Meeting 6

Quasi Experimental Designs

Quasi-Experimental Design Framework
Interpreting Results

Interrupted Time Series

Regression Discontinuity

Matching

Examples from Readings

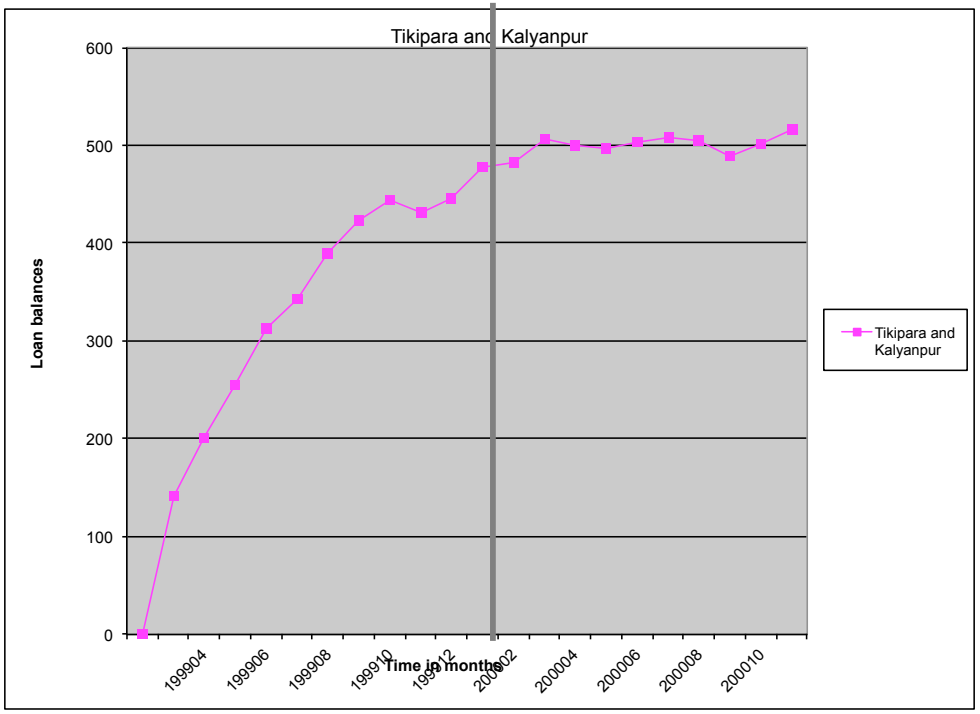
Quasi-experiments

- What if one cannot randomly assign units to the program under study?

Example: estimate the impact of a microcredit

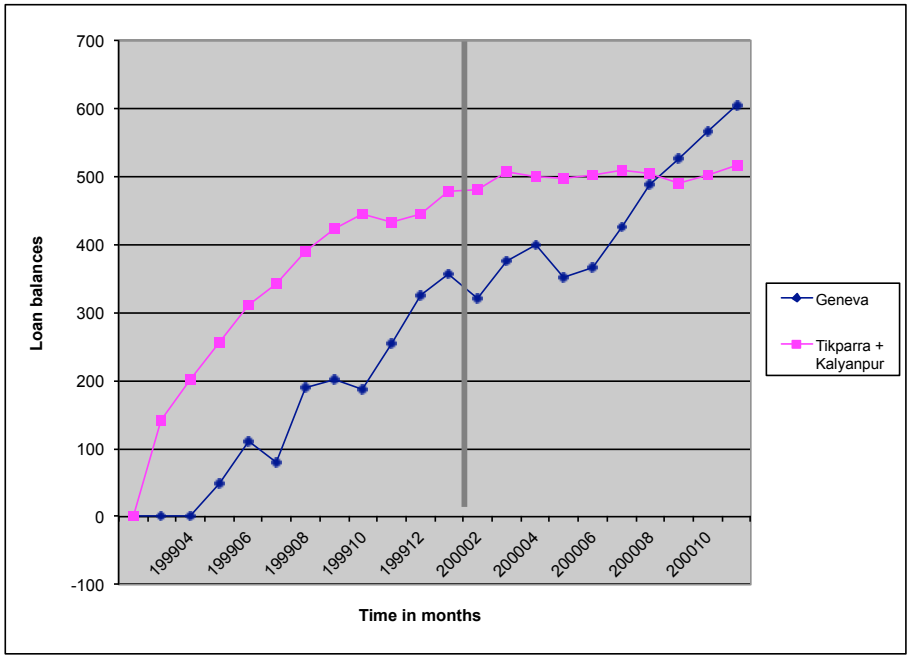
- Suppose we want to estimate the impact of changing the interest rate on borrowing in a microcredit.
- Based on previous lecture we would like to randomize the interest rate. What?!
  - Actually, some banks might let you do that.
  - But not many.

An idea



- Use interest rate change in two branches at an MFI.
- Seems like interest rate change had a big impact.
- But what else could have happened in between?
- Perhaps it was simply maturing naturally?
- Or other events?

A refinement

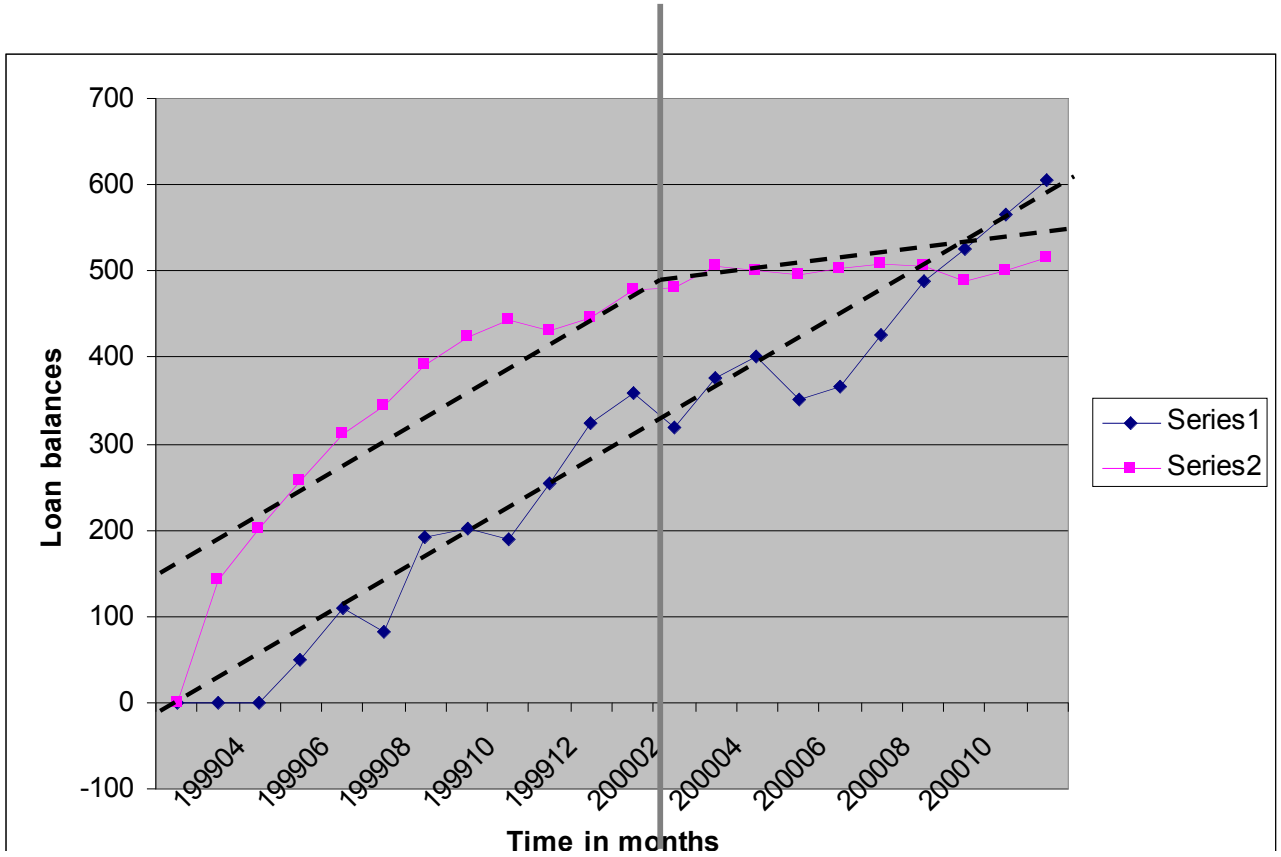


- There was no interest rate change in Geneva branch.

