Introduction to Evaluation Research(ers) for Journalists

Presented by Angeli Kirk
EASST Launch Conference
Kampala, Uganda
July 12-14, 2012

What is Impact Evaluation (IE)?

- Impact evaluation assesses how a social program affects the wellbeing or welfare of individuals, households, communities, or businesses.

- Impact evaluation aims to identify effects that can be *attributed* to the intervention being evaluated, separate from any effects from other observed or unobserved factors – *causality*.

Impact Evaluation: for what and for whom?

- Different stakeholders have different objectives, and this will influence what and how to evaluate.
  - IE for accountability – for the financiers (donors / government): what impact do you get for your money?
  - What do you get which is *due* to the program?
    - Huge demand for this sort of evaluation
  - Which type of program achieves the most impact within given budget constraints?

Impact evaluation: for what and for whom?

- Impact evaluation for designing / improving programs for program management:
  - Less of whether a program works or not but more how to improve some components.
  - Examples:
    - Should credit be given to individuals or to groups?
    - Are bi-weekly loan repayments necessary?

Impact evaluation: for what and for whom?

- Impact evaluation for knowledge accumulation – for academics, policymakers, and the world
  - Potential research areas
    - Role of incentive payments for service providers (health, education, agricultural extension)?
    - How important credit constraint and risk aversion are relative to each other in explaining low adoption rates?
    - Has privatization of “public” services negatively affected the poor?

Session Goals

- Familiarity with key impact evaluation vocabulary and concepts
- Familiarity with basic impact evaluation approaches

- Interest to learn more!

Key Concepts & Terminology

- Development
  - Growth, GDP
- Counterfactual, Comparison group
- Baseline, Endline
- Observable, Unobservable characteristics
- Bias
  - Selection bias, Omitted variable bias
  - Spillovers and contamination
- Significance
  - Statistical, Practical
- Internal, External Validity

Development

- Multidimensional
    - Growth, productivity
    - Living standards
    - Health outcomes
        - Infant mortality
        - Life expectancy
    - Education
        - Years of schooling
        - Test scores
    - Other

Growth & GDP

- Growth: Increases in goods and services produced
  - That is, increases in GDP
- Gross Domestic Product (GDP): Official value of all goods and services produced within a country over one year
- GDP per capita
  - National GDP divided by population
  - Not a measure of standard of living, or average income
    - Wealth distribution
    - Non-market transactions
    - Underground economy

Which Development Impacts?

- *Outcomes of interest* depend on purpose of the intervention!
- Wellbeing at the individual level can be captured by income & consumption, health outcomes, or both
- At the community level, poverty levels or growth rates may be appropriate, depending on the question.
- Sometimes we may use more easily-measurable characteristics as *proxies* for the wellbeing characteristics we truly care about.
    - School attendance as education, human capital
    - Daily meals, vaccinations as health, human capital

Impact

- Impact: The difference between state of the world in the presence of an intervention compared to what the state of the world *would have been in the absence* of the intervention.

Counterfactual

- Counterfactual: What would have happened to the participants in a program in the absence of the intervention
  - We cannot observe someone with and without the program *at the same moment in time*
  - The counterfactual cannot be observed from the treatment group; can only be inferred from a comparison group.

Comparison group

- Comparison group: group drawn from the population not assigned to participate in the intervention
  - Used to infer the counterfactual
  - Called the "control group" in a randomized experiment
  - The treatment group is *compared* to the comparison group to *estimate* the true impact of the program
  - The closer the comparison group is to the true counterfactual, the more accurate our estimates of impact
- The goal: to find a comparison group that is like the treatment group in every way, *except* for exposure to the intervention/treatment/program

Observables, Unobservables

- Observable characteristics
  - Can be measured for a study
  - OR, Have been measured for a study
- Unobservable characteristics
  - Cannot be measured for a study
  - OR, Have not been measured for a study
  - If it could be measured but isn't, it's "unobservable"
  - If information is unavailable for the purpose of analysis, it's "unobservable"

Baseline, Endline data

- Baseline data: collected prior to program implementation
  - Collected for treatment and comparison
  - Ideally, baseline data show treatment and comparison groups to have the same characteristics
  - Used to show impacts for subgroups, according to pre-program characteristics

- Endline data: data collected after implementation
  - Collected for treatment and comparison
  - Used (with baseline) to estimate impacts

Statistical Significance

- Findings unlikely to have occurred due to random chance
  - Observed effect (difference between groups) is due to actual relationship between factors, not the result of sampling error
  - Acceptably small chance of a "false positive"
  - Reflects sample size, variability of observed traits

- Example. *Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities*
  - "Students who were dewormed had significantly higher school attendance rates."
  - It is unlikely that the estimated relationship was due to chance. Instead, we believe that there was a true positive relationship between deworming and attendance.

