The counterfactual

Howard White

International Initiative for Impact Evaluation

Write down a definition of impact evaluation

So, what is impact evaluation?

$$
\begin{aligned}
RR_{observed} &= RR_{adjusted} \\
&= RR_{adjusted} \times \frac{RR_{unadjusted}}{RR_{unadjusted}} \\
&= RR_{unadjusted} \times \frac{RR_{adjusted}}{RR_{unadjusted}} \\
&= RR_{unadjusted} \times \varepsilon_{statistical\ assumptions} \\
&= \frac{a^*_{14}/b^*_{14}}{e^*_{04}/f^*_{04}} \times \varepsilon_{statistical\ assumptions} \\
&= \frac{A_1/B_1}{A_0/B_0} \times \frac{A_0/B_0}{E_0/F_0} \times \frac{\alpha_{11}/\beta_{11}}{\gamma_{01}/\delta_{01}} \times \frac{\alpha_{12}/\beta_{12}}{\gamma_{02}/\delta_{02}} \times \\
&\quad \frac{\alpha_{13}/\beta_{13}}{\gamma_{03}/\delta_{03}} \times \frac{\alpha_{14}/\beta_{14}}{\gamma_{04}/\delta_{04}} \times \frac{\frac{a^*_{14}/b^*_{14}}{e^*_{04}/f^*_{04}}}{\frac{a_{14}/b_{14}}{e_{04}/f_{04}}} \times \varepsilon_{statistical\ assumptions} \\
&= RR_{causal} \times \varepsilon_{confounding} \times \varepsilon_{losses} \times \varepsilon_{sampling} \times \varepsilon_{nonresponse} \times \\
&\quad \varepsilon_{missing\ data} \times \varepsilon_{measurement} \times \varepsilon_{statistical\ assumptions} \cdot
\end{aligned}
$$

What is impact evaluation?

Impact evaluations answer the question as to what extent the intervention being evaluated altered the state of the world

= the … dicator with the interv… red to what it would have been in the absence …

We can see this

But we can’t see this

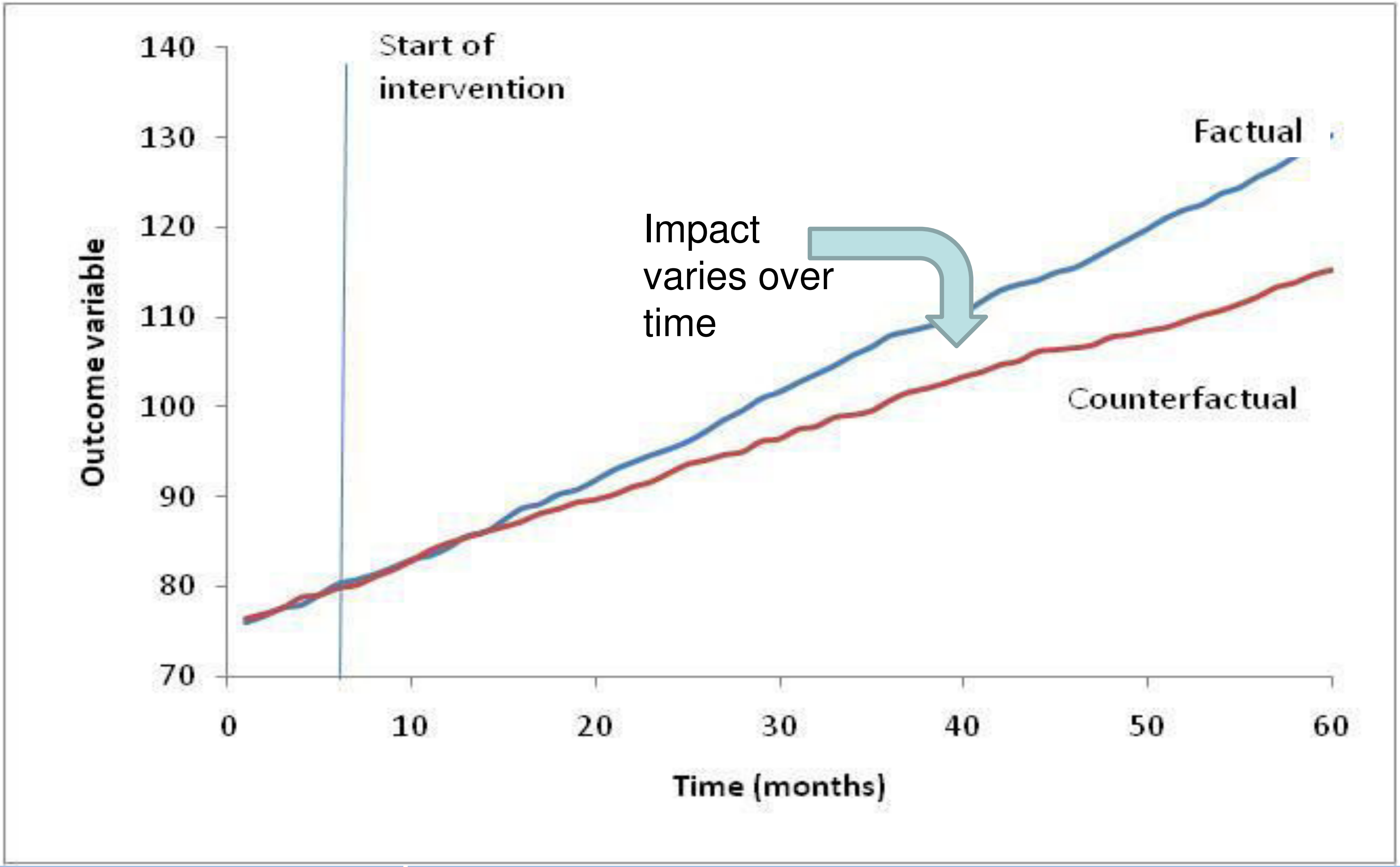
$= Y_t(1) - Y_t(0)$

So we use a comparison group

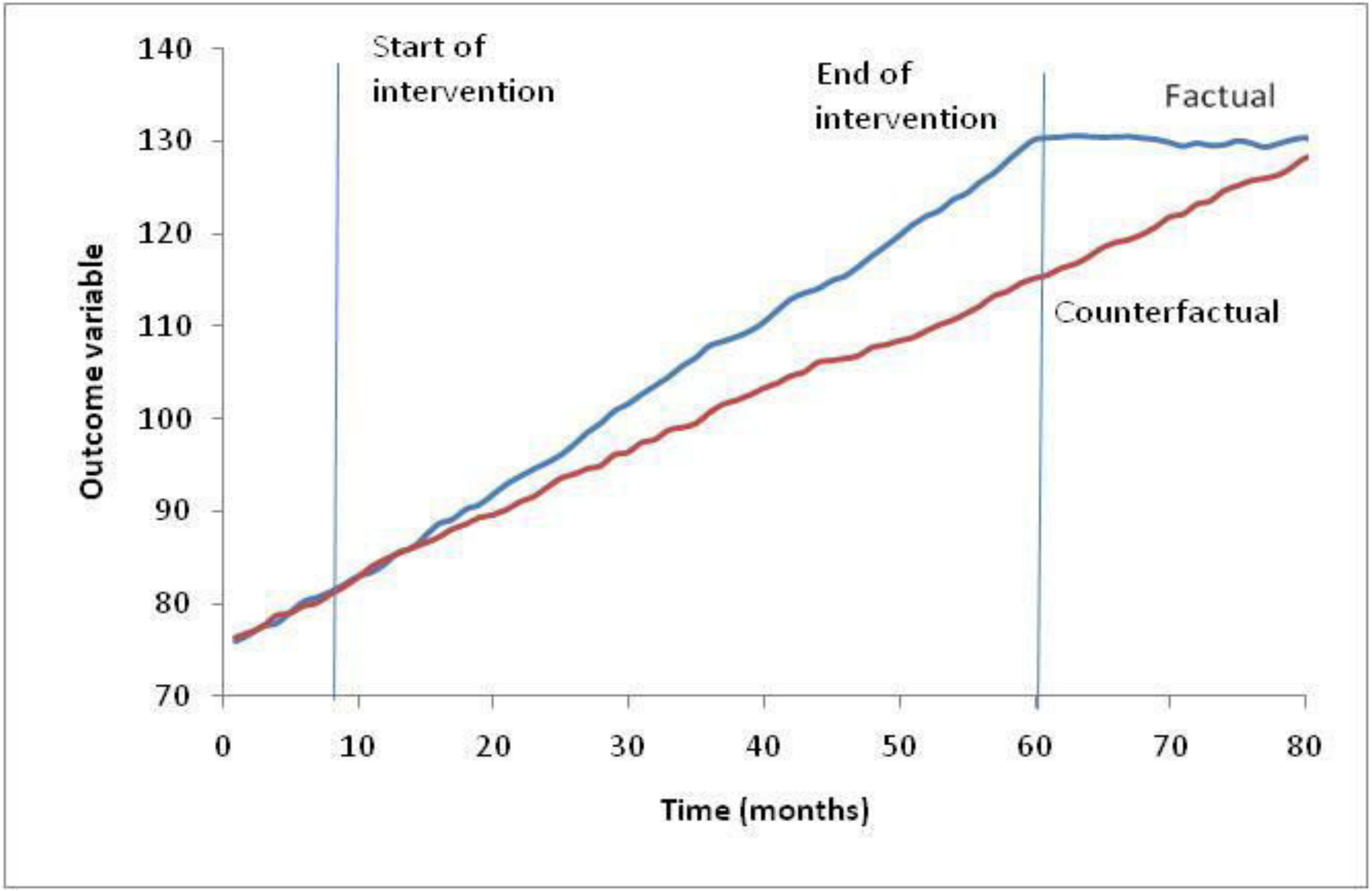
Terminology refresher

- Counterfactual
- Comparison group
- Control group

The attribution problem: factual and counterfactual



… and is it sustainable?



When to do an impact evaluation?

It all depends on how long you need to see an impact:

- Supplementary feeding for pregnancy weight gain – less than nine months
- Learning outcomes, lifetime earning?
- What has been the impact of the French revolution?"It is too early to say" *Zhou Enlai*
- Yahoo – randomly assign 100,000 hits to a modified design of home page and get results in <u>one hour</u>

- So where does the counterfactual come from?
- Most usual is to use a comparison group of similar people / households / schools / firms…

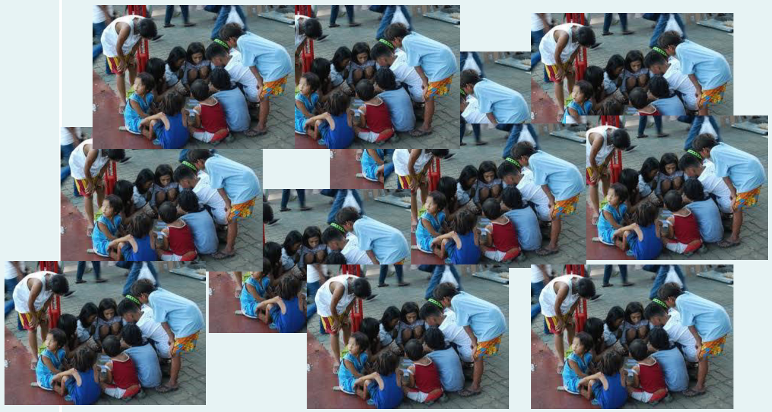
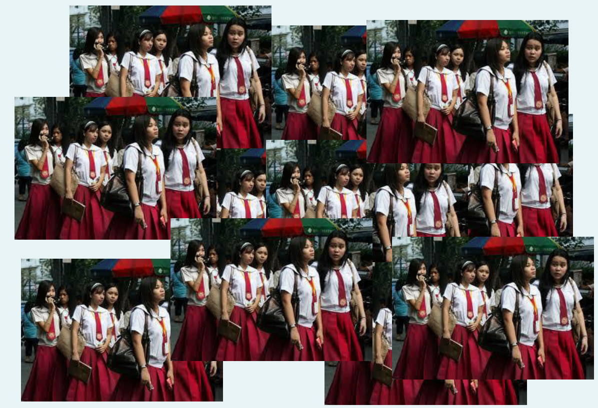
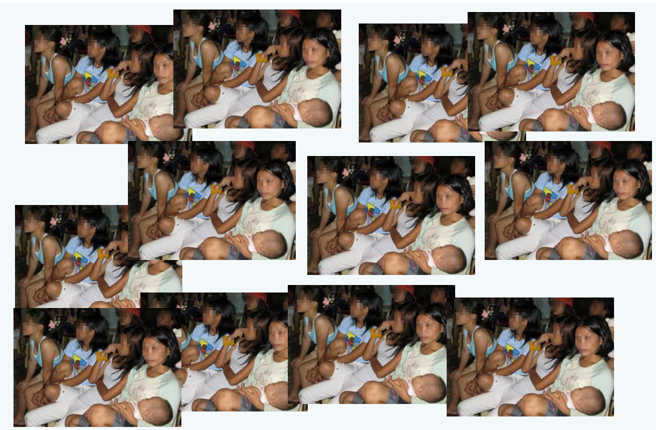
The core of large n designs

	Before	After
Project		
Comparison		

Large n

- *n* is the number of units of assignment, e.g. schools, villages, sub-districts (the unit of assignment can be different from the treatment unit and unit of analysis)
- If *n* is large then we create treatment (project) and comparison groups which are identical prior to the intervention…
  - And use statistical analysis to assess post-intervention differences between treatment and comparison: we say these differences are caused by the intervention

So in fact

	Before	After
Project		
Comparison		

What do we need to measure impact?

Agricultural extension in Uganda

Robusta coffee yield kg/ha

	Before	After
Project (treatment)		720
comparison		

The majority of evaluations have just this information …
which means we can say absolutely nothing about impact

Before versus after single difference comparison
Before versus after = 720 – 620 = 100

	Before	After
Project (treatment)	620	720
comparison		

This 'before versus after' approach is outcome monitoring, which has become popular recently. Outcome monitoring has its place, but it is not impact evaluation

Post-treatment comparison comparison

Single difference = 720 – 680 = 40

	Before	After
Project (treatment)		720
comparison		680

But we don't know if they were similar before… though there are ways of doing this (statistical matching = quasi-experimental approaches)

Double difference =

(720-6… = 60

Benefits of ex ante designs:

- Baseline data
- Better comparison group (including possible RCT

It's never too early to start your impact evaluation

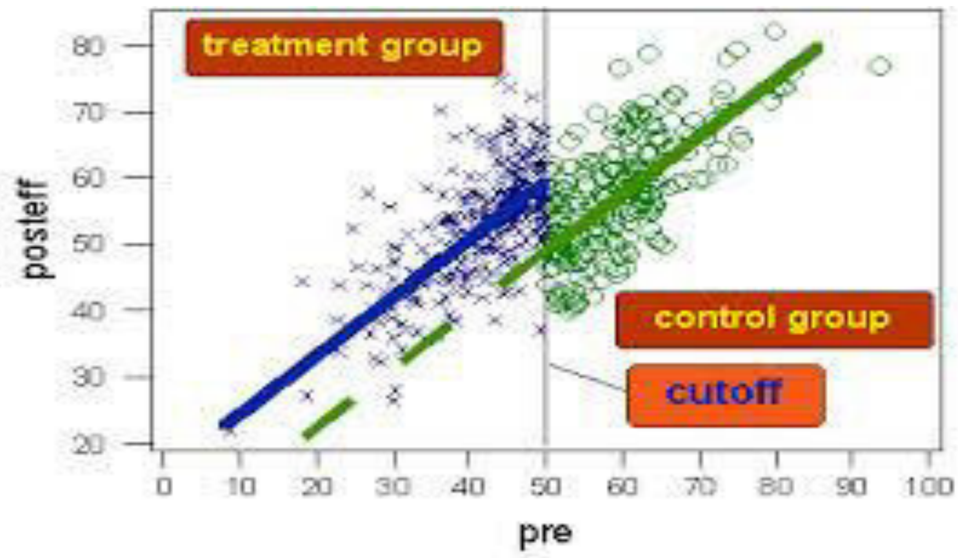
Proj…
comp…

Conclus… comparison… impact evaluatio… matching). SO WE NEED BASELINE DATA FROM PROJECT AND COMPARISON AREAS

Where does the comparison group come from?

Experimental:

- Randomized control trials
- Natural experiments



Non-experimental:

- Quasi-experimental (statistical matching)
- Other statistical methods (e.g. instrumental variables)

We will learn more about these methods in future lectures

Exercise

Complete the table below using <u>one selected outcome indicator for your intervention</u>

- Before versus after
- Ex post single difference
- Double difference

What conclusions can you draw about (i) the programme and (ii) methods?

	Before	**After**
Project		
Comparison		

Please visit: www.3ieimpact.org/

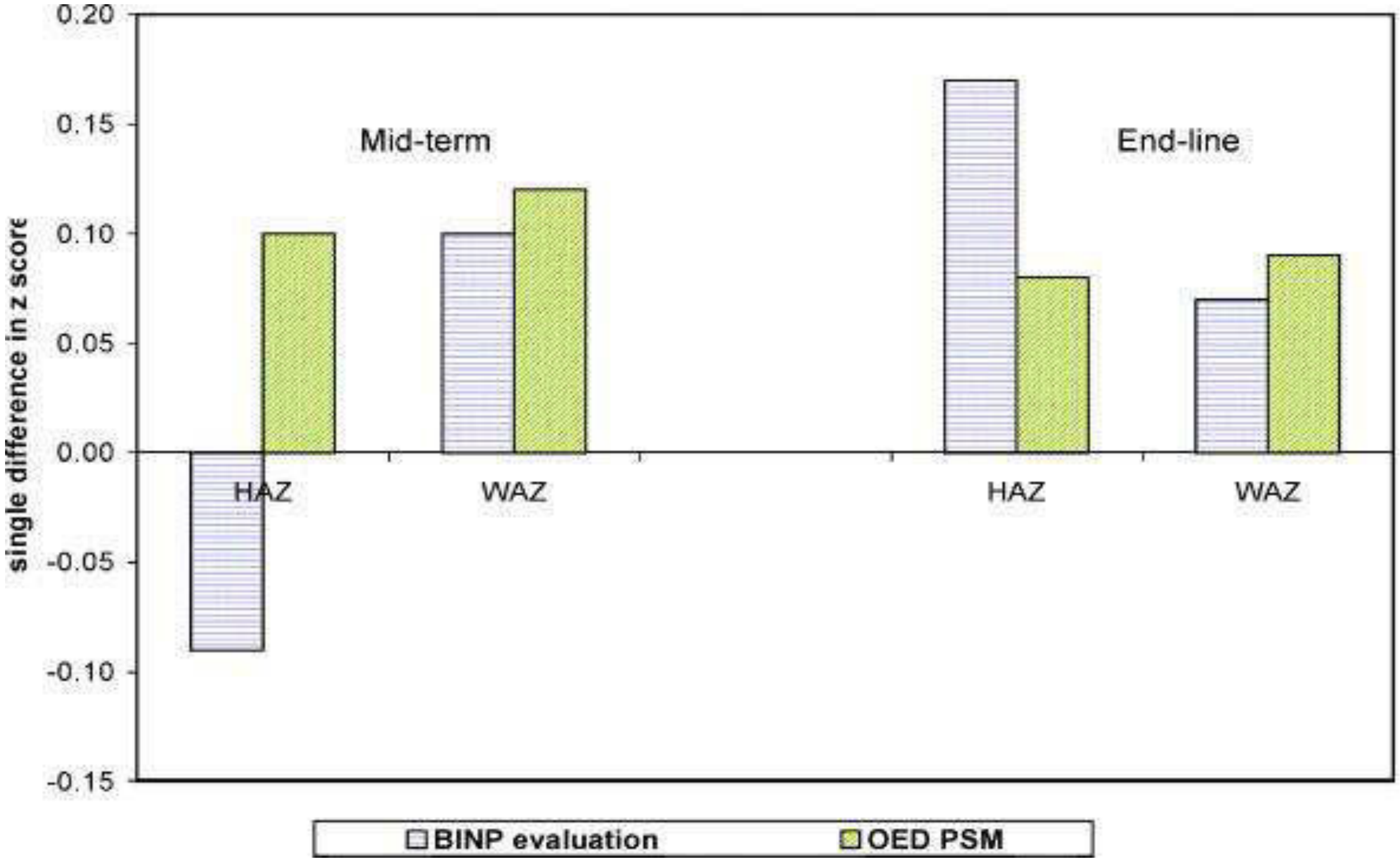
Theory-based impact evaluation

Howard White
International Initiative for Impact Evaluation

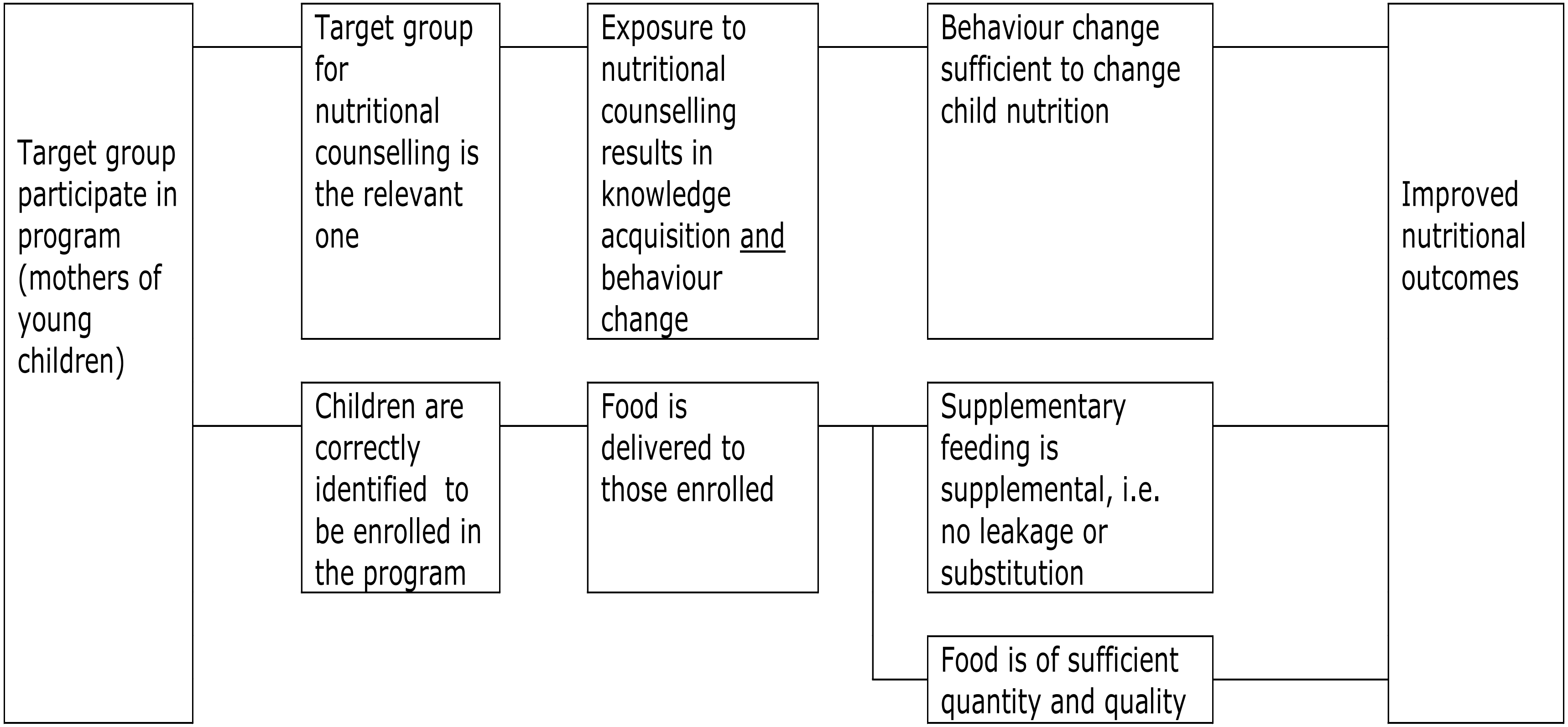
Impact evaluation: an example

The case of the Bangladesh Integrated Nutrition Project (BINP)

