Meeting 5

Causality and Impact Evaluation
Example
Cause and Effect
Threats to Internal Validity
The Magic of Random Assignment
Further Examples

Examples

- Microcredit is a revolutionary idea. Does it work?
- Millennium development goals: reasonable ideas. But do they work?

Examples (cont’d)

- We want to improve school attendance (in the belief that attendance leads to learning).
- We implement a program to provide incentives to children for attending (prizes based on an attendance target).
- How can we evaluate its impact?

Challenge of impact assessment: establishing causality

- Outcomes could be produced by something(s) other or in addition to than the program.
  - By selection: of who gets the program or not (e.g., neediest).
  - By omitted variable bias: correlation of treatment with other factors which in turn have an impact on the program (most motivated sign up).
  - By reverse causation: changes in the outcome cause people to select into or out of treatment (those who think they will benefit seek out treatment).
- Impact evaluation:
  - Establish the effect of program service receipt on relevant mediators, output, and outcome measures.
  - Estimate changes brought about by the program above and beyond those resulting from other processes and events affecting the phenomena of interest.
  - Estimate what their status *would* have been if they had *not* received program services (i.e., counterfactual state of affairs).
  - Alternative explanations for outcomes (x causes y; what else could cause y?)

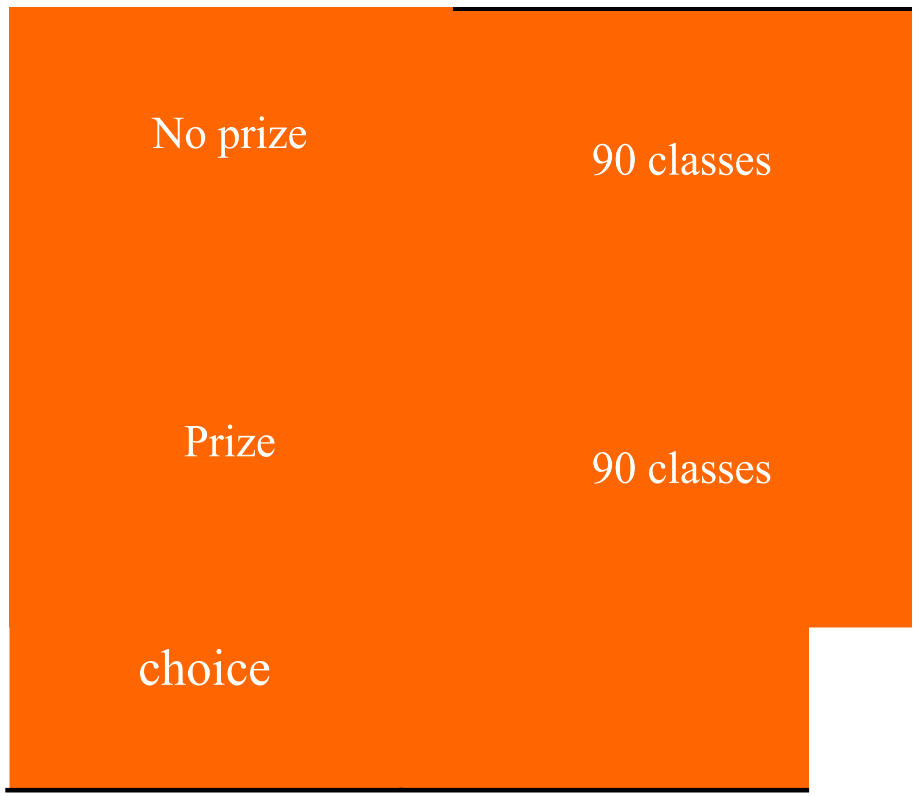
What is an experiment?

- An experiment refers to a randomized control trial.
- Traditionally done in labs where you ensure through the controlled setting of the lab that all subjects are treated identically, except a randomized treatment administered to some subjects vs a control treatment to others.
- Now also done in the field, where you can’t control background factors as much but where you can randomly assign the treatment.

Why do experiments work?

- By randomly assigning the treatment in a lab you guarantee that the only difference between treatment and control groups is the receipt of treatment *and* that this is not linked in any way to background characteristics or outcomes.
  - Kills off selection, omitted variables bias, and reverse causality.
  - Guarantees that the treatment and control groups are *on average* identical along both observable and unobservable dimensions.
    - Observable, e.g., prior income, health conditions, school…
    - Unobervable: e.g., motivation, risk attitudes, parents

A simple idea



Example

- But we are concerned that prizes can change the nature of people's motivation (intrinsic to extrinsic).
- Psychology suggests that external motivation can be more effective if people believe in their effort.
- Change the curriculum to emphasize malleable rather than fixed intelligence.
  - Fixed: I'm smart or not.
  - Malleable: if I study I can become smarter.