Average loan balances, *Safe*Save

A refinement: testing assumptions

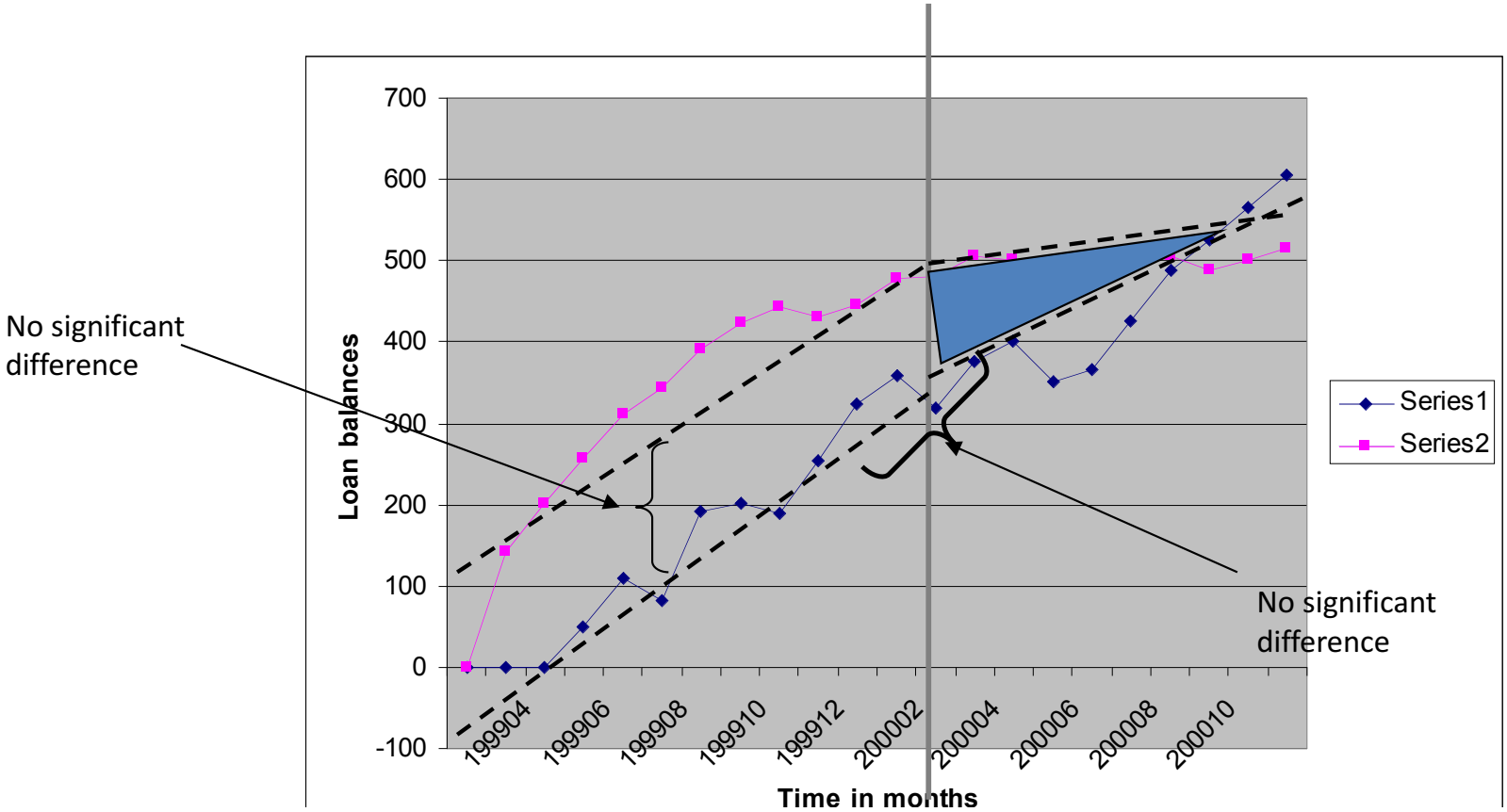
Common linear trend in pre-period (Intra-ocular test)



Average loan balances, *Safe*Save

A refinement: testing assumptions 2

Common linear trend in pre-period (statistical test)



Average loan balances, *Safe*Save

Example: communication skills

- We want to evaluate the effect of communication skills on medical students.
- We have a treatment in mind (enhanced clerkship emphasizing communication).
- Problem: each cohort must get the same clerkship, so randomization isn't an option.
- Problem: we have only one medical school, so no comparison group we can draw on.

Macy Cohort Study of Communication

	1999	2000	2001	2002
Comparison	2nd Year	3rd Year	4th Year	PGY 1
Class of 2001	**OSCE PRE TEST**	**Traditional Clerkships**	**OSCE POST TEST**	
Intervention	1st Year	2nd Year	3nd Year	4th Year
Class of 2002		**OSCE PRE TEST**	**Enhanced Clerkships**	**OSCE POST TEST**

OSCE=objective structured clinical examination

Quasi-experimental design

- All have independent variable (manipulated variable)=PROGRAM
- Usually have a pre-test (or proxy pre-test)
- Then divide into two types
  - Those with comparison group
  - Those without (reflexive)

Reflexive designs

- Designs with no control group.
  - Sometimes only thing you can do
    - Full coverage or highly saturated programs
    - Volunteer bias so no appropriate comparison group
    - Easy and cheap
  - Should always try to build in some other ways to strengthen this inherently weak design
- One group designs
  - Cannot rule out history and maturation by design
    - Subjects act as their own controls
      - Either simple pre-post
      - Or longitudinal panel design

Reflexive designs: what to do when you can't collect baseline data?

- Proxy pre-test
- Retrospective

Reflexive designs: proxy pre-test

- No opportunity for a true pre test, so…
  - Find or create a proxy
    - Archived data
      - Routinely collected indicators
    - Retrospective
      - Recollect back and report on pre-status
      - Perceived competence associated with training programs
        - » Retro pre / post design can be more accurate than
        - » True pre / post design
        - » (because training teaches you what you don't know)
  - Evaluation of condom program in NYC
    - Proxy pre: Kids just entering school (don't know about condom program yet)
    - Proxy pre: Ask kids what knew about/whether using condoms in beginning of their freshmen year

Reflexive designs

- Pre-test and Post-test single group design
  - What happens from pre to post could be due to:
    - History
    - Maturation
    - Testing
    - Instrumentation
    - Mortality/Attrition
    - Regression (to the mean)

Multiple group designs

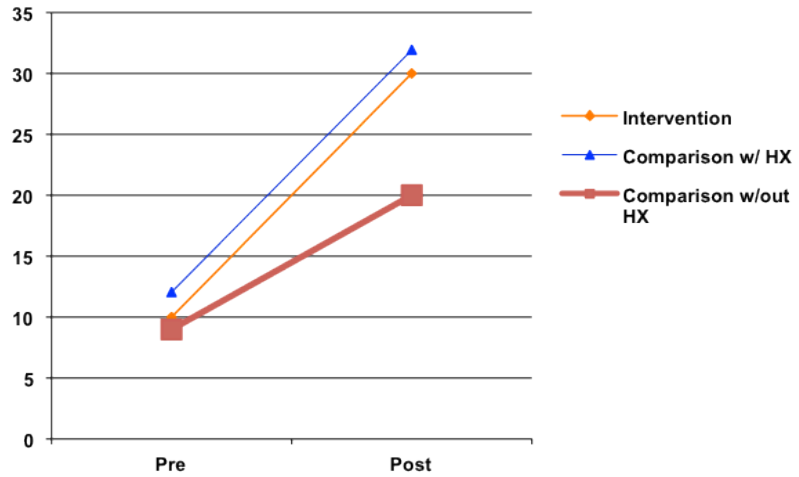
- Comparing two groups for relative outcomes
  - Key independent variable issue: Comparability of groups at outset
    - If comparable, only difference is program
    - Post-test differences can be attributed to the program
- *The* multiple group threat to independent variable is selection
  - Multiple group threats to independent variable parallel single group
  - But in multiple group, threats depend on ***selection***
  - If threats occurred equally (no selection) across multiple groups – not a problem
    - Groups are still equal
    - Only difference is still program

Selection interacts with all other threats

- Single group threat=History (outside events, interfering events)
- Multiple group threat=Selection-History
  - Outside events impact one group but not the other
  - Serving to make the groups unequal
- Random assignment randomly distributes these threats across the two groups = no selection bias
  - Note: Key point RA doesn't eliminate threats it spreads them evenly across the treatment and comparison.

