something)
- Take up of existing program is not complete
  - Encouragement design (provide information or incentive for some to sign up)

Impact Evaluation Designs: Quasi-experimental

- Quasi-experimental designs
  - Pre-post or interrupted time series
  - Matched comparison group
  - Regression discontinuity
- Quasi-experimental methods can be informative, but conclusions depend on the quality of comparisons
  - Not all comparisons are equal
  - Some comparisons provide information, BUT may hide confounding factors

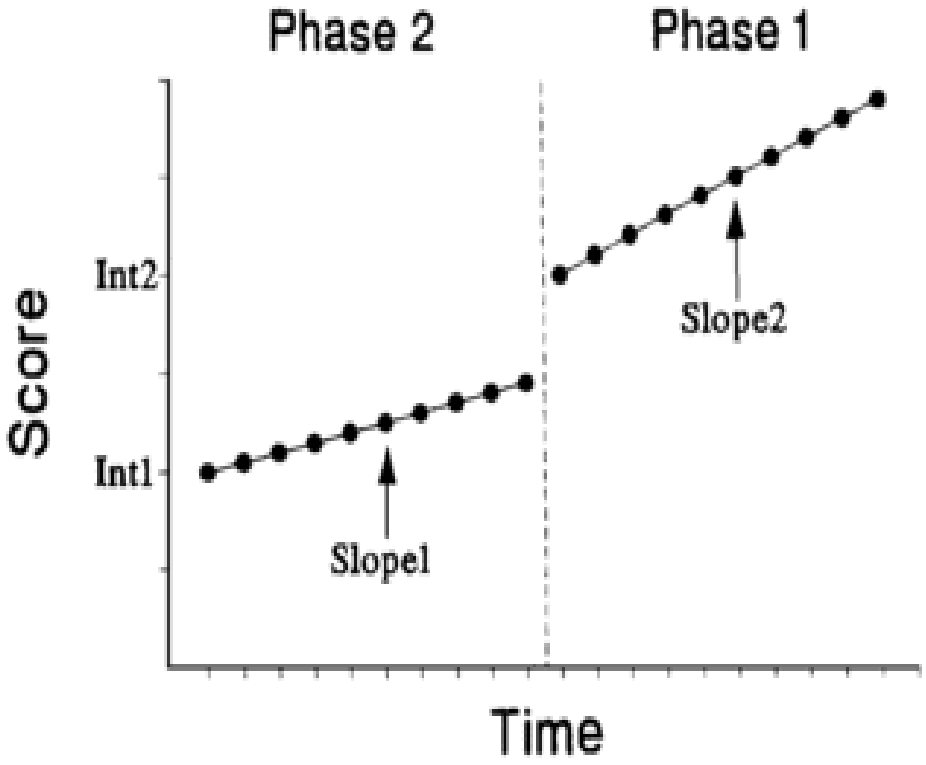
Quasi-experimental: Pre-Post Design

- Observe only the treatment group
- We don't know what would have happened in the absence of the intervention



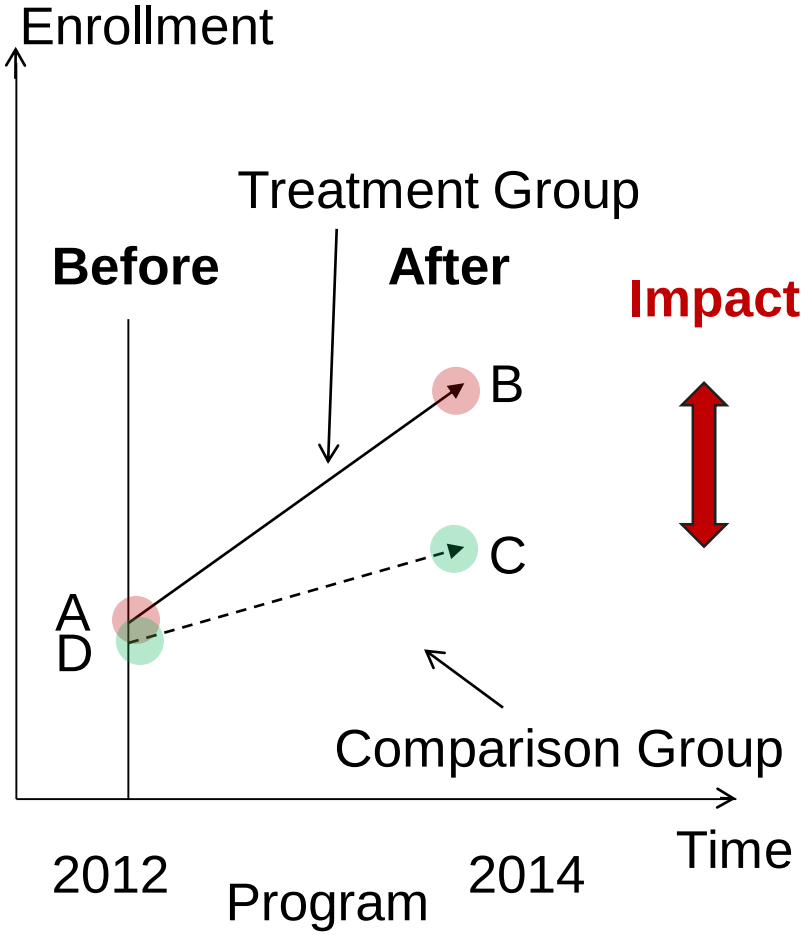
Quasi-experimental: Interrupted Time Series Design