Example

	Fixed intelligence (standard) curriculum	Malleable intelligence ("treatment") curriculum
No prize	30 classes	30 classes
Prize	30 classes	30 classes
Prize with choice	30 classes	30 classes

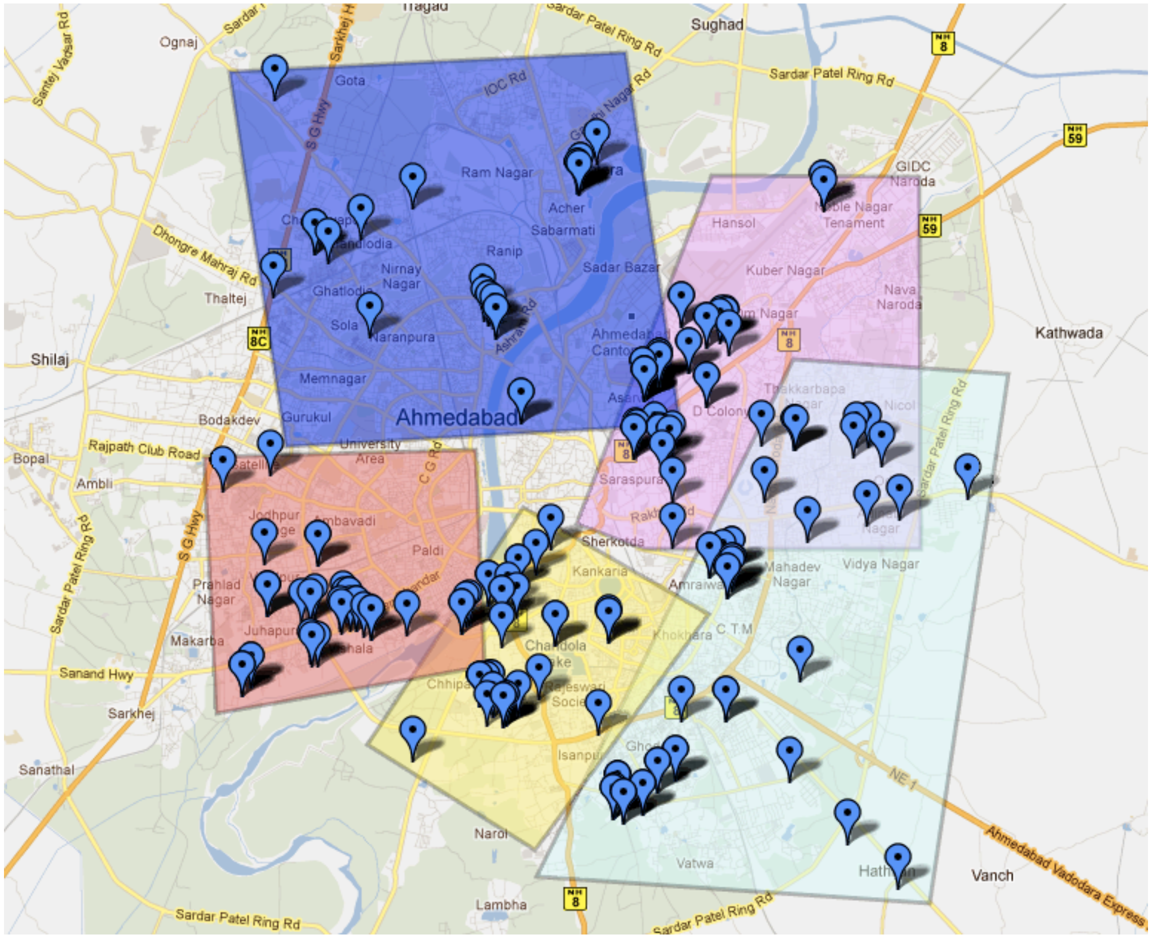
Example

- Now we also become interested in the idea that we should treat the parents too.
- And also that perhaps parents need to, but don't, believe that education is valuable.

		Malleable intelligence curriculum		
	Fixed intelligence curriculum	Classroom only	+ parent treatment	+ parent treatment + returns to education treatment
No prize	15 classes	15 classes	15 classes	15 classes
Prize	15 classes	15 classes	15 classes	15 classes
Prize with choice	15 classes	15 classes	15 classes	15 classes

Example

- But we have to worry about treatment interference (or spillovers)



Basic elements of research design

- Time: Randomization occurs before treatment; treatment occurs occurs before outputs / outcomes / impacts you want to measure.
- Programs or treatments: the alternative programs you will offer.
- Units (groups or individuals): subjects exposed to the treatments(s).
- Observations: What you observe post-treatment.

Unit of analysis

- Randomizing at individual level is usually best
  - More cases (true randomness)
  - More independence (not nested)
- Problem of randomizing intact groups
  - Fewer cases (less likely to be truly random)
  - Units within groups not independent
    - Ecological fallacy: Making inferences about individuals when it's really their ecology (institution, social group)

Design notation

- X = Program, Cause, Treatment
- O = Observation (Measure, Data)
- R = Random Assignment
- N = Non-Equivalent Comparison Groups
- Multiple Horizontal Lines = Groups
- Multiple Vertical Markers = Time Points

R	O	X	O	
R		O		O

What makes an evaluation flawed?