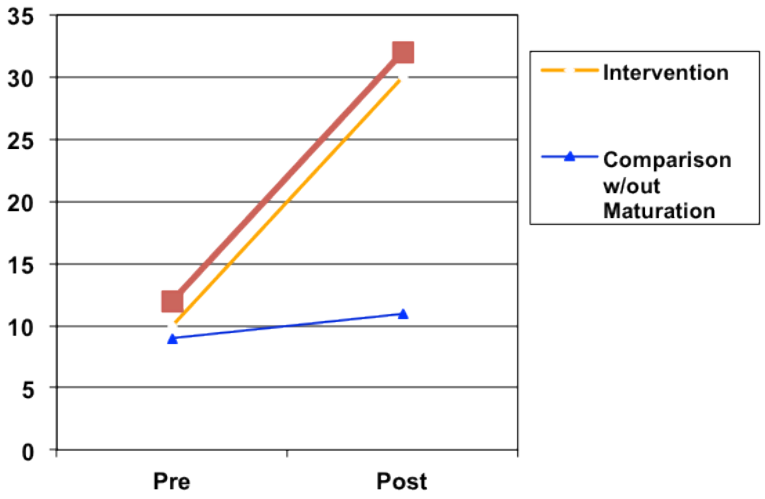
Selection-history

- Any event/force/influence that occurs between pre-test and post-test that groups experience differently



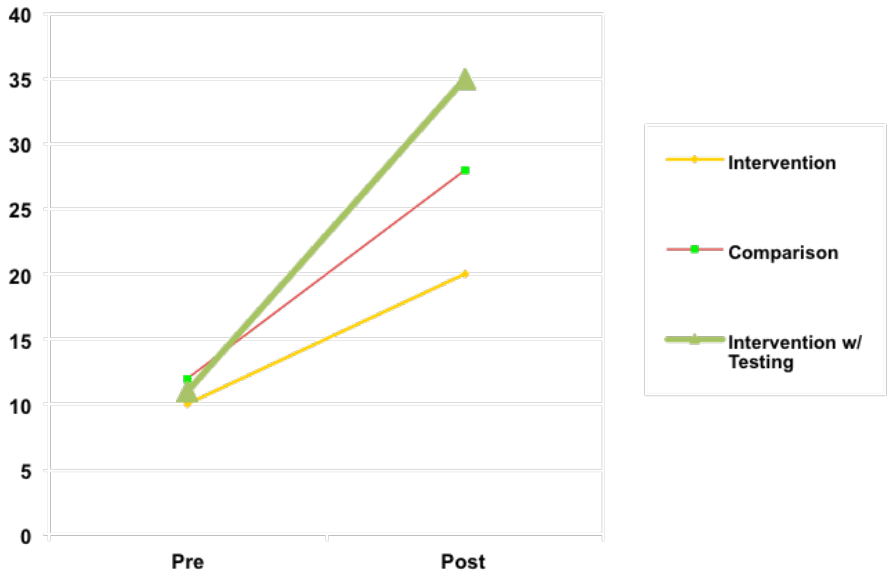
Selection-maturation

- Differential rates of normal growth between pretest and posttest



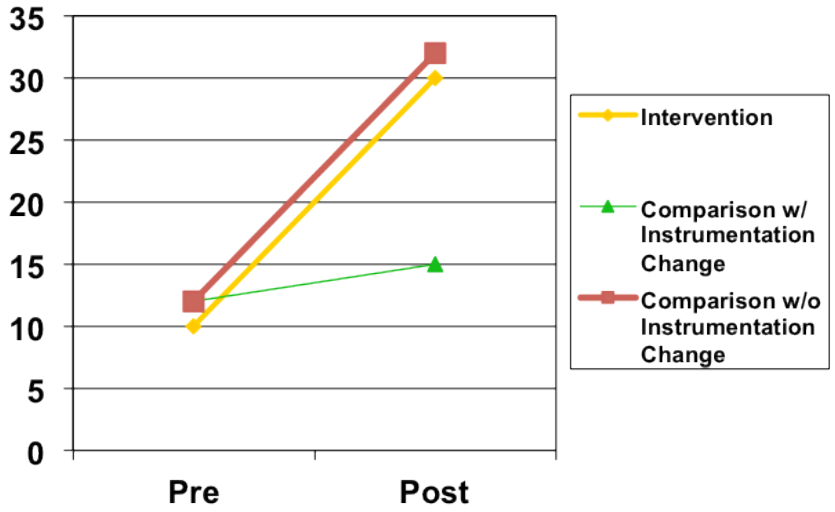
Selection-testing

- Taking the pre-test (have pre-test data collected) has a differential effect on post-test scores



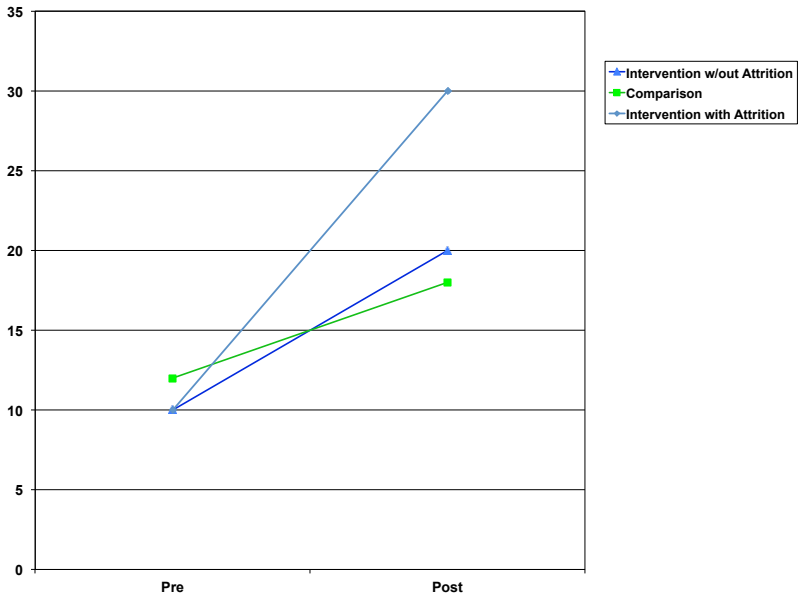
Selection-instrumentation

- For one group, the way the tests are administered or scored or designed changes



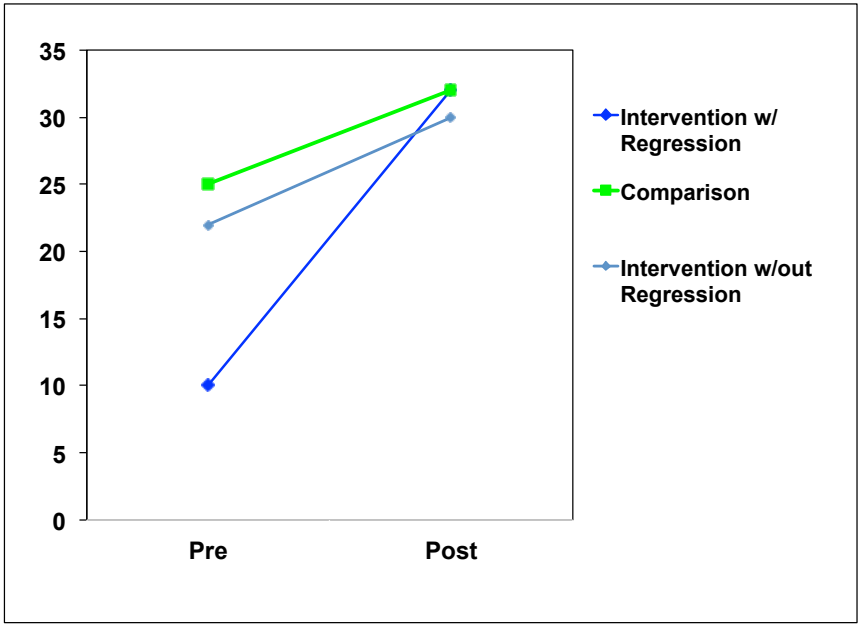
Selection-attrition

- Non-random drop-out: More of a particular kind of unit drops out of one group compared with the other group