- Observe treatment group at multiple points in time before and after intervention
- Multiple measures allow us to rule out many confounding variables
- The key problem is the effect of external events not related to the intervention

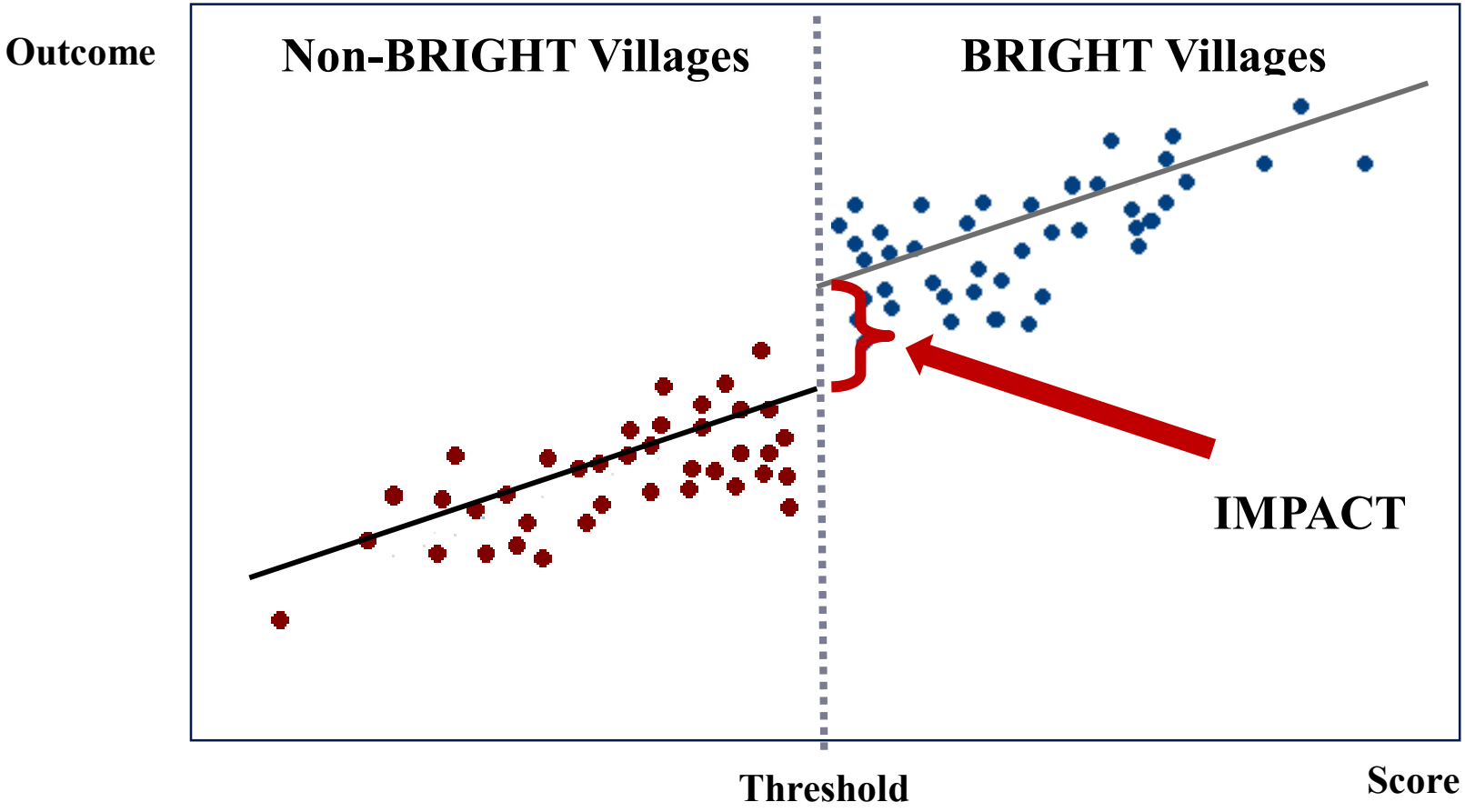


