SESRI Policy & Program Evaluation Workshop

Doha, Qatar

January 19-22, 2015

Outline: Session 3

- Review of Randomized Control Trials (RCTs)
- Evaluation Design Concerns:
  - Selection Bias
    - Omitted Variable Bias
    - Spurious Relationships
    - Alternative Explanations
- Internal / external validity
- Introduction to Quasi-Experimental Designs:
  - Pre-post Comparisons with Control Group
  - Interrupted Time Series
  - Matching using Propensity Score

Recap: RCTs

- Powerful research design that relies on random assignment into treatment and control groups to create a compelling counterfactual.
- Members of treatment and control group have equal likelihood of receiving the treatment.
- Creates groups that are "equal in expectation."
- Differences in outcomes can be attributed to the effect of the treatment
- Allows strong causal inference.

What problems can random assignment solve?

- Randomization corrects for specification error, such as selection bias. This includes:
  - Omitted variable bias
  - Spurious relationships
  - Alternative explanations

Selection Bias

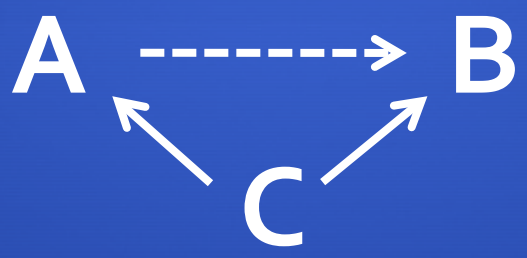
- Individuals who choose to participate in a program may be different in important ways from individuals who do not participate.
- Back to our financial literacy program example: does the program/treatment cause people to save more?
  - Without random assignment, people who are more likely to save may be more likely to participate in the program, so it may seem like the program leads them to save, while in fact they were more likely to do so in any case.
    - Here we are confounding the effect of individuals' initial propensity to save with their participation in the program
  - With random assignment, people who are more likely to save are equally expected to be represented in treatment and control groups, so we can estimate the true effect of the program.

Types of Selection Bias

- Omitted Variable Bias
  - Attributing the outcome effect to the program or policy when in fact it is caused by some other factor that is not accounted for in the design or model
- Spurious Relationships
  - Spurious relationships occur if we identify a relationship between A and B when in reality a third variable, C, is influencing both of them.
- Alternative Explanations
  - Any reason that individuals in the treatment group are different from the control group and therefore we cannot attribute the effects solely to the program or intervention

Example

- If we think a financial literacy intervention (A) affects savings behavior (B), we are interested in the causal relationship between A and B.
- It may be, however, that age (C) affects both participation in the intervention and savings behavior. Perhaps older individuals are more likely to save AND participate in the intervention.



- In this case, we might incorrectly estimate the program effect.

Correcting for Selection Bias

- With randomization, we can account for alternative explanations for the causal relationship between A—>B because the treatment and control group are *equal in expectation* of receiving the treatment
- Without randomization, we can correct for selection bias by including statistical controls in our models (e.g., controlling for potential confounding variables such as age, gender, prior savings behavior, etc.)

Internal vs. External Validity

- Internal Validity
  - The extent to which the researcher can establish without a doubt that A → B
  - RCTs tend to prioritize internal validity
- External Validity
  - The extent to which a causal relationship holds over variations in people, contexts, treatments and outcomes
  - Also referred to as *generalizability*
- Neither is "good" or "bad" – it depends on what the needs of the program and its stakeholders are, including the evaluator
- Tension between internal and external validity is always something the researcher / evaluator must contend with

Threats to Internal Validity: A Checklist

- Ambiguous Temporal Precedence
  - Does A precede B?
- Selection
  - Are treatment and control groups systematically different except receipt of intervention?
- History
  - Did something else occur simultaneously with program implementation that might affect outcomes?
- Maturation
  - Might the outcome have occurred without program due to natural changes?
- Regression
  - Are the participants extreme cases or outliers?
- Attrition
  - Did participants drop out of the study in a systematic way?
- Testing
  - Was the same test used for pre- and post measurement?
- Instrumentation
  - Are the measure used (e.g., survey, interview protocol, test) appropriate?
- Additive and Interactive Effects of Threats to Internal Validity
  - Is something else happening that complicates the relationship between the cause and effect?

Quasi-experimental Research Designs

- Random assignment of treatment is extremely rare
  - Often researcher is asked to evaluate program after program has already been implemented, so researcher cannot control assignment into treatment
  - In this case, the evaluator may have to use existing data to try to approximate causal inference
- How can an evaluator establish a counterfactual in the absence of random assignment?
- Use of quasi-experimental designs can provide leverage to isolate treatment effects.