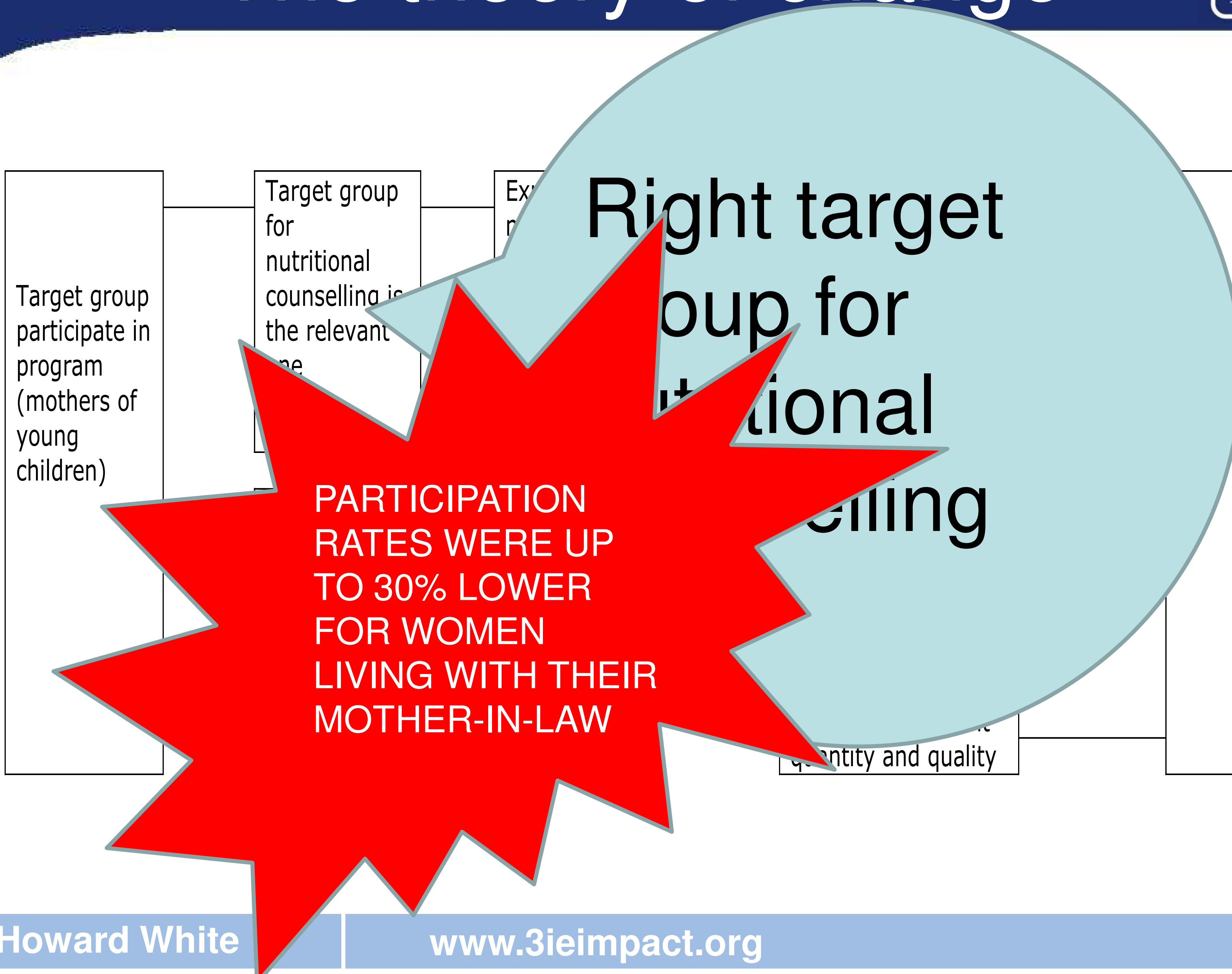
Comparison of impact estimates



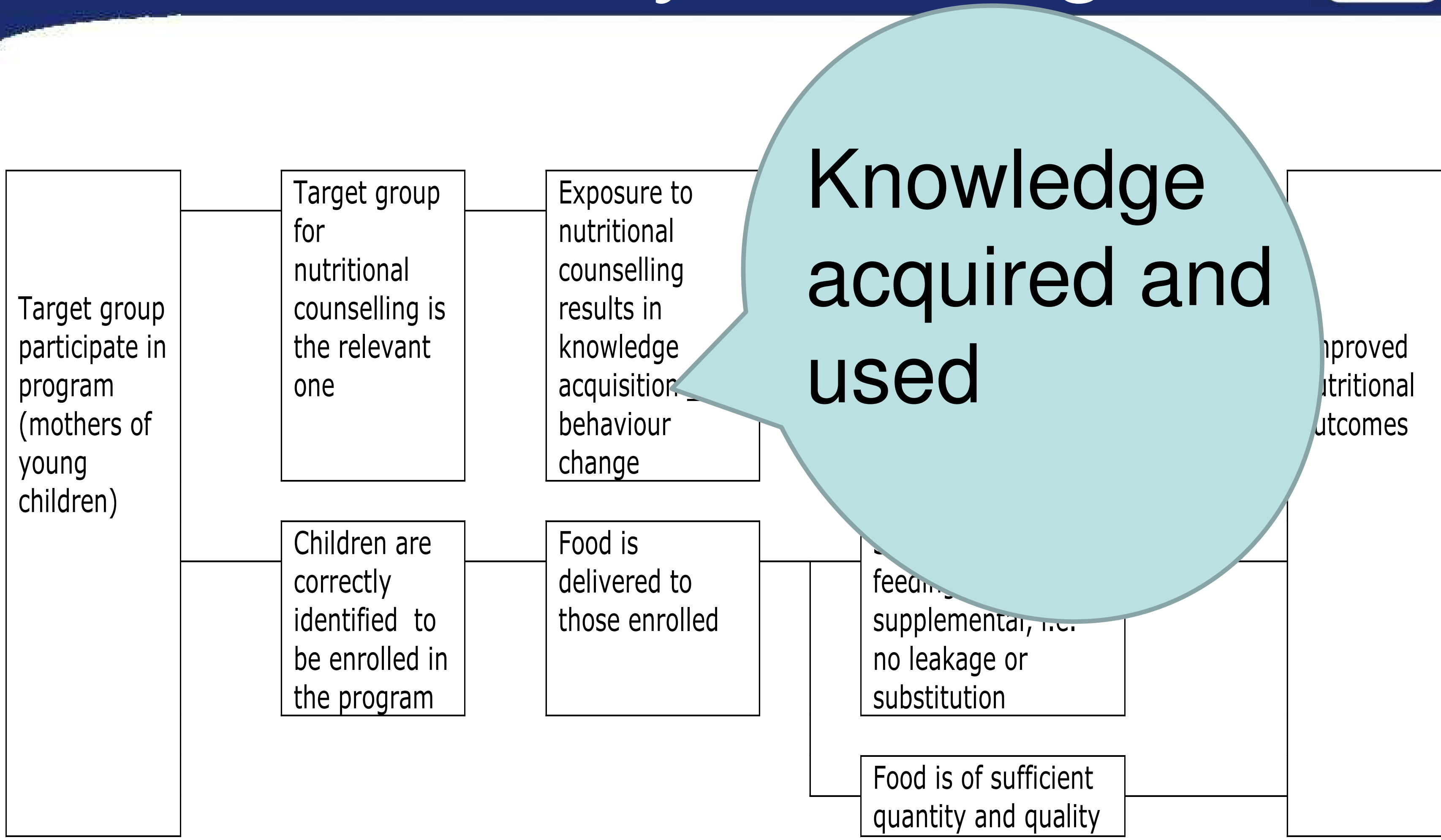
Summary of theory



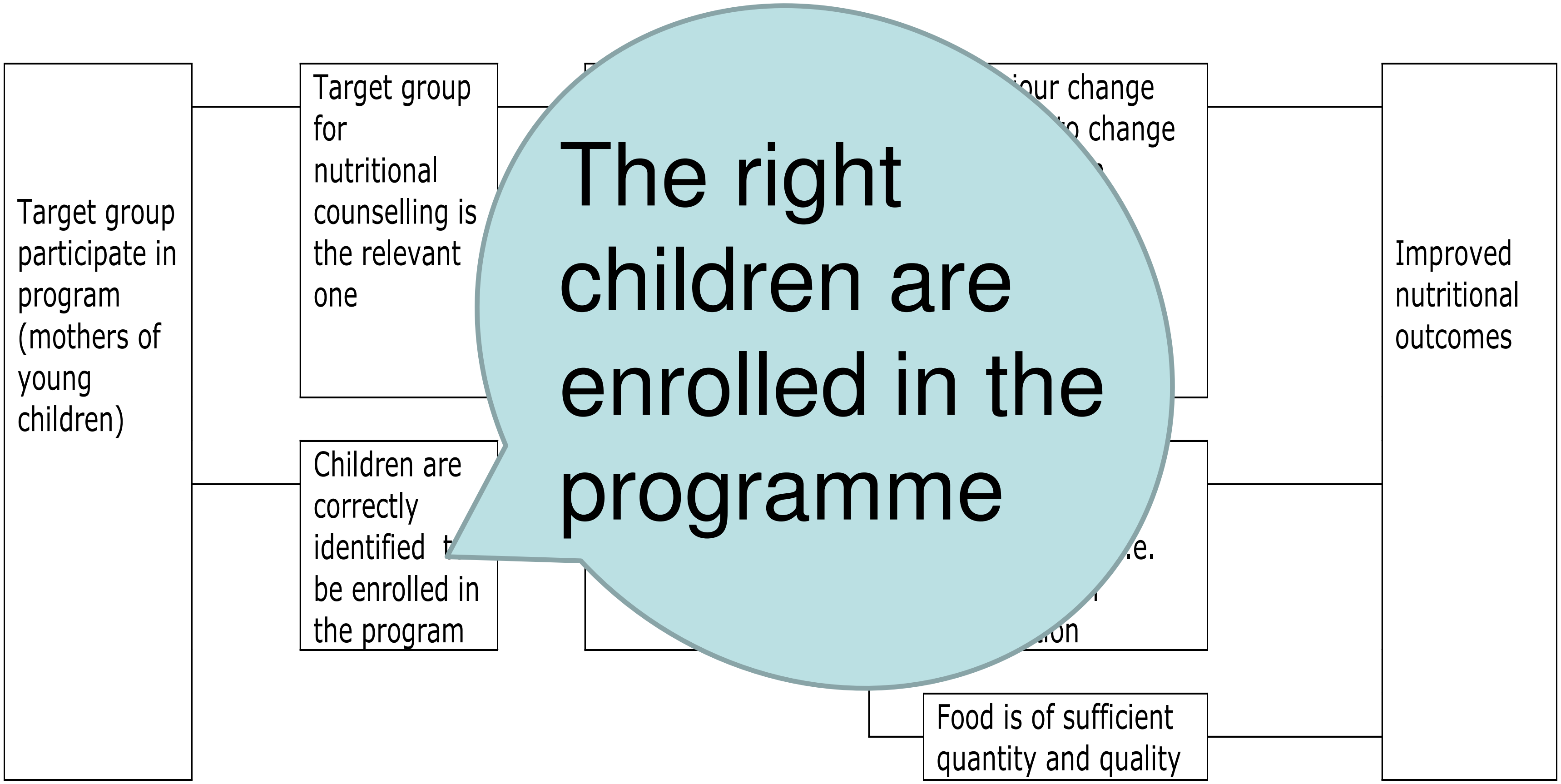
The theory of change



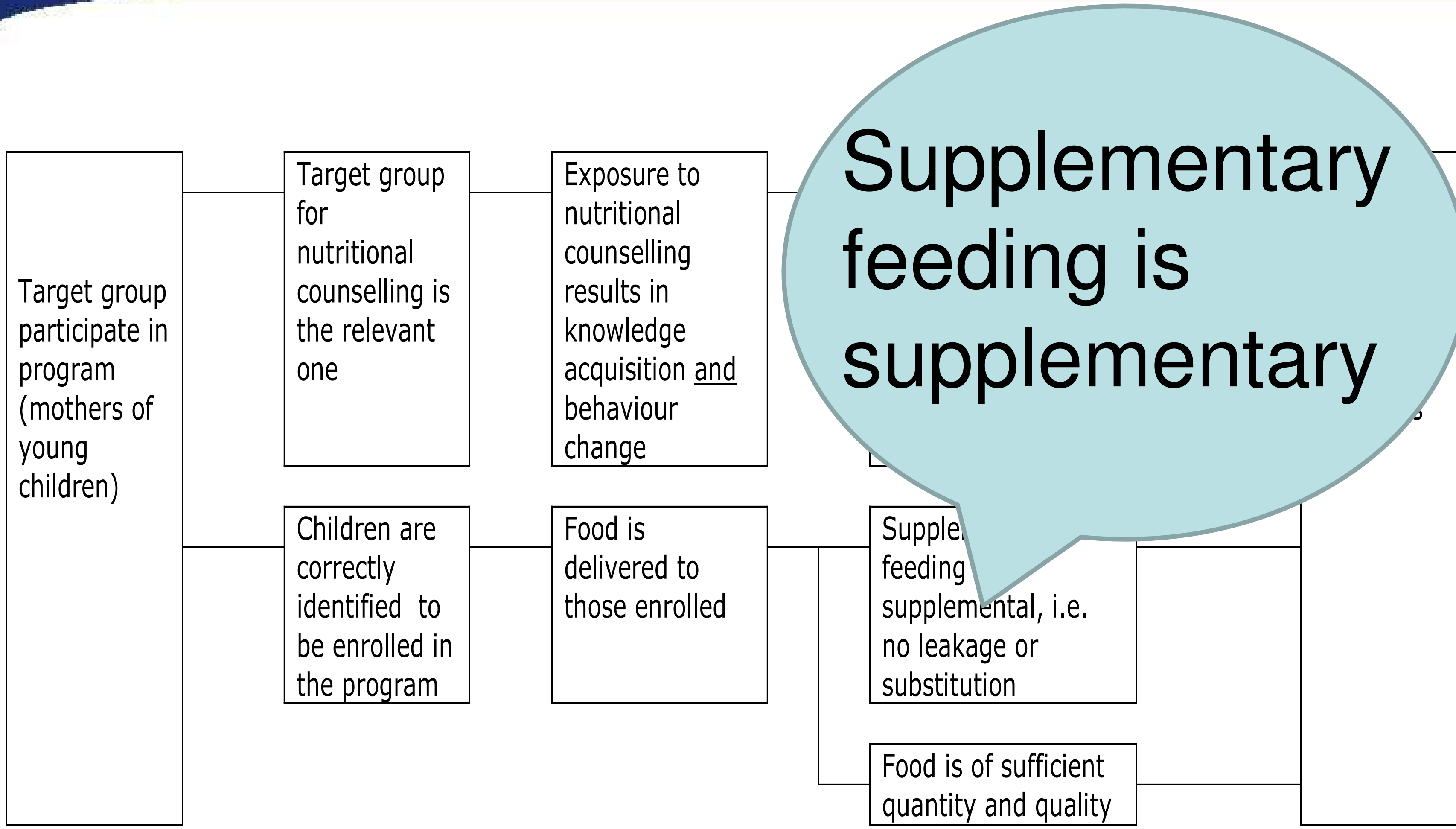
The theory of change



The theory of change



The theory of change



Lessons from BINP

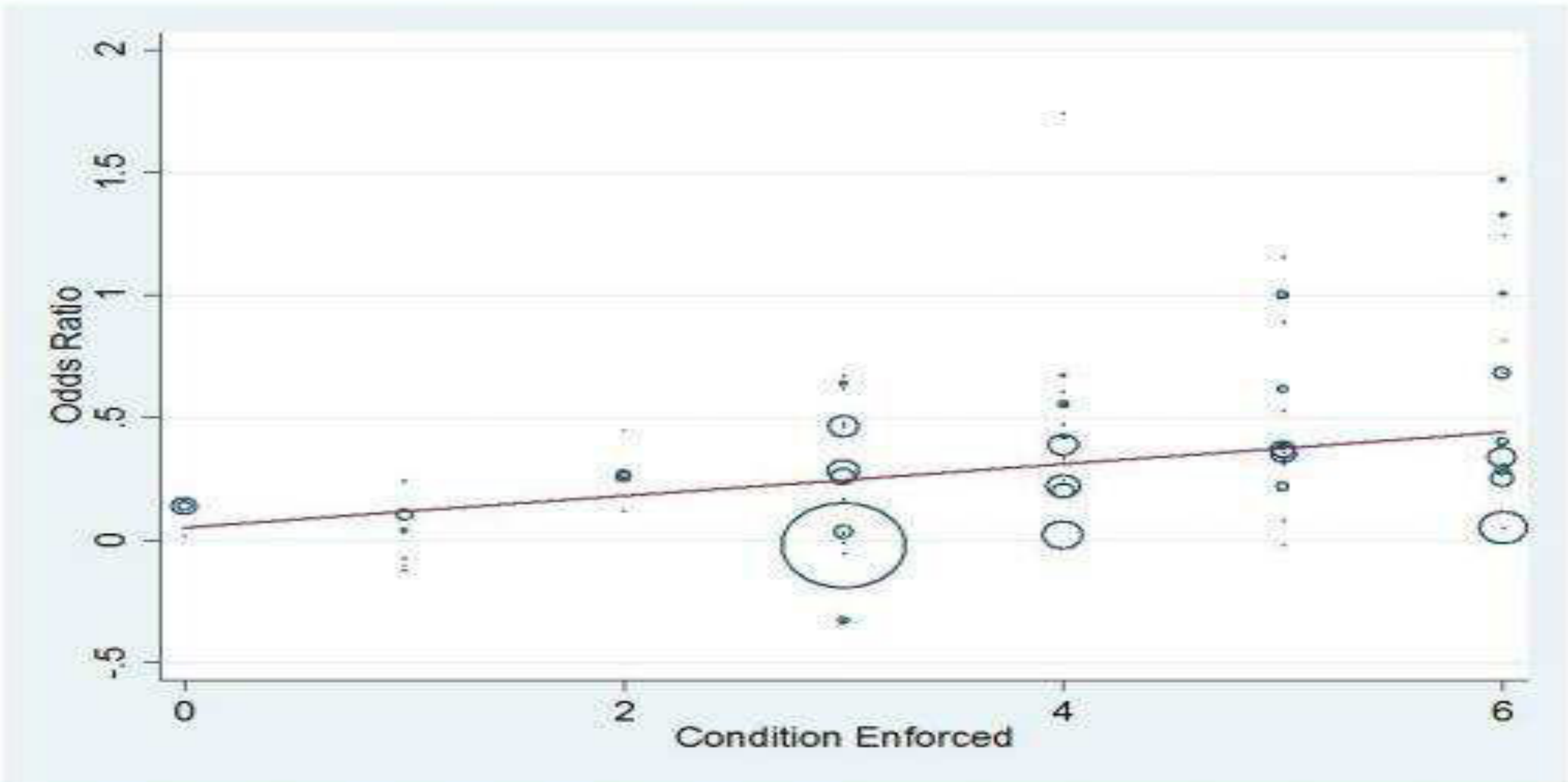
- Apparent successes can turn out to be failures
- Out
  in
-
  s
  mo
- But the a
- And independence can matter

Outcome monitoring cannot tell us about results – what difference we made… Only impact evaluation can do that

And theory leads to more nuanced questions

- E.g. conditional cash transfer second generation questions:
  - Conditions or not?
  - What sort of conditions?
  - Who to give money to?
  - How to give the money?
  - When and how often to give money?