Practical Significance

- An observed effect or an observed difference is practically significant if its size is great enough “to matter” in the real world
- Example: A scholarship program increases average school attendance by an additional year vs. increases average test scores from 70% to 71%
- Sometimes called “economic significance”

Bias

- Bias: Statistical bias that causes systematic deviation of the estimates of impact from the true impact

- Selection bias: Estimation bias because of systematic nonprogram differences between treatment, comparison
  - Selection: Because we *select* treatment individuals who are unlike comparison individuals (targeting, program placement)
  - Self-selection: Because individuals who *self-select* into participation are unlike those do not (example: motivation)

- Unobservable characteristics are important here!

Bias, continued

- Omitted variable bias: Statistical bias that occurs when certain characteristics (often unobservable) – which correspond with a variable of interest (say, treatment) *and* also affect the measured outcome variable – are omitted from a regression analysis.
- Because they are not included as controls/covariates in the analysis, one incorrectly attributes the measured impact solely to the program.

Validity

- Internal Validity
  - The study findings are true within the tested sample.
  - *Randomized assignment* into treatment and control can help ensure internal validity.

- External Validity
  - The study findings are true for a wider population.
  - The use of a random sample of the population helps ensure external validity – that the findings represent the wider population from which the sample is drawn.

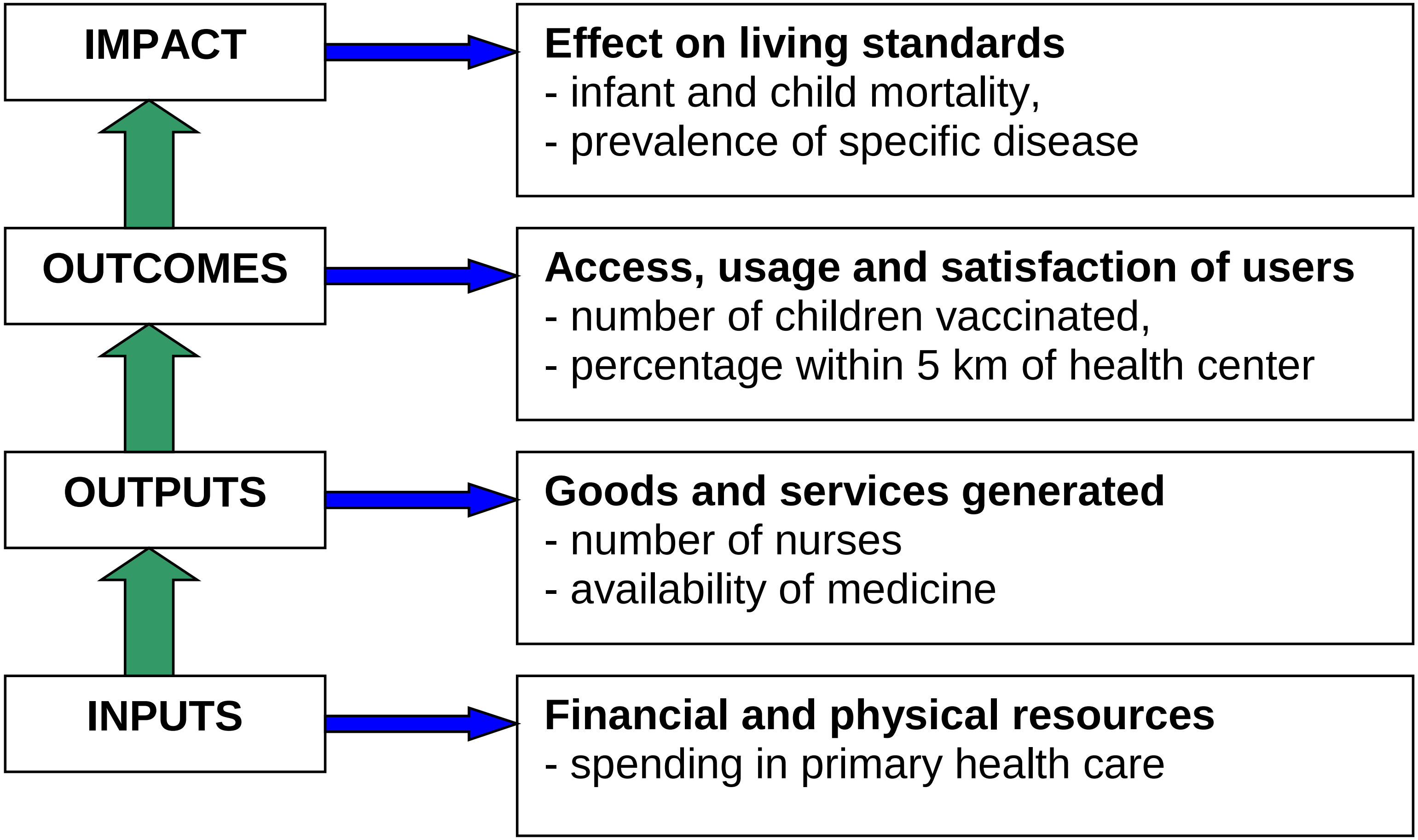
- Using a sufficiently large sample sample contributes to both internal and external validity.