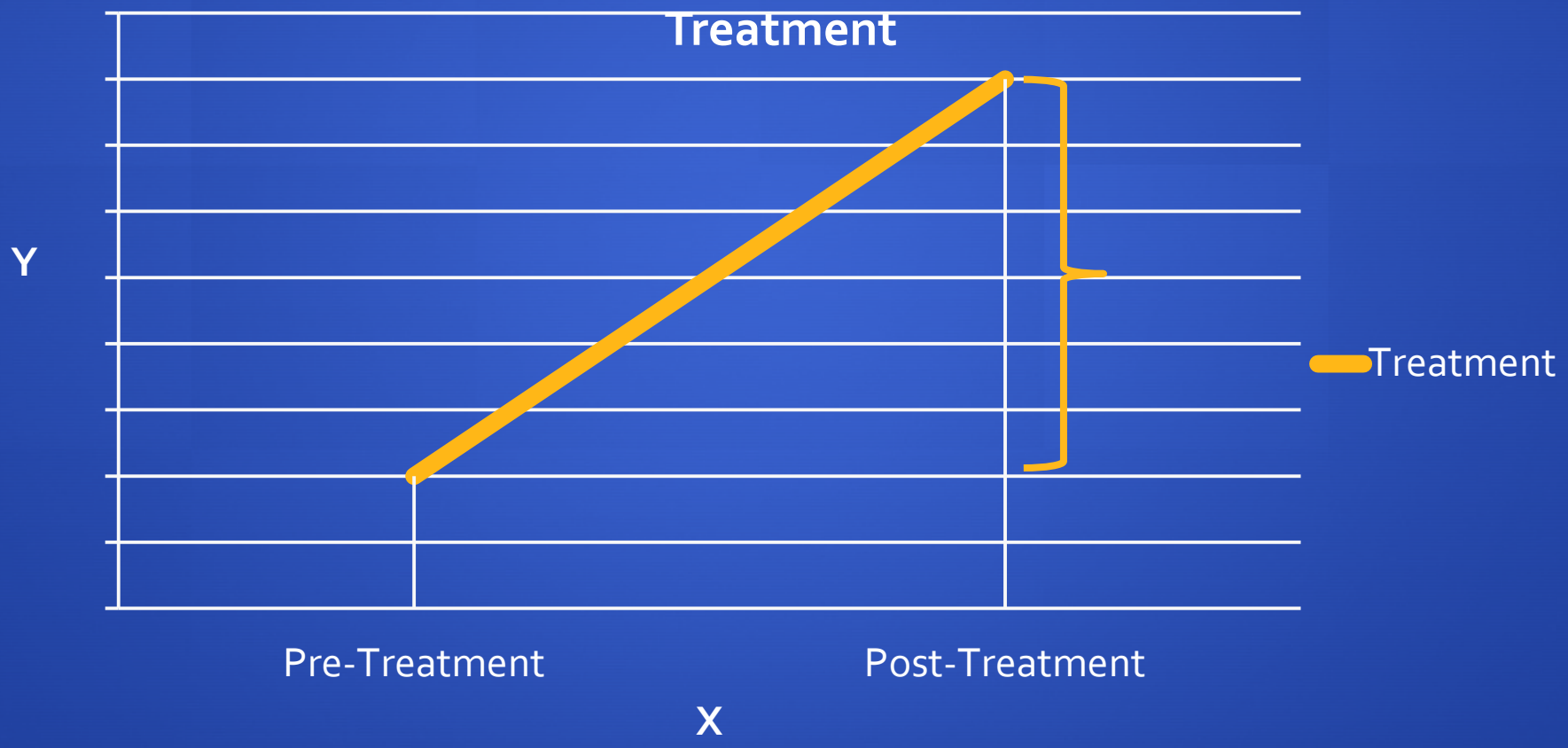
Rigorous Quasi-Experimental Designs

- Numerous quasi-experimental designs
  - Vary depending on how the control or comparison group is constructed, what comparisons are made
- We will focus on three common quasi-experimental designs that do not require the evaluator to administer the treatment: pre-post comparison with control group, interrupted time series, and matching.

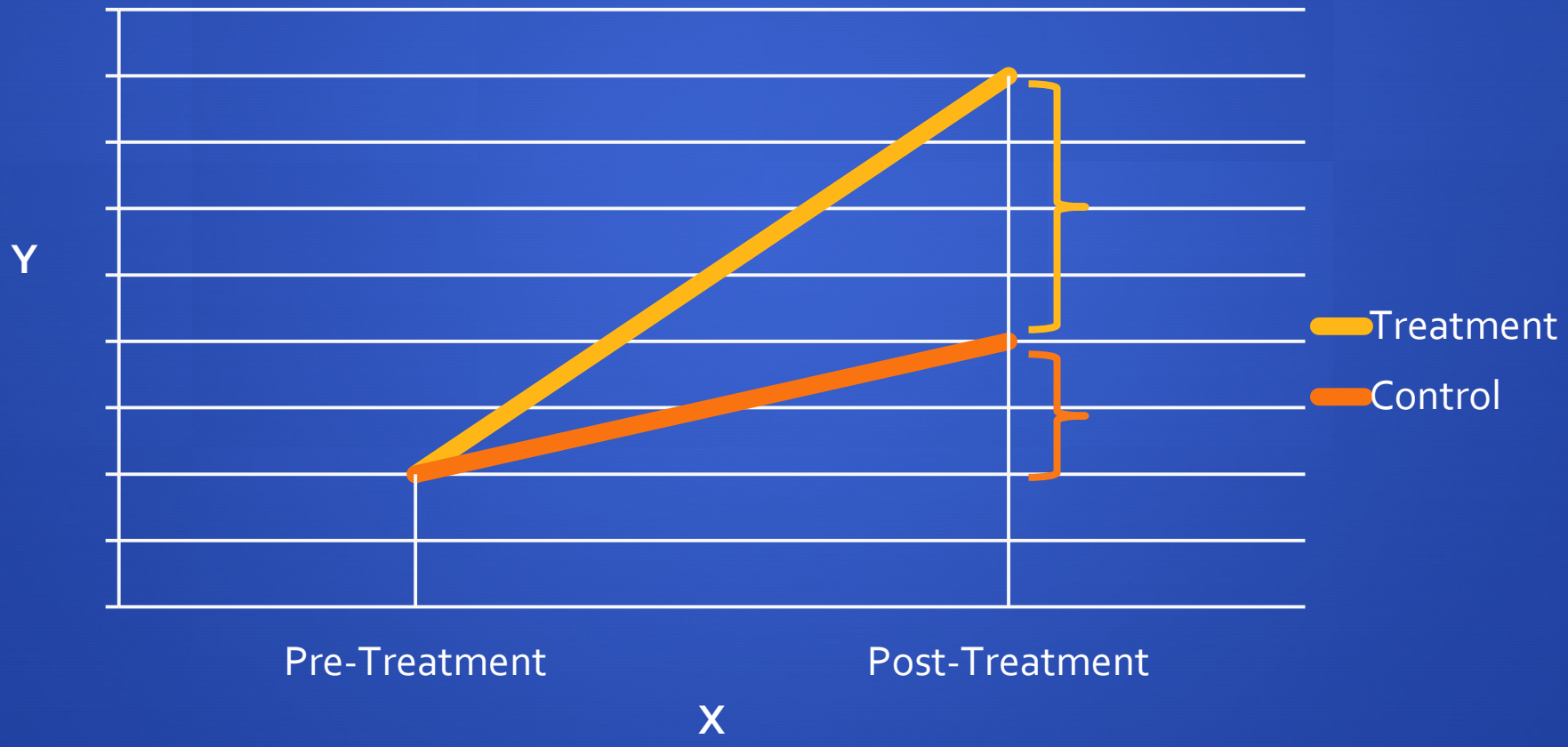
Pre-post comparison with control group

- Two-group design, each subject has two data points.
- Estimated effect is the difference in differences.
- Advantages over pre-post with no control group
  - May capture omitted variables, alternative explanations that affect both groups
- Limitations
  - Limited measures of outcomes
  - Difficulties in finding an appropriate control group
  - Threats to validity: maturity, seasonality, history