Conditionality

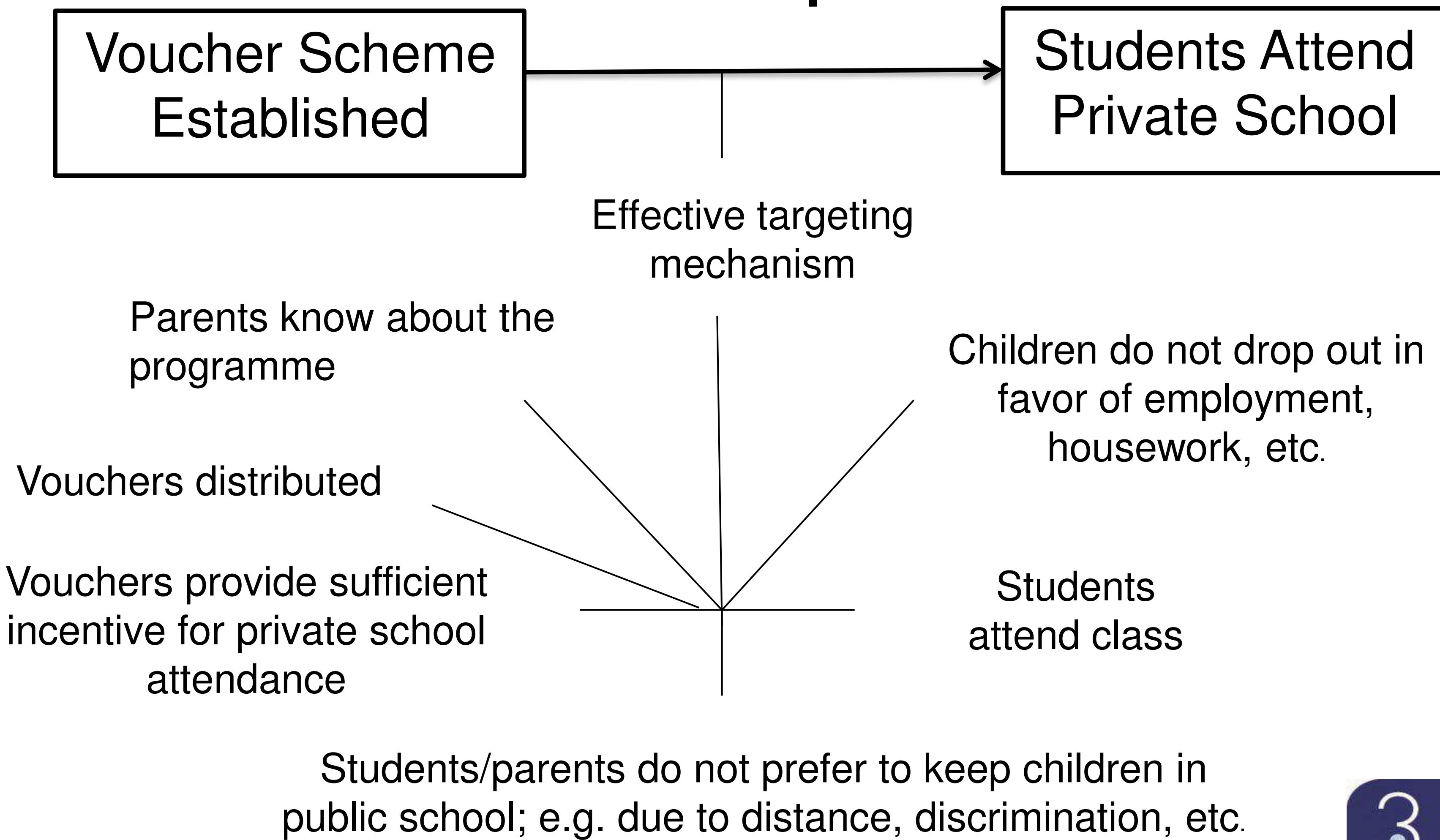


Children 60% more likely to be in school with conditionality which is monitored and enforced compared to no conditions

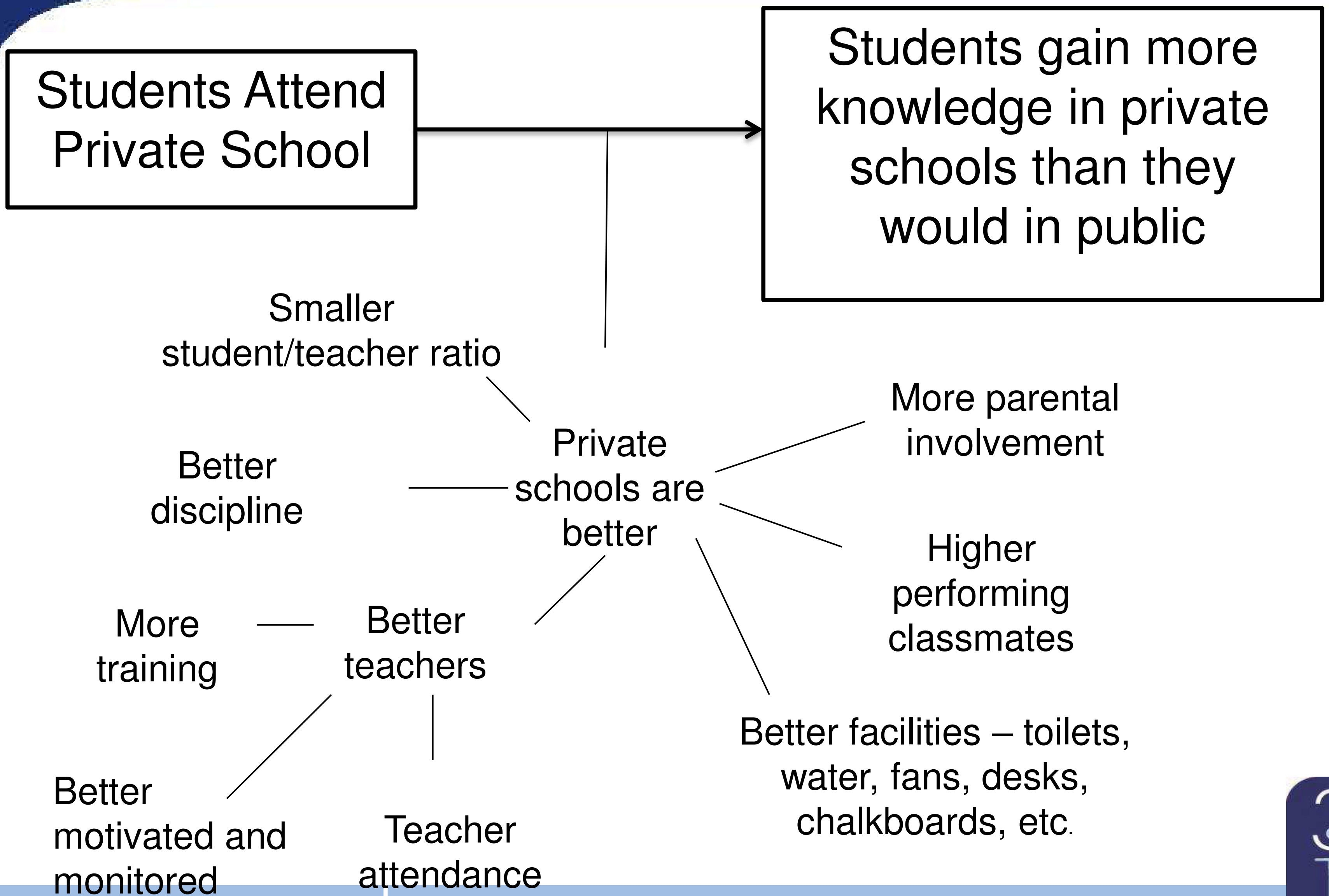
Theory of Change: School vouchers

Voucher Scheme Established → Students Attend Private School → Students gain more knowledge in private schools than they would in public → Higher test scores → Better life (income etc.)

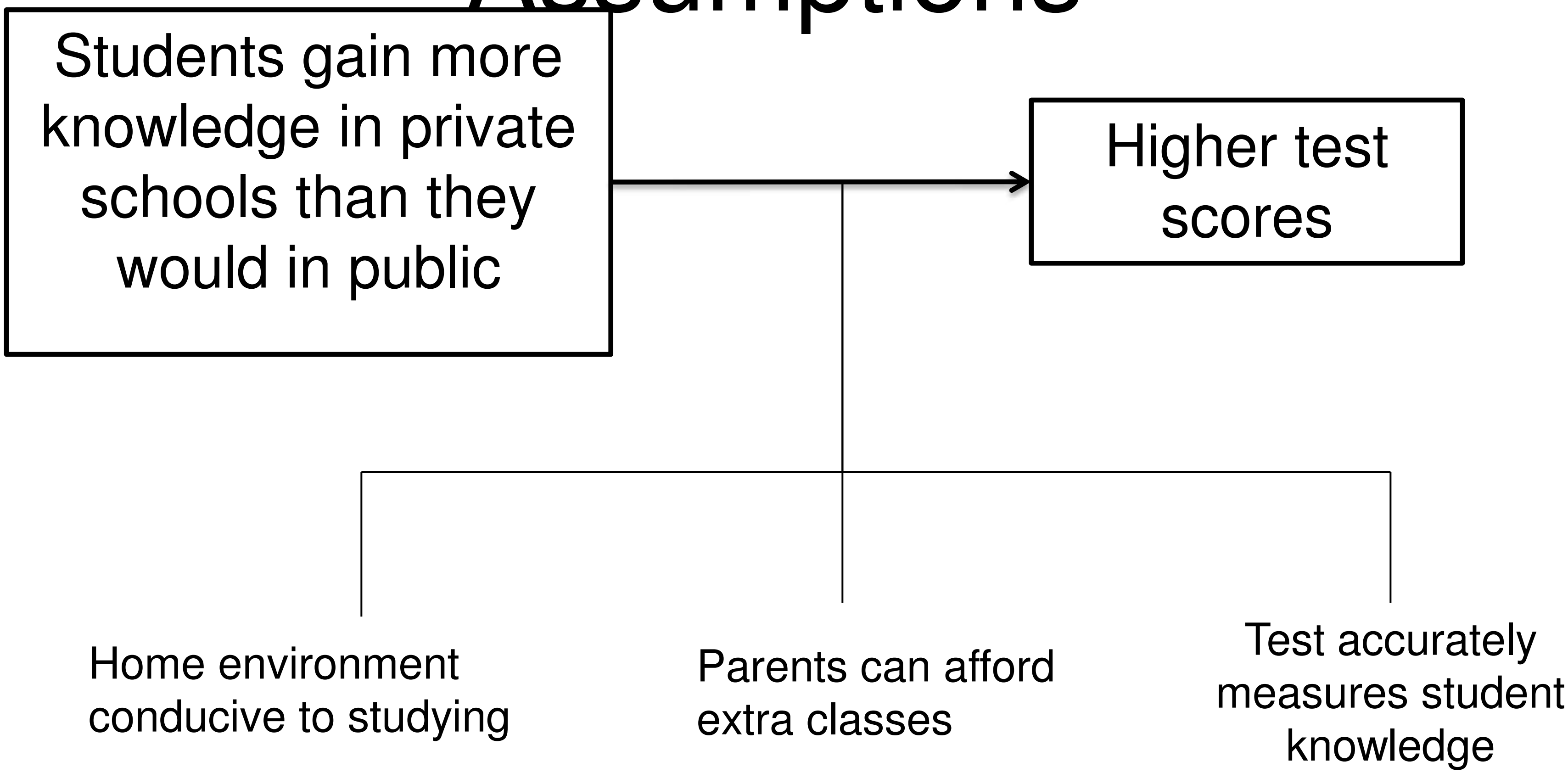
Possible Transmission Routes and Assumptions



