

# Performance Evaluation of Policies and Programmes

*Adam B. Jaffe*

*Director, Motu Economic and Public Policy Research*

Treasury/MBIE Seminar

27 August 2013

# Background

- Evidence-based policy, etc.
- Skepticism?



# The Problem

- A policy or a programme is like a new drug. We would like to know if it is effective, and how its effectiveness compares to alternatives.
- With a drug, it is not enough that the patient gets better. With a policy, it is not enough that the policy goal is met.
- Want to measure the **treatment effect**, i.e. how the state of the policy objectives compares to what it would have been without the policy.



# We'd like to know...

- Magnitude of impacts (“outputs” and “outcomes”)
- Magnitude of impacts relative to resources required (cost-effectiveness)
- Relative effectiveness of different instruments or approaches
- Relative effectiveness in different contexts (conditional cost-effectiveness)



# Examples

- Health service delivery modes
- Scholarships
- Tax subsidies
- Regulations
- Grant programs
- .....



# Analytical Issues

- Outputs and outcomes that are hard to measure
- Long and/or uncertain lags between action and outcomes
- Characterizing the unobserved “but for” world
  - **Selection bias in programme participation**
- Others I will not say much about:
  - Incremental versus average impact
  - General equilibrium effects



# Thought on metrics

- Quantify where possible, but...
- Non-quantifiable doesn't mean unimportant
- Multiple metrics
- Tradeoff between comparability and precision
- Almost always proxy or indicator rather than “true” variable
  - Measurement (random) error
  - Behavioral changes in response to evaluation
- Long/uncertain lags → ongoing evaluation



# Isolating the Treatment Effect

- Typically, start by comparing performance of treated group before and after the treatment
- Issues
  - Placebo effect
  - Regression to the mean
  - Sectoral trends
- Compare change in treated group to change in “control group”