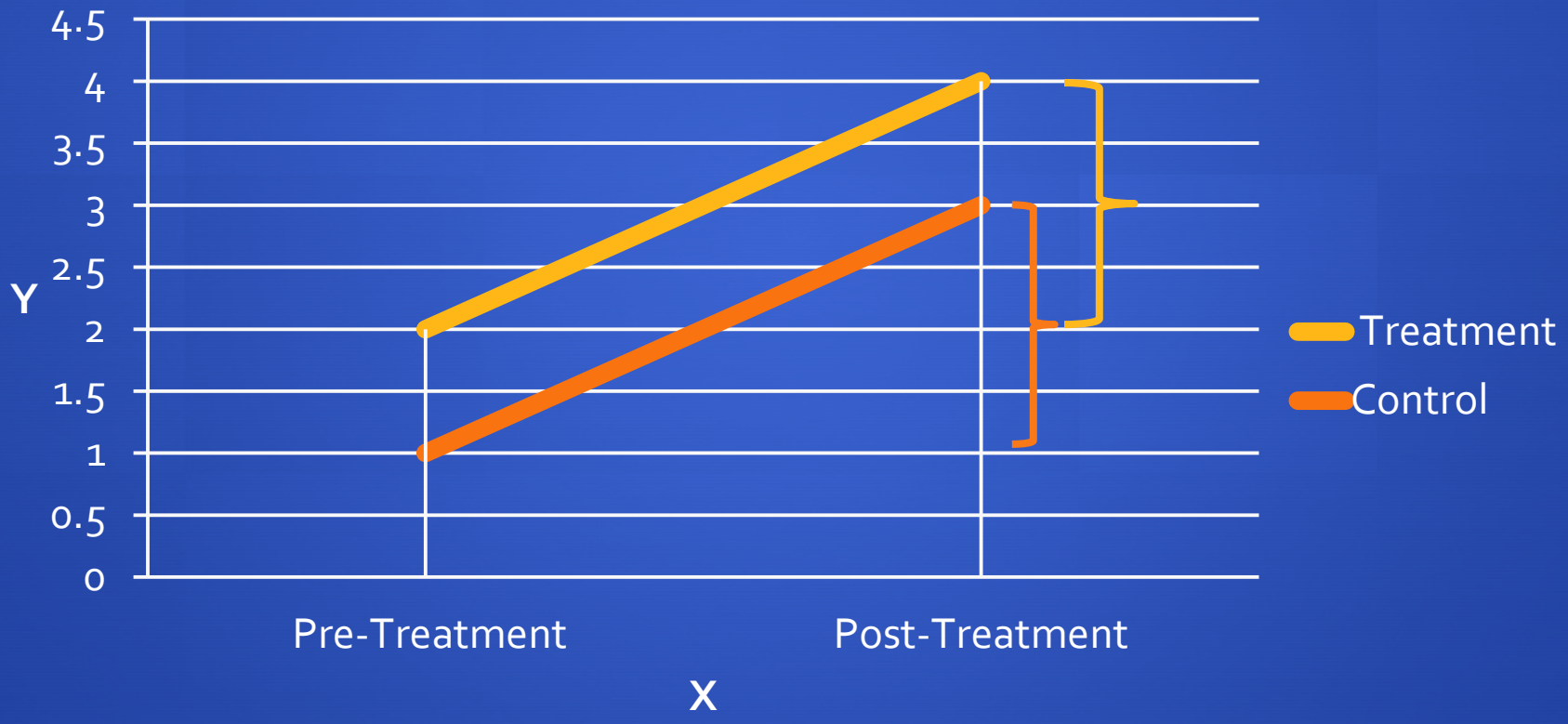
Pre-post comparison without control group