IE Versus Other Tools

- The key distinction between impact evaluation and other monitoring & evaluation tools is the focus on discerning the impact of the program *from all other influencing factors* ("confounding effects") using econometric and statistical techniques.

- Note of caution: some of the vocabulary is the same, used differently!

Monitoring and IE



IMPACT

Effect on living standards
- infant and child mortality,
- prevalence of specific disease

OUTCOMES

Access, usage and satisfaction of users
- number of children vaccinated,
- percentage within 5 km of health center

OUTPUTS

Goods and services generated
- number of nurses
- availability of medicine

INPUTS

Financial and physical resources
- spending in primary health care

Example: providing fertilizer to farmers

- The intervention: provide fertilizer to farmers in a poor region of a country (call it region A)
    - Program targets poor areas
    - Farmers have to enroll at the local extension office to receive the fertilizer
    - Program ran from 2002 to 2004
    - We have data on yields for farmers in the poor region and another region (region B) for both years

Example

- We observe that the farmers we provide fertilizer to have a *decrease* in yields from 2002 to 2004

Did the program not work?

Example

- We observe that the farmers we provide fertilizer to have a *decrease* in yields from 2002 to 2004

Did the program not work?

- We don’t know. Further study reveals there was a national drought, and everyone’s yields went down (failure of the “reflexive comparison”)

Not IE: Before/After Comparisons

- Why not compare individuals before and after – the “reflexive comparison”?
- Problem: Other things change over time, too
  - Many factors affect crop yield in a given year
  - The rest of the world changes and you are not sure what was caused by the program and what by the rest of the world.

Example: providing fertilizer to farmers

- Next, we compare the farmers in the program region to those in another region. We find that our “treatment” farmers have a larger decline than those in region B.

Did the program have a negative impact?

Example: providing fertilizer to farmers

- Next, we compare the farmers in the program region to those in another region. We find that our “treatment” farmers have a larger decline than those in region B.

Did the program have a negative impact?

- We don’t know (program placement, selection bias)
  - Farmers in region B have better quality soil (unobservable)
  - Farmers in the other region have more irrigation, which is key in this drought year (observable)

Example: providing fertilizer to farmers

- Now consider the farmers within region A, where the fertilizer program was offered.
- Compare “treated” (enrolled) farmers with their (unenrolled) neighbors. We think the soil is roughly the same.
- We observe that treatment farmers’ yields decline by less than comparison farmers.

Did the program work?

Example: providing fertilizer to farmers

- Now consider the farmers within region A, where the fertilizer program was offered.
- Compare "treated" (enrolled) farmers with their (unenrolled) neighbors. We think the soil is roughly the same.
- We observe that treatment farmers' yields decline by less than comparison farmers.