Quasi-experimental: Matched Comparison Group Design

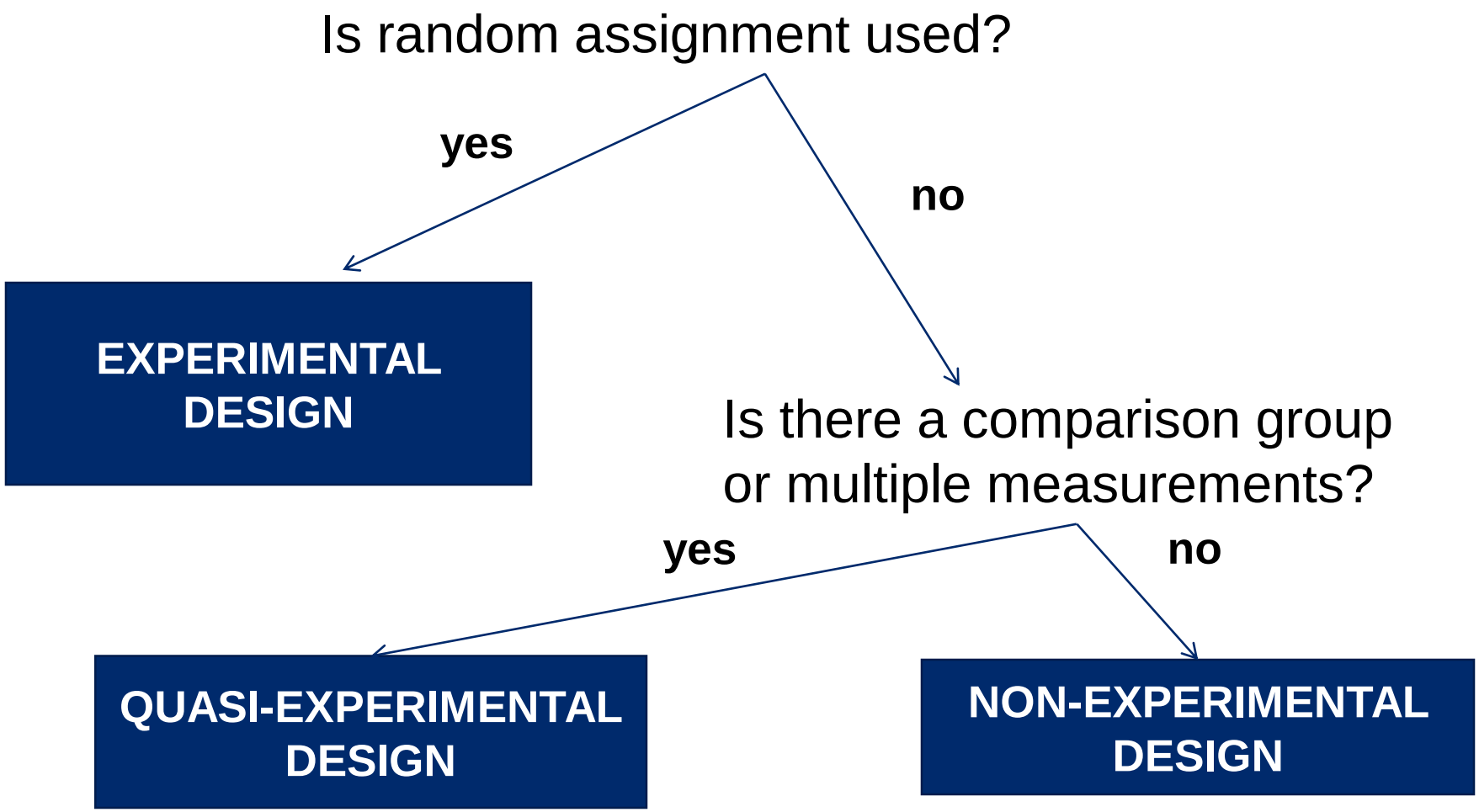
- Use baseline data to match groups
- Because it is not possible to account for every characteristic, there is less confidence that the groups are similar.