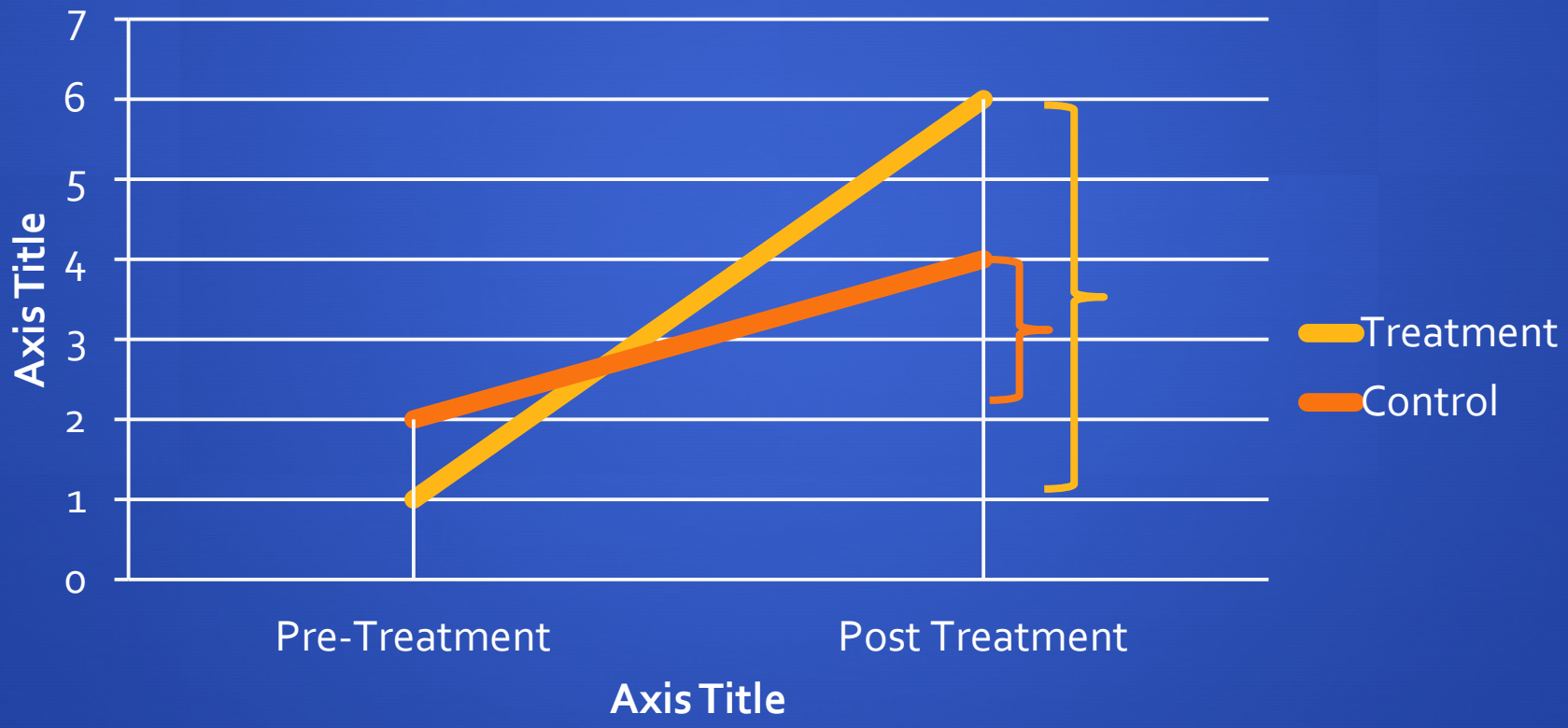
Pre-post comparison with control group



Pre-post comparison with control group



Pre-post comparison with control group



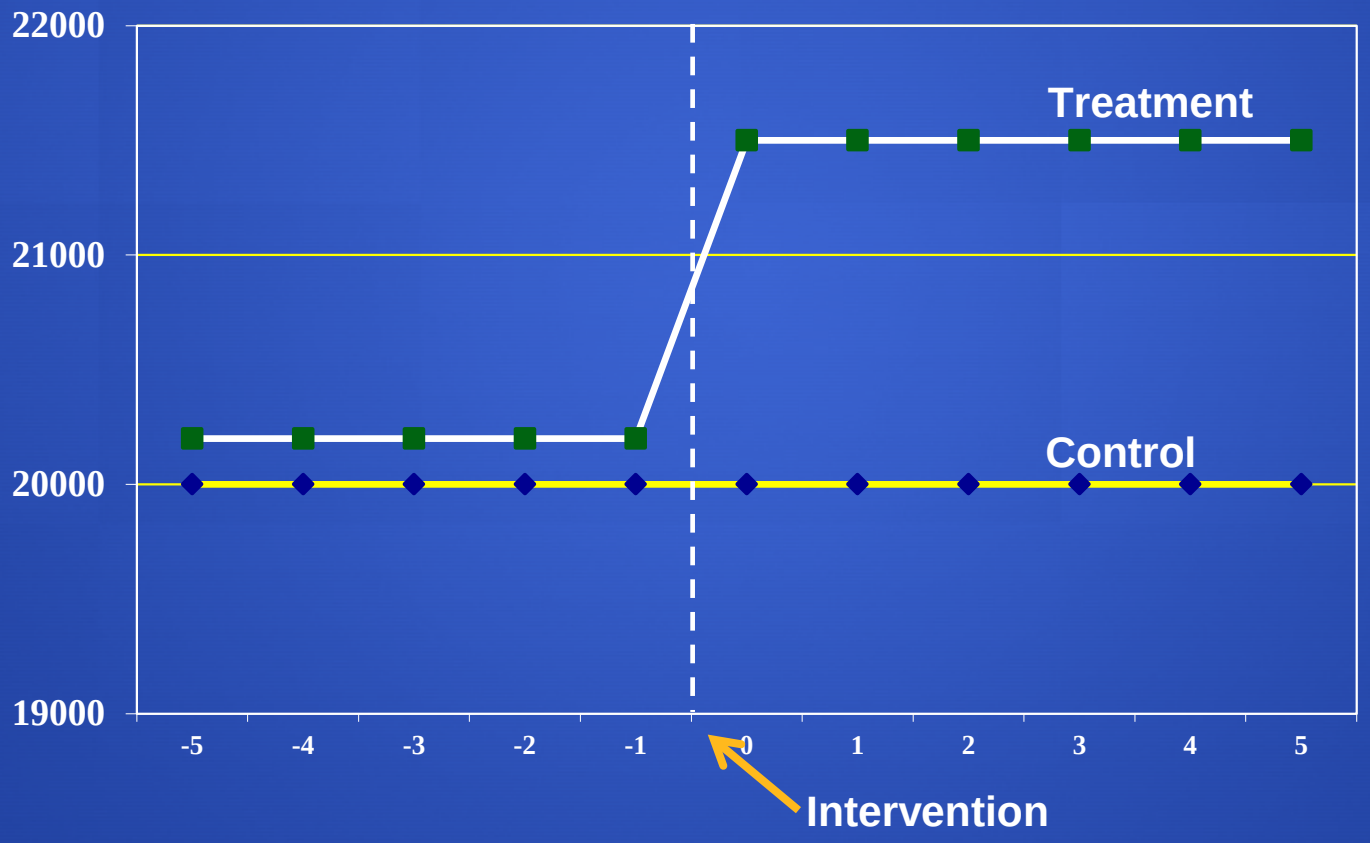
Pre-post comparison with control group

- How do we select a control group?
  - Remember that we may have constraints such as stakeholder input, data availability, etc.
  - A good strategy is to think of your *ideal experiment* and then try to approximate that given your constraints to the best of your ability
- General strategy: select control group based on values of a variable that is correlated with the outcome.
  - Ad hoc choice of group that should look like the treated (e.g., other migrant groups in neighboring states)
  - Statistically adjust for various factors (Z) within a regression
  - Formal matching based on pre-intervention characteristics of treated and controls (e.g., pre-intervention earnings)

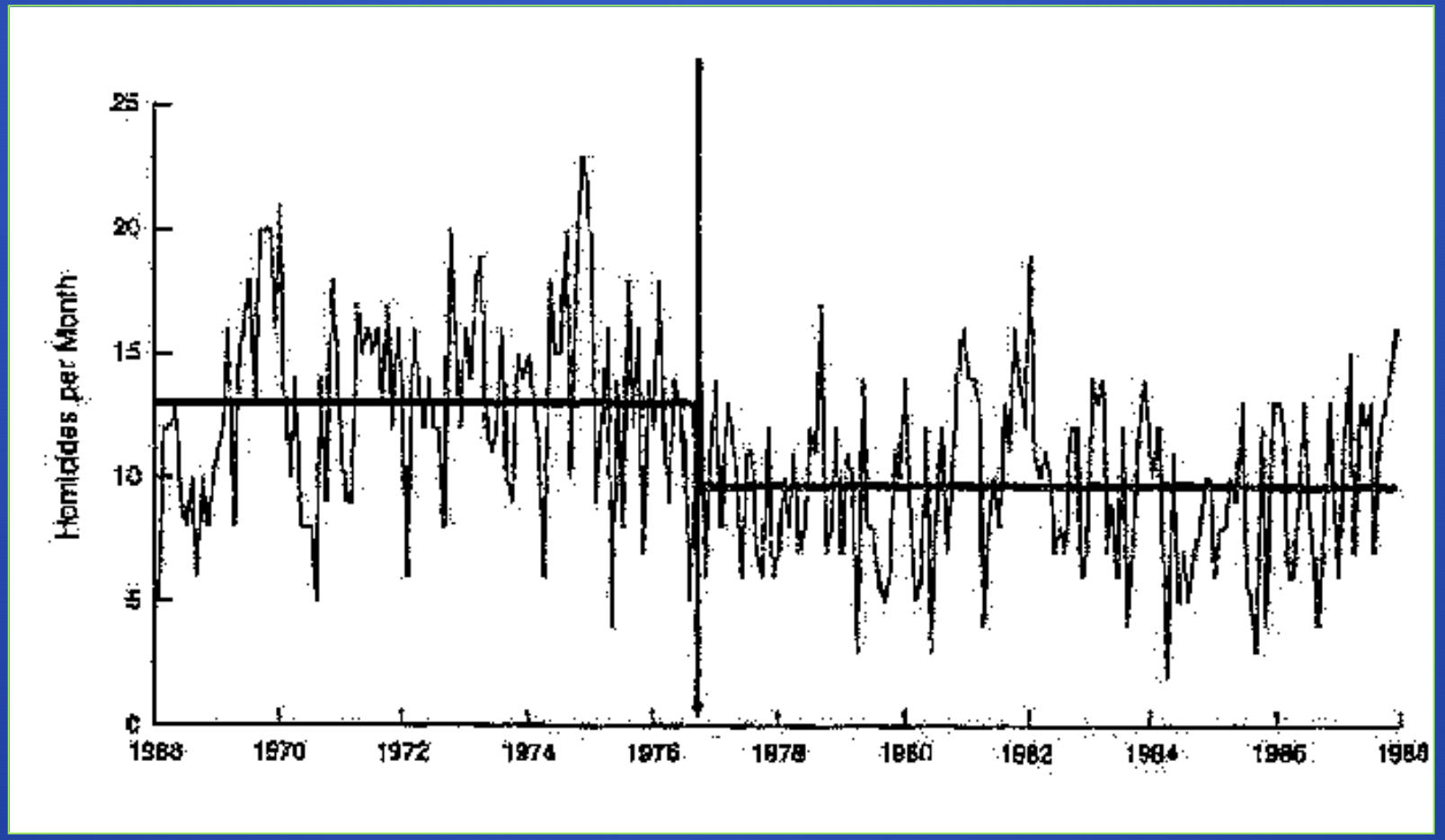
Interrupted Time Series

- Time series: numerous observations made on the same outcome consecutively over time
- Interrupted: treatment creates a hypothesized breaking point for the time series
- Control group: sometimes – helps with threats to validity
- Extension of the pre-/post design: more than one observation before & after treatment
- Counterfactual
  - If no control group, extrapolate pre-treatment trend
  - If control group, compare pre-post changes within treatment group to pre-post changes in control group