1. Fails to accurately measure the outcomes
   - If don't have good measures – how to know that the "constructs" really changed (or didn't change).
2. Fails to rule out alternative explanations
   - Internal Validity
     - Degree to which the design allows us to attribute the results/findings to the program
3. Fails to establish counterfactual
   - Comparing information about outcomes for program participants with estimates of what their outcomes would have been had they not participated
4. Fails to link outcomes to program
   - Rossi definition of Program Effect (Program Impact):
     - Change in target population that has been brought about by the program
     - If no program, the effect would not appear

- A well run experiment solves problems 2-4 (although a badly run experiments can create it's own problems of internal validity…)

How to interpret a negative impact

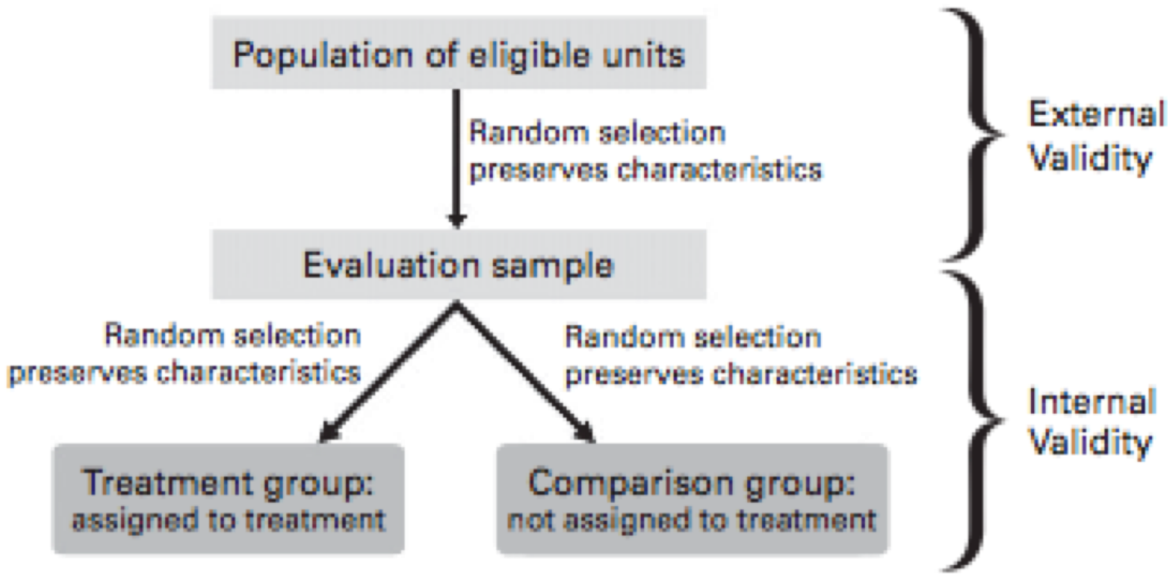
- Evaluation flawed.
- Program (impact) theory flawed.
- Process (implementation) theory flawed.
- Actual implementation flawed.
  - Good program, bad Evaluation vs
  - Good evaluation, bad program

Internal vs external validity

- Internal Validity
  - Accuracy of the experimental conclusions
    - Is the manipulated variables (the program) the only possible cause of the observed outcome?
    - Would the effects have occurred without the program?
- External Validity (Generalizability)
  - Inferences about whether the causal relationship holds over variation in persons, settings, treatments, and measurement variables
    - Do my results apply only to the people, settings, situations in my study? (SCC: units, treatments, observations, settings)
- Construct Validity
  - Inferences about the degree to which the units, treatments, observations, settings on which data are collected *accurately represent* the higher-order constructs they are supposed to represent.

Internal vs. External validity

Figure 4.2 Random Sampling and Randomized Assignment of Treatment



Threats to internal validity

TABLE 2.4 Threats to Internal Validity: Reasons Why Inferences That the Relationship Between Two Variables Is Causal May Be Incorrect

1. *Ambiguous Temporal Precedence:* Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.
2. *Selection:* Systematic differences over conditions in respondent characteristics that could also cause the observed effect.
3. *History:* Events occurring concurrently with treatment could cause the observed effect.
4. *Maturation:* Naturally occurring changes over time could be confused with a treatment effect.
5. *Regression:* When units are selected for their extreme scores, they will often have less extreme scores on other variables, an occurrence that can be confused with a treatment effect.
6. *Attrition:* Loss of respondents to treatment or to measurement can produce artifactual effects if that loss is systematically correlated with conditions.
7. *Testing:* Exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confused with a treatment effect.
8. *Instrumentation:* The nature of a measure may change over time or conditions in a way that could be confused with a treatment effect.
9. *Additive and Interactive Effects of Threats to Internal Validity:* The impact of a threat can be added to that of another threat or may depend on the level of another threat.

A well-run experiment solves 1-5, 6 (unless differential attrition), and 8 (somewhat) but not 7 & 9.