Quasi-experimental: Regression Discontinuity



Types of Designs



W. Trochim, *Research Methods Knowledge Base*, 2006

Other considerations

- Budget
- Feasibility
- Usefulness
- Time

EGRA Sample Design

Session Objectives

- Understand best practices of research design for pilot and full data collection, including:
  - Principles of sampling for EGRA
  - Practical implications of different types of research design

What is a sample?

Who do you want to generalize to?

The Theoretical Population

What population can you get access to?

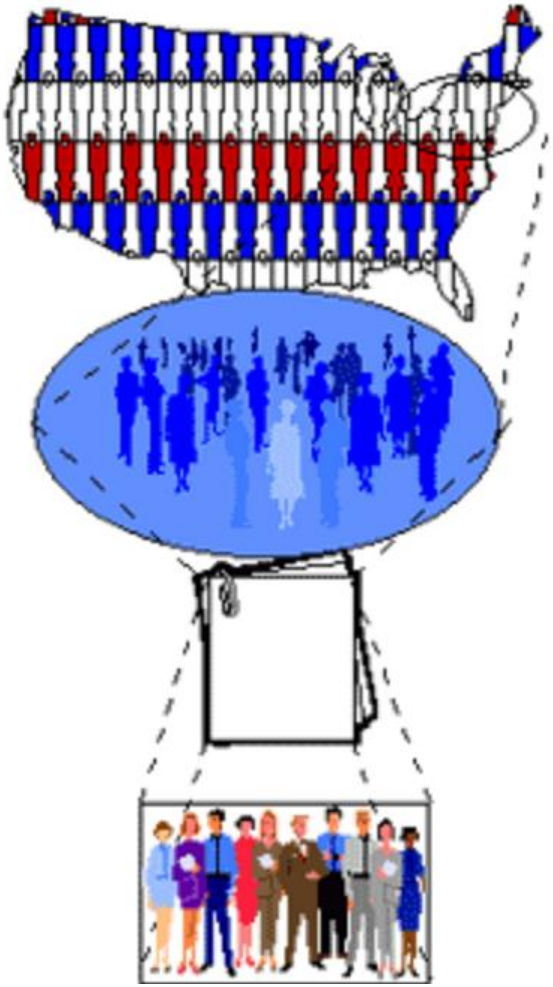
The Study Population

How can you get access to them?

The Sampling Frame

Who is in your study?