Interrupted Time Series

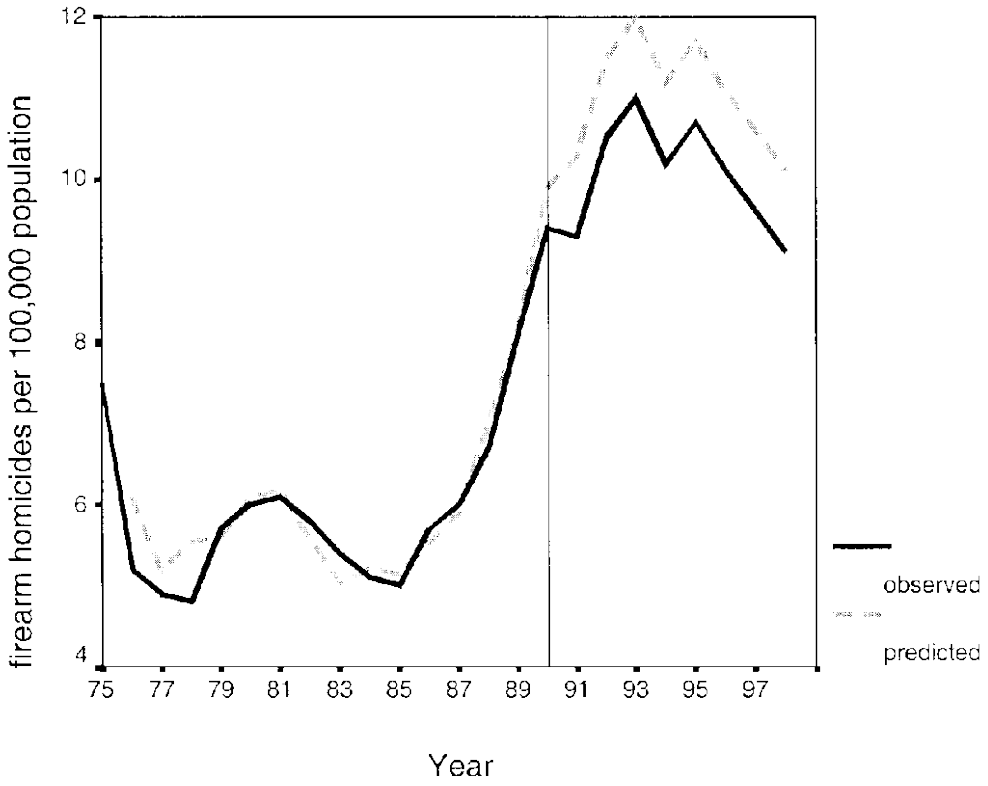


Interrupted Time Series



Number of Homicides per Month by Firearms in D.C.

Potential Comparison Group for ITS: Neighboring State of Maryland



Interrupted Time Series

- Most effective when:
  - Intervention is quick and well documented
    - If intervention is spread out over many months or years, may not know when to expect to see effects
    - Greater chance of history or omitted variable
  - Theory implies an immediate effect
    - If effect is not expected to manifest itself for many years, then greater chance of history or omitted variables creating bias
    - Example – effects of restrictions on cigarette advertising on lung cancer
  - Multiple observations both pre- and post-treatment
- Threats to validity: same as pre-post w/ control

Matching

- Matching seeks to create pairs of treated and untreated subjects that have a similar probability of receiving the treatment, even though only one actually does.
- Useful when intervention already implemented, but you have lots of data on the general population, including baseline measures of outcome.

Matching

- Propensity Score
  - One approach to matching
  - PS is the predicted probability that person receives treatment based on all available pre-intervention variables
  - Propensity score gives you a single variable that summarizes each person along the dimension that you are most interested in – i.e., the probability of being treated
  - 2-stage process
    - What factors predict probability of treatment?
    - What is an individual's probability of treatment?
  - Match each treated case with a control case whose PS is similar

Matching

- Strengths
  - Selection bias: compares change in outcomes for treated subjects with change in outcomes of control subjects that have a similar (but not identical) probability of treatment
  - Directly compares otherwise similar members of the treatment and control group (using average treatment effect (ATE) or similar statistics)
- Weaknesses
  - Quality of control group depends critically on how the propensity score is constructed. Are all relevant factors included in the model?
  - Researcher decides what is included in model, which may lead to misspecification
  - Limited pre- and post-treatment observations

GROUP EXERCISE

- Reimagine the Qatar financial literacy training RCT as an ITS or a pre-post matching design. What kind of data would you need? What biases might this design not protect against? Are there any advantages?
- Use the "Threats to Internal Validity" checklist to critique and improve on your design.

SESRI Policy & Program Evaluation Workshop

Doha, Qatar

January 19-22, 2015

Outline: Session 4

- Measurement
- Types and Sources of Data
- Operationalization
- Measurement error
- Reliability and validity
- Measurement theory
  - Random error
  - Bias

The Process of Operationalization: Research on Remittances

We are interested in studying remittances from a policy perspective:

Who sends them, in what amount , for what purpose?

How can we increase savings behavior among foreign workers and their families?

What are the aggregate flows of money between countries and how do they change over time?

The Process of Operationalization: Research on Remittances

What is a remittance?

The transmission of money to a foreign place

How will we know one when we see it?

A receipt for a transmission from a financial institution at either end of the transaction

Self-reports of transmissions

Aggregate monetary flows between countries

What about informal transmission of goods and products?

The Process of Operationalization: Research on Remittances

What is a remittance?

The transmission of money to a foreign place

What is the unit for which we can observe or measure remittances?

A receipt for a transmission from a financial institution at either end of the transaction: a transaction

Self-reports of transmissions: an individual

Aggregate monetary flows between countries: national-level data for a time period (How much per month or year)

The Process of Operationalization: Research on Remittances

The difference between description and analysis or explanation.

What is the average size of a remittance in Qatar?

H: A savings education program will increase the average size of remittances.