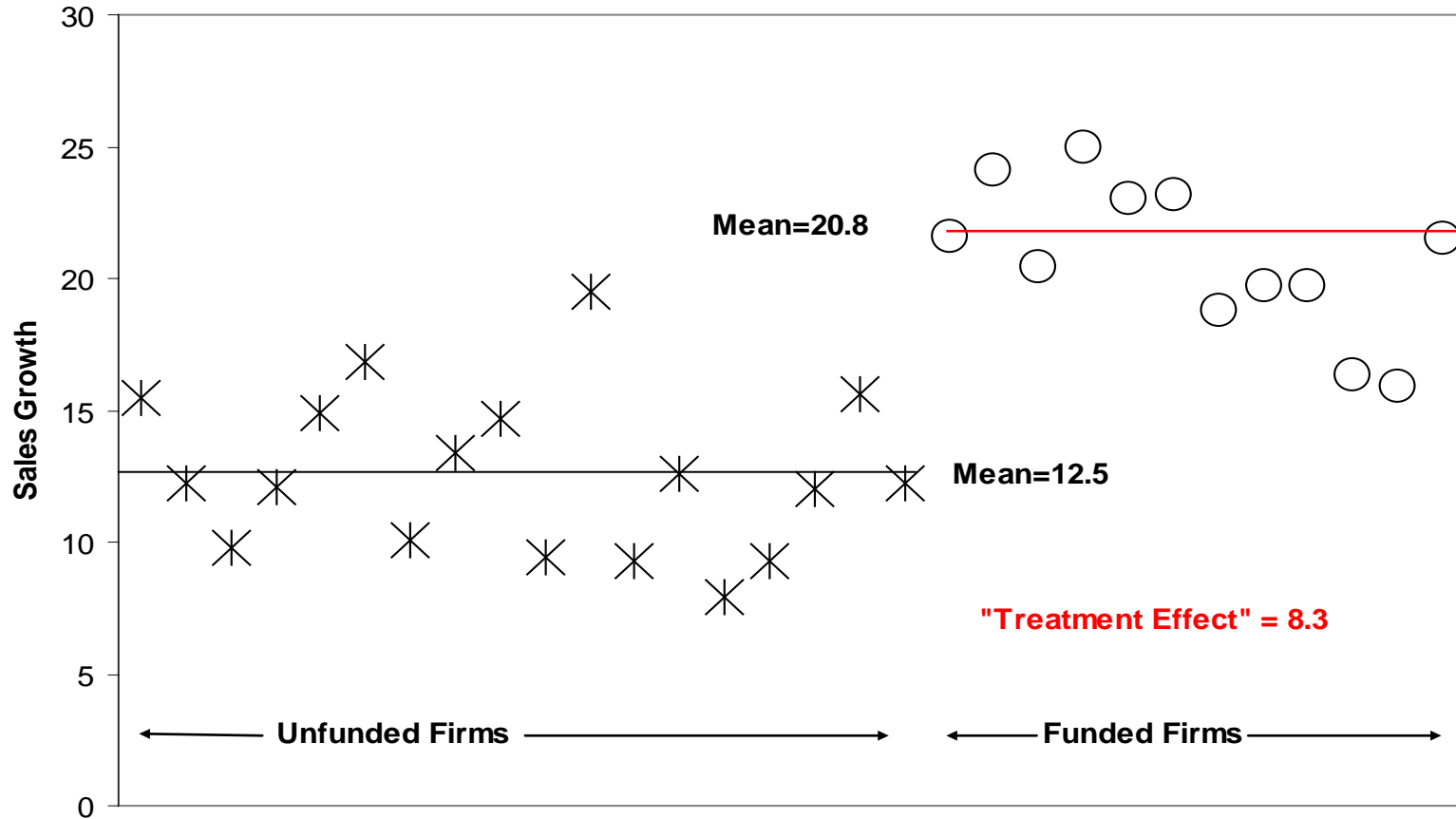


# “Difference in difference” approach

- “Gold Standard is DID with Random Assignment (“RA”) to treatment group and control group



# Hypothetical Comparison of Mean Sales Growth for Funded and Unfunded Firms Ignoring Selection Bias



# Selection Bias

- Frequently, government program provides assistance to some individuals or firms but not to others
- Makes those not provided assistance a natural control group, but...
- Programme targets are chosen on the basis of need (unemployed; under-achieving students), or expectation of success (scholarships; research grants)
- Creates selection bias in difference-in-difference analysis