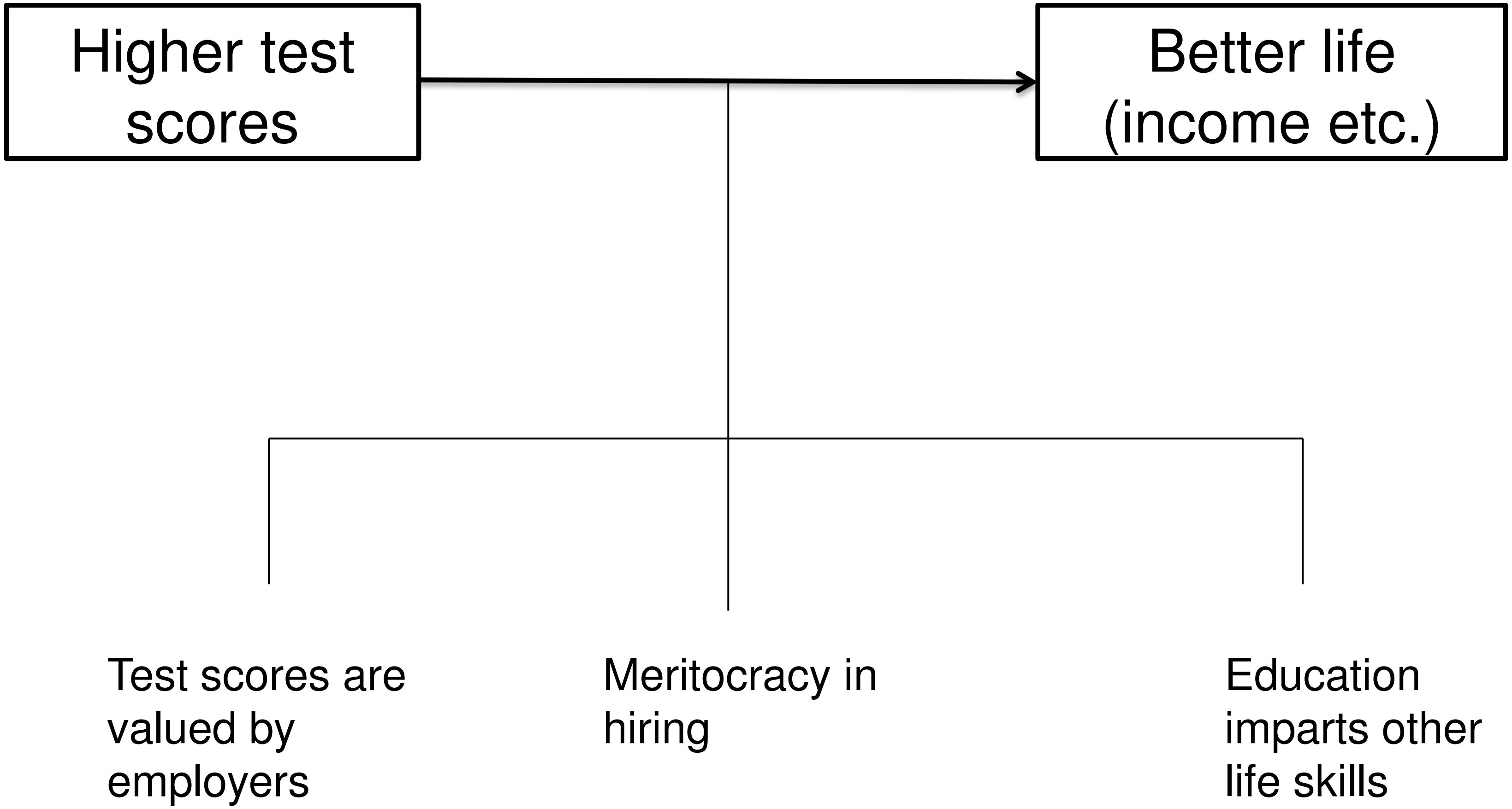
Possible Transmission Routes



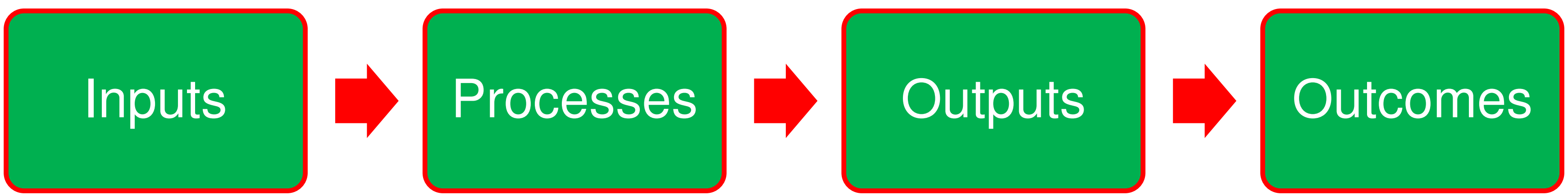
Transmission Routes and Assumptions



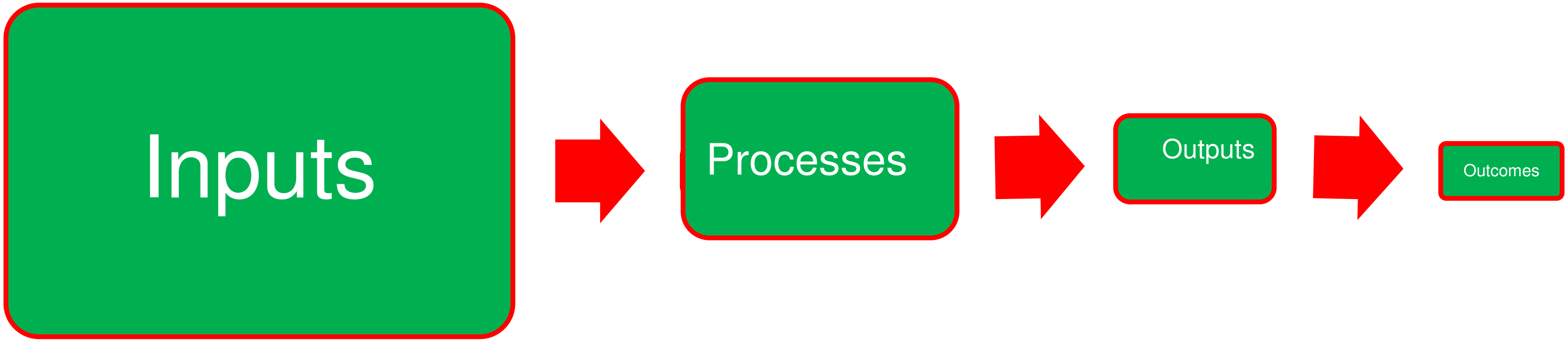
Theory of Change

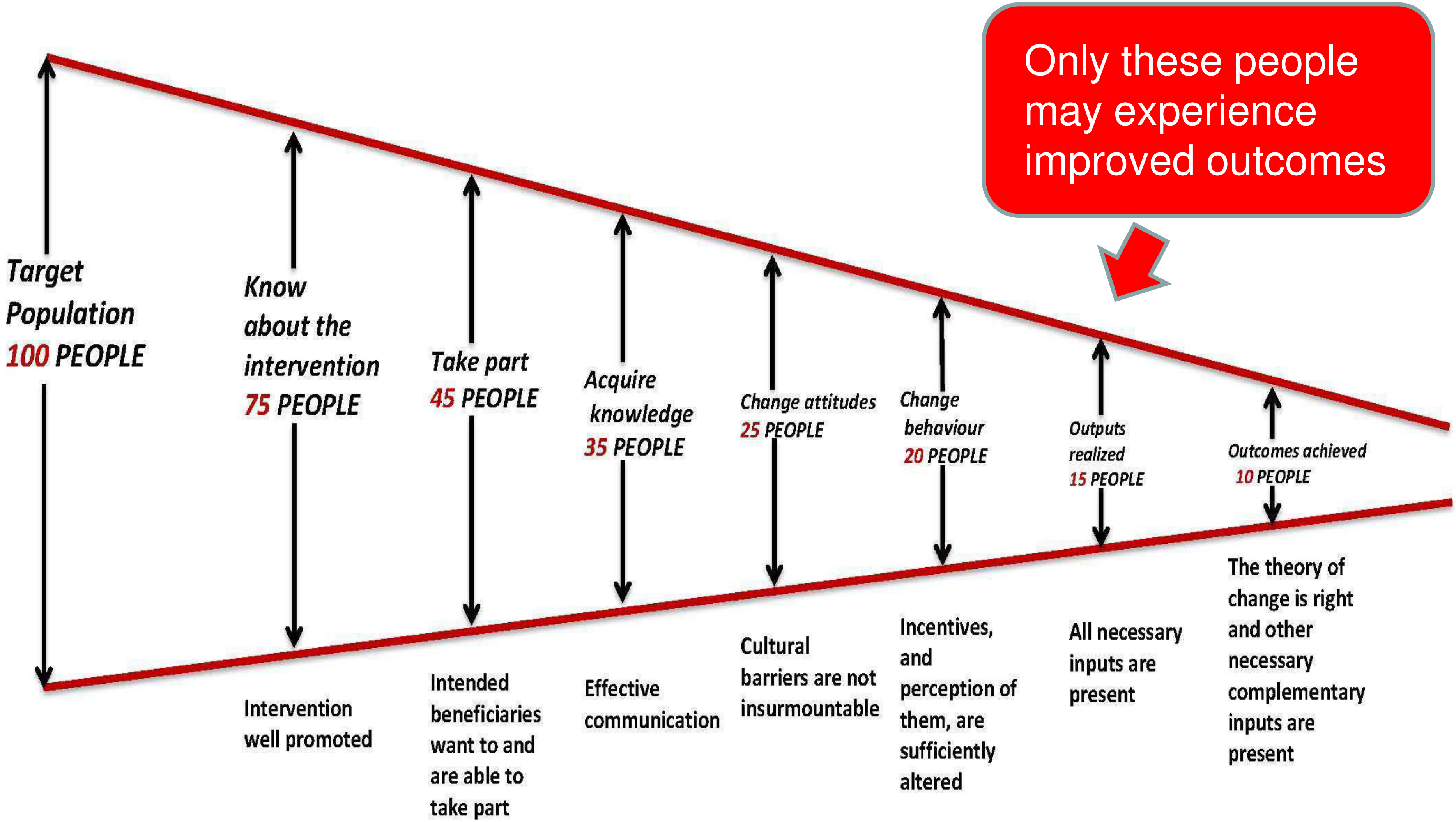


A typical theory of change



What it really looks like

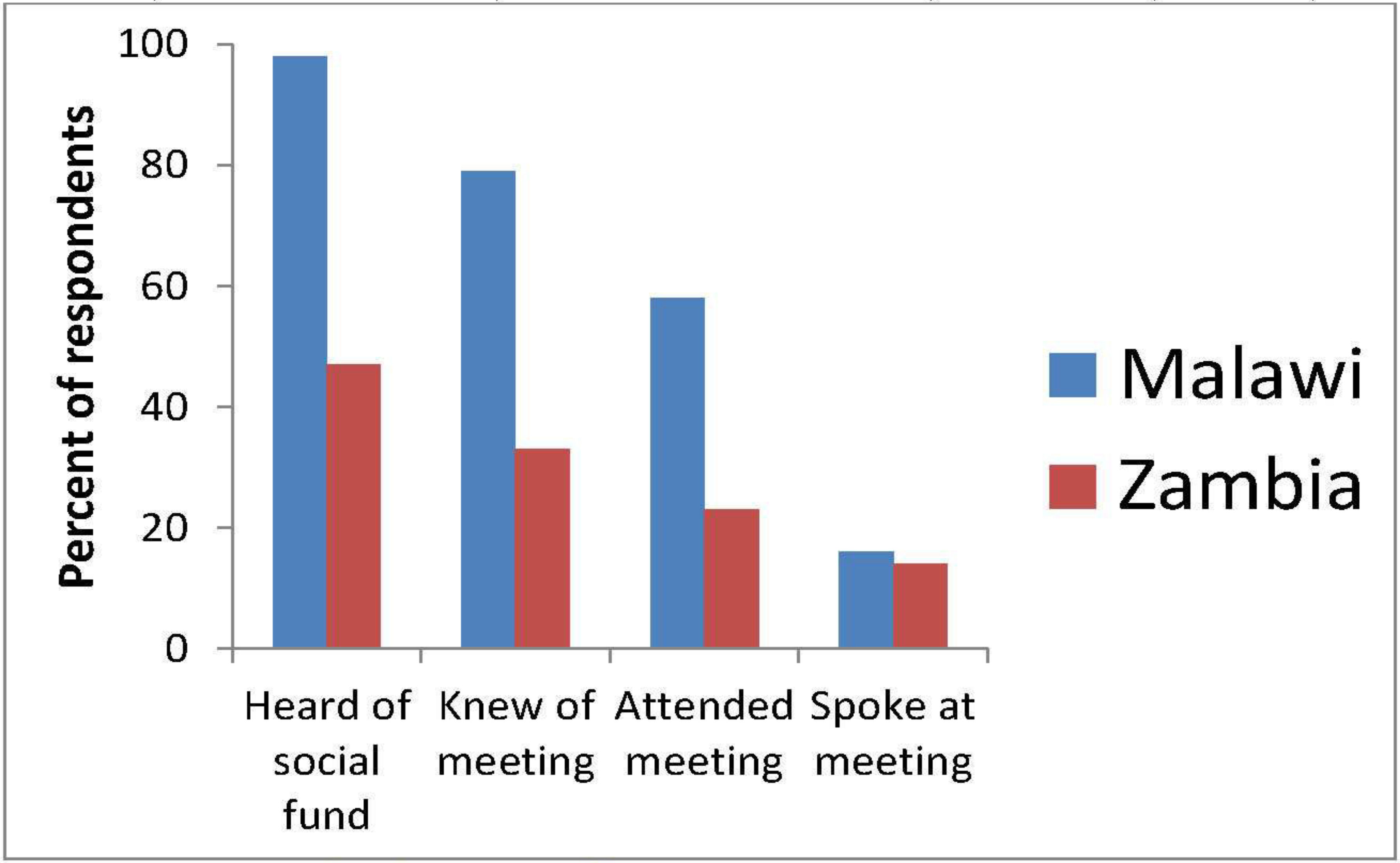




Funnel of Attrition

An example from social funds

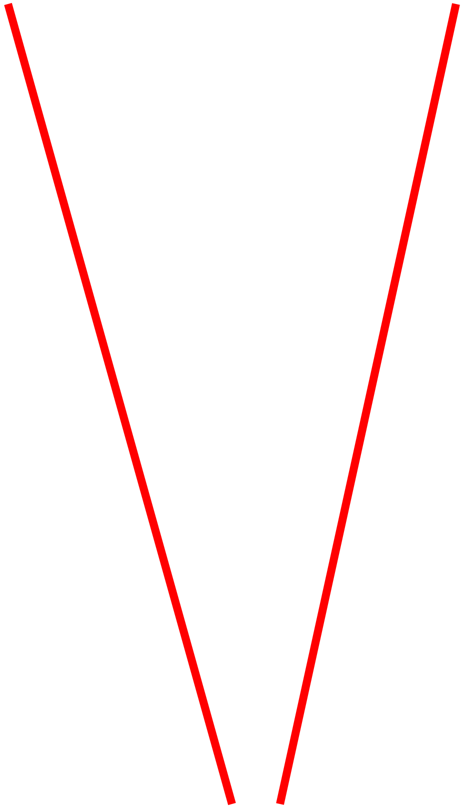
The value of the indicator at each step in the causal chain is necessarily lower than the previous step



Source: data from *Social Funds: Assessing Effectiveness*, World Bank, 2002.

The funnel operates within steps in the causal chain

- Show up
- Attend
- Stay awake
- Pay attention
- Understand
- Agree
- Absorb
- Retain
- Act



Many interventions fall at the first hurdle

- Free male circumcision: 25% if free down to just 10% with partial subsidy
- Pre-school in Mexico, fewer than 10% of parents who registered actually took part
- Insurance schemes typically less than 10% take up

And participation declines over time

- ½ households stopped using improved cookstoves by 8 month follow up survey
- Water treatment: fewer than 1/3 households using filters in Cambodia and pasteurising in Kenya after 3-4 years.. And only 5% disinfecting in Guatemala after just one year

The need for formative research

Texting:

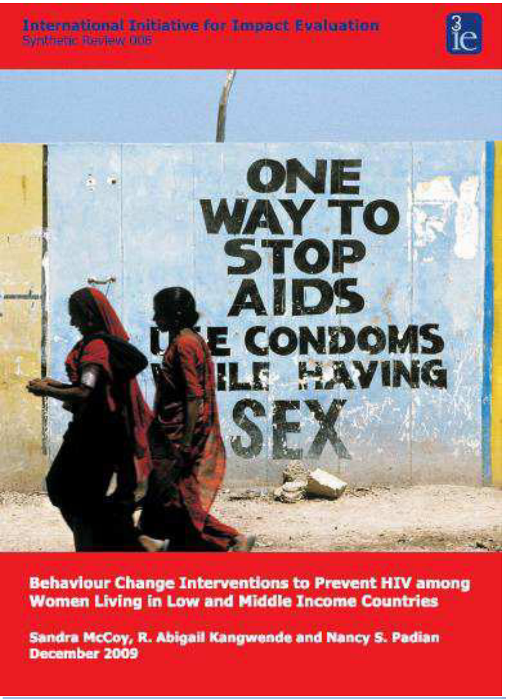
- Parliamentarians
- Banking
- TB

Examples of weak links

Ghana cookstoves

Improving hygiene in catering facilities in UK

3ie: Improving lives through impact evaluation



Thank you

Visit www.3ieimpact.org

Randomized control trials

Howard White, 3ie

Establish the counterfactual using a comparison group

So what?

- Comparison groups are nothing new
- What is new is attention to threats to validity of comparison group from
  - Selection bias
  - Contamination
  - Spill over effects (e.g. from FFS)

The problem of selection bias

- Program participants are not chosen at random, but selected through
  - Program placement
  - Self selection
- This is a problem if the correlates of selection are also correlated with the outcomes of interest, since those participating would do better (or worse) than others regardless of the intervention

Selection bias from program placement

- A productivity enhancement programme is targeted at poor and marginal farmers
- These farmers have less land and other assets like capital, literacy, access to labour and so on… so their outcomes (productivity) will be lower than that of non-participants, maybe even with the project
- Hence productivity for project farmers will be lower than the average for other farmers
- The comparison group has to be drawn from a group of similarly deprived farmers

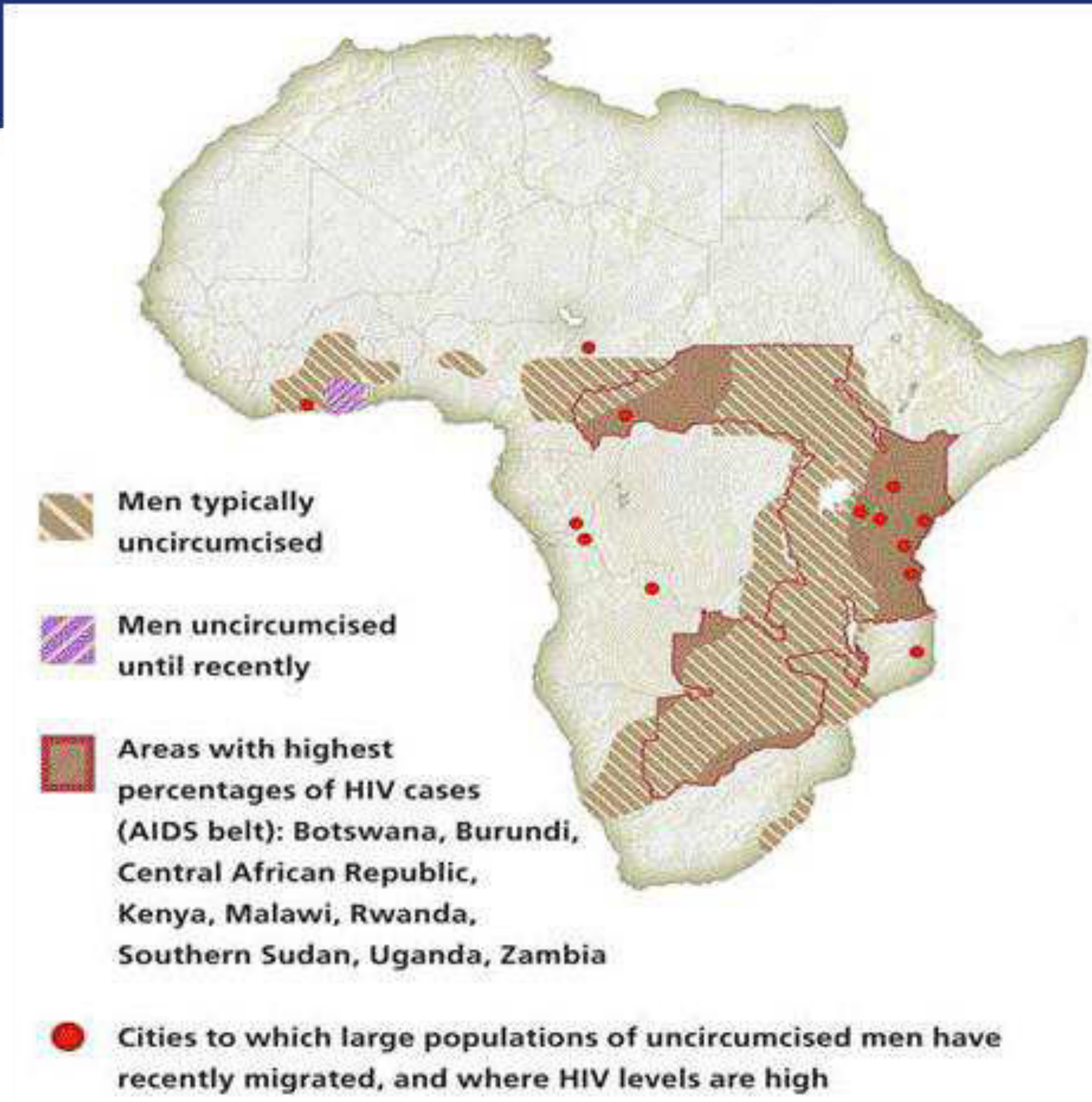
Selection bias from self-selection

- A farmer field school programme recruits farmers from a community on a voluntary basis
- But those farmers who join are likely to be ‘more progressive, i.e. more interested in changing practices
- So those farmers who join the programme are more likely to adopt new practices and have better outcomes than those who don’t join… even in the absence of the programme

And it may be that those communities in the programme may be better performing than non-programme communities as a result of either self-selection or progamme placement

Examples of selection bias

- Hospital delivery in Bangladesh (0.115 vs 0.067)
- Secondary education and teenage pregnancy in Zambia
- Male circumcision and HIV/AIDS in Africa



HIV/AIDs and circumcision: geographical overlay

Main point

There is 'selection' in who benefits from nearly all interventions. So need to get a comparison group which has the same characteristics as those selected for the intervention.

Randomization (RCTs)

- Randomization addresses the problem of selection bias by the random allocation of the treatment

- Unit of assignment may not be the same level as the unit of analysis, e.g.
  - Randomize across villages but measure individual learning outcomes
  - Randomize across sub-districts but measure village-level outcomes

Some RCTs

Gujarat pollution

Zambian hairdressers

Some more RCTs

Computer-assisted learning, China

Early marriage, India

Quiz

For each of these four examples, what is:

- The unit of assignment
- The unit of treatment
- The unit of analysis?




When can we randomize?

- When there is ‘over subscription’ (and we can generate over subscription through a raised threshold)
- When a programme will be rolled out over time
- Using an encouragement design for a universally available but not universally adopted intervention

Some different ways to randomize

Pipeline

Raised threshold

Prior matching, e.g. **matched pairs** can reduce necessary sample size

By analogy, could expand eligible area and randomize within that

Matched pairs randomization

Prior matching, e.g. **matched pairs** can reduce necessary sample size

20 villages in eligible sample, e.g.

- 2 much larger than others
- 2 very close to town
- 2 different ethnic group

In these pairs, one is treated and one control, hence making balance more likely

More ways to randomize

Don’t need randomize across whole eligible population

Just use these guys
for the RCT

Encouragement design

No universal scheme is universally adopted.

There are three groups: (a) always adopt, (b) never adopt, and (c) may adopt with encouragement

An encouragement design provides an incentive to group (c) to adopt in treatment versus no incentive in control

Different types of design

Don't need a 'no treatment' control

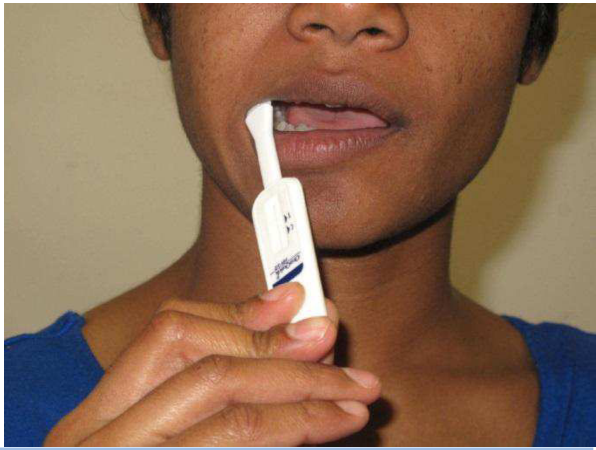
In medicine the control gets the standard practice of care ie the existing treatment. This comparison is often the one of most interest to policy makers

So everyone can get basic package, with some addition in the control to 'make it work better'

Factorial Design

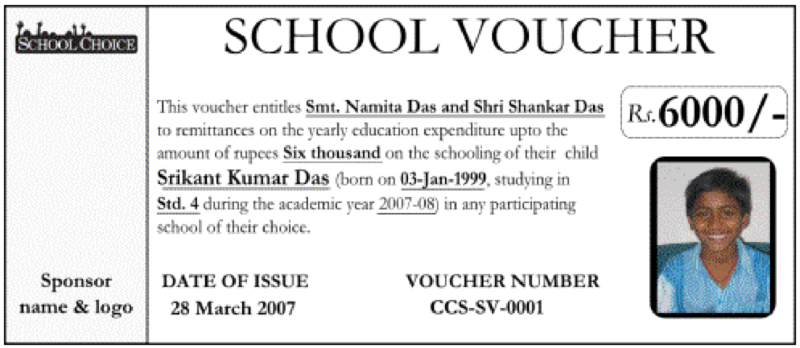
Rapid IE

- Low cost (<US$100k) impact evaluation in 6-12 months
- How is that possible?
  - Simple RCT i.e. individual level randomization
  - Measure outputs or intermediate outcomes (e.g. adoption)



What sort of things can we ramdomize at individual level?

- Vouchers



SCHOOL CHOICE

SCHOOL VOUCHER

This voucher entitles Smt. Namita Das and Shri Shankar Das to remittances on the yearly education expenditure upto the amount of rupees Six thousand on the schooling of their child Srikant Kumar Das (born on 03-Jan-1999, studying in Std. 4 during the academic year 2007-08) in any participating school of their choice.

Rs.6000/-

Sponsor name & logo

DATE OF ISSUE
28 March 2007

VOUCHER NUMBER
CCS-SV-0001

SHS

- Information

Refresher

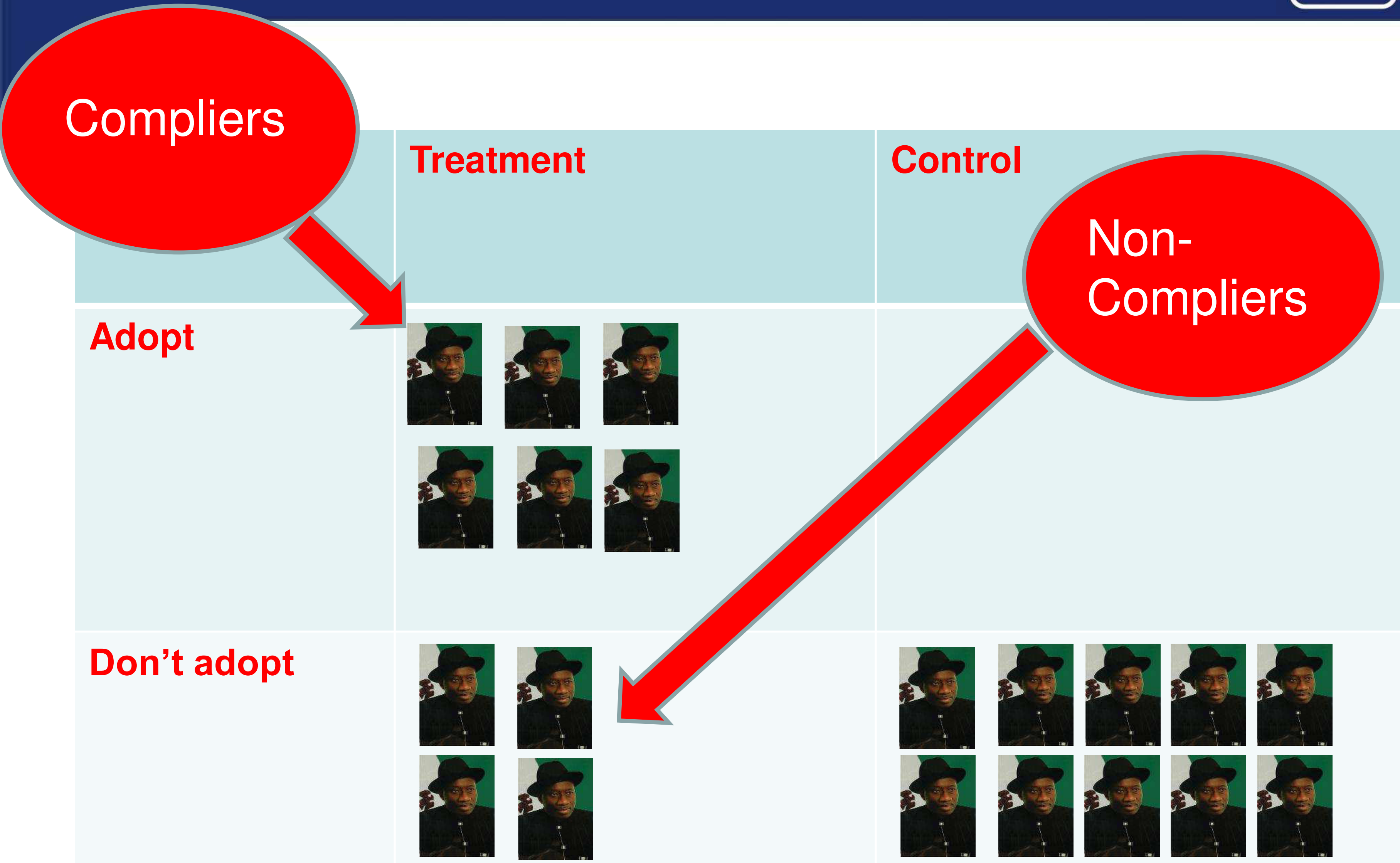
- Simple RCT (can be stratified sampling)
- Cluster RCT
- Pre-matching e.g. matched pair randomization
- Pipeline randomization

Types of treatment effect

Intention to treat effect (ITT): the total impact averaged over all those targeted by the intervention

Treated of treated effect (ToT): the impact just on those who actually take part

Compliance and treatment effects



Calculating ITT and ToT

- Total income in treatment = 200
- Total income in control = 140
- Ex-post single difference = 200- 140 = 60
- ITT = 60 / 10 = 6
- ToT 60 / 6 = 10

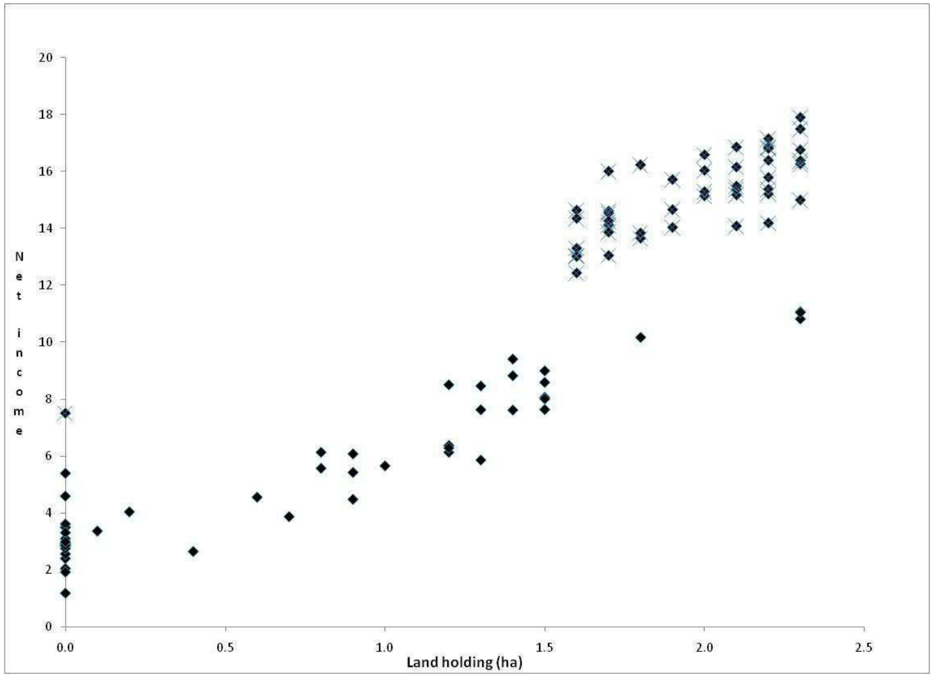
Intention to treat effect is ‘diluted’ by non-compliance (remember the funnel)

Which measures true impact?