The Sample



*W. Trochim, *Research Methods Knowledge Base*, 2006

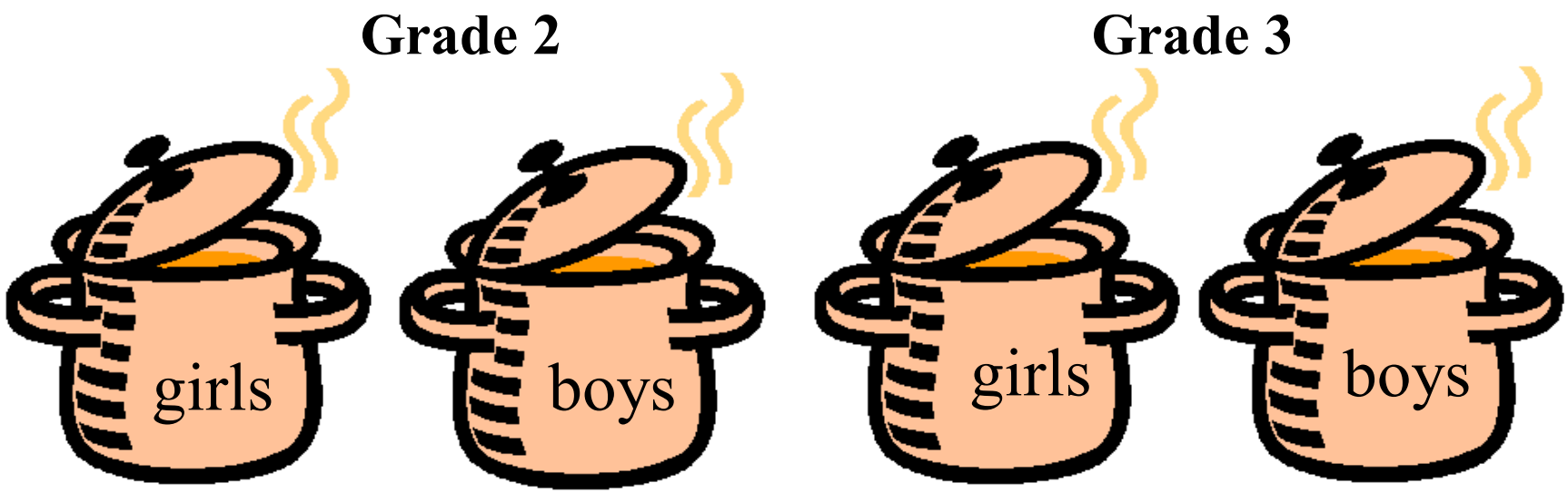
If life is like eating a box of chocolates…. sampling is like ladling from a pot of soup.

You don’t have to drink the whole pot of soup to know what it tastes like, but you do have to stir it well to make sure you get a good taste. In other words, you need to sample properly to ensure you get an accurate picture.

For example, if you want to know children's reading abilities in Grade 2 and Grade 3, you need to "ladle from two pots of soup"—i.e., sample children from each class.

Grade 2 **Grade 3**

What if you want to know how girls in grade 2 compare with boys in grade 2? And what about boys versus girls in grade 3?



Considerations for Sampling

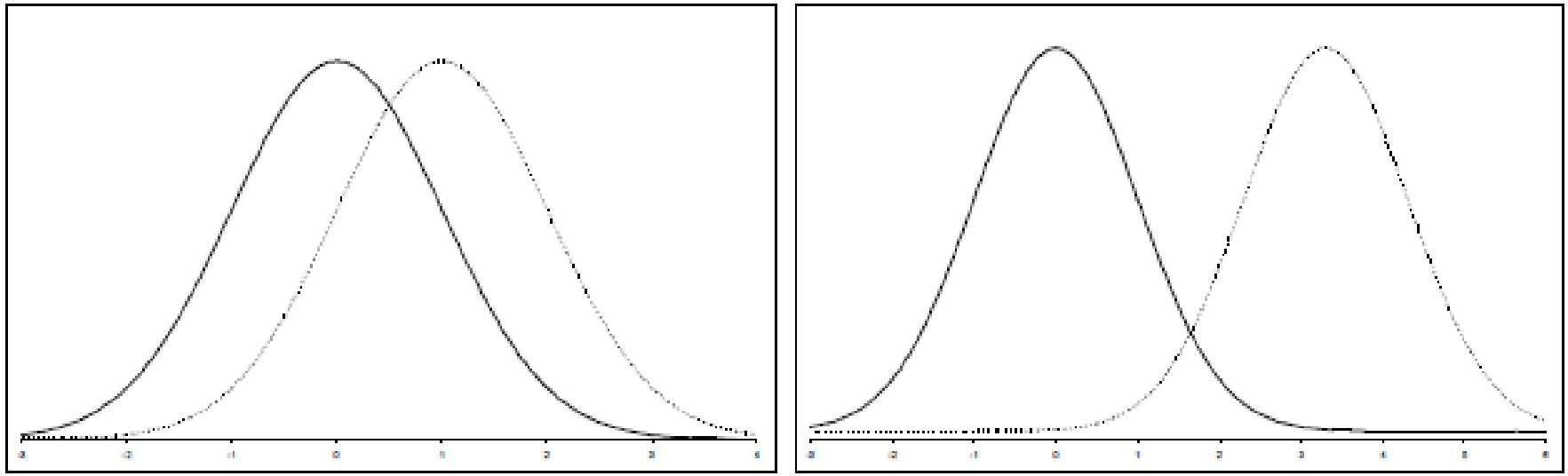
- Cluster versus non-cluster
- Type of data analysis
- Alpha
- Power (1 – beta)
- Effect size