Selection-regression

- Different rates of regression to the mean
- Mean is 35



Strengthening QE designs

- Common variations for strengthening Q-Es
  - Multiple periods of assessment
    - Multiple pre-tests or post-tests or longer term follow-ups
  - Multiple comparison groups or program groups
    - Contemporaneous
    - Another point in time (cohort)
  - Multiple or manipulated “treatments”
    - Variations in dosage, type, order of treatment
    - Interrupted (introduce and remove)

Comparison groups

- Before the program starts!
- Importance of comparability
  - Know what makes groups comparable.
  - How do people enter/select into groups?
  - Pre-test on pre-period DV (outcome)
    - Establishing similarity of groups
  - Other factors known to be related to DV
    - Establishing similarity
  - Demographics o.k. (descriptive)
    - But doesn't necessarily mean that the groups are similar on the key factors (the ones that will shape outcomes)
      - E.g., motivation

Finding comparison groups

- Naturally existing comparison groups
  - Men vs. women in the labor market?
- Matching or constructing equivalent controls
  - Individual matching
    - Need large pool to match from
    - Need to know what to match on
    - Need good data on those factors
  - Aggregate matching
    - Easier (matching distributions of characteristics of units)
    - Can throw out units to equalize
  - Only as good as what you know about what makes the groups different
- Statistical control (linear regression)
  - Adjust the outcome score on the basis of "nuisance" factors
  - Examine effect of program on DV over and above what the covariates have already explained in the DV's variation
  - Model the selection process and incorporate that into analyses as a control variable
- Normative or generic controls
  - Use data from broader population to compare

Finding comparison groups

- Different geographical locations
- Waiting lists
- Similar units within organization
- Ineligibles
- Eligibles who chose not to participate
- Staggered implementation of programs
- Expansion of programs
- Samples from earlier or later in time (cohort)

What to do when you can't find a comparison group?

- You can't find a comparison group?
  - Cohort
  - Separate Samples
  - Static Group Comparison

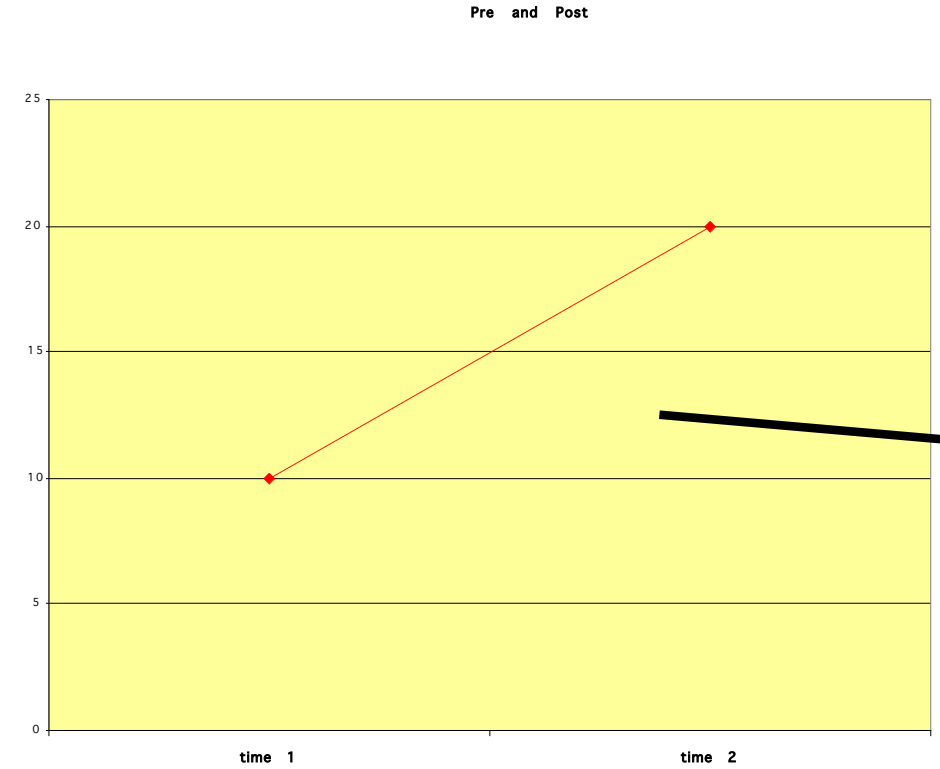
Static group comparison

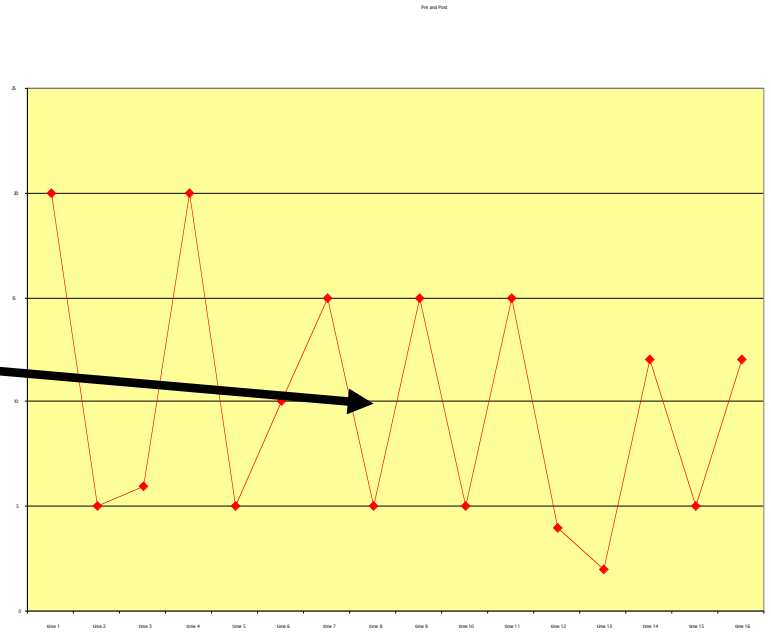
- Essentially, a very non-equivalent design
- Can't find/create a good comparison group
  - So you use what's available, even if it's not good
    - Education Program
      - Compare to other schools but other schools have much less educated teachers
        - » Use anyway, because that's all you have