ATE vs LATE

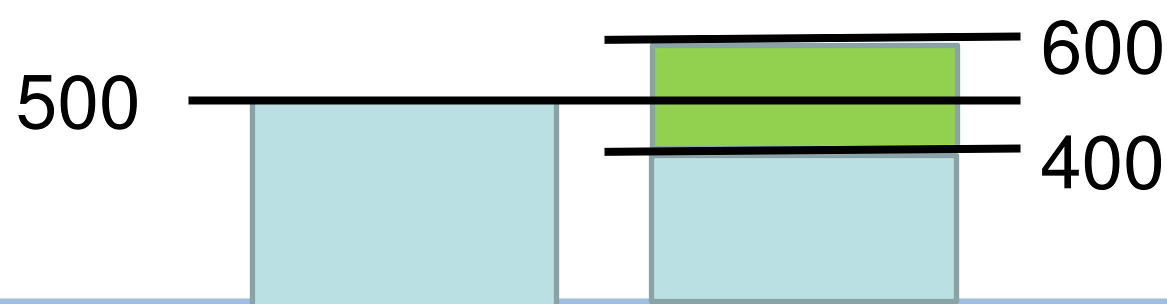
- ATE = average treatment effect
- Can also do sub-group analysis (have to allow for this in your power calculations, and mean you will probably use stratification)
- LATE = Local average treatment effect: treatment effect is just for those for whom you are measuring impact

Examples of LATE



RDD

‘Caliper raised threshold’



Dealing with 'cross-overs'

	Treatment	Control
Always adopt		
Adopt if offered		
Never adopt		

Cross-overs

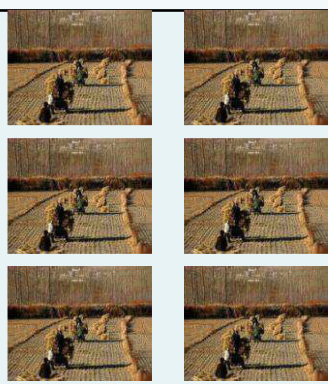
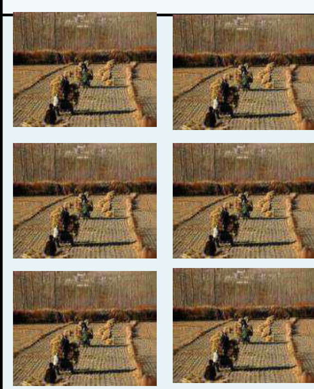
Dealing with ‘cross-overs’

- Y(T) = 400 Y(C) = 200
- Impact = 400 – 200 = 200
- Change in take up = 4
- ITT = 200/8 = 25
- ToT =200/4 =50

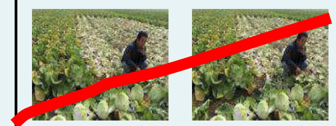
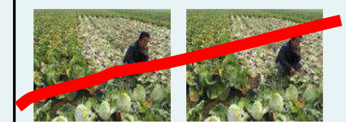
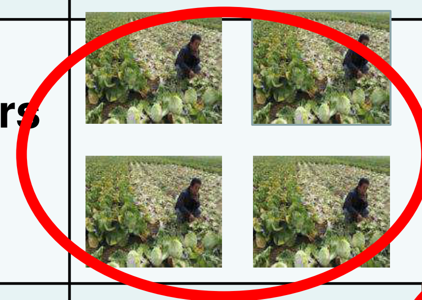
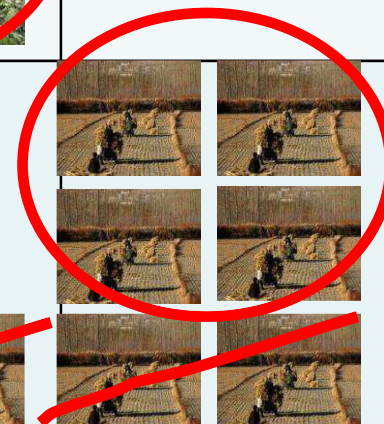
Cross-overs cause ‘under-estimate’ of impact (but not really)

Encouragement design

Before *Income(t,c) = 200*

	Treatment	Control
Adopt		
Don’t Adopt		

After *Income (t) = 240, c=200*

	Treatment	Control
Already adopted		
New adopters		
Don’t adopt		

Encouragement design: calculation

$$Impact\ ToT = \frac{Difference}{n\ x\ proportion\ new\ adopters}$$

$$= \frac{40}{8\ x\ 0.5} = 10$$

Preparing for an RCT

- Has to be an *ex-ante* design
- Has to be politically feasible, and confidence that program managers will maintain integrity of the design
- Perform power calculation to determine sample size (and therefore cost)
- Collect baseline data to:
  - Test quality of the match
  - Conduct difference in difference analysis

Thinking about RCT designs

- What are my
  - Unit of analysis (what outcomes are you measuring?
  - Unit of assignment?
- Do I have sufficient units of assignment (i.e. power calculation)
- How many ‘treatment arms’ will I have?
- What do the comparison group get?
- What sort of spillovers might there be?
- How likely is contamination of treatment or control?
- How much of the programme am I going to randomize and how (e.g. pipeline)?
- Who needs to agree to a RCT? Have they? Cultural factors?

Steps in carrying out an RCT

- Establish outcomes, theory of change, evaluation questions
- Design data collection instruments
- Unit of assignment, treatment and analysis?
- Establish eligibility criteria and eligible population
- Power calculation and draw random sample
- Randomly assign intervention and control
- Conduct baseline
- Check balance
- Endline and impact estimates
- Influence policy

Overcoming resistance to randomization

- There is probably an untreated population anyway
- Need not randomly allocate whole programme just a bit
- Exploit different designs which make less difference to the programme
- Don’t need ‘no treatment’ control
- Randomization is more transparent
- RCTs are not unethical, spending money on programmes that don’t work is unethical

Some issues

- RCTs can’t handle complexity - FALSE
- RCTs are not applicable to all development interventions - TRUE
- RCTs can’t be done for interventions with ‘intangible’ outcomes - FALSE
- RCTs are unethical - FALSE but can be better

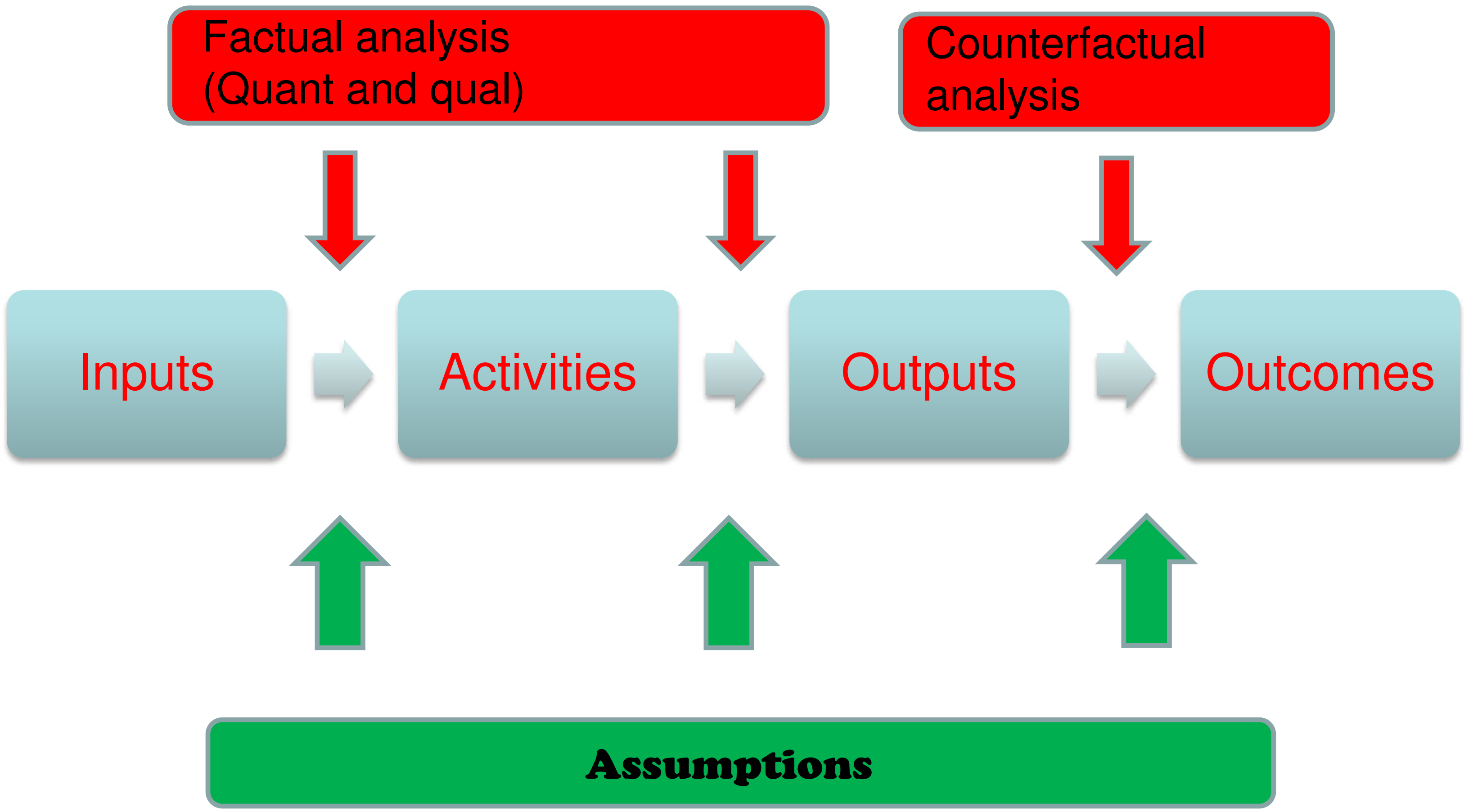
Exercise

- Is your intervention (or any component of it) amenable to randomization?
- What are the unit of assignment, treatment and outcome measurement?

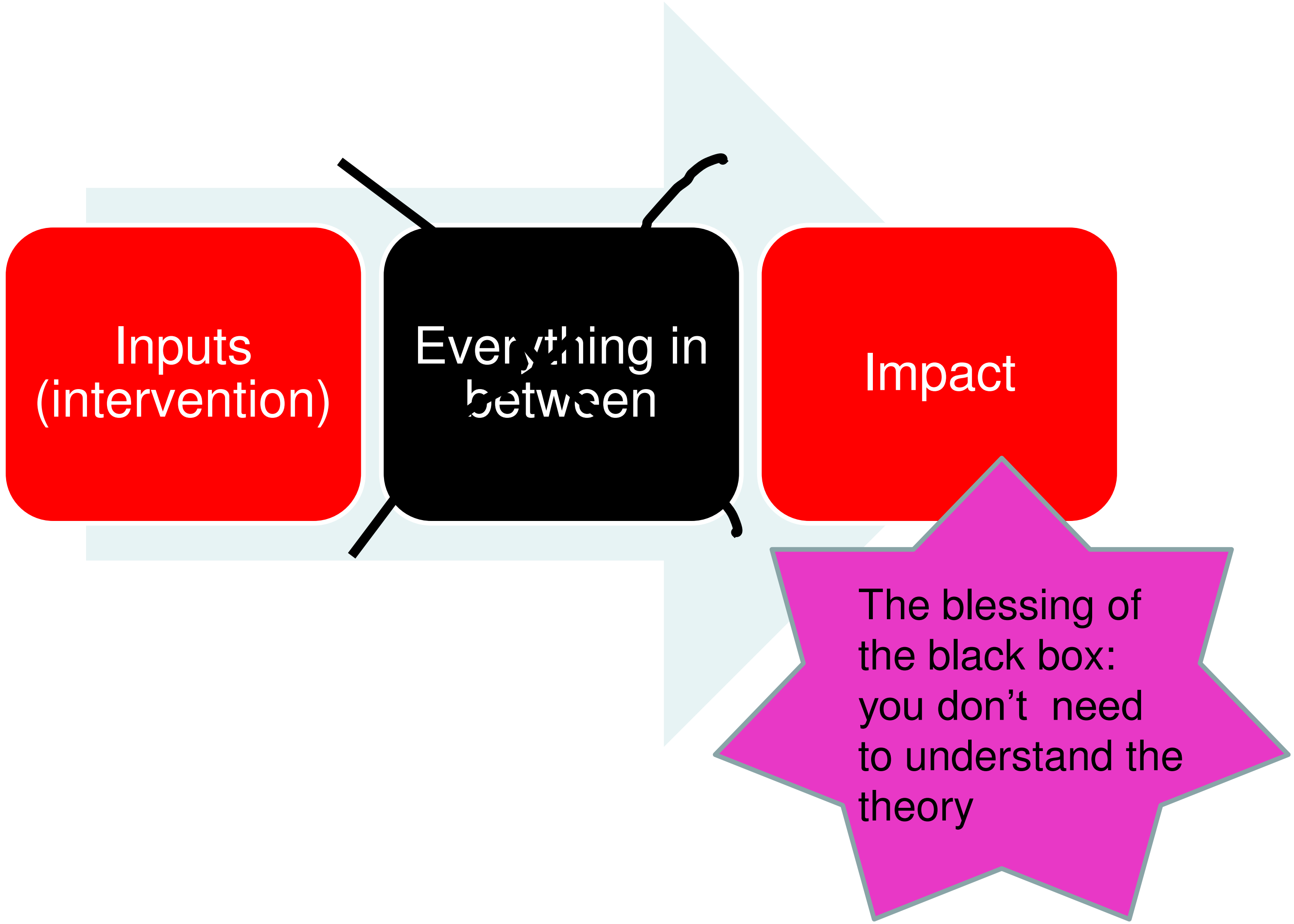
Thank you

Visit www.3ieimpact.org

Understanding where RCTs fit it



An RCT theory of change



Statistical matching and other quasi-experimental designs

Howard White

International Initiative for Impact Evaluation

Quasi-experimental approaches (advantage is can be ex post, but can also be ex ante)

Where o where art thou, baseline?

Where o where art thou, baseline?

- Existing datasets
  - Previous surveys
  - Monitoring data, but no comparison
- Recreating baselines
  - From existing data (e.g. 3ie working paper on Pakistan post-disaster)
  - Using recall: be realistic

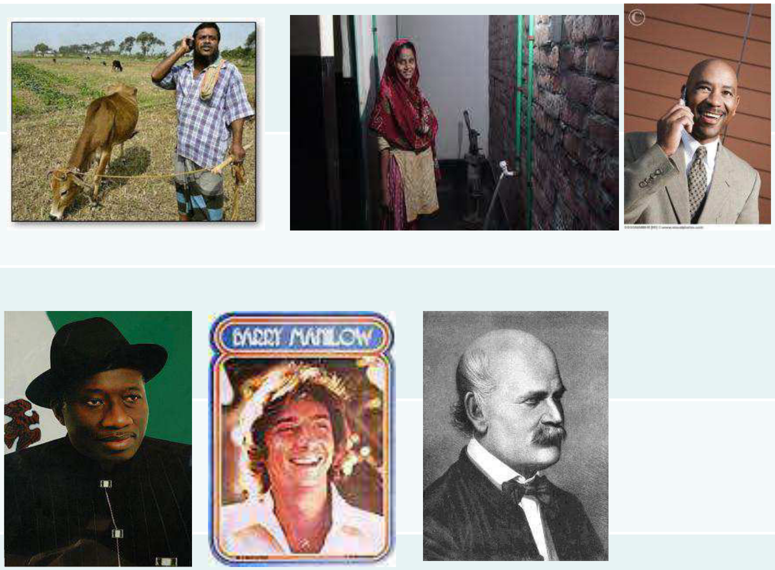
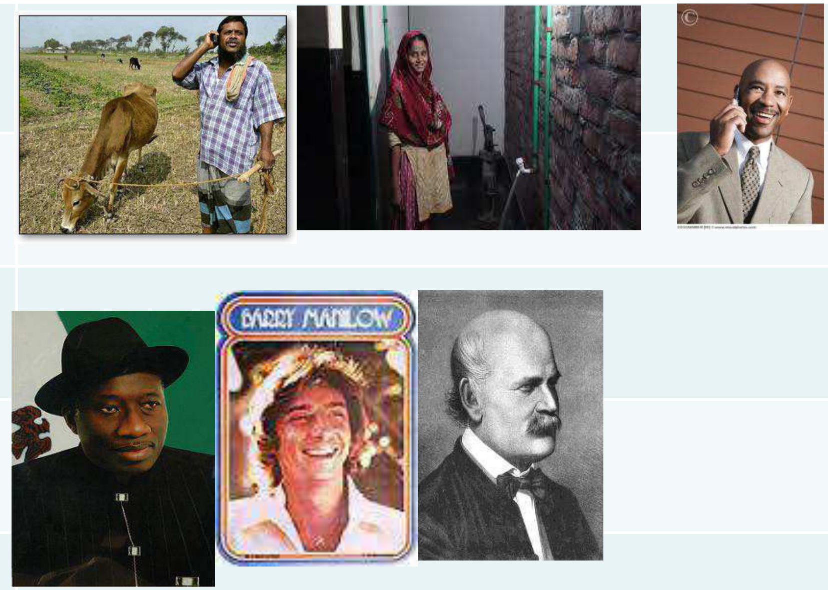
Matching methods

- Quasi-experimental methods (construct a comparison group)
  - Propensity score matching (PSM)
  - Regression discontinuity design (RDD)
  - 'Intuitive matching'
- Regression-based
  - Instrumental variables: need to be well-motivated

Difference in difference (DID) often listed as a method, but DID best done with matching

Propensity score matching

- Need someone with all the same age, education, religion etc.

Treatment	Comparison
	

Propensity score matching

- But, matching on a single number calculated as a weighted average of these characteristics gives the same result and matching individually on every characteristic – this is the basis of propensity score matching
- The weights are given by the 'participation equation', that is a probit equation of whether a person participates in the project or not

$$PART = \beta_0 + \beta_1 AGE + \beta_2 EDUC + \beta_3 ASSETS + \ ...$$

Propensity score matching: what you need

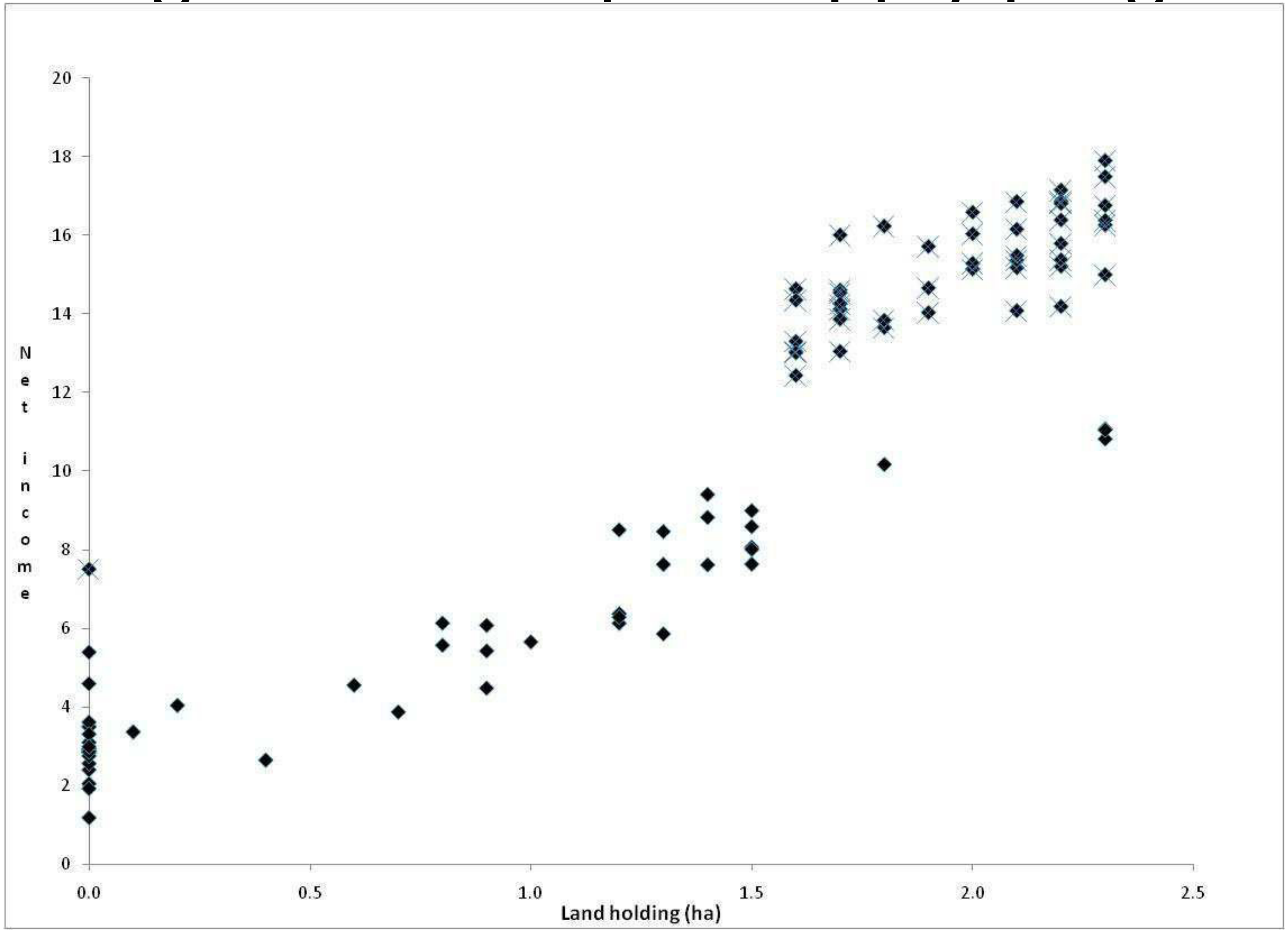
- Can be based on ex post single difference, though double difference is better
- Need common survey for treatment and potential comparison, or survey with common sections for matching variables and outcomes

Propensity score matching:

Example of matching: water supply in Nepal

Variable	Before matching	After matching
Rural resident	Treatment: 29% Comparison: 78%	Treatment: 33% Comparison: 38%
Richest wealth quintile	Treatment: 46% Comparison: 2%	Treatment: 39% Comparison: 36%
H/h higher education	Treatment: 21% Comparison: 4%	Treatment: 17% Comparison: 17%
Outcome (diarrhea incidence children<2)	Treatment: 18% Comparison: 23% OR = 1.28	Treatment: 15% Comparison: 23% OR = 1.53

Regression discontinuity: an example – agricultural input supply program



Naïve impact estimates

- Total = income(treatment) – income(comparison) = 9.6
- Agricultural h/h only = 7.7
- But there is a clear link between net income and land holdings
- And it turns out that the program targeted those households with at least 1.5 ha of land (you can see this in graph)
- So selection bias is a real issue, as the treatment group would have been better off in absence of program, so single difference estimate is upward bias