$\Big\}$ => sample cell size

FINAL SAMPLE = Sample cell size * number of disaggregation categories

Effect Size

$$\text{Effect Size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{Standard Deviation}}$$



(a) (b)

Source: Cole, R. 2002. *It's the Effect Size, Stupid*.
http://www.leeds.ac.uk/educol/documents/00002182.htm

Cohen's Standard	Effect Size	Percentile Standing	Percent of Nonoverlap
LARGE	0.8	79	47.4%
	0.7	76	43.0%
	0.6	73	38.2%
MEDIUM	0.5	69	33.0%
	0.4	66	27.4%
	0.3	62	21.3%
SMALL	0.2	58	14.7%
	0.1	54	7.7%
	0.0	50	0%

Source: UCCS Dr. Lee A.Becker http://www.uccs.edu/lbecker/effect-size.html

Calculating the Size of a Sample Cell: Non-clustered

Analysis	t-test (difference between the means)
Effect size	.25 (18% non-overlap)
Alpha (type I error)	.05 (one measure)
Power (1 - beta)	80%
Sample size	**101 individuals**

Effect size of .25 => about 18% non-overlap in normal populations

Calculating the Size of a Sample Cell: Non-clustered

Analysis	t-test (difference between the means)
Effect size	.25 (18% non-overlap)
Alpha (type I error)	**.005 (ten measures)**
Power (1 - beta)	80%
Sample size	**191 individuals**

Effect size of .25 => about 18% non-overlap in normal populations

Calculating the Size of a Sample Cell: Non-clustered

Analysis	t-test (difference between the means)
Effect size	**.15 (10% non-overlap)**
Alpha (type I error)	.005 (ten measures)
Power (1 - beta)	80%
Sample size	**523 individuals**

Effect size of .15 => about 10% non-overlap in normal populations

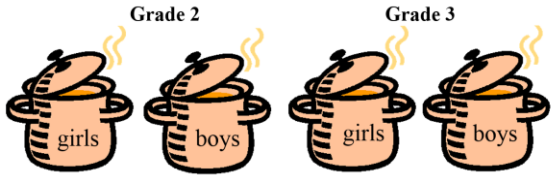
Calculating the Size of a Sample Cell: Non-clustered

Analysis	**regression**
Effect size	.15 (10% non-overlap)
Alpha (type I error)	.005 (ten measures)
Power (1 - beta)	80%
Sample size	**1103 individuals**

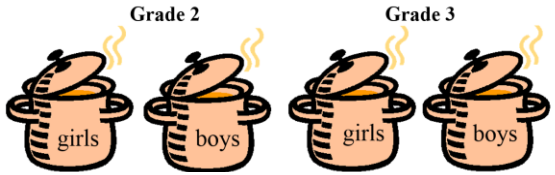
Effect size of .15 => about 10% non-overlap in normal populations