Further threats to internal validity

- Differential attrition
  - People may drop out of treatment differentially.
- Social experience of being in an experiment
  - Diffusion of Treatment (Contamination)
  - Compensatory Rivalry
  - Compensatory Equalization
  - Resentful Demoralization
- Generalizability
  - Artificiality of Situation
    - Able to do RA
    - Process of doing RA/Experiment (changes program)
    - Enough controls (expensive – affects external validity)
    - When RA (representativeness of sample – external validity)

Internal validity: ruling out alternative explanations

- By design (random assignment)
- By preventive action (experimental design)
  - If worried about drop-outs, use incentives
  - If worried that control group will become resentful, provide alternative program
- By argument
  - Assess plausibility of alternative explanations
    - Using evidence from literature, previous studies, logic
  - A priori vs A posteriori
- By analysis
  - Statistically control for alternative explanations
    - Good measures of the right alternative explanations
    - Valid means of statistically controlling for them

Other strategies for enhancing internal validity

- Expand across time - measurements
  - Pretest
  - Posttests
- Expand, vary treatment
  - Add and remove
  - Partition into different levels/types
    - Sensitivity (Dosage)
- Expand measurements
  - More and better outcome measures
- Add groups

→ Applies to non-random assignment as well.

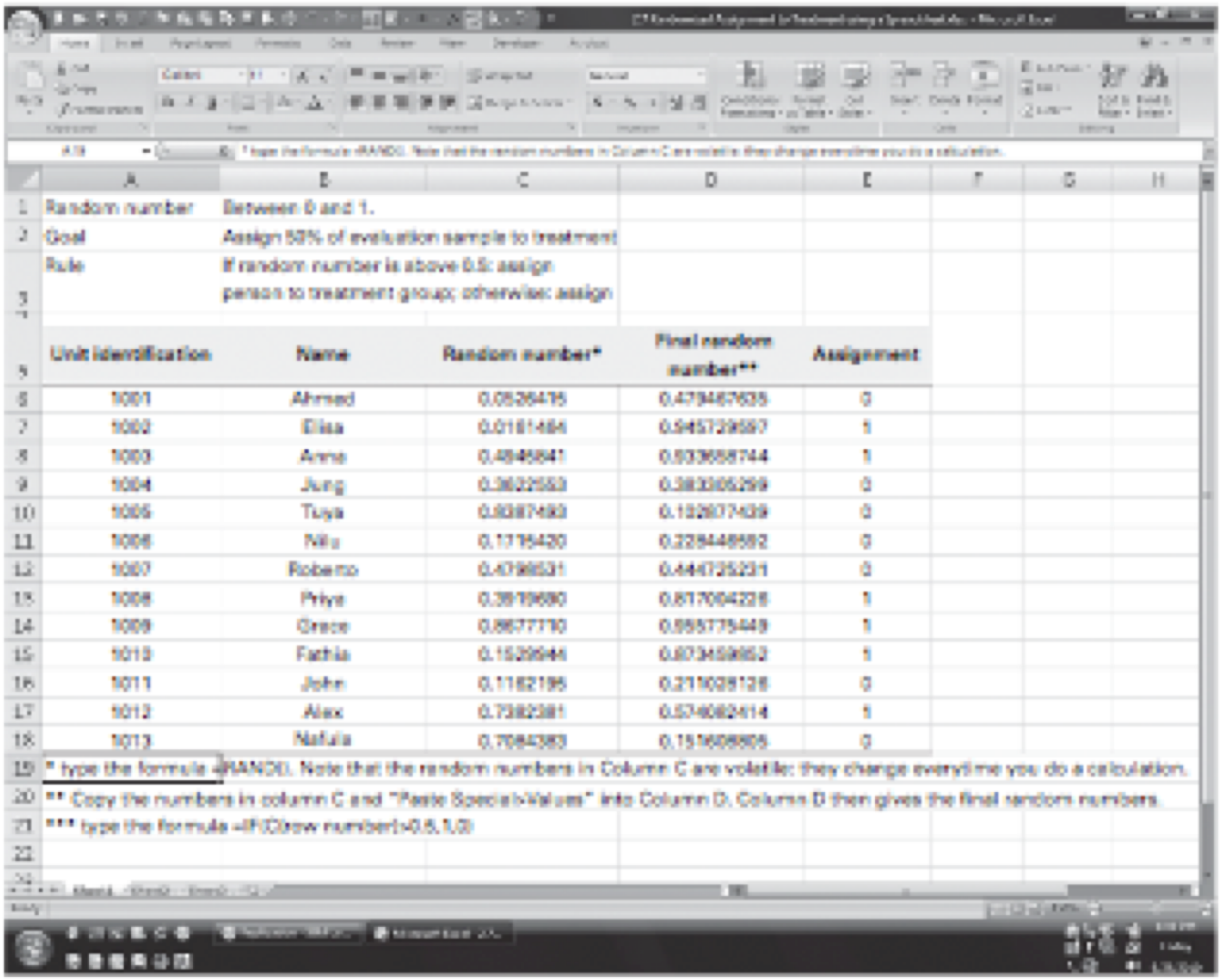
Further threats and solutions for internal validity

- Refusal rates (non-compliance): may need agreement to be randomly assigned – subjects may refuse or not comply.
  - Two options for analysis: intent-to-treat analysis (take intended assignment to treatment as the de facto treatment) and/or scale intent-to-treat effect by differential participation rate.
    - E.g., assign 50% to treatment and control.
    - Of treatment group 80% (or 40 percent of total sample) comply, likewise in the the control.
    - Just compare treatment vs control accepting that 20% of the treatment group was untreated and 20% of control group was treatment. E.g., average wages in intended treatment group ($800) – average wages intended control group ($200)=intent-to-treat effect ($600).
    - But we know that percent actually treated is 40% in the intended-to-treat group and 10% in the intended-not-to-treat.
    - Scale intent to treat effect by the difference: $600/(0.4-0.1)=$2000 is the effect of the actual treatment.