Regression discontinuity

- Where there is a ‘threshold allocation rule’ for program participation, then we can estimate impact by comparing outcomes for those just above and below the threshold (as these groups are very similar)
- We can do that by estimating a regression with a dummy for the threshold value (and possibly also a slope dummy) – see graph
- In our case the impact estimate is 4.5, which is much less than that from the naïve estimates (less than half)
- Where threshold is not perfectly applied use ‘fuzzy RDD’

Instrumental variables

- Want a variable which is correlated with having the intervention but NOT the outcome
- Can be hard to find in practice (random assignment is being treated as an instrument when regression used to get the treatment effect)
- E.g. Duflo paper on dams, uses gradient as instrument

Exercise

- For each evaluation question identify if it is large n or small n
- For the large n studies, for an ex ante design, could you randomize?
- What matching strategy could you use if a quasi-experimental approach

Thank you

Visit www.3ieimpact.org

Power calculations

Howard White, 3ie

Some sampling basics

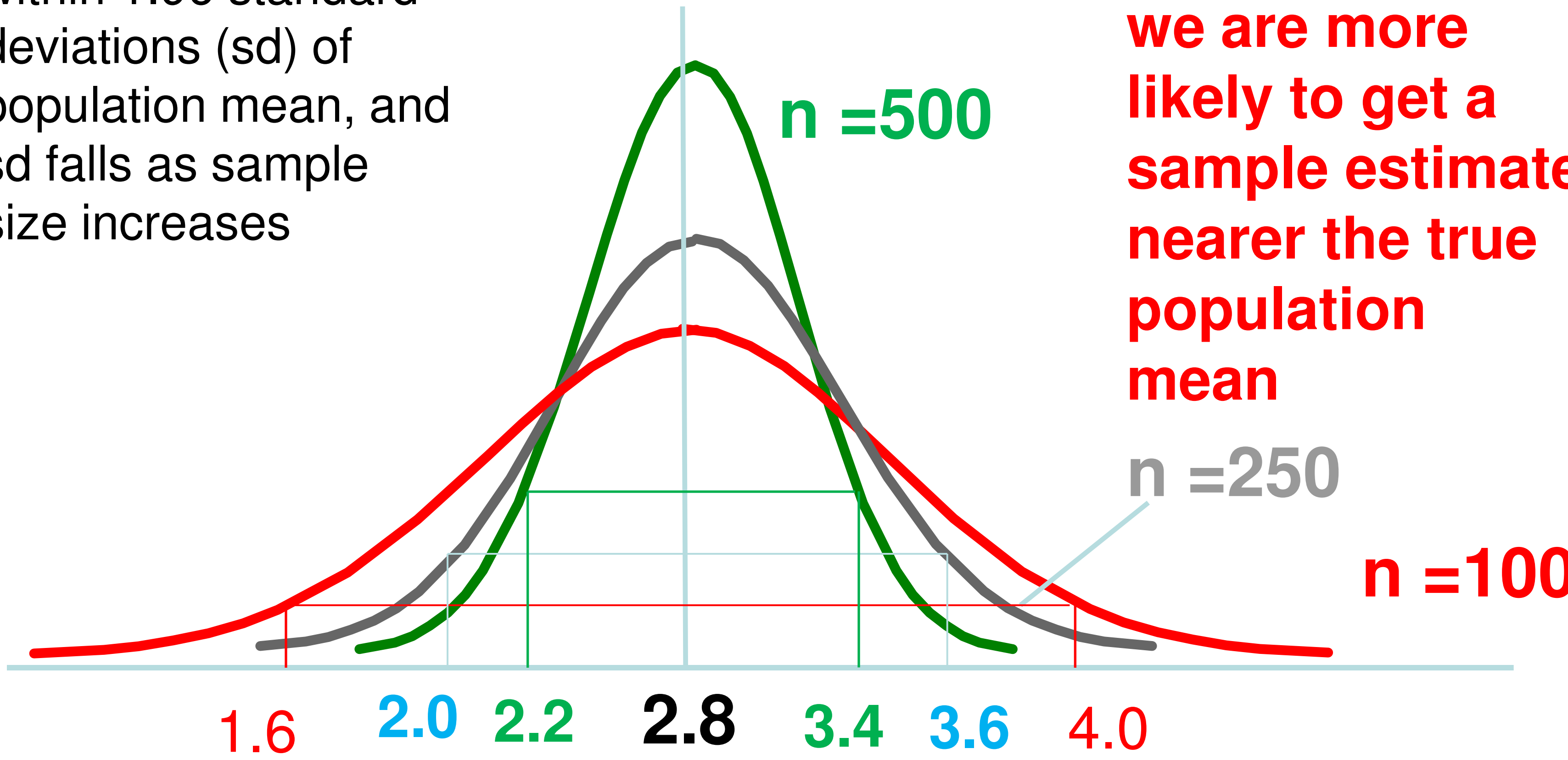
Population mean: the true value of a parameter, i.e. the average weight for age of all children aged under in the region of interest

Sample mean: the average weight for age in a sample drawn from the population

The larger the sample the more likely it is that the sample mean is close to the population mean (provided our sample is a *random* sample)

Distribution of sample means

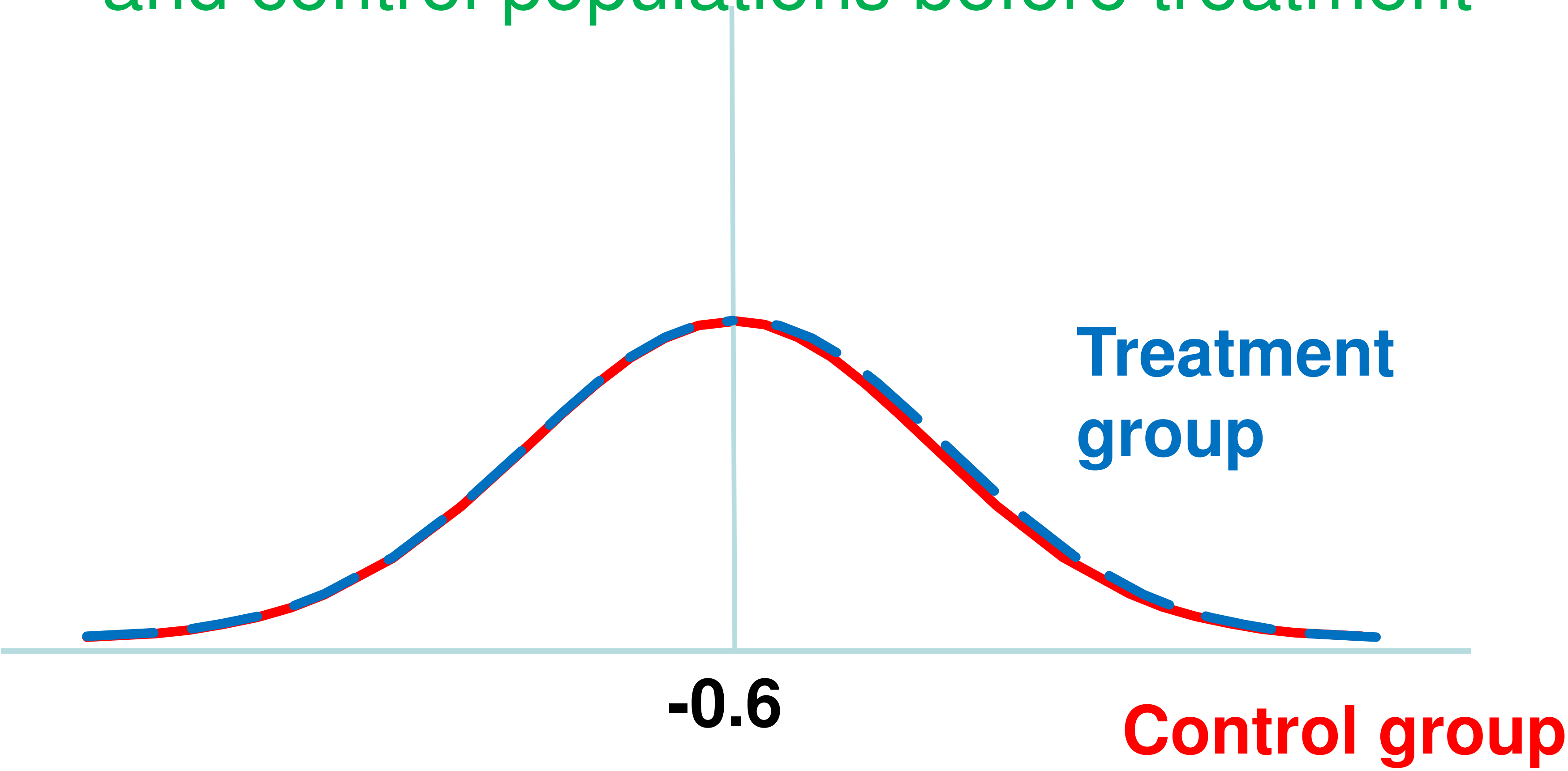
95% of estimates fall within 1.96 standard deviations (sd) of population mean, and sd falls as sample size increases



So as sample size increases we are more likely to get a sample estimate nearer the true population mean

This is the basis for large n designs. The sample is large enough to be representative of the populations, so we are reasonably sure that programme effects we measure are not exceptions

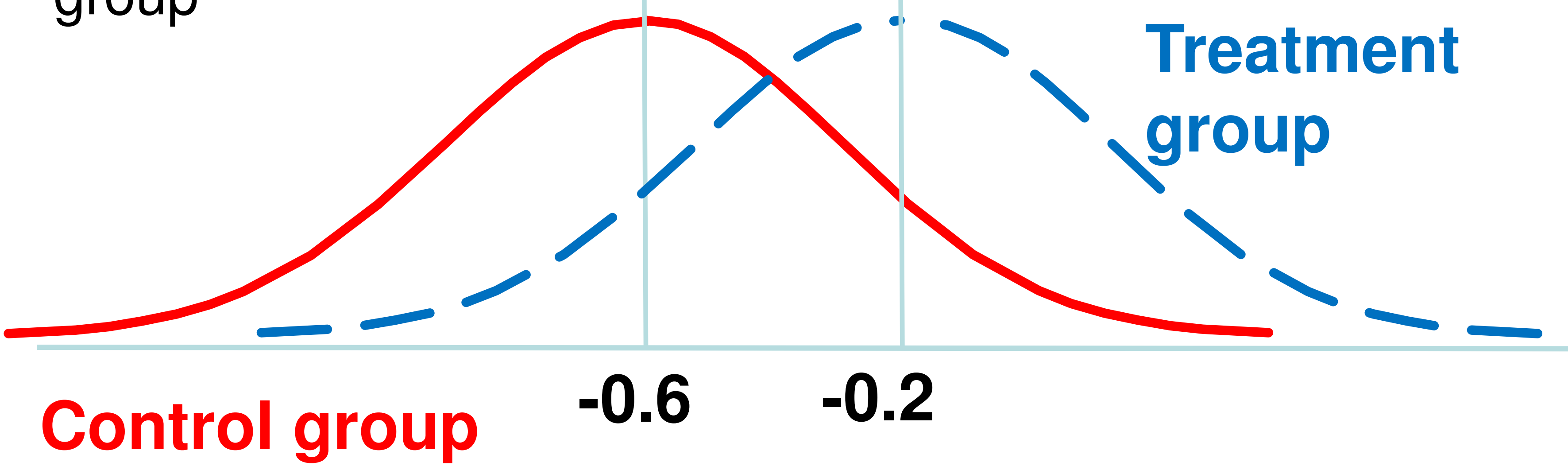
Distribution of WFA z score in the treatment and control populations before treatment



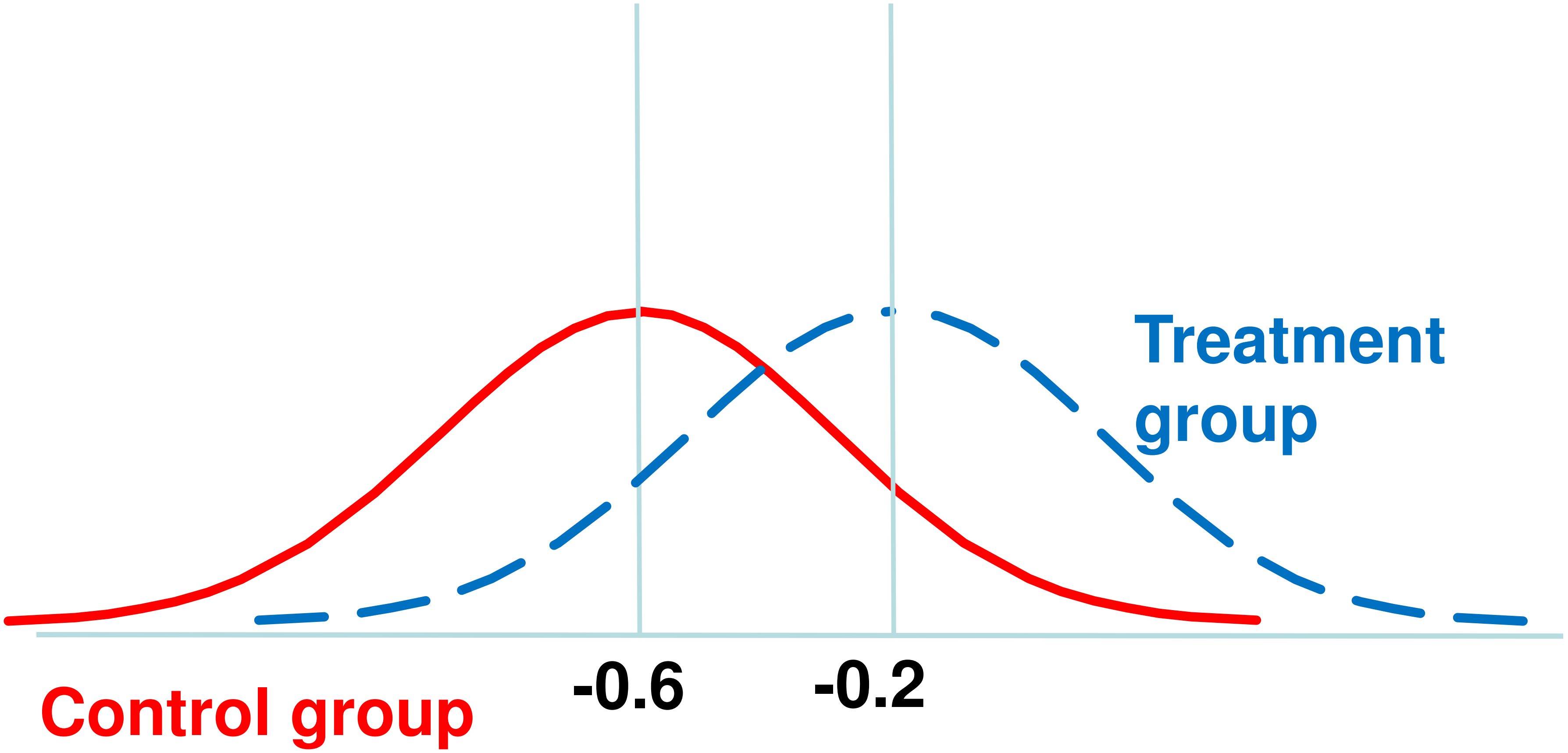
And after treatment

There are quite a lot of children in control group with higher WFA than children in the treatment group

So with too small a sample we may find treatment ineffective, or even harmful



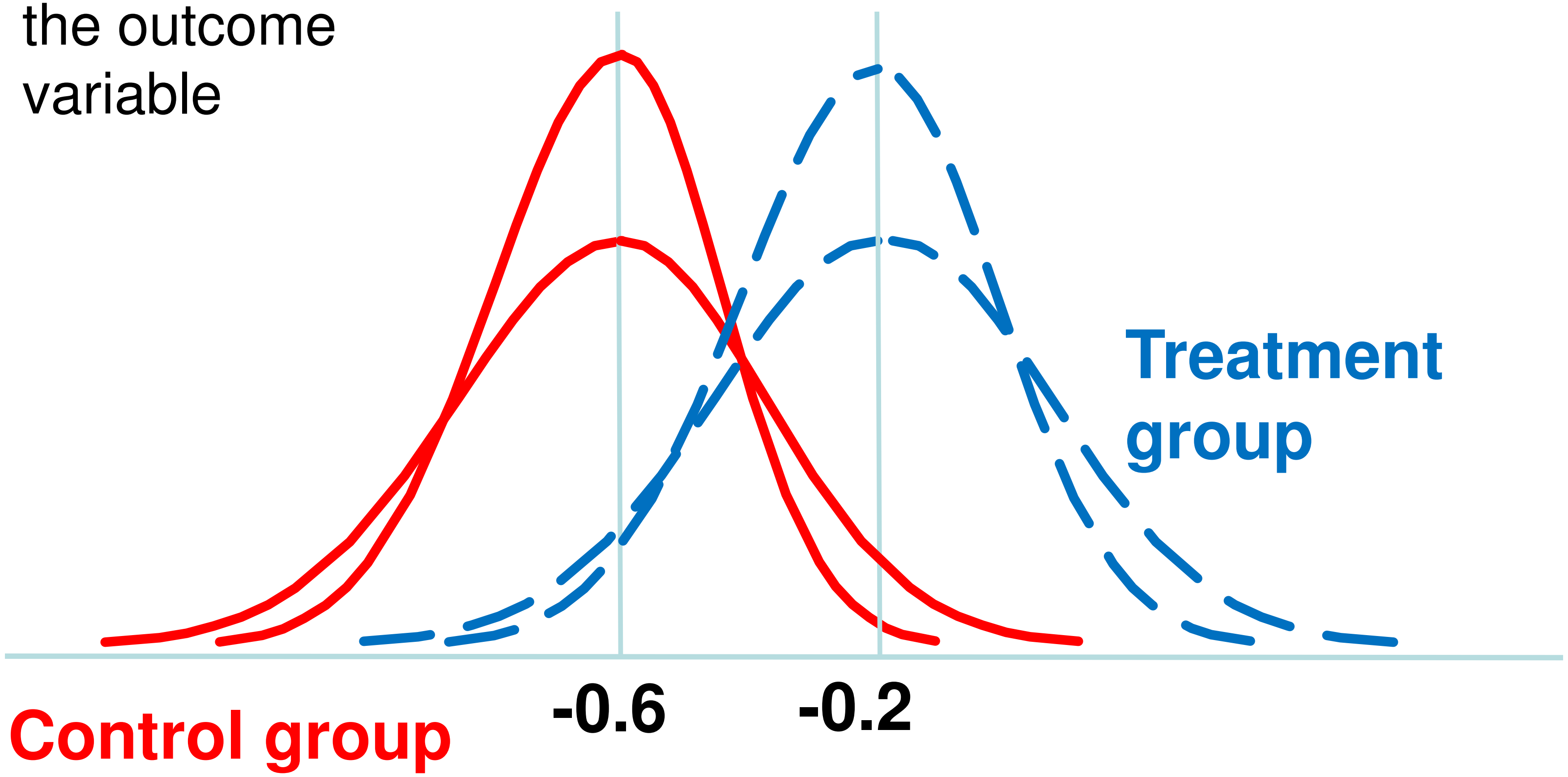
More formally we are testing hypothesis



So how large a sample do we need?

What makes it easier to detect programme impact?

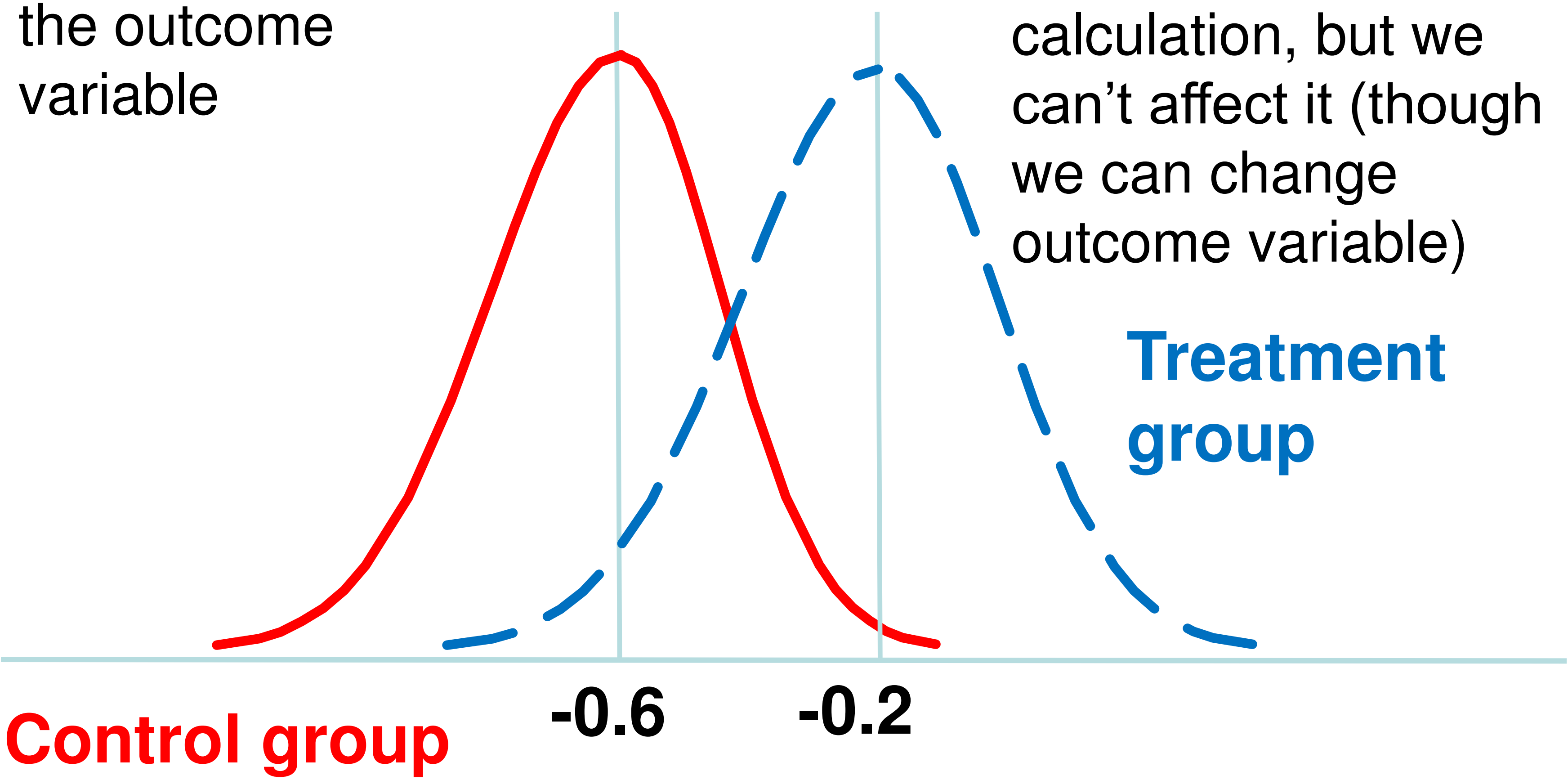
Less variability in the outcome variable



What makes it easier to detect programme impact?