Did the program work?

Example: providing fertilizer to farmers

- Now consider the farmers within region A, where the fertilizer program was offered.
- Compare “treated” (enrolled) farmers with their (unenrolled) neighbors. We think the soil is roughly the same.
- We observe that treatment farmers’ yields decline by less than comparison farmers.

Did the program work?

- We don’t know. (individual unobservables, self-selection bias) Farmers who went to register with the program may have more ability, and thus could manage the drought better than their neighbors while the fertilizer was irrelevant.

Example: providing fertilizer to farmers

- What if we observe no difference between the two groups?

Did the program not work?

Example: providing fertilizer to farmers

- What if we observe no difference between the two groups?

Did the program not work?

- We don’t know. What little rain there was may have caused the fertilizer to run off onto the neighbors’ fields. (spillover/contamination)

Control

- We need a control/comparison group that will allow us to attribute any change in the “treatment” group to the program (causality)

How to find a good comparison?

- We need a control/comparison group that will allow us to attribute any change in the "treatment" group to the program (causality)
- Instead of using before/after comparisons, we need to use comparison groups to proxy for the counterfactual
- Three core problems in finding suitable comparisons:
  - Programs are targeted
    - Recipients receive intervention for particular reason (selection bias)
  - Participation is voluntary
    - Individuals who participate differ in observable and unobservable ways (selection bias)
  - Spillovers and contamination
    - Individuals "near" the treatment group may be (partially treated)
- Hence, a comparison of participants and an arbitrary group of non-participants can lead to misleading or incorrect results

Counterfactual: Methodology

→We need a comparison group that is as identical in observable and *unobservable* dimensions as possible, to those receiving the program, and a comparison group that will not receive *spillover* benefits.

→Number of techniques:

- →Randomization as “gold standard”
- →Various matching techniques

1. Randomization

- Common in medical trials
- Individuals/communities/firms are randomly assigned to participation (“treatment”)
- With sufficient sample size, treatment and control groups have the same distribution of characteristics at baseline
- Counterfactual: randomized-out group
- *Advantages:*
  - Often called the “gold standard”: by design, selection bias is zero on average and mean impact is revealed
  - May be perceived as a fair process of allocation with limited resources

Randomization: Disadvantages

- *Disadvantages*:
  - Ethical issues, political constraints
  - Internal validity: people might not comply with the assignment (selective non-compliance)
  - External validity (generalizability): usually run controlled experiment on a pilot, small scale. Difficult to extrapolate the results to a larger population.

When to Randomize

- If funds are insufficient to treat all eligible recipients
  - Randomization can be a fair and transparent approach
- If program will be rolled out in separate phases due to administrative or budget constraints
  - If randomized into phases, early participants can be compared to late participants
- The program is administered at the individual, household or community level
  - Higher aggregation of implementation difficult
    - example: national-level policy, roads and infrastructure
- Program will be scaled-up
  - Learning what works is very valuable, merits randomizing some individuals out

Randomization in our example…

- Simple answer: randomize farmers within a community to receive fertilizer...
- Potential problems?
  - Run-off (contamination) so control for this
  - Take-up (what question are we answering)

2. Matching

- *Advantages:*
  - Does not require randomization
- *Disadvantages:*
  - Strong identification assumptions
  - Requires very good quality data: need to control for all factors that influence program placement
  - Requires significantly large sample size to generate comparison group

Matching in our example…

- Using statistical techniques, we match a group of participants with non-participants using characteristics like
  - Gender
  - Household size
  - Education
  - Experience
  - Land size
  - Rainfall to control for drought
  - Trends over time
- Any observable characteristics not affected by fertilizer

Matching in our example…two scenarios

- Scenario 1: We show up afterwards, we can only match (within region) those who got fertilizer with those who did not. Problem?
  - Problem: select on expected gains and/or ability (unobservable)

- Scenario 2: The program is allocated based on historical crop choice and land size. We show up afterwards and match those eligible in region A with those in region B. Problem?
  - Problems: same issues of individual unobservables, but lessened because we compare eligible to potential eligible
  - Now unobservables across regions

3.Regression discontinuity design

- Exploit the rule generating assignment into a program given to individuals only above a given threshold – Assume there is discontinuity in participation but not in counterfactual outcomes
- Counterfactual: individuals just below the cut-off who did not participate

Advantages:

- ▫ Delivers marginal gains from the program around the eligibility cut-off point. Important for program expansion

Disadvantages:

- ▫ Threshold has to be applied in practice, and individuals should not be able manipulate the score used in the program to become eligible.