Disaggregation in Sampling

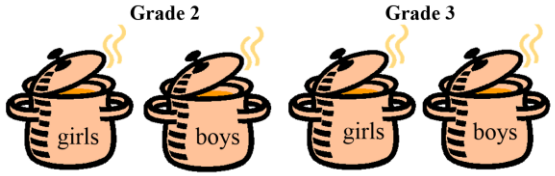
- Province A



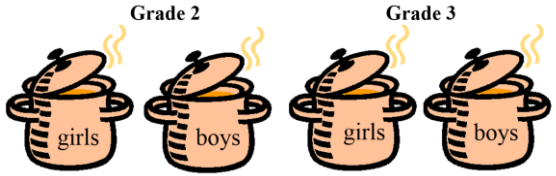
- Province B



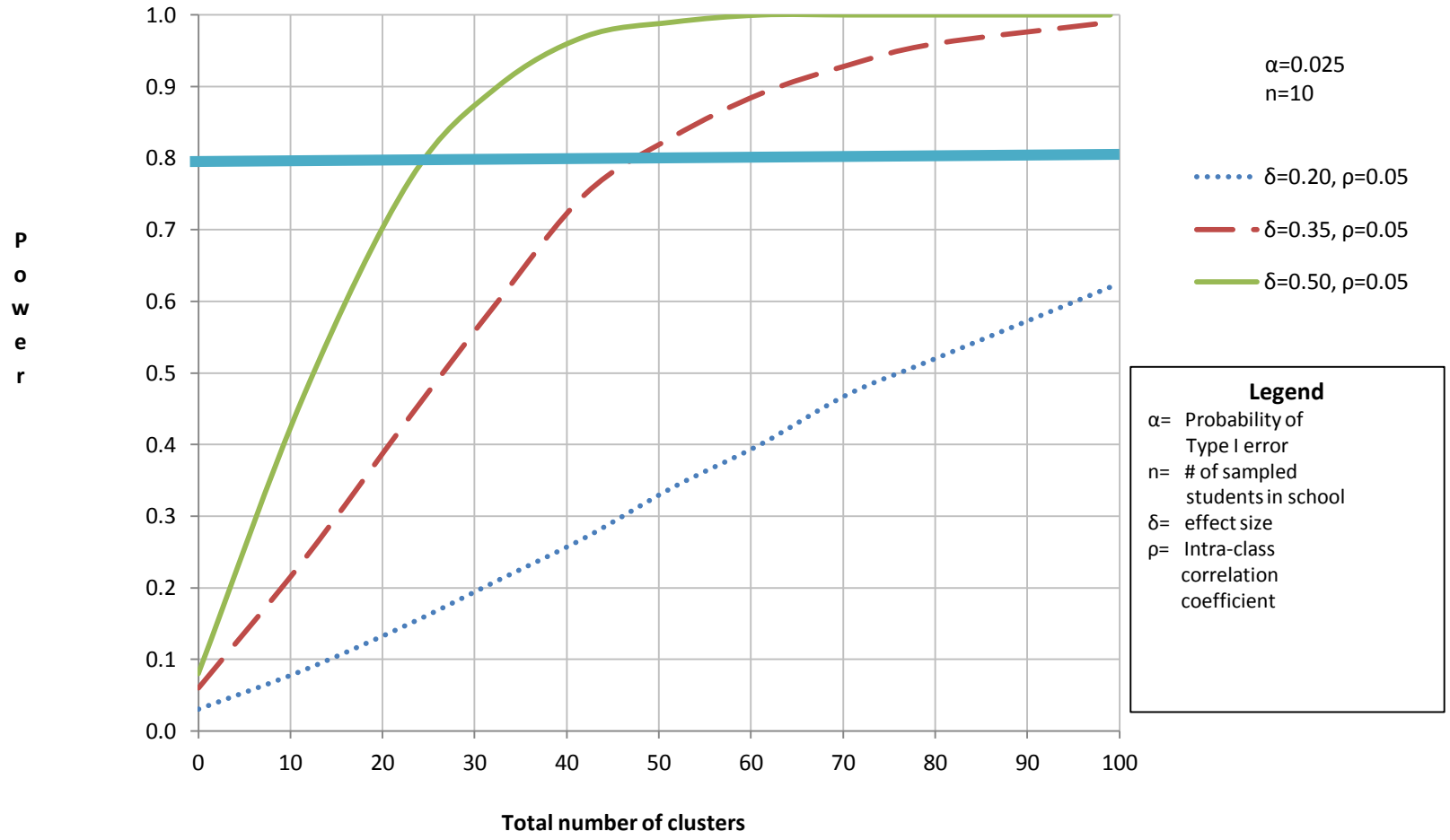
- Province C



- Province D



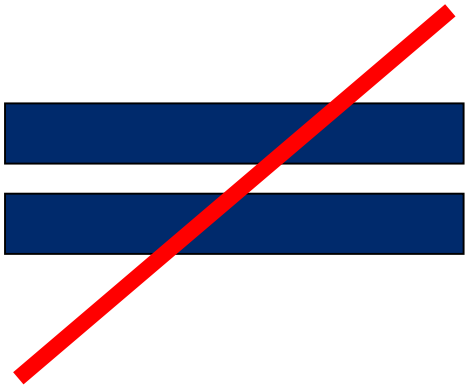
Calculating Size of a Sample Cell: Clustered



Adding rho (intra-class correlation coefficient)

Calculating Size of a Sample Cell: Clustered

Assessing 10 students in 100 schools



as assessing 20 students in 50 schools

Students tend to vary more between schools than within, so sampling more schools is usually preferable to sampling more students within the same school to obtain a target sample size.