Further threats and solutions for internal validity

- Not allowed to randomly assign (e.g., for entitlement programs).
  - Can randomly "promote" the treatment among a random set of individuals, and not promote it among others.
  - Will work if take-up of the entitlement program is <100%.
  - But then analysis is like non-compliance – take into account differential participation in treatment with and without random promotion.

How do you actually randomize?



Random number	Between 0 and 1.
Goal	Assign 50% of evaluation sample to treatment
Rule	If random number is above 0.5: assign person to treatment group; otherwise: assign

Unit identification	Name	Random number*	Final random number**	Assignement
1001	Ahmed	0.0526415	0.479467635	0
1002	Elisa	0.0161464	0.945729597	1
1003	Anna	0.4846841	0.933658744	1
1004	Jung	0.3622553	0.383305299	0
1005	Tuya	0.8387493	0.132877439	0
1006	Nilu	0.1715420	0.228448582	0
1007	Roberto	0.4798531	0.444725231	0
1008	Priya	0.3819680	0.817004228	1
1009	Grace	0.8677710	0.555775449	1
1010	Fathia	0.1529944	0.873459852	1
1011	John	0.1162196	0.211028128	0
1012	Alex	0.7382381	0.574082414	1
1013	Nafula	0.7084383	0.151608805	0

* type the formula =RAND(). Note that the random numbers in Column C are volatile: they change everytime you do a calculation.

** Copy the numbers in column C and "Paste Special>Values" into Column D. Column D then gives the final random numbers.

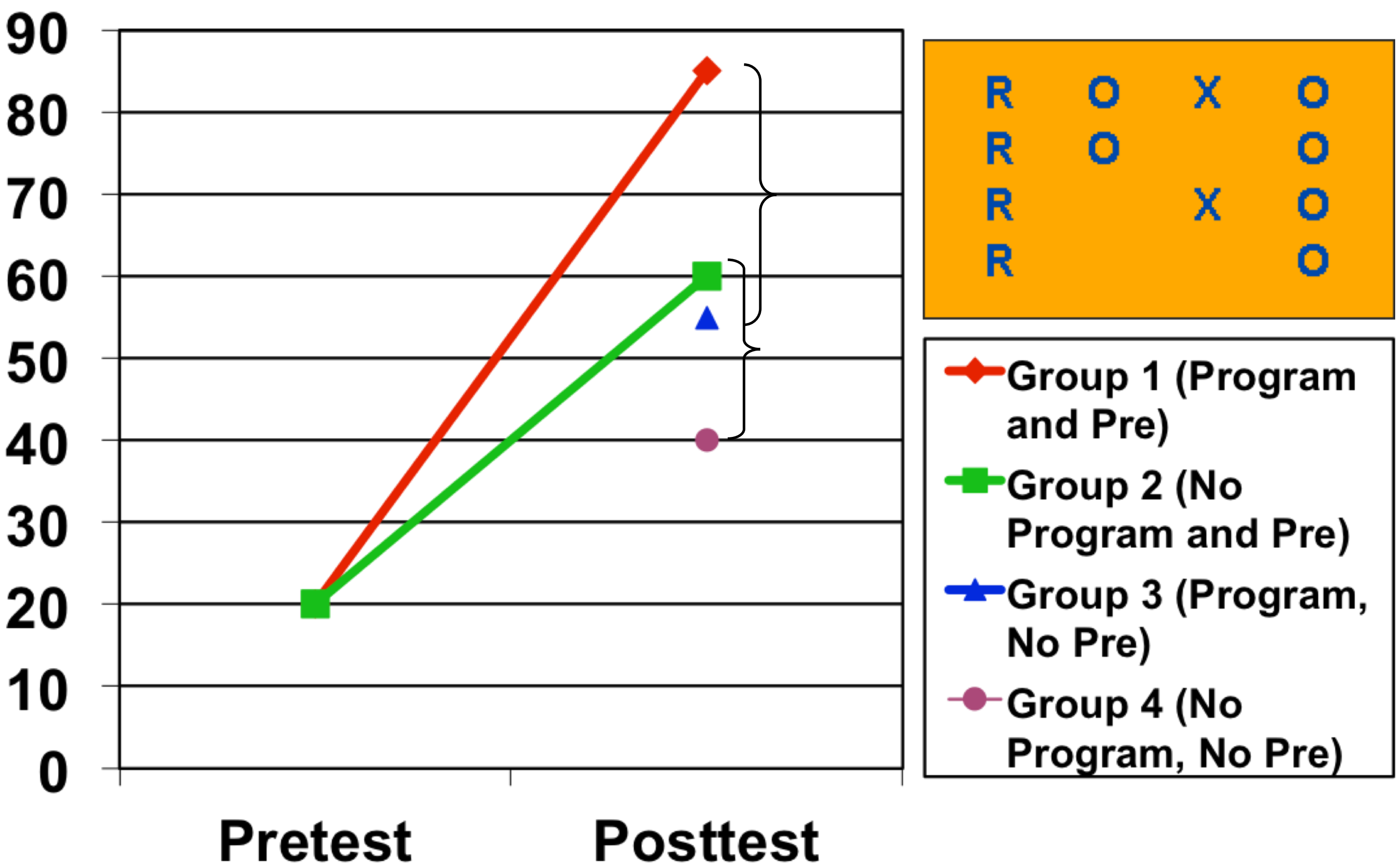
*** type the formula =IF(C{row number}>0.5,1,0)