Time series designs

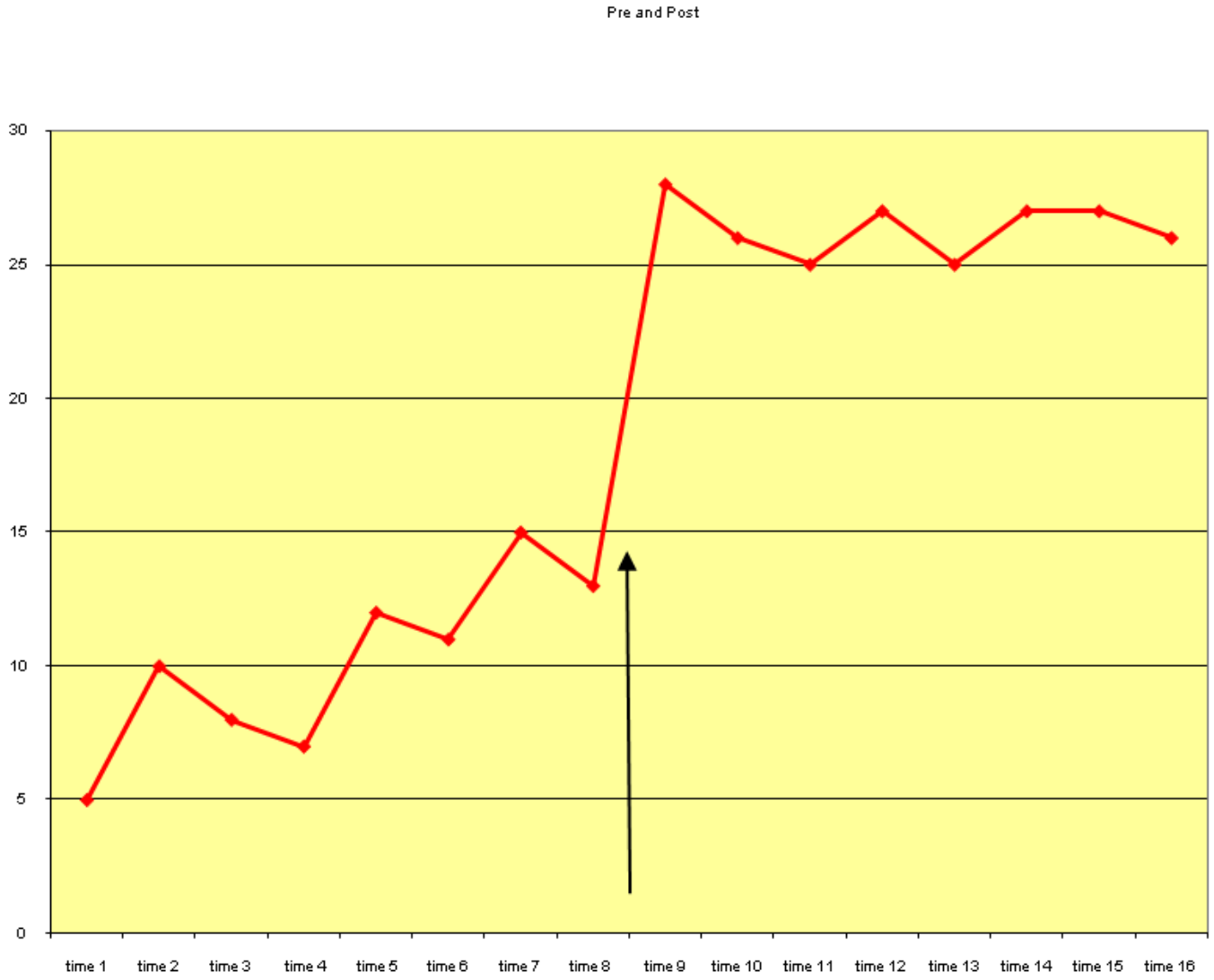
- Measures taken repeatedly
  - So that normal variability can be fully captured
  - When a new policy is introduced
    - Variability due to program effect should stand out
  - Often used to evaluate policies
    - Changes that effect entire populations
- More time points the better
  - Instead of increasing sample size
    - Increase number of assessments
- Best used when
  - Testing and instrumentation not likely to be problems

Pre-post vs time series

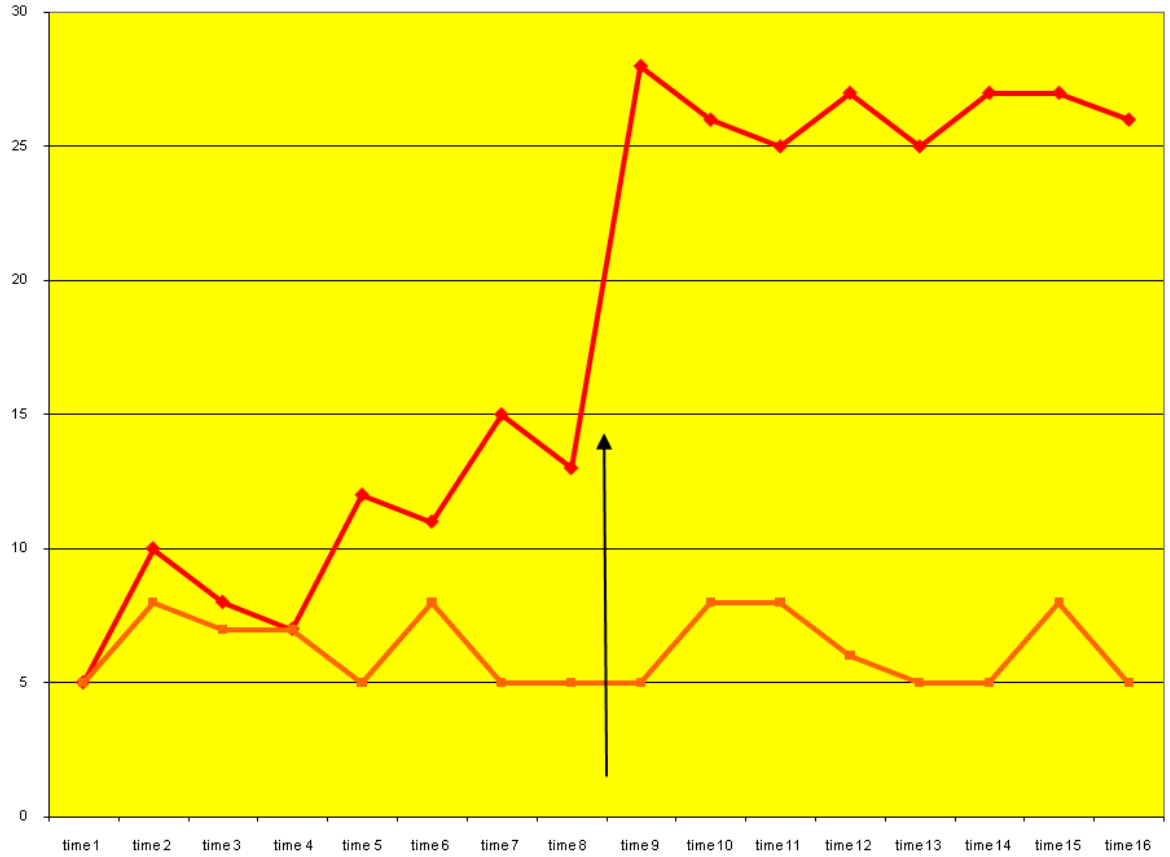




Interrupted time series



Interrupted time series with comparison group



Dis/advantages of time series

- Detailed knowledge of circumstances surrounding the program change
  - Can detail effects of history, maturation etc.
  - Can show whether diffusion of intervention to comparison is occurring
- Often can be "eye-balled"
  - Although statistical methods useful for correcting for time effects (repeated measures across time)
- Helps detect differences between groups that are related to slope (decay, lags, etc.)
- Can use systematic replication to refine and tweak intervention
- Need lots of data points.
- Can controls for broad (macro) events.
- Sensitive to changes in instrumentation.
- Sometimes population changes when policy changes
  - Data after policy/program change comes from different people than before

Example of pre-post / multi-group / time series

- Dehejia, Morduch, and Montgomery

- What is the interest elasticity of loan demand in a microcredit setting?

Introduction

Difficulties measuring elasticities

- Evidence is mostly anecdotal and non-causal.
  - Exceptions: Karlan and Zinman / Gross and Souleles
- Main difficulty: the schedule of interest rates seldom varies within a given program
  - When it does change, it does so for everyone.
  - Thus, it can be hard to disentangle the effect of the interest rate change from broader changes occurring simultaneously (e.g., macroeconomic shocks).

Introduction

Difficulties measuring elasticities

- It may be possible to compare clients of different institutions who face different interest rates at any given moment
  - but then researchers face the question of why some customers selected one institution and why others selected another.
- Also difficult to disentangle the effects of non-price differences among programs.