Sampling: Stratification

	Grade 2	Grade 3	Total
Public schools	400	400	800
Private schools	400	400	800
Non-formal schools	400	400	800
Total	1,200	1,200	**2,400 pupils**

Weights are used in data analysis to compensate for disproportionate representation of different ***types*** of clusters.

Best Practice Guidance: Sample Size Per Cell

- Recommendation/Best Practice: 400 pupils per cell is a good starting point (i.e. reporting statistically significant results).
- Actual cell size can vary, depending on the level of variance in student performance, and desired precision.
- EGRA data available in multiple countries and can be used to calculate precise cell size.

Additional Considerations

- Amount of time available to collect data
- Distance to cover to reach schools (may need to cluster sample)
- Number of assessors you can hire
- Amount of equipment you can purchase
- BUDGET available

Best Practice: Engage with the Ministry of Education to obtain necessary information about schools (e.g., location, number of teachers per grade, number of students per grade, and schedule).

Select a Sampling Method Based on Your Research Design

Census Sample

- *Sample all individuals in a population*
  - Provides the most comprehensive information
  - Most studies do not have the resources to use this method

Convenience Sample

- *Sample a subset of individuals based on a factor that is convenient (e.g., location or prior relationship)*
  - By definition, method is convenient
  - Risks introducing bias into the sample

Random Sample

- *Sample is selected by a random method*
  - Is the most rigorous method since it minimizes sampling bias
  - Size of sample is determined based on various factors

General EGRA Sampling Guidance

- There are **no rules of thumb**. The ***purpose*** of assessment will drive both research design and sampling.
- Sampling design is a ***compromise*** between the desire to increase precision and budgetary considerations.
- Increase in a sample size improves ***accuracy*** of population estimates.
- Expected ***effect size*** has the largest impact on sample size
- When possible, engage a statistician with good knowledge of research objectives for computing a sample

Refer to EGRA Toolkit annex for more detailed sampling guidance.

Sampling for Different Research Designs: Cross-sectional Design

- Consists of one or more samples drawn from the population(s) at one point in time.
- Allows researchers to describe characteristics of the population(s) and to identify correlations between characteristics at that point in time.

Example:

EGRA administered once to a random sample of students in a nationally representative sample of schools in order to get a snapshot of student performance at that time.

Brainstorming Activity:
What can we learn from a cross-sectional EGRA study? What can we use data for? What are the implications for sampling?

Sampling for Different Research Designs: Experimental and Quasi-experimental Designs

- What changes do we see ***as a result*** of the intervention?
- Does the intervention affect groups of students differently?
- What moderating/mediating factors affect results?

Example: EGRA administered to a random sample of students in a representative sample of schools at baseline and to another random sample of students at endline in order to monitor change in overall performance over time as the result of an intervention.

Brainstorming Activity:
What can we learn from repeated cross-sectional EGRA studies? What can we use data for? What are the implications for sampling?

Research Design and Sampling: Practical Takeaways

- Research design structures
  - EGRA administration timing
  - EGRA administration intervals
  - Sample design
  - Type of data analysis
  - What we we learn
- Every design has limitations
- There are no “good” or “bad” designs but there are more appropriate or less appropriate designs for the ***purpose***
- Budgets will impact designs

Exercise: Research Design and Sampling Activity

Handout 2.1: Participant Sampling Activity

Handout 2.1: Participant Sampling Activity

Handout 2.1: Participant Sampling Activity

Session 2: EGRA Research Design and Sampling

Your organization has been asked to conduct an EGRA in the country of Readistan. Various stakeholders are interested in learning the following:

- Reading levels for all 6 grades of public government schools
- Information as to whether children in private schools are doing better than children in public schools
- Information on how children in all 5 regions of the country are doing
- Insight into how children are performing in 3 languages

Discuss the following:

1. What is the purpose of the EGRA activity based on the information desired?
2. What are the research questions?
3. What will the sample size be based on the information desired by different stakeholders? How many pupils would need to be sampled and assessed? (assume 100 schools of each type in each region, 10 students per grade per school)

Questions to When Determining EGRA Design

- What is the main ***purpose*** of the study? What other purposes will data be used for?
- What are the ***limitations*** of the selected design?
- Does the proposed EGRA study description clearly explain the purpose, the process and the limitations?
- What ***effect size*** was used in sampling calculations? What does it mean in terms of comparisons between groups?
- What ***disaggregation*** categories were used to create a sample?