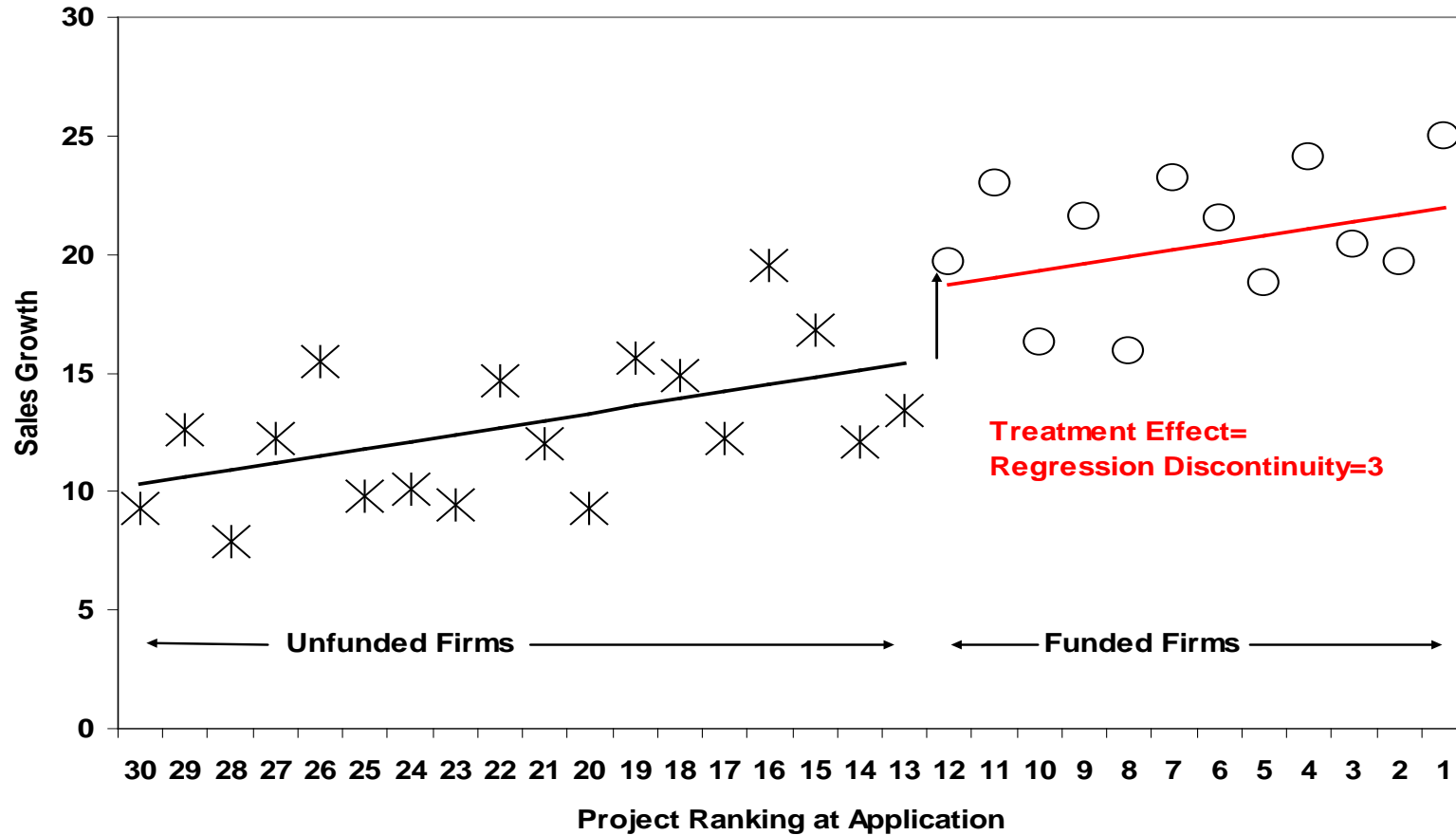


# Regression Discontinuity (“RD”) Approach to Selection Bias

- Retain information on ranking used to select individuals or firms for participation in the program
- Use this measure of qualification or need as regressor in explaining subsequent success of treated and untreated groups
- Dummy variable for program participation then captures treatment effect *after* controlling for selection effect



# Hypothetical Comparison of Mean Sales Growth for Funded and Unfunded Firms Controlling for Selection Bias via Project Ranking at Application



# Regression Discontinuity (“RD”) Approach to Selectivity Bias

- Statistically controls for the source of non-random difference between the treated and untreated groups
- Works for positive or negative selection effect
- Requires retention of information about criteria for selection
- Requires ability to measure success of both treated and untreated individuals/firms
- Note: if the selection criteria are not, in fact, correlated with success, then slope will be zero but RD measure of treatment effect is still unbiased