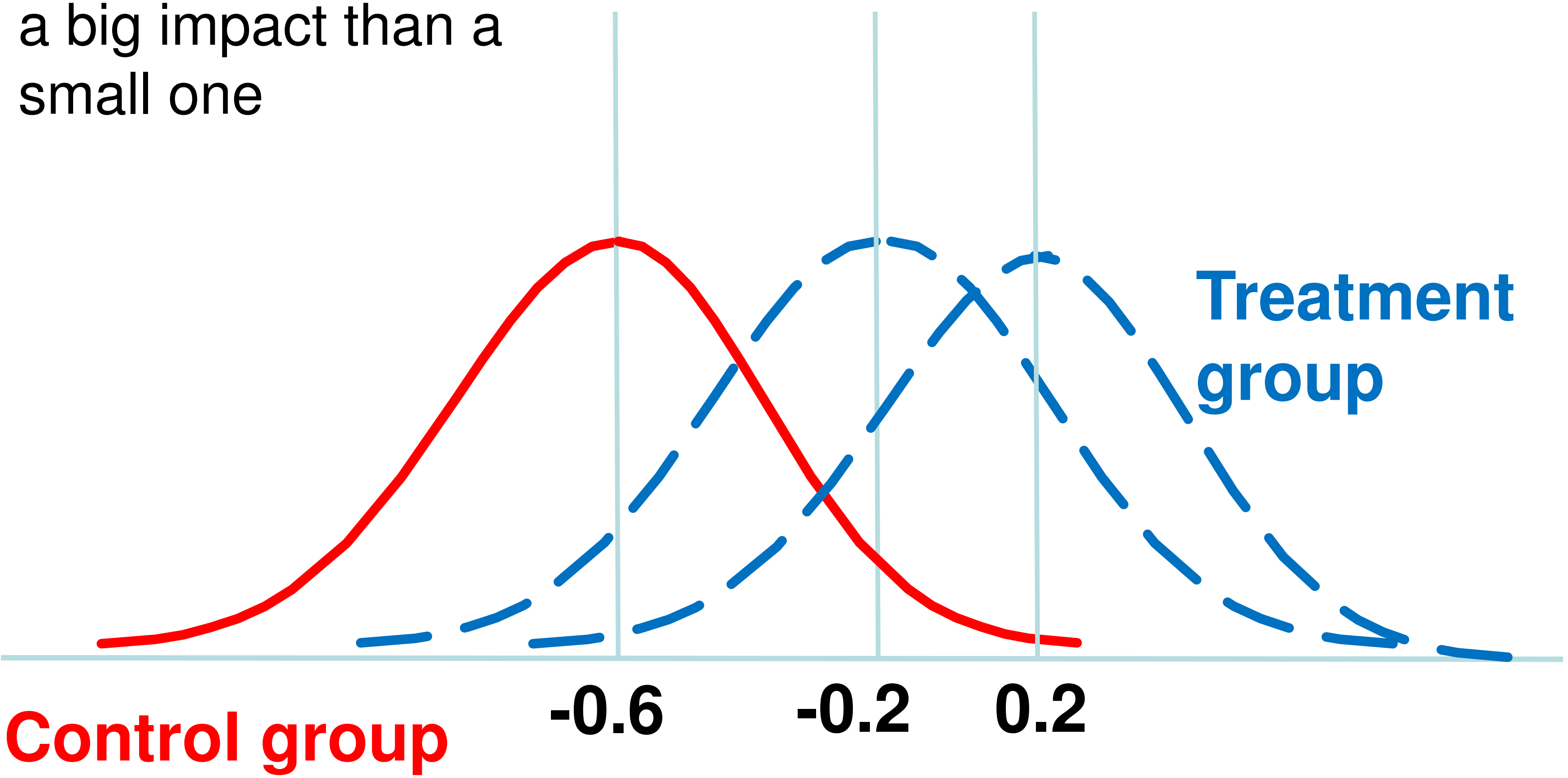
Less variability in the outcome variable

So we need to know that for our power calculation, but we can't affect it (though we can change outcome variable)



What makes it easier to detect programme impact?

It's easier to detect a big impact than a small one

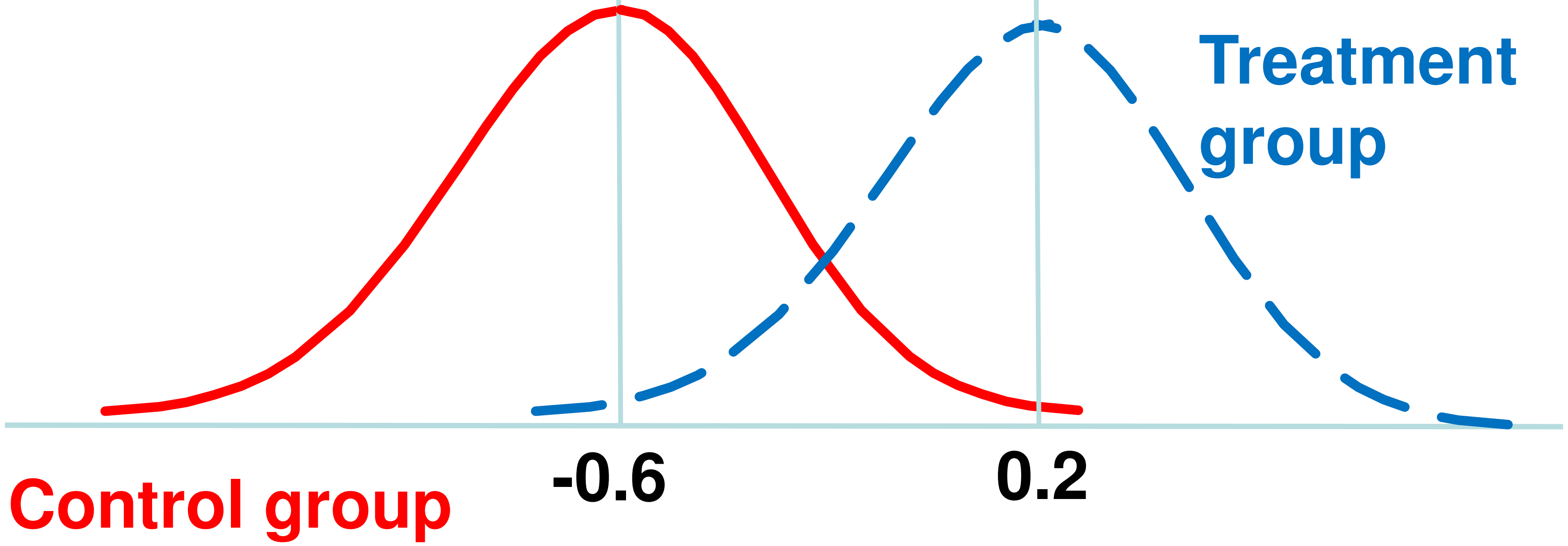


What makes it easier to detect programme impact?

It's easier to detect a big impact than a small one

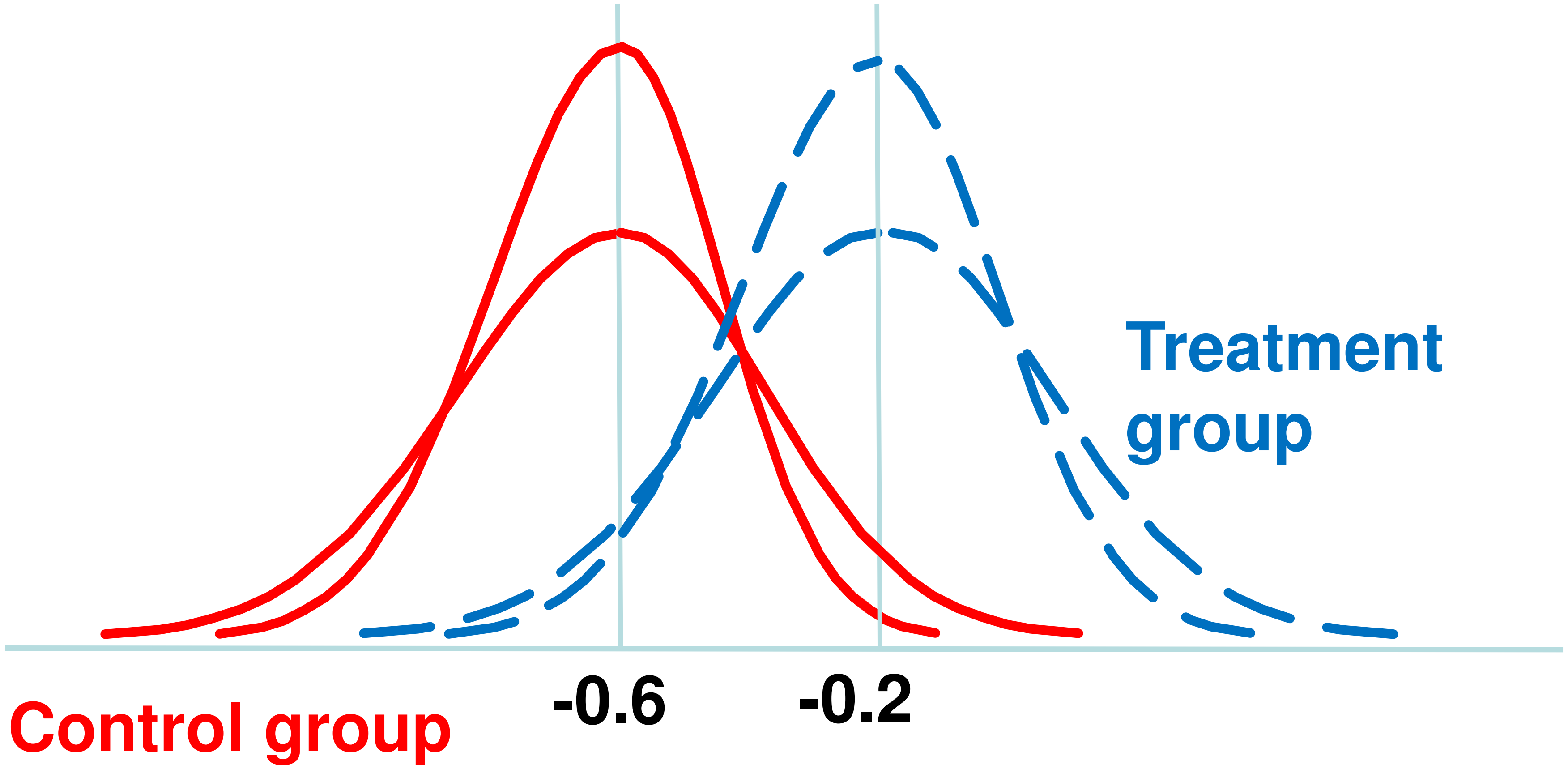
Policy makers determine the minimum effect we want to observe to make programme worthwhile

The bigger this is, the more likely we are to be able to measure impact



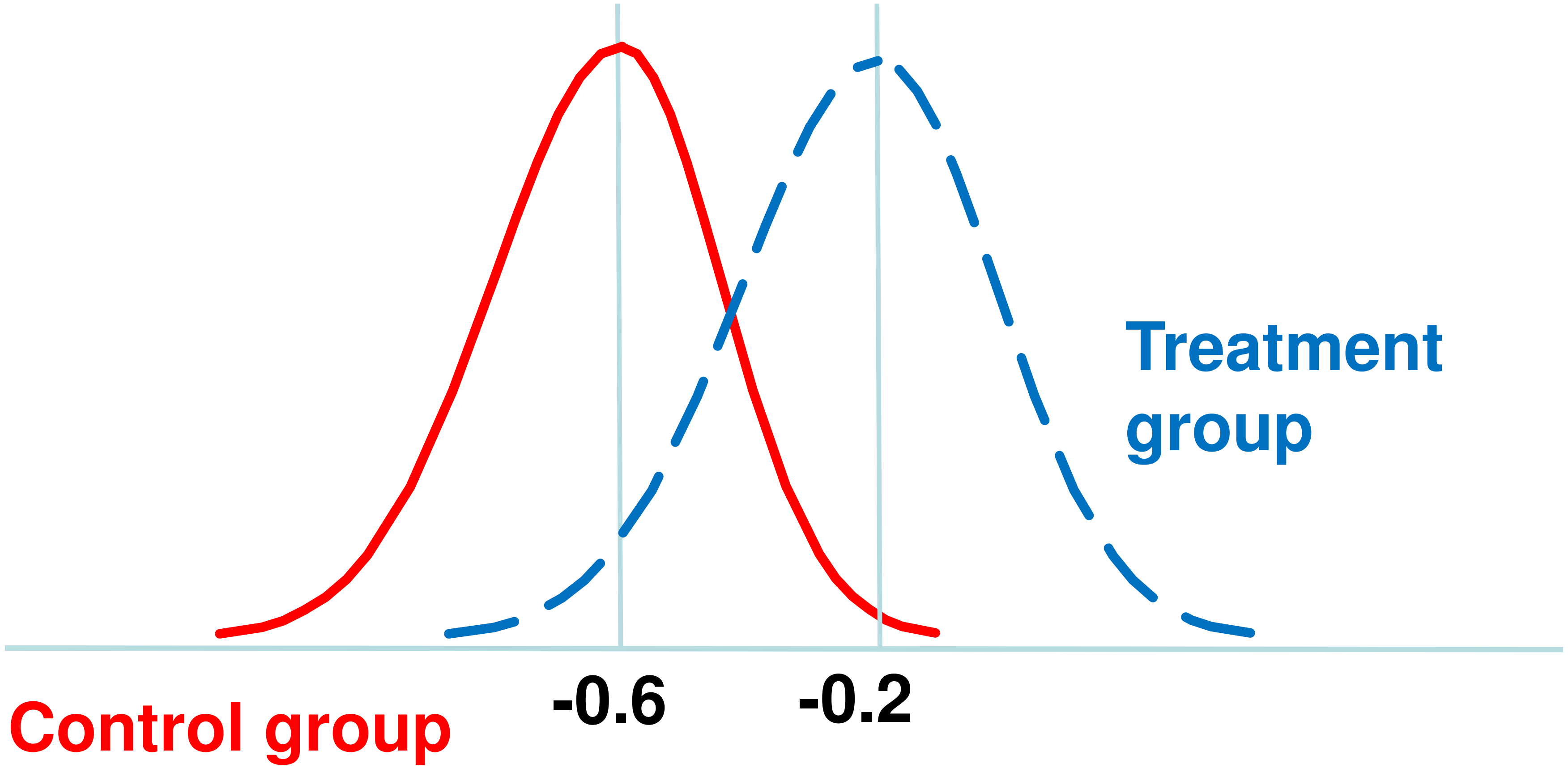
What makes it easier to detect programme impact?

A larger sample



What makes it easier to detect programme impact?

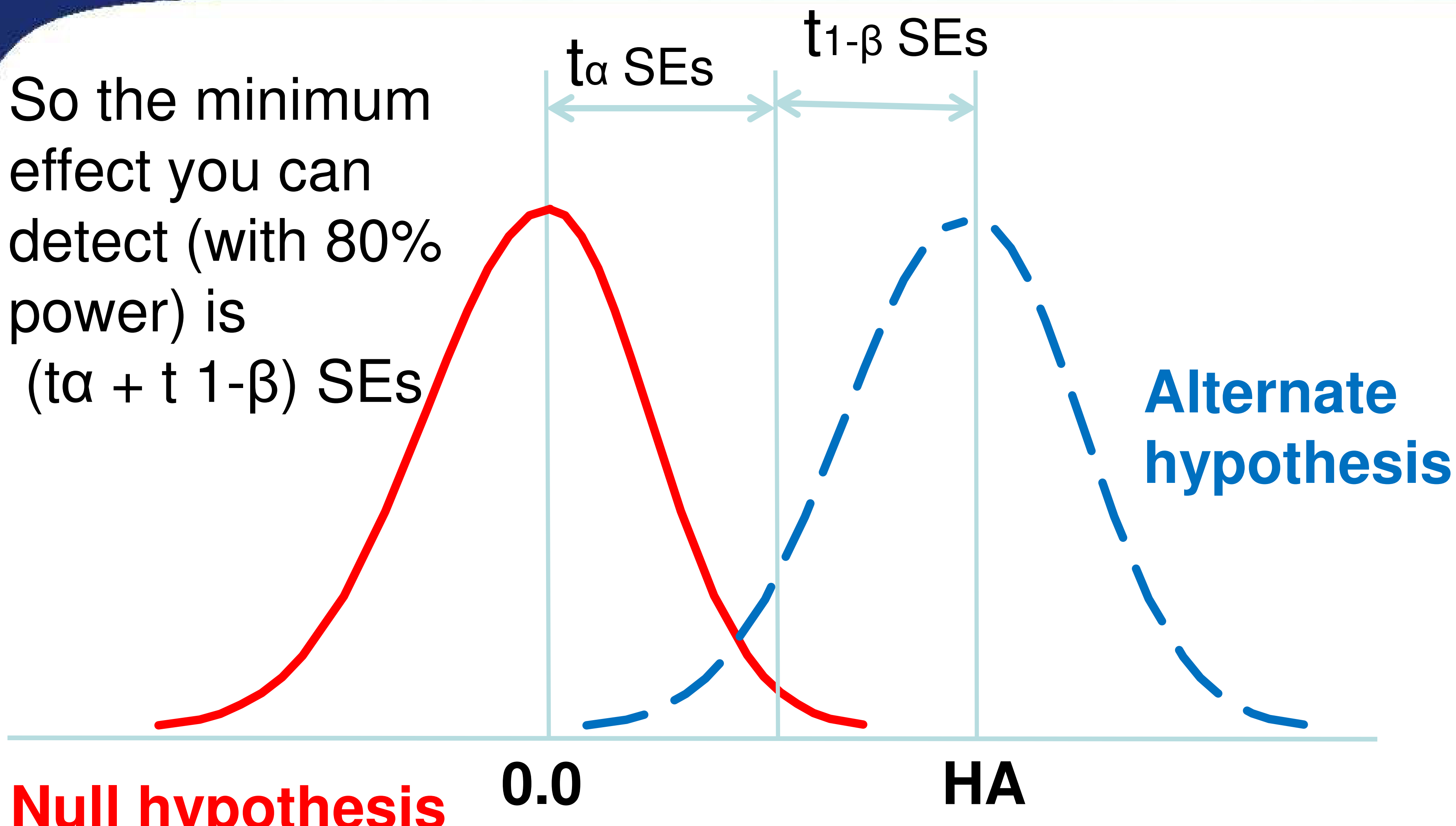
A larger sample



More formally

How far apart do the distributions need to be?

So the minimum effect you can detect (with 80% power) is
($t_\alpha + t_{1-\beta}$) SEs



Standard error of impact estimator

$$SE = \sqrt{\frac{\sigma^2}{n_t} + \frac{\sigma^2}{n_c}} = \sigma\sqrt{\frac{1}{P\,(1-P)n}}$$

- The noisier your outcome indicator, the harder it is to detect an effect
- We need an estimate of σ_y from another data source (as we haven't collected our own data yet)

So…

$$MDE = (t_{\alpha} + t_{1-\beta})\,\sigma_y \sqrt{\frac{1}{P(1-P)n}}$$

MDE is the Minimum Detectable Effect, that is the smallest effect you can expect to find with these sample characteristics. So you want MDE to be as SMALL as possible so you can find small effects.

Equal treatment and control samples

$$MDE = (t_{\alpha} + t_{1-\beta})\,\sigma_y \sqrt{\frac{1}{P(1-P)n}}$$

And obviously increasing n helps

MDE = f[1/P(1-P)]

$\delta(MDE)/\delta P = (1-P) - P = 1 - 2P = 0 \Rightarrow P = ½$

$\delta^2(MDE)/\delta P^2 = -2$ so maximize MDE

Do your own power calculation

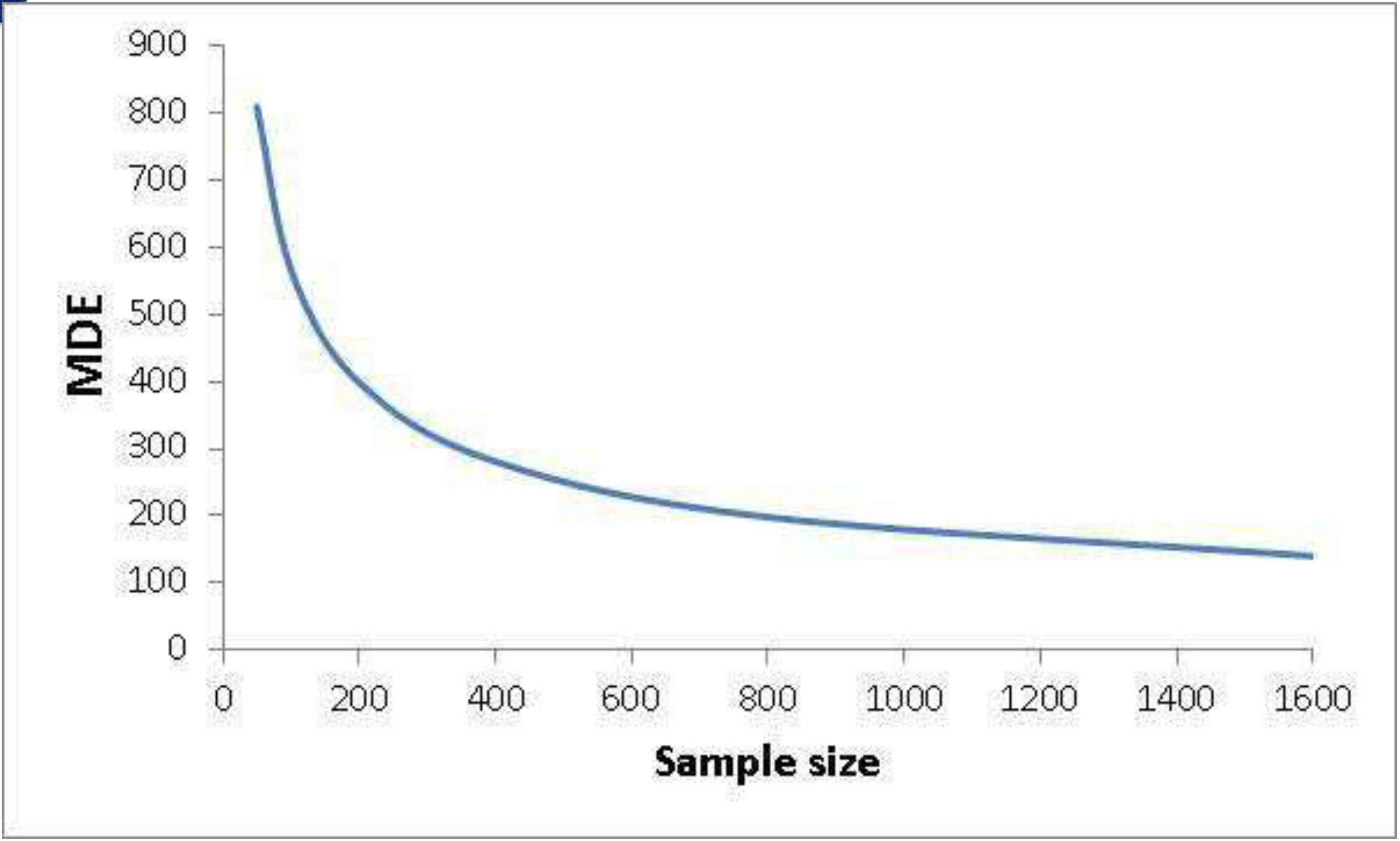
Power calculation

σy =	400	n	100
		p	0.5
alpha	0.975		
Power (1-beta)	0.8	MDE	226
t(alpha)	1.98	MDE size	0.57
t(beta)	0.85		

Here you have to type in the formula

use the t.inv(...) command

Increasing sample size has a decreasing effect on MDE



An Excel Exercise

- Average income in project area is Rs. 5,000 per month
- Using state data from national household income and expenditure survey $\sigma y = 1{,}000$
- What sample size do we need to detect a 5% increase in monthly income?
- The poverty line is Rs7,500. What sample size do we need if reaching that is the MDE?
- What is the risk of taking the goal of lifting people out of poverty for our power calculation?