Randomization check

- If you are analyzing the data, then check that the randomization worked. How?
  - Confirm that all measurable variables are balanced across treatment and controls groups.
    - And hope the same is true for the unmeasured…
- If you are the evaluator / designing the experiment:
  - Confirm that your proposed randomization balances pre-treatment / baseline characteristics.
    - What to do if it does not? Re-randomize or ex post adjustment.

Examples

Solomon four group design



Field and Pande

- Most microcredits require a steady flow of repayments. In theory flexible repayment schedules are better for the client (can time payments efficiently), but MFI’s claim that steady repayment imposes financial discipline.
- But key incentive is probably the dynamic one: want to borrow again.

The experiment

- Field and Pande randomize whether borrowers had weekly payments or monthly payments (but still weekly meetings with the group).
- Otherwise classic Grameen-type loan (joint liability, weekly group meetings, women).

Results

Table 1: Repayment Schedule and Loan Default

	Full loan repaid					
	within 60 weeks		within fifty six weeks		within fifty four weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
Weekly payment	-0.012	-0.016	-0.009	-0.013	0.011	0.010
	(0.022)	(0.022)	(0.022)	(0.023)	(0.028)	(0.029)
Monthly payment, weekly meeting	-0.005	-0.005	-0.012	-0.012	-0.042	-0.038
	(0.014)	(0.014)	(0.017)	(0.017)	(0.040)	(0.040)
Control variables	No	Yes	No	Yes	No	Yes
Observations	1017	1005	1018	1006	1028	1016
Mean value, monthly payment, monthly meeting	0.987		0.985		0.964	
	(0.112)		(0.122)		(0.185)	

Results

Table 2: Repayment Schedule and Client Delinquency

	Ever late payment		Average number of days past due		Rate of absence at meetings	
	(1)	(2)	(3)	(4)	(5)	(6)
Weekly payment	0.017	0.016	0.012	0.011	-0.0003	-0.0003
	(0.013)	(0.012)	(0.011)	(0.011)	(0.0003)	(0.0003)
Monthly payment, weekly meeting	0.010	0.010	0.011	0.013	-0.0006	-0.0007
	(0.011)	(0.011)	(0.018)	(0.021)	(0.0006)	(0.0007)
Control variables	No	Yes	No	Yes	No	Yes
Observations	966	966	966	966	966	966
Mean value, monthly payment, monthly meeting	0.0081		0.009		0.0005	
	(0.0045)		(0.0070)		(0.0005)	

Group liability

- Gine and Karlan tackle group liability.
- Look at a bank in the Phillipines that took away this feature.
- Treatment is some exposure to individual liability.