The Process of Operationalization

Deciding on the Units of Measurement and Units of Analysis, i.e. defining how the variables will be measured, observed, or formed

All the variables must be measured for the same units of analysis, especially when evaluating a hypothesis

Deciding on which research design will be used to collect the data

What Does This Mean in Data Terms? A Hypothetical Data Matrix

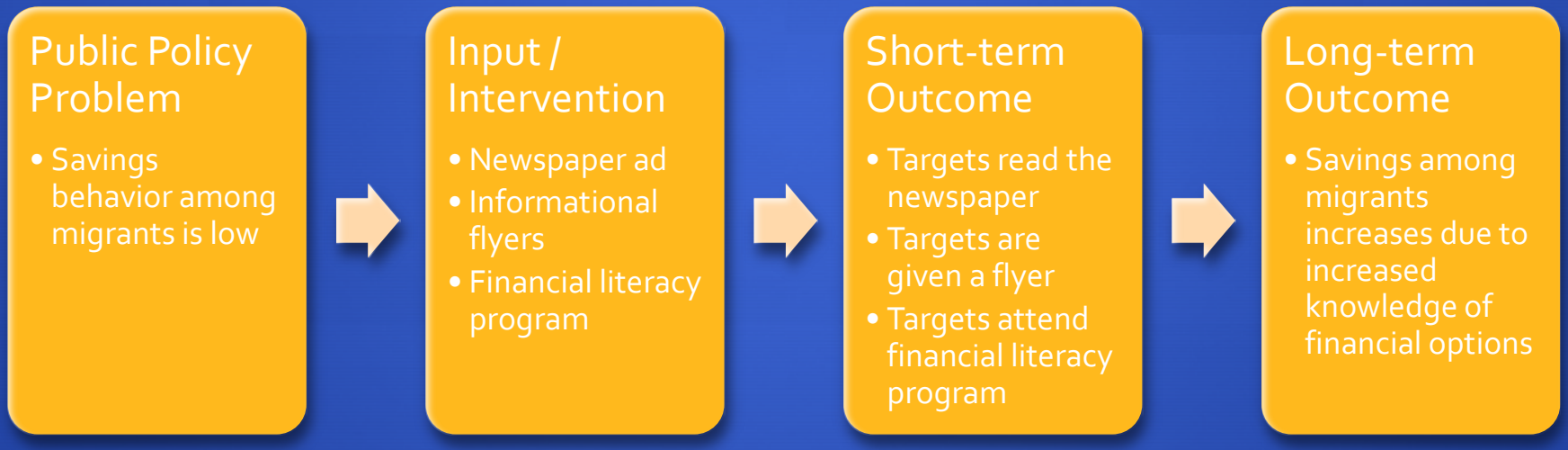
V1	Sex	Age	Treat	Amt. Remittances ($)
Resp1	M	25	1	880
Resp2	M	37	0	400
Resp3	M	30	0	285
Resp4	M	28	1	750
Resp5	M	40	0	1000

What Does This Mean in Data Terms? A Hypothetical Data Matrix

Employing Country	Amount Sent ($M) to India	Egypt	Sri Lanka	Pakistan
Qatar	15,700	931.0	613.2	1,700
Kuwait	2,900	2,200	823.4	485.9
Oman	2,600	231.3	132	283.9
Bahrain	759.6	187.8	-----	160.9

What is missing here?

Causal Diagram of Hypothesis



What Are the Basic Design Considerations?

- Can we develop baseline measures on remittances before the savings education program starts?
- Can we create a panel study/longitudinal data file with repeated measures over time?
- Can we develop an appropriate control group(s)?
- How many different units of analysis can we use?

A Digression: Ways of Collecting Data

SOURCES OF ERROR IN MEASUREMENT

Random errors are due to chance fluctuations and average to zero

In general, they contribute to imprecision

Systematic errors are not due to chance and they have a direction or "bias"

They can raise concerns about either reliability or validity

Thinking about Measurement

For any measure, we can think about the observation consisting of a true score, plus some error.

Observed Value = True Value + Error

Since the error can be either random or systematic or both:

$$\text{Observed Value} = \text{True Value} + \left[\text{Random Error} + \text{Systematic Error}\right]$$

The Concept of “Education about ways to save money”

H: Education about ways to save money will increase saving behavior.

How would we measure it?

You must agree on the units. What is a “unit” of education and of saving behavior?

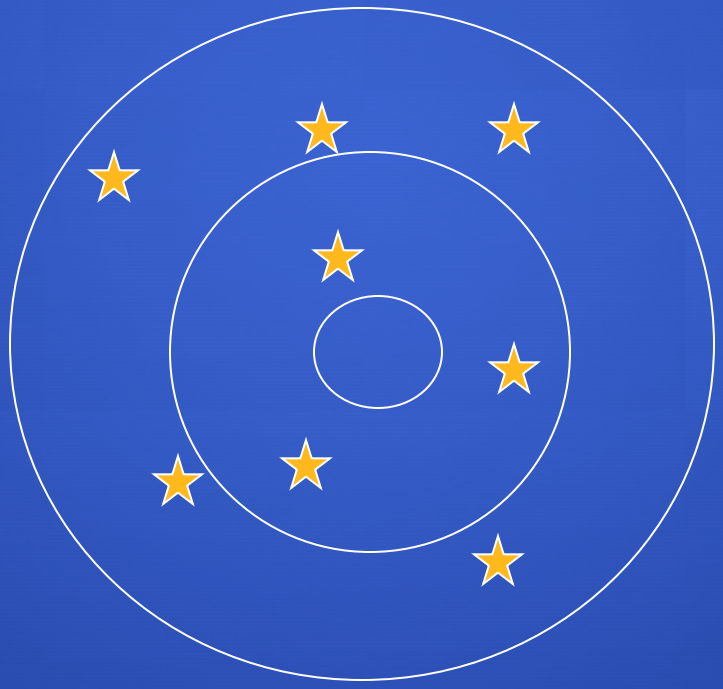
RELIABILITY and VALIDITY Refer to Possible Measurement Errors

Reliability refers to how consistent or precise the measurement is

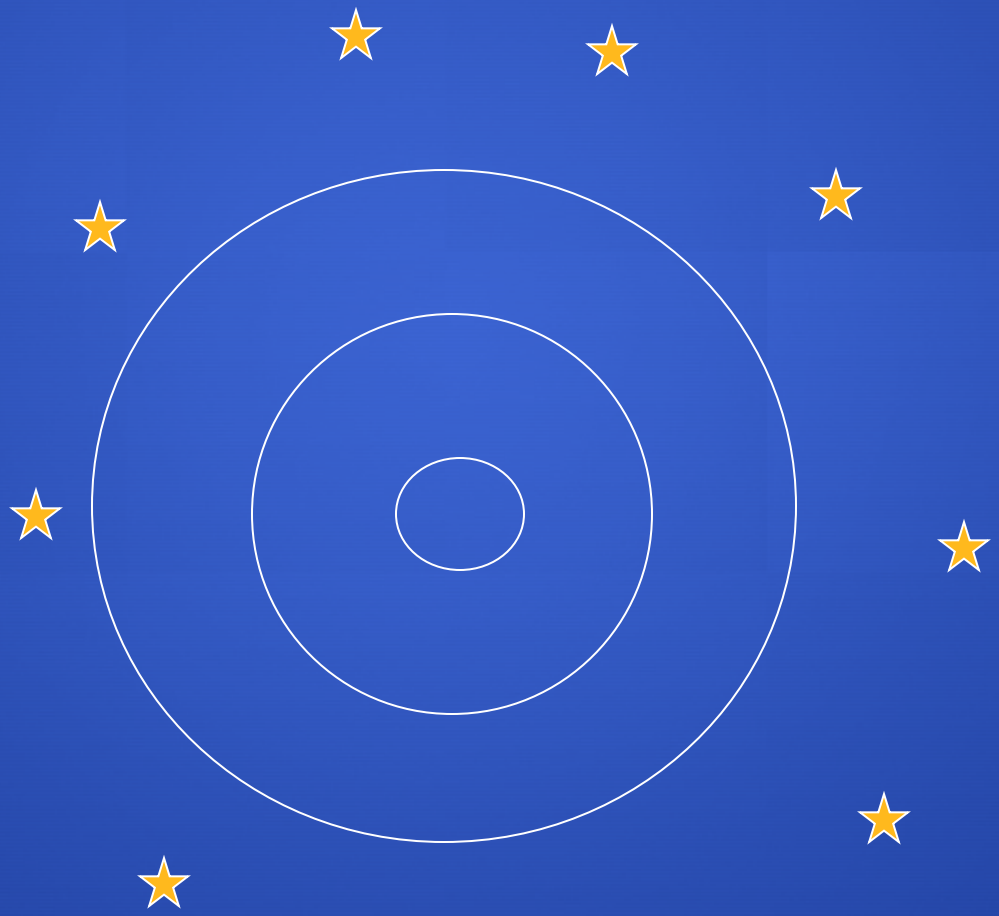
Validity refers to whether we are measuring what we think we are (the concept)