Cluster designs

$$MDE = (t_\alpha + t_{1-\beta})\sqrt{\frac{\rho}{P(1-P)J} + \frac{1-\rho}{P(1-P)Jn}}$$

- **t refers to number of clusters, i.e. J-2 degrees of freedom**
- ***ρ* is intra-cluster correlation coefficient.**
- **Number of clusters drives power, not no. of observations in a cluster**

t Table

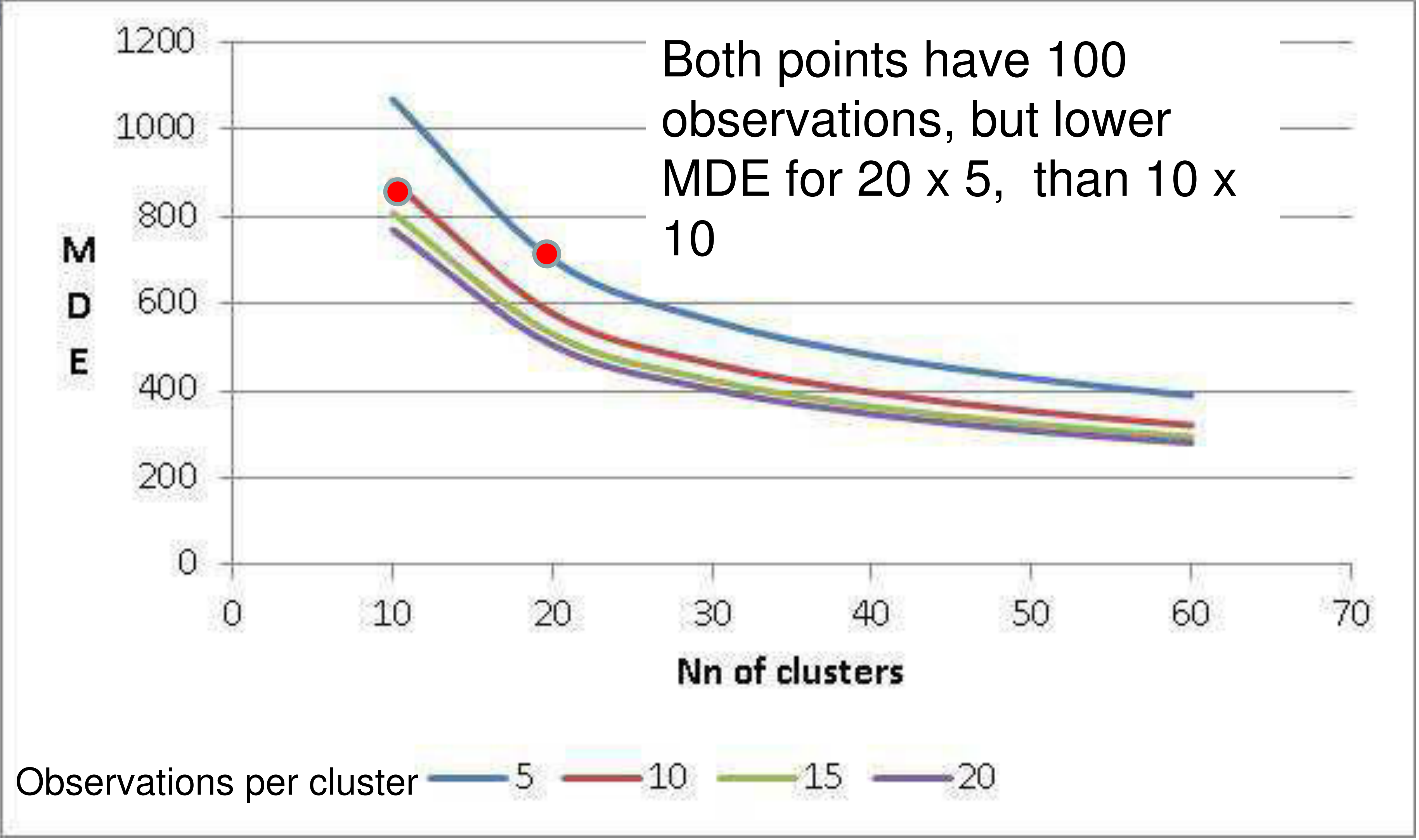
cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	[illegible]	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	[illegible]	[illegible]	2.776	3.747	[illegible]	[illegible]	8.610
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.015	2.571	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.943	2.447	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.729	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
20	[illegible]	[illegible]	[illegible]	[illegible]	1.325	1.725	2.086	[illegible]	[illegible]	[illegible]	[illegible]
21	0.000	[illegible]	[illegible]	1.063	1.323	1.721	2.080	[illegible]	[illegible]	[illegible]	[illegible]
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	[illegible]	[illegible]	3.792

MDE will be very large if n is low

That is there needs to be a huge impact for you to be able to detect it

Rho

- We want variation *within* clusters
- So a <u>lower</u> value of ρ is better
- If there is no variation it is as if each cluster is just one observation
- You need to use existing data to get a value of ρ, which will usually be in the range 0.15- 0.25



Both points have 100 observations, but lower MDE for 20 x 5, than 10 x 10
1200
1000
800
600
400
200
0
M
D
E
0
10
20
30
40
50
60
70
Nn of clusters
Observations per cluster
5
10
15
20

Remember

- We can increase power by covariate matching e.g. matched power randomization
- The formula for the power calculation varies with the design – see the 3ie Power Calculation Spreadsheet

And we need to increase sample size for

- Households which can’t be located
- Or aren’t in
- Or refuse
- Or return unusable data
- Or don’t comply with treatment

Rule of thumb is to add 20%

Exercise: How many clusters and total observations do you need?

MANAGING IMPACT EVALUATIONS

$$e^{i\pi} + 1 = 0$$

$$e^{iu} = \cos(u) + i \sin(u)$$

$$\gamma = \lim_{n \to \infty}\left(1 + \frac{1}{2} \ldots + \frac{1}{n} - \log(n)\right)$$

$$V - E + F = 2$$

$$S - I = \sum_{k=1}^{p} \frac{B_{2k}}{(2k)!}\left(f^{(2k-1)}(n) - f^{(2k-1)}(0)\right) + R$$

Howard White

International Initiative for Impact Evaluation

What to ‘impact evaluate?’

- Different stuff
  - Pilot and innovative programs
  - Innovative programs
- Established stuff
  - Representative programs
  - Important (flagship) programs
- Look to fill gaps

What do IE managers need to know?

- If an IE is needed and viable
- Your role as champion
- The importance of ex ante designs with baseline (building evaluation into design)
  - Funding issues
- The importance of a credible design with a strong team (and how to recognize that)
  - Help on design
- Ensure management feedback loops

Issues in managing IEs

- What team to commission?
- Different objective functions of managers and study teams
- Project management buy-in
- Trade-offs
  - On time
  - On richness of study design

THE
QUARTERLY
JOURNAL OF
ECONOMICS
FOUNDED 1886

Volume CXXI   February 2006   Issue 1

ARTICLES

PUBLISHED FOR HARVARD UNIVERSITY
BY THE MIT PRESS, CAMBRIDGE, MASSACHUSETTS

QJE 121(1) 1-349 (2006)
ISSN 0033-5533
No. 484

Overview on data collection

- Baseline, midterm and endline
- Treatment and comparison
- Process data
- Capture contagion and spillovers
- Quant and qual
- Different levels (e.g. facility data, worker data) – link the data
- Multiple data sources

Costs largely driven by large survey so additional rounds increase costs (marginal costs of increasing sample size are not so great)

Data used in BINP study

- Project evaluation data (three rounds)
- Save the Children evaluation
- Helen Keller Nutritional Surveillance Survey
- DHS (one round)
- Project reports
- Anthropological studies of village life
- Action research (focus groups, CNP survey)

Piggybacking

- Use of existing survey
- Add
  - Oversample project areas
  - Additional module(s)
- Lead time is longer, not shorter
- But probably higher quality data and less effort in managing data collection

Some study costs

- IADB vocational training studies: US$20,000 each
- IEG BINP study US$40,000-60,000
- IEG rural electrification study US$120,000
- IEG Ghana education study US$500,000
- Average 3ie study US$300,000 +
- Average 3ie study in Africa with two rounds of surveys; US$500,000 +

Some timelines

- Ex post 12-18 months
- Ex ante:
  - lead time for survey design 3-6 months
  - Post-survey to first impact estimates 6-9 months
  - Report writing and consultation 3-6 months
  - Then wait 5 years

Budget and timeline

- Ex post or ex ante
- Existing data or new data
- How many rounds of data collection?
- How large is sample?
- When is it sensible to estimate impact?

Exercise

- Propose for your intervention
  - Team composition
  - Management structure (quality assurance)
  - Timeline for impact evaluation
  - Budget

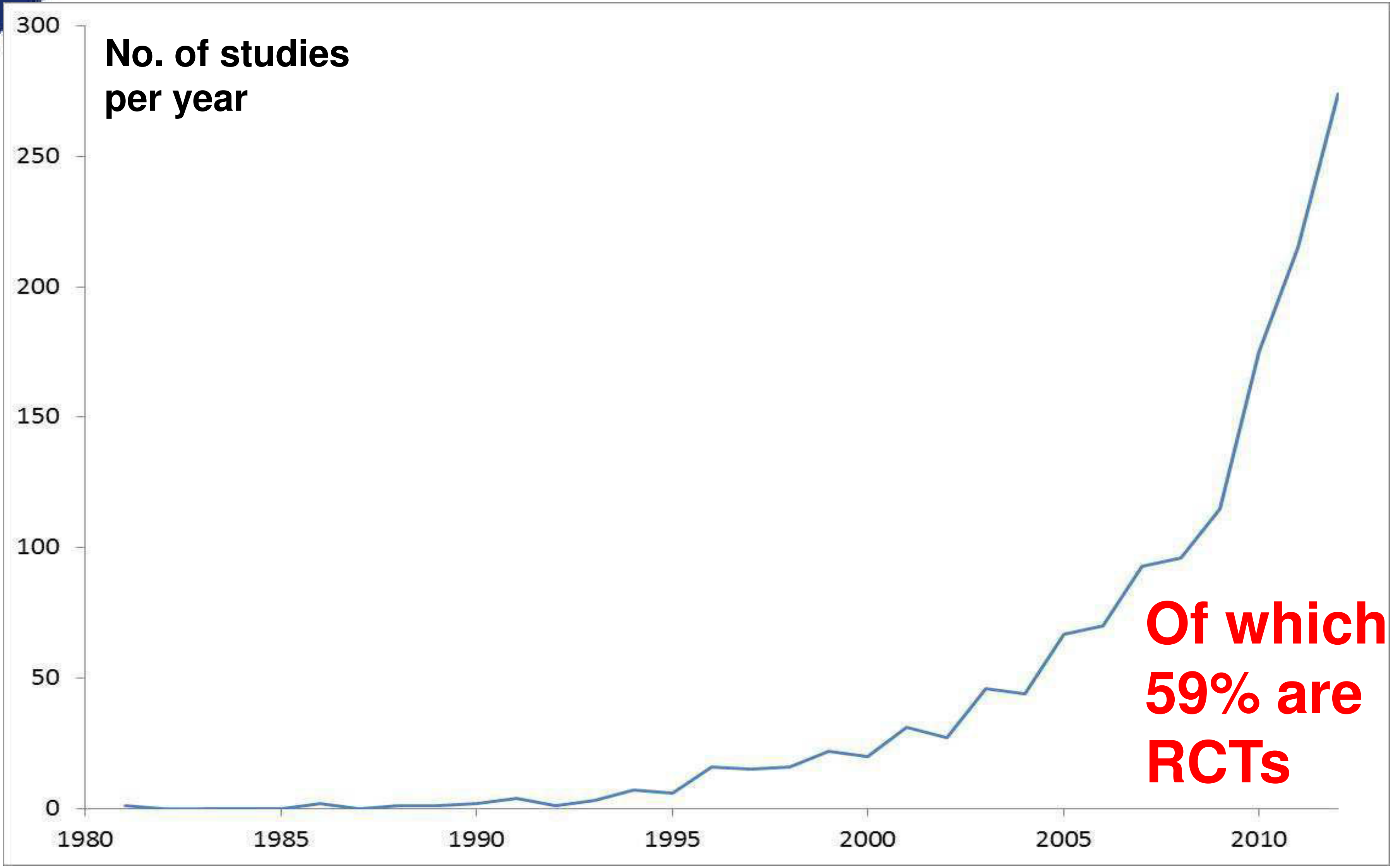
Thank you

Visit www.3ieimpact.org

Using impact evaluations for better policies and programmes, and better lives

Howard White
International Initiative for Impact Evaluation

The rapid growth in impact evaluations



Lesson One

Rigorous impact evaluations can and have yielded evidence which can be, and has been, used by policy makers for better policies and programmes

What should evidence be used for?

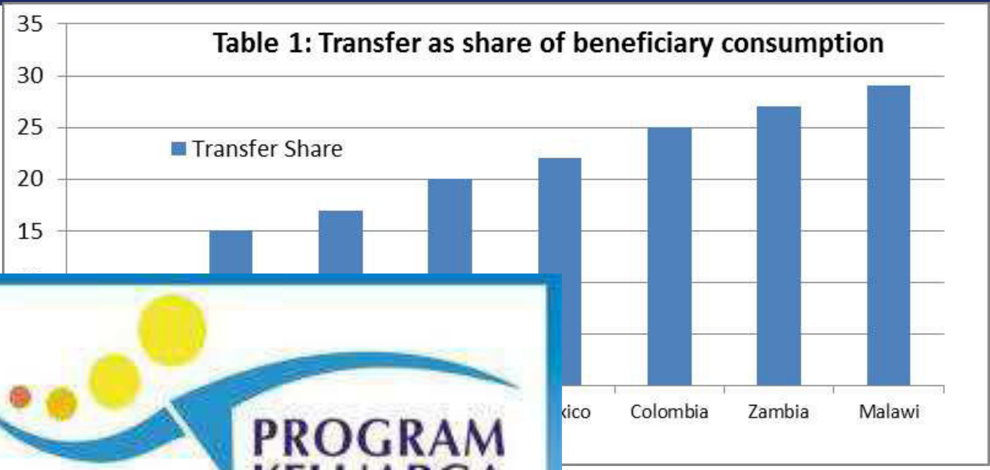
Going to scale

- Oportunidades (Mexico): national and international
- Pre-school (Mozambique)
- School-based nutrition (China)

What should evidence be used for?

Changing policy

- LEAP (Ghana): raise amount of the transfer
- PKH (Indonesia): revise targeting mechanism
- Irrigation (West Bengal): ease access for small farmers



What should evidence be used for?

Pilot to learn what works

- Cookstoves (Ghana)
- Wage subsidy (South Africa)

And close what doesn't

Lesson Two

Design studies to answer second generation (policy-relevant) questions

Second generation impact questions

Conditional Cash Transfers (CCTs)

Computer Assisted Learning (CAL)

And including formative research and evaluation

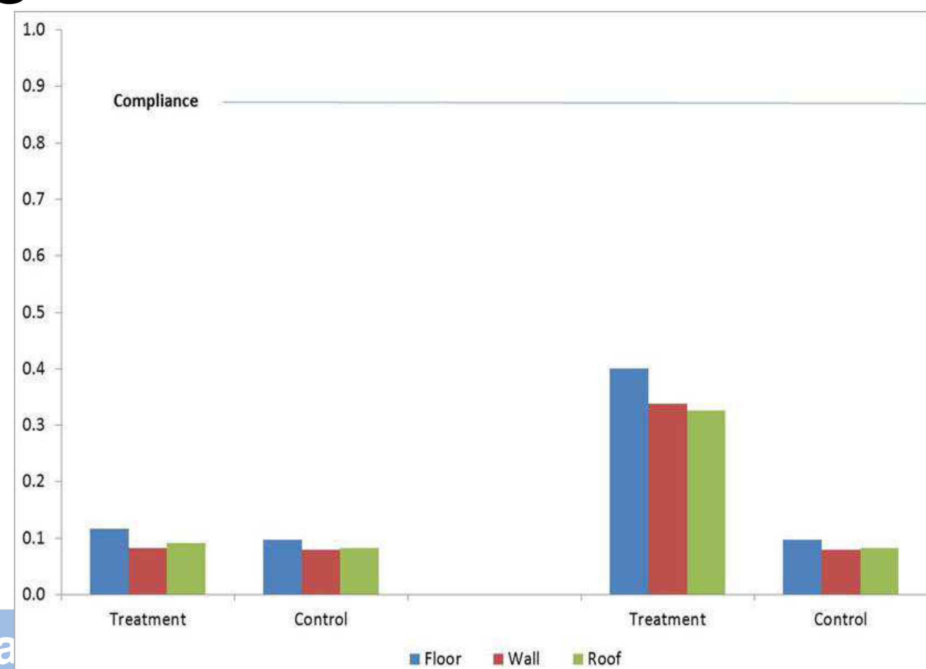
Texting:

- Parliamentarians
- Banking
- TB

Lesson three

Credible identification matters, but it is not being a RCT which makes an impact evaluation a gold standard, that also requires paying attention to context and answering the policy question of interest

The cult of significance



Lesson Four

The competing incentives of researchers and policy-makers needs careful management

What you want

What researchers want

Lesson Five

Policy influence is about both the product and the process

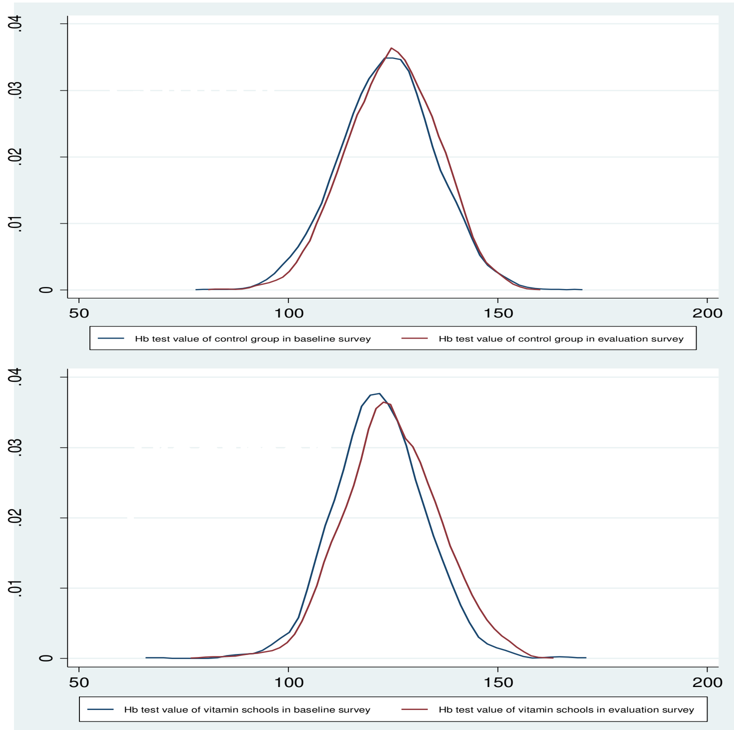
Getting the process right

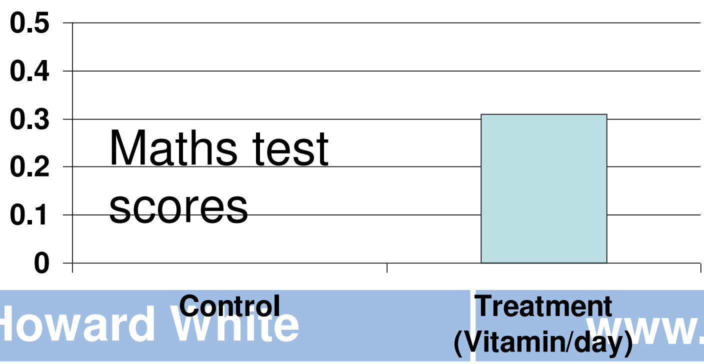
- Plan stakeholder engagement
- And do it from the start
- And monitor how you do it
- And present it right

Multivitamins to tackle anemia

Presenting results: the power of the anecdote

Anemia





For just 4 cents a day Wang went from being a C student to a B student

Please visit: www.3ieimpact.org/

Improving Lives with Impact Evaluation