Data

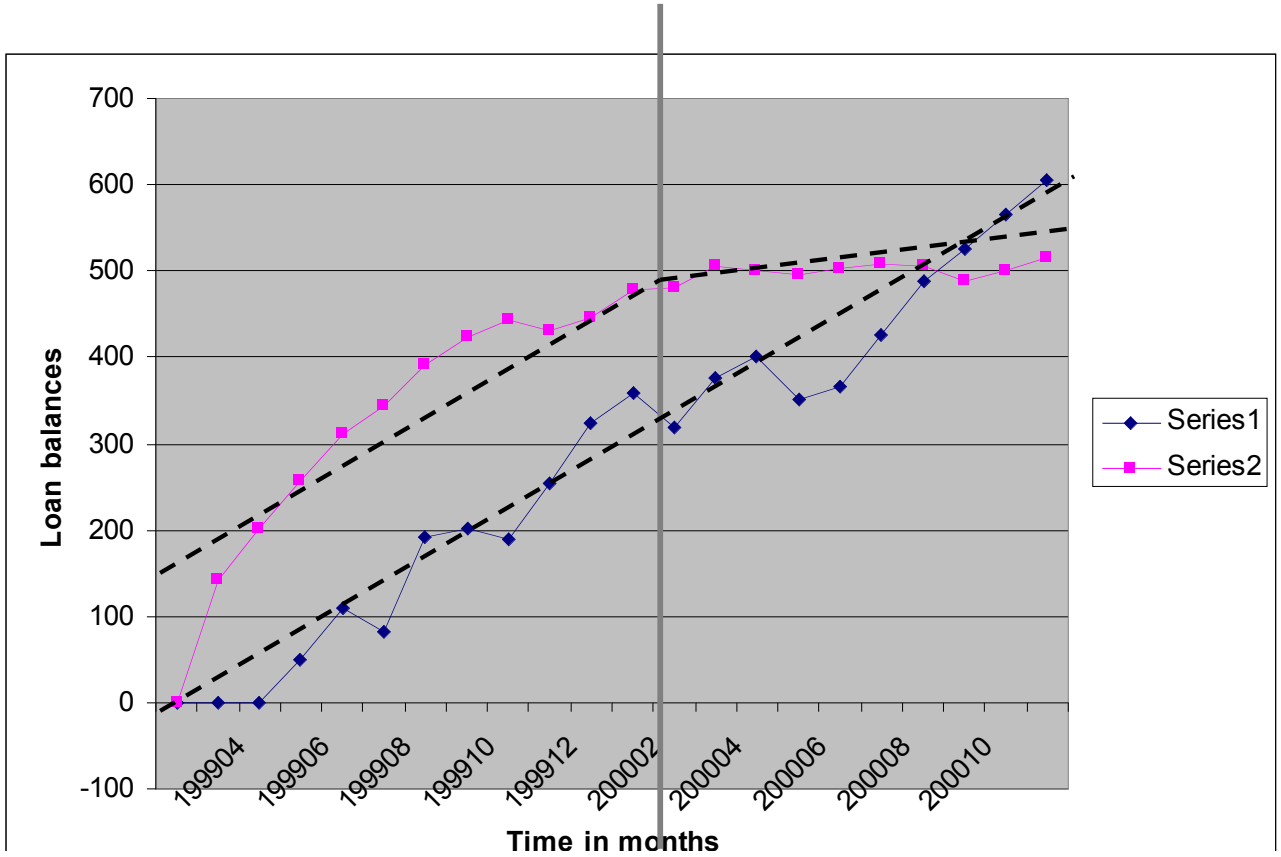
- Most of the analysis focuses on 68,037 month-customer observations between January 1999 and January 2001.
- They reflect data on 5147 customers, not all of whom are part of the program during the entire period.
- The change in the interest rate occurs midway through the sample, in February 2000.

QE evaluation design

- The analysis below compares changes in Tikkapara and Kalyanpur to ongoing conditions in Geneva.
- Identification is based on two assumptions:
  - Unanticipated between-branch variation interest rates allows us to time the interest rate change.
  - Common trend across branches prior to interest rate increase.

Identification

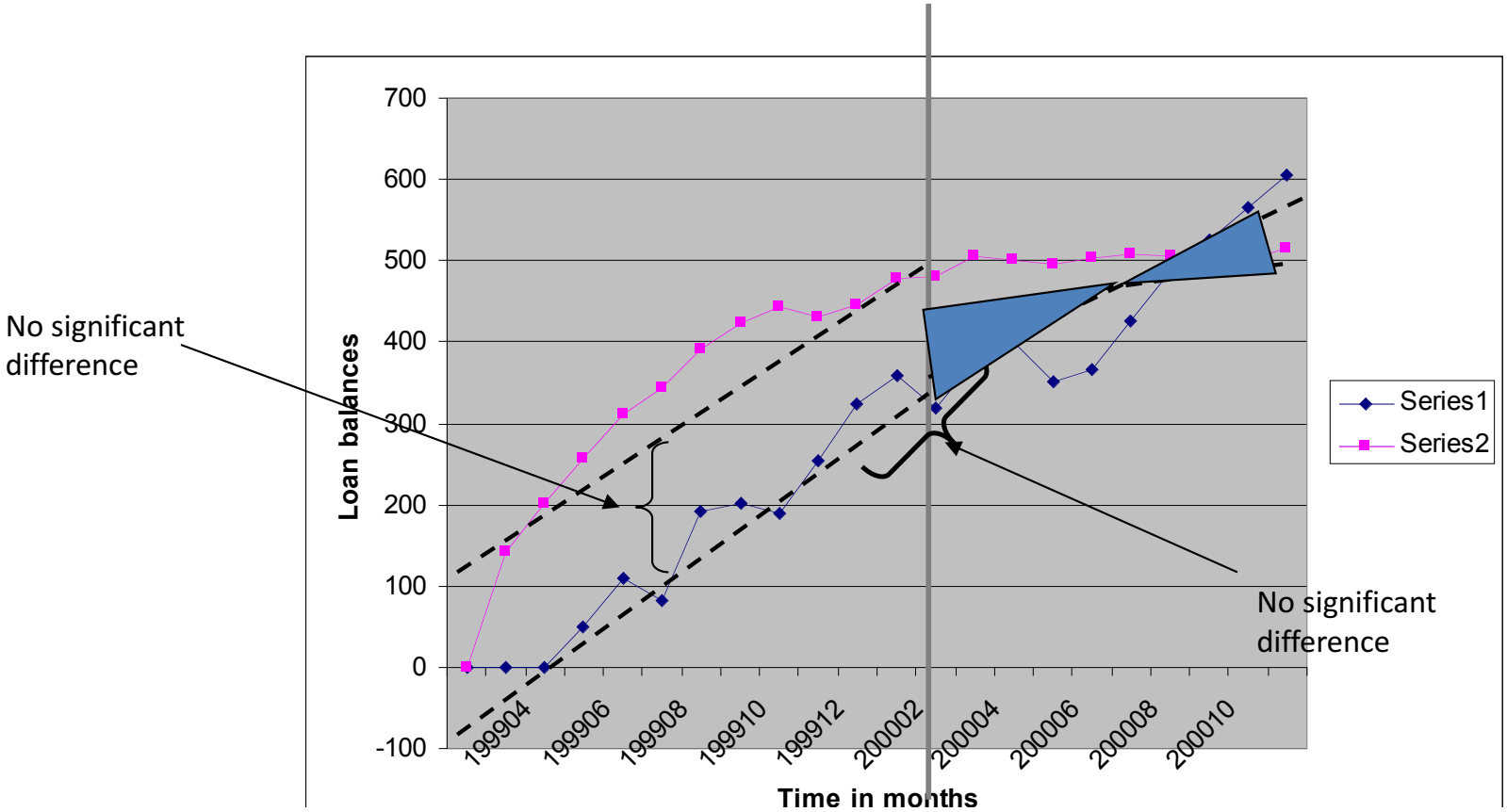
Common linear trend in pre-period (Intra-ocular test)



Average loan balances, *Safe*Save

Identification

Common linear trend in pre-period (regression test)



Average loan balances, *Safe*Save

Identification

Regression framework

Difference in difference estimator:

$$yit = \beta_0 + \beta_1 Treated_i + \beta_2 Post_t + \beta_3 Treated \times Post + \varepsilon_{it}$$

β_3 gives the impact of interest:

The change in borrowing comparing before and after the change in Tikkapara and Kalyanpur, relative to the contemporaneous change in Geneva.

Identification

Refinements/Extensions

1. Control for borrower characteristics, including age and length of time in the program (why?).
2. Include a full set of month-year dummies rather than simply controlling for time effects with a before-versus-after dummy (why?).
3. Allow for trend differences between the treated and comparison groups, in addition to a shift in the level of borrowing.

1. Include account fixed effects; this then controls for all non-time-varying differences between borrowers across districts.
2. Expand basic set-up to consider the heterogeneity of responses (along dimensions such as gender, wealth, and age).