Successful randomization

Table 1: Summary Statistics

	All	Control	Treatment	p-value on t-test of difference: (2) - (3)	Treatment Wave 1	Treatment Wave 2	Treatment Wave 3	p-value on F-test for (5), (6) and (7)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Center Performance, pre-intervention (Aug 2004)								
Total number of active accounts	20.224	20.262	20.182	0.964	20.727	18.666	20.756	0.914
	(0.884)	(1.245)	(1.263)		(2.649)	(2.684)	(1.663)	
Number of new clients	3.159	3.641	2.644	0.190	2.800	1.350	3.209	0.274
(May-Aug 2004)	(0.380)	(0.594)	(0.460)		(1.459)	(0.509)	(0.655)	
Number of dropout clients	1.603	1.551	1.658	0.802	1.000	0.700	2.256	0.124
(May-Aug 2004)	(0.211)	(0.212)	(0.374)		(0.298)	(0.179)	(0.612)	
Retention	0.904	0.900	0.909	0.685	0.944	0.949	0.883	0.282
(May-Aug 2004)	(0.012)	(0.017)	(0.016)		(0.019)	(0.017)	(0.024)	
Proportion of missed weeks over cycle	0.060	0.054	0.068	0.332	0.113	0.054	0.063	0.264
(May-Aug 2004)	(0.007)	(0.009)	(0.011)		(0.049)	(0.016)	(0.013)	
Pastdue (maturity) / Scheduled total amortization due (in 100s)	0.092	0.000	0.193	0.258	0.005	0.329	0.000	0.397
	(0.085)	(0.000)	(0.178)		(0.005)	(0.304)	(0.000)	
Pastdue (30d) / Scheduled total amortization due (in 100s)	0.001	0.000	0.001	0.298	0.005	0.000	0.000	0.082
	(0.001)	(0.000)	(0.001)		(0.005)	(0.000)	(0.000)	
Pastdue (90d) / Scheduled total amortization due (in 100s)	0.000	0.000	0.000	--	0.000	0.000	0.000	--
	(0.000)	(0.000)	(0.000)		(0.000)	(0.000)	(0.000)	
Total loan amount	122,922.4	124,142.9	121,590.9	0.853	110,636.4	108,500.0	130,377.8	0.771
	(6868.4)	(10580.5)	(8616.4)		(17828.1)	(15613.8)	(12075.5)	
Average Loan size	6,033.2	5,996.1	6,073.7	0.806	5,196.8	6,030.0	6,308.5	0.425
	(157.5)	(220.6)	(226.2)		(473.2)	(410.0)	(312.4)	
Number of active centers, August 2004	161	85	76		11	21	44	
Number of centers in the sample	169	88	81		11	24	46	
B. Individual-level Performance, pre-intervention (Aug 2004)								
Proportion of missed weeks over cycle	0.062	0.059	0.065	0.324	0.083	0.065	0.059	0.185
	(0.003)	(0.004)	(0.005)		(0.016)	(0.008)	(0.005)	
Indicator for having at least one missed week	0.483	0.467	0.501	0.190	0.343	0.557	0.537	0.000
	(0.013)	(0.018)	(0.019)		(0.040)	(0.045)	(0.024)	
Proportion of past due balance, at maturity date	0.080	0.040	0.125	0.439	0.000	0.062	0.184	0.674
	(0.055)	(0.022)	(0.115)		(0.000)	(0.055)	(0.184)	
Past due balance, 30 days past maturity date (binary)	0.001	0.000	0.001	0.286	0.000	0.008	0.000	0.010
	(0.001)	(0.000)	(0.001)		(0.000)	(0.008)	(0.000)	
Total excess savings	319,924.5	286,583.4	357,940.0	0.625	223,869.7	216,725.5	441,811.5	0.740
	(72780.0)	(82775.0)	(123967.1)		(74987.2)	(57842.1)	(197449.3)	
Loan amount	6,107.2	6,143.6	6,069.1	0.570	5,558.4	5,772.7	6,368.7	0.003
	(65.5)	(93.1)	(92.2)		(180.3)	(193.7)	(125.5)	
Number of active clients, August 2004	3,285	1,708	1577		298	394	885	

Standard errors in parentheses. In Panel A, the number of active centers is less than 169 in August 2004 because there are 8 centers that started after the first conversion and added to the sample. T-statistics reported in column (4) is the probability of (column (2) - column (3)) being zero. F-statistics in Column (8) is from a regression of the outcome variable of interest on a set of indicator variables for each of the treatment waves. The exchange rate at the time of the experiment was 52 pesos = US$1.

Gine and Karlan results

Table 2: Loan-level Impact on Default, Savings, and Loan Size by Conversion Waves
OLS

Dependent Variable:	Proportion of missed weeks	Indicator for having at least one missed week	Proportion of past due balance, at maturity date	Past due balance, 30 days past maturity date (binary)	Total excess savings	Loan Size
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Baseline clients						
Treatment	-0.010	-0.023	-0.128	0.001	-242.696	-643.713**
	(0.016)	(0.041)	(0.122)	(0.002)	(165.222)	(322.439)
Observations	14333	14333	14333	14333	14332	14333
R-squared	0.18	0.20	0.06	0.03	0.31	0.26
Mean of dependent variable	0.075	0.075	0.220	0.002	6844.599	6844.401
Panel B: New clients						
Treatment	0.000	-0.010	-0.001	-0.001	-342.842	-735.826***
	(0.010)	(0.036)	(0.002)	(0.003)	(255.235)	(215.034)
Observations	6049	6049	6049	6049	6046	6049
R-squared	0.02	0.05	0.01	0.01	0.04	0.05
Mean of dependent variable	0.069	0.385	0.008	0.006	5284.816	5284.345

Gine and Karlan results

Table 4: Center-level Performance

OLS, Probit

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Center performance						
Dependent variable:	Proportion of missed weeks	Pastdue (at maturity) / Scheduled total amortization due	Pastdue (30d) / Scheduled total amortization due	Pastdue (90d) / Scheduled total amortization due	Total loan amount	Average loan amount
Specification:	OLS	OLS	OLS	OLS	OLS	OLS
Treatment	-0.013	-0.487	-0.379	-0.330	8,194.497*	-156.631
	(0.008)	(0.347)	(0.344)	(0.345)	(4,552.822)	(166.569)
Mean of dependent variable	0.07	0.35	0.28	0.21	98387.23	5418.58
Observations	1907	1941	1941	1941	2507	2507
Number of centers	169	169	169	169	169	169
R-squared	0.05	0.01	0.01	0.01	0.22	0.20
Panel B: Entry and dropout decisions						
Dependent variable:	Active accounts	Retention rate	New accounts	Number of dropouts	Dissolved center	
Specification:	OLS	OLS	OLS	OLS	OLS	Probit
Treatment	2.974***	0.032*	1.487***	0.197	-0.013	-0.137*
	(0.608)	(0.017)	(0.399)	(0.275)	(0.016)	(0.078)
Mean of dependent variable	15.36	0.80	2.51	3.16	0.03	0.37
Observations	2507	2017	2017	2017	2017	169
Number of centers	169	169	169	169	169	
R-squared	0.25	0.29	0.07	0.19	0.07	