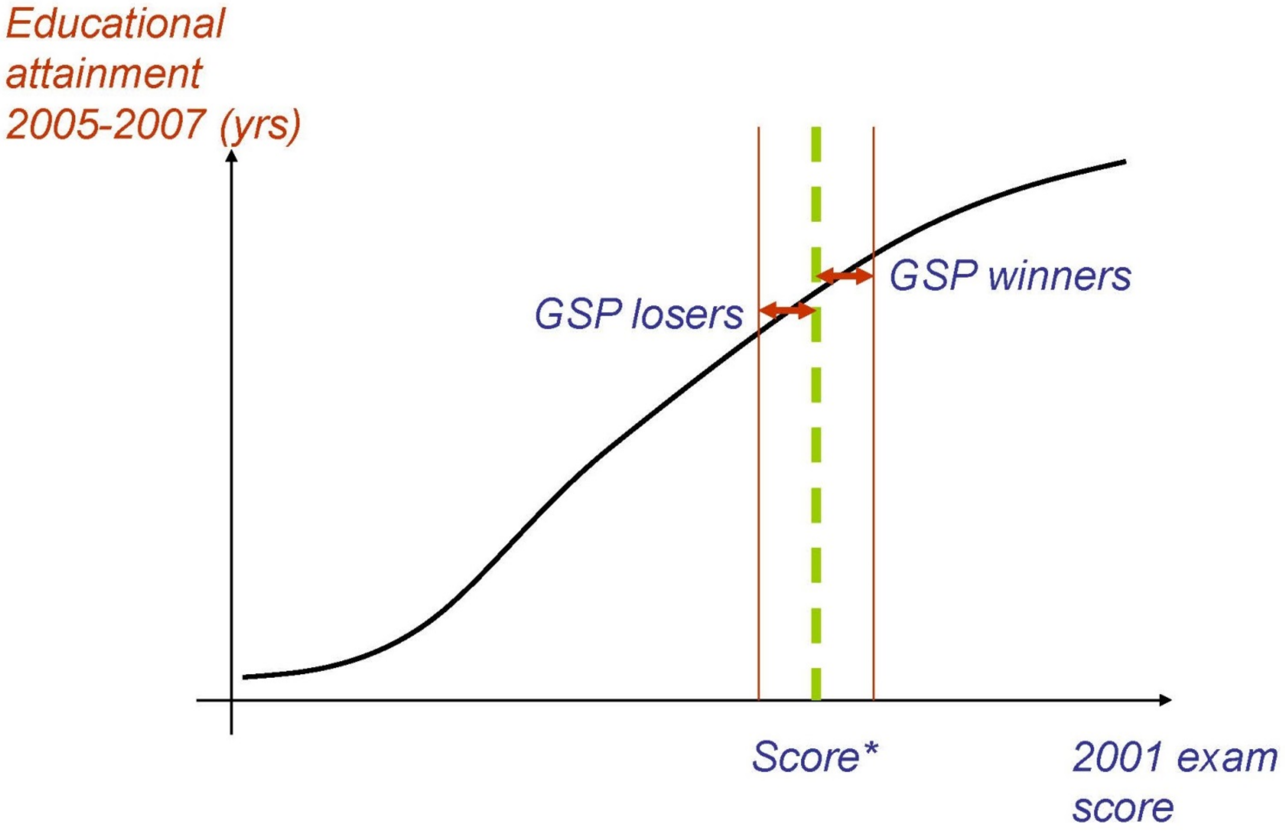
RDD in our example…

- Back to the eligibility criteria: land size and crop history
- We use those right below the cut-off and compare them with those right above…
- Problems:
  - How well enforced was the rule?
  - Can the rule be manipulated?
  - Local effect