Results on elasticities

Table 1: Summary statistics January 1999 - January 2001

(monetary values in 1985-86 taka)	Tikkapara and Kalyanpur	Geneva
% Women	64	67
Age	27	28
Savings deposit	43 ($2)	44 ($2)
Savings deposit if >0	65 ($3)	50 ($2)
Savings balance	579 ($27)	217 ($10)
Default rate on loans	0.058	0.02

Results on elasticities

Table 1: Summary statistics January 1999 - January 2001

(monetary values in 1985-86 taka)	Tikkapara and Kalyanpur	Geneva
Initial loan amount	1384 ($64)	891 ($41)
Length of loan cycle (months)	17.5 (14.5)	16.0 (16.3)
Loan balances	434 ($20)	480 ($22)
Loan balances if > 0	1051 ($48)	816 ($37)
Repayments	65.6 ($3)	103 ($5)
Repayments if > 0	201 ($9)	405 ($19)

Results on elasticities

Table 2: Testing trend differences and D-i-D assumptions

	(1)	(2)	(3)	(4)
Sample	TIKA	GE	TIKA+GE	TIKA+GE
Specification	linear trend	linear trend	linear trend	linear trend with controls
Constant	97.8***	-114**	-114**	-165***
	[8.94]	[46.2]	[56.9]	[54.7]
Treated			212***	35.8
			[57.6]	[54.9]
Trend	35.3***	38.4***	38.4***	31.4***
	[1.19]	[4.82]	[5.94]	[5.66]
Treated×Trend			-3.09	-7.99
			[6.05]	[5.76]
Post	375***	-24.4	-24.4	38.6
	[26.6]	[58.5]	[72.0]	[68.6]
Treated×Post			400***	330***
			[76.4]	[72.8]
Trend×Post	-33.7***	-5.45	-5.45	-8.67
	[1.84]	[5.20]	[6.40]	[6.09]
Treated×Post×Trend			-28.3***	-24.4***
			[6.64]	[6.32]
Age				1.87***
				[0.22]
Time in Program				23.0***
				[0.30]
Observations	49,551	10,955	60,506	60,506
R^2	0.034	0.056	0.036	0.13
RMSE	635	498	613	584

Results

Table 3: Diffs in Diffs

	(1)	(2)	(3)	(4)
Specification	Diffs-in-diffs	Diffs-in-diffs with controls for borrower characteristics	Diffs-in-diffs with controls, month-year dummies	Diffs-in-diffs balanced sample, month-year dummies
Constant	242***	121***	-167.7***	-179.1***
	[15.4]	[15.8]	(722.1)	(39.4)
Treated	93.6***	-105***	-41.3***	176.8***
	[15.9]	[15.3]	(15.5)	(43.0)
Post	273***	197***	503.8***	735.4***
	[16.5]	[15.6]	(22.6)	(56.2)
Treated×Post	-92.7***	-157***	-209.1***	-357.4***
	[17.4]	[16.5]	(16.6)	(42.3)
Age		1.95***	2.0***	0.2
		[0.22]	(0.2)	(0.2)
Time in Program		23.9***	23.0***	30.2***
		[0.27]	(0.3)	(2.1)
Implied interest rate elasticity of borrowing	-0.40	-0.68	-0.91	-0.71
Pre-treatment mean LHS	329	329	329	651
Post-treatment mean LHS	515	515	515	1093
Observations	68,037	68,037	68,037	25,926
R^2	0.020	0.12	0.13	0.24

Conclusions

- *Safe*Save's balance sheets show that the interest rate increase coincided with break-even status, though timing is unclear.

- At the same time, our estimates show considerable sensitivity to the interest rate increase among borrowers.

- We estimate elasticities in the range of approximately -0.73 to -1.12, with our preferred estimate being at the upper end of this range.

Complementary methods: statistical control

- Example: Evaluation of Residential Energy Conservation Programs in Minnesota
  - Program Goal: Reduce energy use
  - But…lots of things influence energy use
    - Weather
    - Household size
    - Square footage of house
    - Number of floors
    - Age of house
    - Gas hot water
    - Gas prices
- Point of statistical control: Control for these "nuisance factors" in order to isolate impact of program
  - If you could RA households, all above differences would "wash out"
- Separate out program effects from "covariate" effects
- *In practice, run a regression*.

Caution when using statistical control

- Most means of exerting statistical control depend upon a linear relationship between the covariates and the outcome variables
- Perfectly measured covariates are rare
- May have not measured key covariates
- Ruling out alternative explanations after the fact via statistical control
  - "adjustment" is a slippery slope

Other designs: multiple pre-post test measures

- Sequence of data points measured typically at successive times spaced at uniform time intervals.
- Help identify patterns and persistence of effects.

Treatment	$O_1\ O_2\ O_3\ O_4$	X	$O_5\ O_6\ O_7\ O_8$
Comparison	$O_1\ O_2\ O_3\ O_4$		$O_5\ O_6\ O_7\ O_8$

Other designs: switching replications

R	O	X	O		O
R	O		O	X	O

Other designs: cohort designs

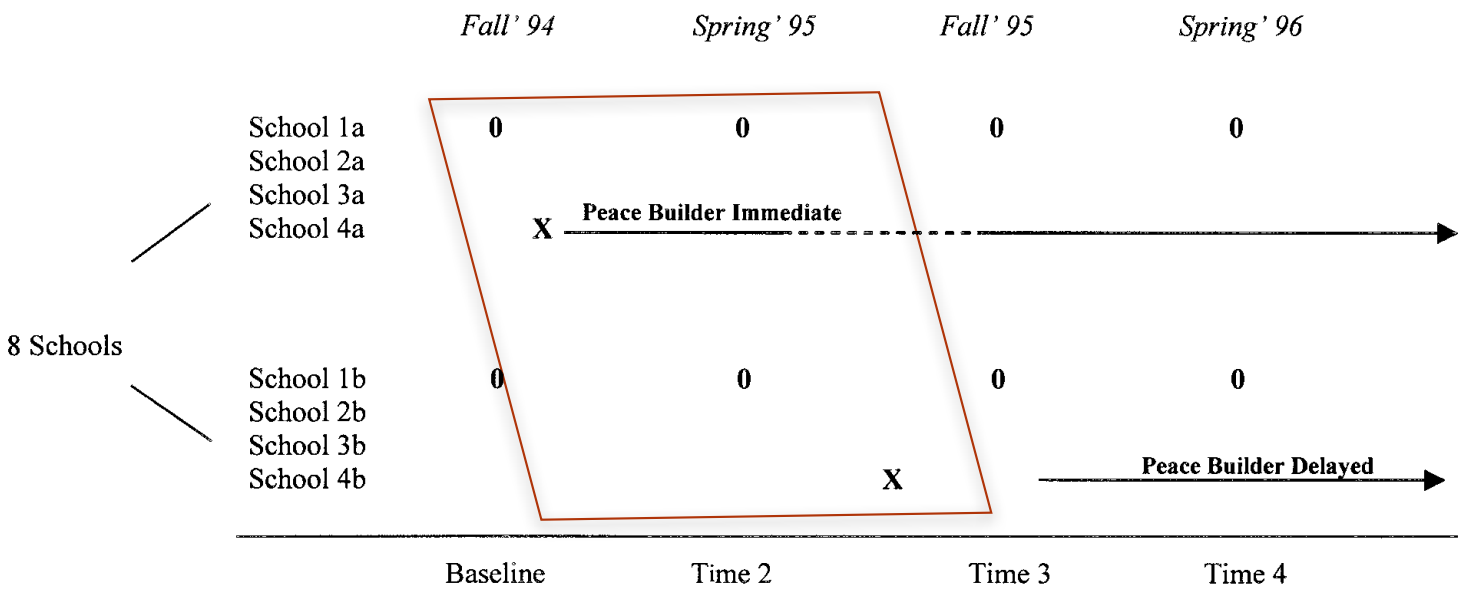
$N_{\text{Class of 2000}}$	O	O		
$N_{\text{Class of 2001}}$		O	X	O

Macy Cohort Study of Communication

	1999	2000	2001	2002
Comparison	2nd Year	3rd Year	4th Year	PGY 1
Class of 2001	**OSCE PRE TEST**	**Traditional Clerkships**	**OSCE POST TEST**	
Intervention	1st Year	2nd Year	3nd Year	4th Year
Class of 2002		**OSCE PRE TEST**	**Enhanced Clerkships**	**OSCE POST TEST**