Reliable, Not Valid

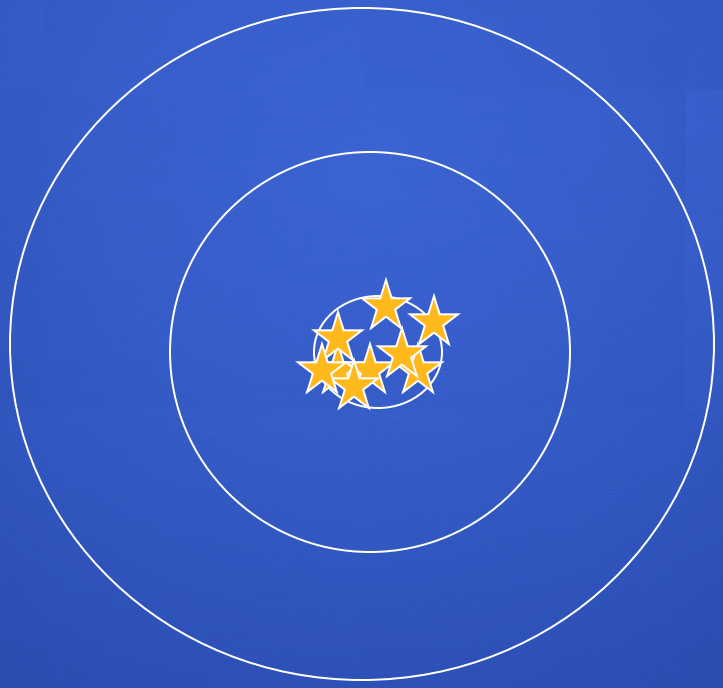
Valid, Not Reliable



Not Valid, Not Reliable



Valid and Reliable



Reliability is concerned with precision and consistency in measurement.

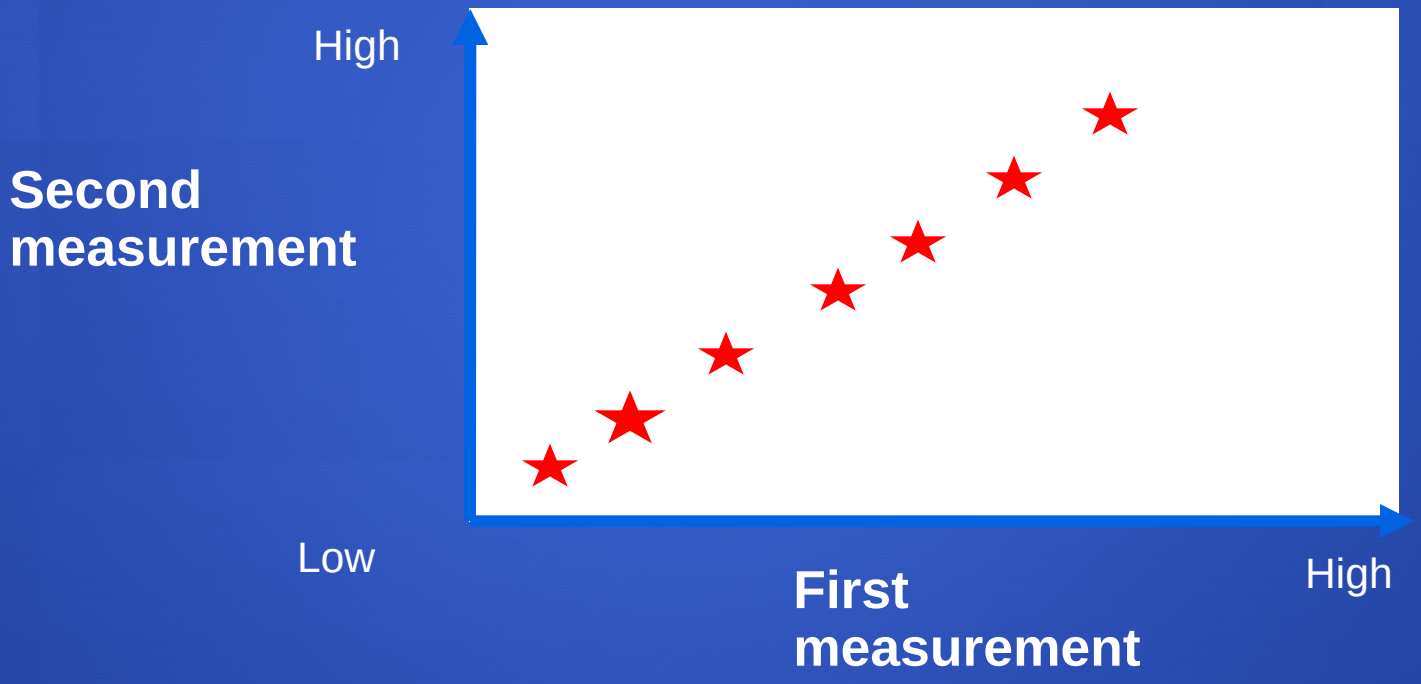
It can involve a relationship between repeated measurements of the same concept, in the form of a hypothesis:

Measurement Theory

The Measurement Hypothesis:

Repeated measurement of the same concept using the same operationalization should return the same value, on average.

Expected Observations with Repeated Measurement



What Does This Mean in Data Terms? A Hypothetical Data Matrix

ID	X_{t1}	X_{t2}	V_4	V_5
Resp1	Yes	Yes	4	7
Resp2	Yes	Yes	3	1
Resp3	No	No	2	4
Resp4	No	No	1	2
Resp5	Yes	No	3	4

STANDARD ASSUMPTION ABOUT RANDOM ERRORS

The only difference between two measures of the same concept should be random error

Random errors:

1. Are both positive AND negative

2. Sum to zero

Standard Assumptions about Systematic Errors (Bias)

1. They tend to be either positive OR negative.

2. They DO NOT sum to 0.

   You are asked a question in a survey and one week later you are asked the same question. If the survey question and its responses are valid and reliable, your answers should be the same.

Correlation

Since reliability involves a relationship, it can be characterized by a statistical measure of association such as a correlation coefficient.