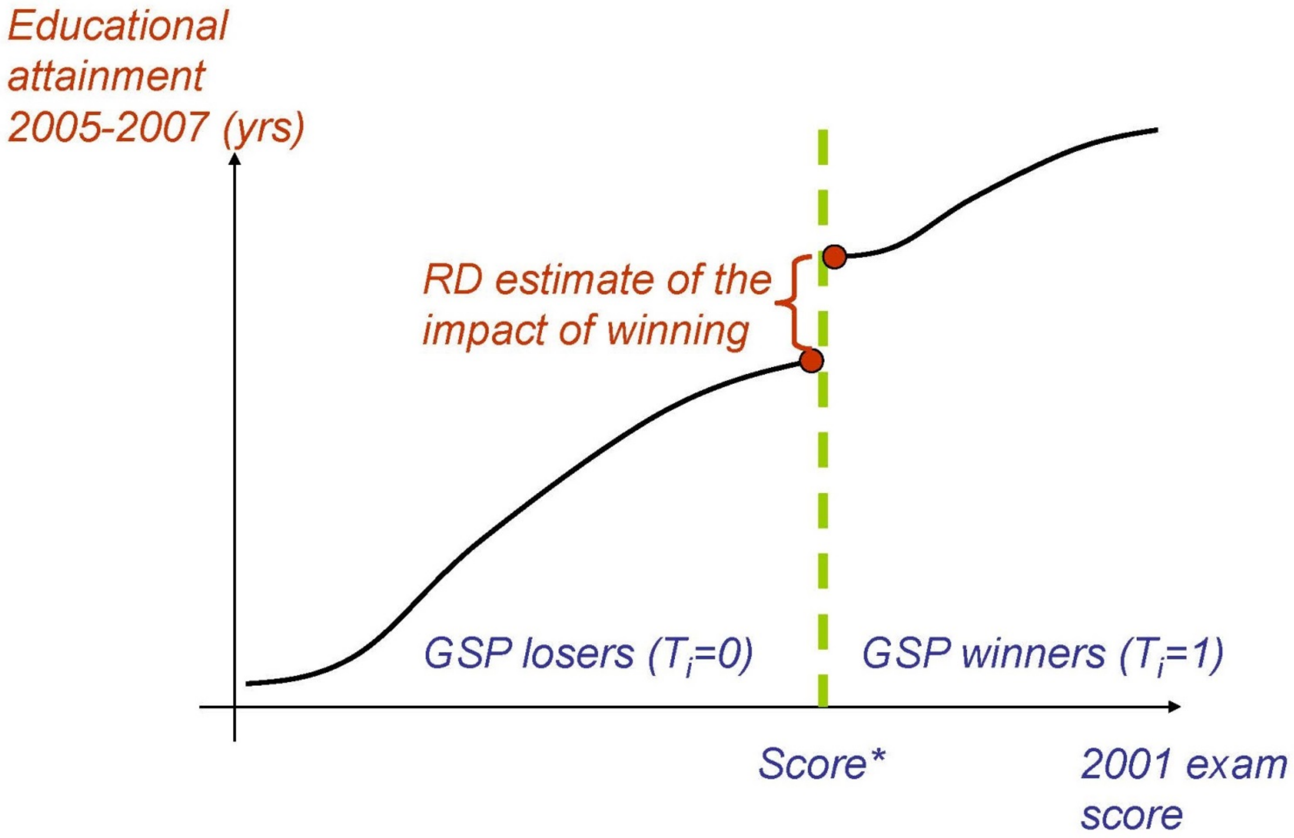
RD Example – Scholarship Program

- Girls Scholarship Program (GSP)
  - 2001-2002
  - Busia, Kenya
  - girls above threshold score received scholarship
    - were part of the treatment group
  - girls below threshold score did not receive scholarship
    - formed the control group

RD Example – Scholarship Program



RD Example – Scholarship Program


Educational attainment 2005-2007 (yrs)
RD estimate of the impact of winning
GSP losers (T_i=0)
GSP winners (T_i=1)
Score*
2001 exam score

Discussion example:
building a control group for irrigation

- Scenario: we have a project to extend existing reaches and build some new canal
- An initial analysis shows that farmers who are newly irrigated have increased yield… was the project a success?
- What is the evaluation question?
- What is a logical comparison group and method?