OSCE=objective structured clinical examination

Delayed design



0 = Data Collection **X = Intervention**

Delayed design

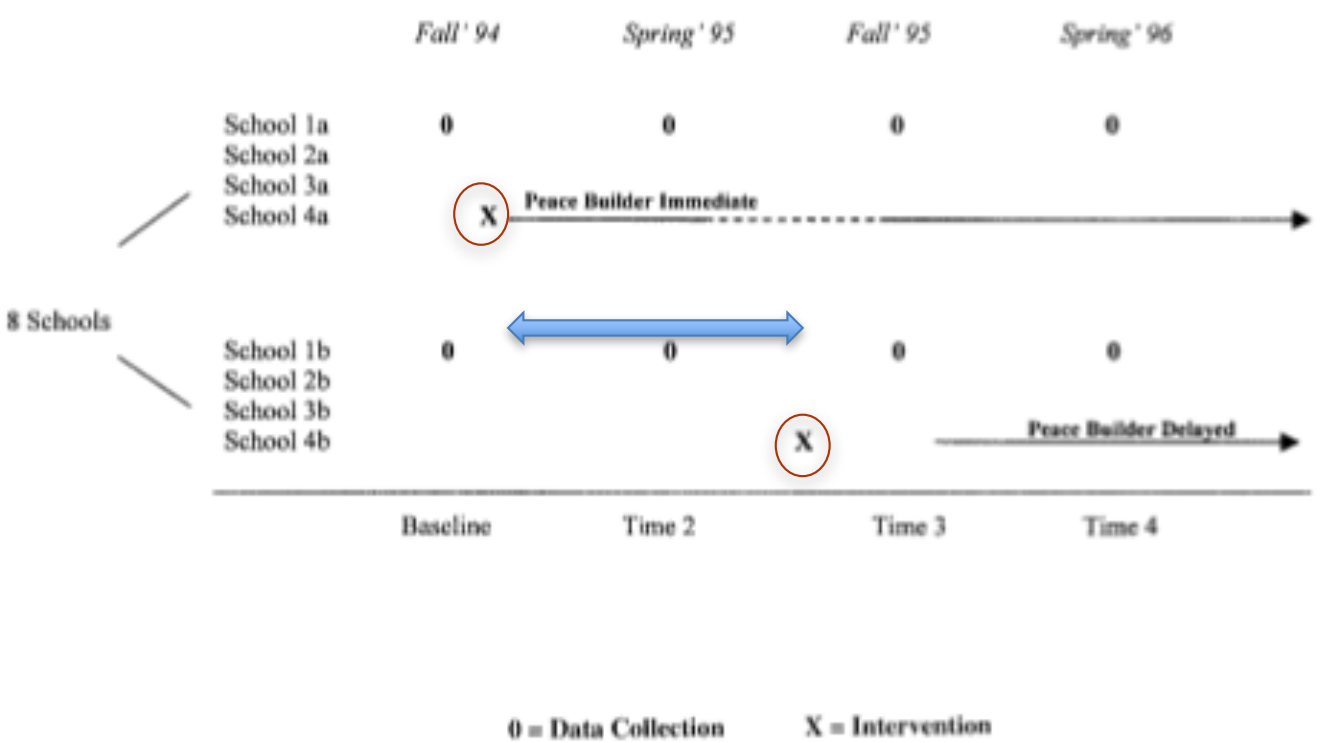


Figure 1. Overview of project design, data collection, and intervention schedule.

Miguel and Kremer deworming experiment

TABLE VIII

SCHOOL PARTICIPATION, SCHOOL-LEVEL DATA[a]

	Group 1 (25 schools)	Group 2 (25 schools)	Group 3 (25 schools)		
Panel A:					
First year post-treatment (May 1998 to March 1999)	*1st Year Treatment*	*Comparison*	*Comparison*	*Group 1 − (Groups 2 & 3)*	*Group 2 − Group 3*
Girls <13 years, and all boys	0.841	0.731	0.767	0.093***	−0.037
				(0.031)	(0.036)
Girls ≥13 years	0.864	0.803	0.811	0.057**	−0.008
				(0.029)	(0.034)
Preschool, Grade 1, Grade 2 in early 1998	0.795	0.688	0.703	0.100***	−0.018
				(0.037)	(0.043)
Grade 3, Grade 4, Grade 5 in early 1998	0.880	0.789	0.831	0.070***	−0.043
				(0.024)	(0.029)
Grade 6, Grade 7, Grade 8 in early 1998	0.934	0.858	0.892	0.059***	−0.034
				(0.021)	(0.026)
Recorded as "dropped out" in early 1998	0.064	0.050	0.030	0.022	0.020
				(0.018)	(0.017)
Females[b]	0.855	0.771	0.789	0.076***	−0.018
				(0.027)	(0.032)
Males	0.844	0.736	0.780	0.088***	−0.044
				(0.031)	(0.037)
Panel B:					
Second year post-treatment (March to November 1999)	*2nd Year Treatment*	*1st Year Treatment*	*Comparison*	*Group 1 − Group 3*	*Group 2 − Group 3*
Girls <13 years, and all boys	0.713	0.717	0.663	0.050*	0.055*
				(0.028)	(0.028)
Girls ≥14 years[c]	0.627	0.649	0.588	0.039	0.061*
				(0.035)	(0.035)
Preschool, Grade 1, Grade 2 in early 1998	0.692	0.726	0.641	0.051	0.085**
				(0.034)	(0.034)
Grade 3, Grade 4, Grade 5 in early 1998	0.750	0.774	0.725	0.025	0.049**
				(0.023)	(0.023)
Grade 6, Grade 7, Grade 8 in early 1998	0.770	0.777	0.751	0.020	0.026
				(0.027)	(0.028)
Recorded as "dropped out" in early 1998	0.176	0.129	0.056	0.120*	0.073
				(0.063)	(0.053)
Females[b]	0.716	0.746	0.648	0.067**	0.098***
				(0.027)	(0.027)
Males	0.698	0.695	0.655	0.043	0.041
				(0.028)	(0.029)

Separate samples

- Separate samples
  - Units use for “pre” not same as use for “post”
  - Customer satisfaction survey in agency
    - Agency has rotating client based
      - Sample clients at pre in March
      - Sample clients in September (Post)
      - Not necessarily the same clients but still compare pre-satisfaction levels with post-satisfaction levels
  - Limitation
    - Average change, not individual change
    - Non equivalence between samples can be a problem
      - Time differences (seasonal effects, history)
      - Sampling

Summary

Evaluation Design	*Start of Project (pretest) T_1*	*Project Intervention (continues on to end of project)*	*Midterm Evaluation or Several Observations during Implementation T_2*	*End of Project (posttest) T_3*	*Follow-up after Project Operating for Some Time (ex-post) T_4*	*The Stage of the Project Cycle at which Each Evaluation Design can Begin to be Used*
TWO STRONGEST EVALUATION DESIGNS						
1. *Comprehensive longitudinal design with pre-, midterm, post- and ex-post observations on the project and comparison groups.* This is the methodologically strongest design but also the most expensive and time-consuming. Permits assessment of the process of project implementation as well as trend analysis. Random assignment of subjects is rarely possible, so this and following designs normally use comparison groups selected to match the project group as closely as possible.	P_1 C_1	X	P_2 C_2	P_3 C_3	P_4 C_4	Start
2. *Pretest-posttest project and comparison groups.* For most purposes, this is the best available design when the evaluation can begin at the start of the project with a reasonable budget and no particular constraints on access to data or use of a comparison group.	P_1 C_1	X		P_2 C_2		Start
FIVE LESS ROBUST EVALUATION DESIGNS						
3. *Truncated longitudinal pretest-posttest project and comparison group design.* Project and comparison groups observed at two or more points during project implementation, but evaluation does not begin until the project is underway. Evaluation often starts as part of midterm review.		X	P_1 C_1	P_2 C_2		Midterm
4. *Pretest-posttest project group combined with posttest analysis of project and comparison group.* No baseline data collected on comparison group.	P_1	X		P_2 C_1		Start
5. *Posttest project and comparison groups.* No baseline or midterm data collected.		X		P_1 C_1		End
6. *Pretest-posttest project group.* No comparison group.	P_1	X		P_2		Start
7. *Posttest project group only.* No baseline project data or comparison group. This is the weakest QUANT design but very widely used because of limited cost and time requirements.		X		P_1		End

Key
T = time during project cycle
P = project participants
C = comparison group