Reliability and Correlation Coefficients

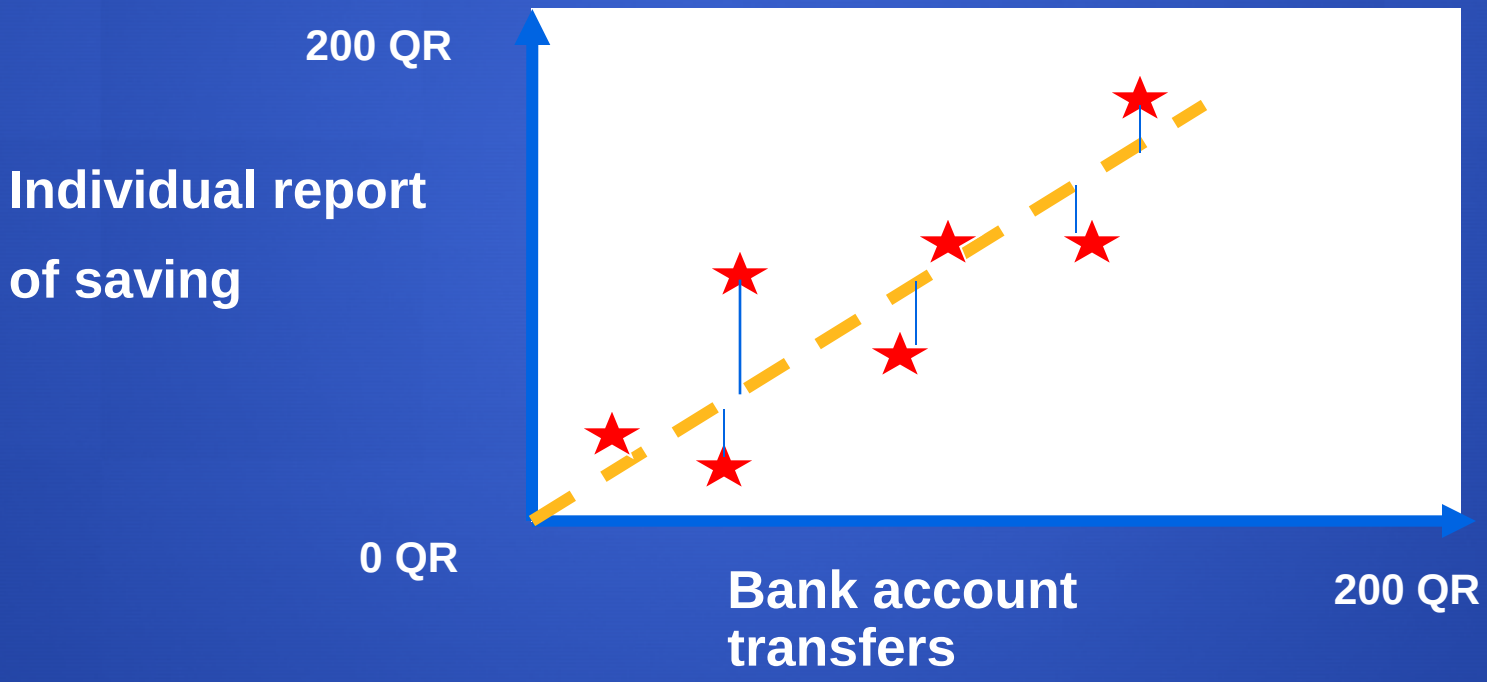
A **correlation coefficient** is a measure of the standardized covariation. It range between –1 and 1.

- Values close to 1 suggest a strong positive association.
- Values close to –1 suggest a strong negative association.
- Values close to 0 suggest no *linear* association.
- Correlation assumes a linear relationship, so it is reduced if the TRUE relationship is not linear.

A Measurement Strategy

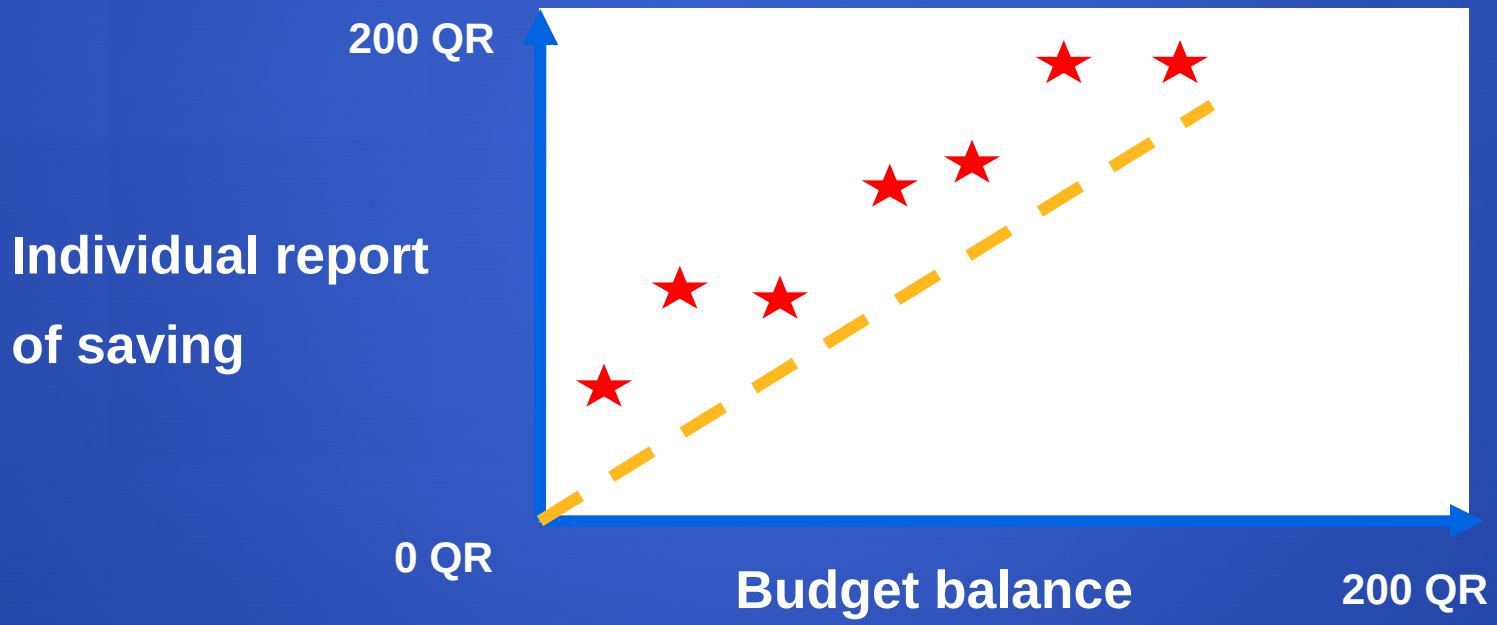
Ask a migrant worker how much of their monthly pay they send home and look at their administrative records for their current account balance.

Repeated Measurement with Random Error



What is the correlation summarizing this relationship likely to be?

Repeated Measurement with Systematic Error (Bias)



What are possible explanations for this observation?

GROUP EXERCISE

- Discuss in groups alternative ways to operationalize the concept of a "financial literacy program" and the implications for the appropriate units of analysis.
- What would the implications of that choice be for measuring the average remittance?