Angrist and Lavy: education example

- A program wants to improve high school matriculation rates (in Israel -- the bagrut) by paying students to take the matriculation exam.
- How can we evaluate this?
- Ideally, the simplest case is randomize some students to the incentive and not others.
- But program administrators might (did) object to this.
- So the scheme was more complex.

Angrist and Lavy: what they did

- Figure out who really needs the program, and assign them for sure.
- Don't offer it to those who clearly don't need it.
- Randomize the middle range.
- But even in the middle range you want those who need the incentive to be more likel to get it.

The design

- Estimate Logit regressions with information from the previous cohort of students: predict the probability of Bagrut certification as a function of **number of Bagrut subject tests they had taken previously** and their **maximum score on these tests**, denoted here by p1i for student i.
- The population of 1302 seniors enrolled in the 1999-2000 school year are enterred into three groups :
- **All** students with a very low probability of Bagrut attainment ($p1i<.053$) were offered the opportunity to earn a bonus. It was inexpensive and politically expedient to offer bonuses to this group, about 15 percent of enrolled seniors in the Southern cohort.
- Students with a very high probability of success were excluded; in particular, we did not offer bonuses to 612 students with $p1i>.66$, about half of seniors.

The design (cont’d)

- The remaining 491 students were potentially eligible.
- Treatment was assigned to these students as a function of family size and father’s education, with students of lower socioeconomic status more likely to be in the treatment group.
- Used the previous cohort of seniors to estimate the probability a student would obtain a Bagrut certificate as a function of family size and father’s schooling,denoted p_{2i}.
- Then randomly assign a high or low threshold to each student. Assigned the incentive if their $p_2 < q_{.22}$, never if $p_2 > q_{.7}$, and in between based on coin toss of Z.

$$T_{ij} = 1[p_{2i} < q_{.22}(j)(1-Z_i) + q_{.7}(j)Z_i]$$

where $q_{.22}(j)$ and $q_{.7}(j)$ are the .22 and .7 quantiles of the p_{2i} distribution in school j.

The design

Table 1: Experimental Design for the Pilot Demonstration

Range for p_{1i}	Range for p_{2i}	Threshold for p_{2i} Low $q_{.22}$	Threshold for p_{2i} High $q_{.7}$	Offered Bonus No	Offered Bonus Yes	Row Totals
		A. All-Treated Sample (p_{1i}<.053)				
$[0, q_{.15}]$		--	--	0	146	
		B. Eligible Sample (.053<p_{1i}<.67)				
$[q_{.15}, q_{.53}]$	$[0, q_{.22}]$	59	64	0	123	123
	$[q_{.22}, q_{.7}]$	127	125	127	125	252
	$[q_{.7}, 1]$	56	58	114	0	114
	Column Totals	241	248	242	247	489
		C. No-treated Sample (p_{1i}>.67)				
$[q_{.53}, 1]$		--	--	612	0	

Basic results

Table 3: Reduced Form Effects in the Pilot Experiment (Eligible Sample)

	All Eligible Pupils				Jewish Eligible Pupils	
	No Covariates	School Covs p_{2i}	School f.e. p_{2i}	p_{1i}, sex, School f.e., P_{2i}	No Covariates	School f.e. p_{2i}
Dependent Variable	(1)	(2)	(3)	(4)	(5)	(6)
Offered	0.521 (0.039)	0.531 (0.030)	0.535 (0.028)	0.535 (0.028)	0.503 (0.041)	0.526 (0.030)
Received Bagrut	-0.003 (0.043)	0.005 (0.042)	0.001 (0.042)	-0.017 (0.039)	0.013 (0.045)	0.014 (0.044)

Independent variable is being offered a high threshold (randomly assigned to Z=1). Dependent variable is whether you were offered the incentive or not (basically everyone with high threshold - half the sample - is offered the incentive) and then whether you eventually matriculated.

Table 4: Results by Sex in the Pilot Experiment

	All eligibles		Random-assignment Sample		No-first-stage Sample	
Dependent Variable	Boys (1)	Girls (2)	Boys (3)	Girls (4)	Boys (5)	Girls (6)
Offered Bonus	0.514 (0.046)	0.540 (0.037)	1	1	0.047 (0.056)	0.057 (0.053)
Received Bagrut	-0.149 (0.063)	0.118 (0.056)	-0.130 (0.097)	0.080 (0.078)	-0.175 (0.089)	0.133 (0.085)
N	200	289	104	148	96	141