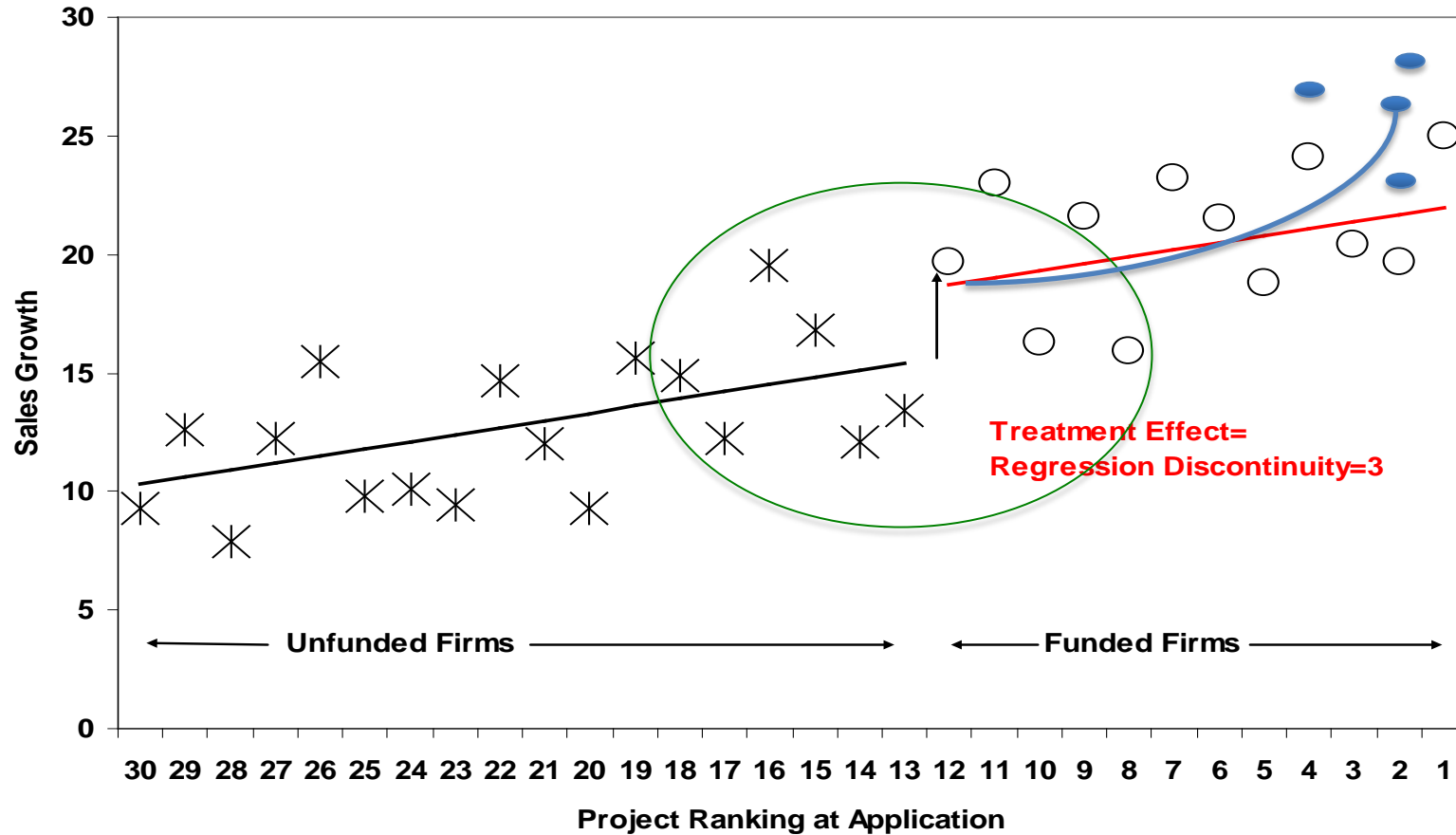


# RD versus Random Assignment

- Both approaches measure the average treatment effect for treated entities
- If the treatment effect were uniform for all entities, then RD reproduces the result of random assignment
- More likely, the magnitude of the treatment effect may be correlated with the selection measure
  - Most appropriate targets may get biggest boost; or
  - Decreasing returns may limit effect for most qualified
- Has implications for potential expansion of program to previously untreated group



# Hypothetical Comparison of Mean Sales Growth for Funded and Unfunded Firms Controlling for Selection Bias via Project Ranking at Application





# RD versus Random RA

- RA always produces unbiased estimate of average effect, but tells you nothing about the underlying variation in efficacy
- Note that in social settings, neither typically deals with placebo effect
- Both methods require tracking of untreated group; not clear which approach makes this easier



# Example of RD Approach

- “Reading First” was a billion-dollar program to introduce new pedagogy, new student evaluation measures, and specific teacher training methods to improve reading performance of 1<sup>st</sup>-3<sup>rd</sup> graders
- Schools were chosen for the program using a ranking index based on poverty rates and fraction of students reading below grade level
- Evaluation was carried out over three years in 248 schools, 125 of which were Reading First Schools



# RD Analysis of Impact of Reading First

Actual

# Public Research Programmes

- Need to track performance of unsuccessful applicants
  - Condition for eligibility to begin with?
  - System of identifiers combined with external data—  
StarMetrics approach
- Outputs and outcomes are hard to measure and subject to measurement response
- Routine/ongoing rather than episodic



# Concluding Thoughts

- Combination of faith and hard-to-measure outcomes
- Accept that some questions are not answerable:
  - Relative effectiveness across policies with incommensurable outcomes
  - Incremental versus marginal
  - GE effects
- Perfect should not be the enemy of good
- But a little knowledge is a dangerous thing
- Long lags as an advantage?



# Advert

## Science and Innovation Policy for New Zeland

Motu Public Policy Seminar

Wednesday 04 September  
Spectrum Theater, BP House

Veronica Jacobsen, Discussant

