

What is impact evaluation, and when and how should we use it?

AusAID
December 15, 2009

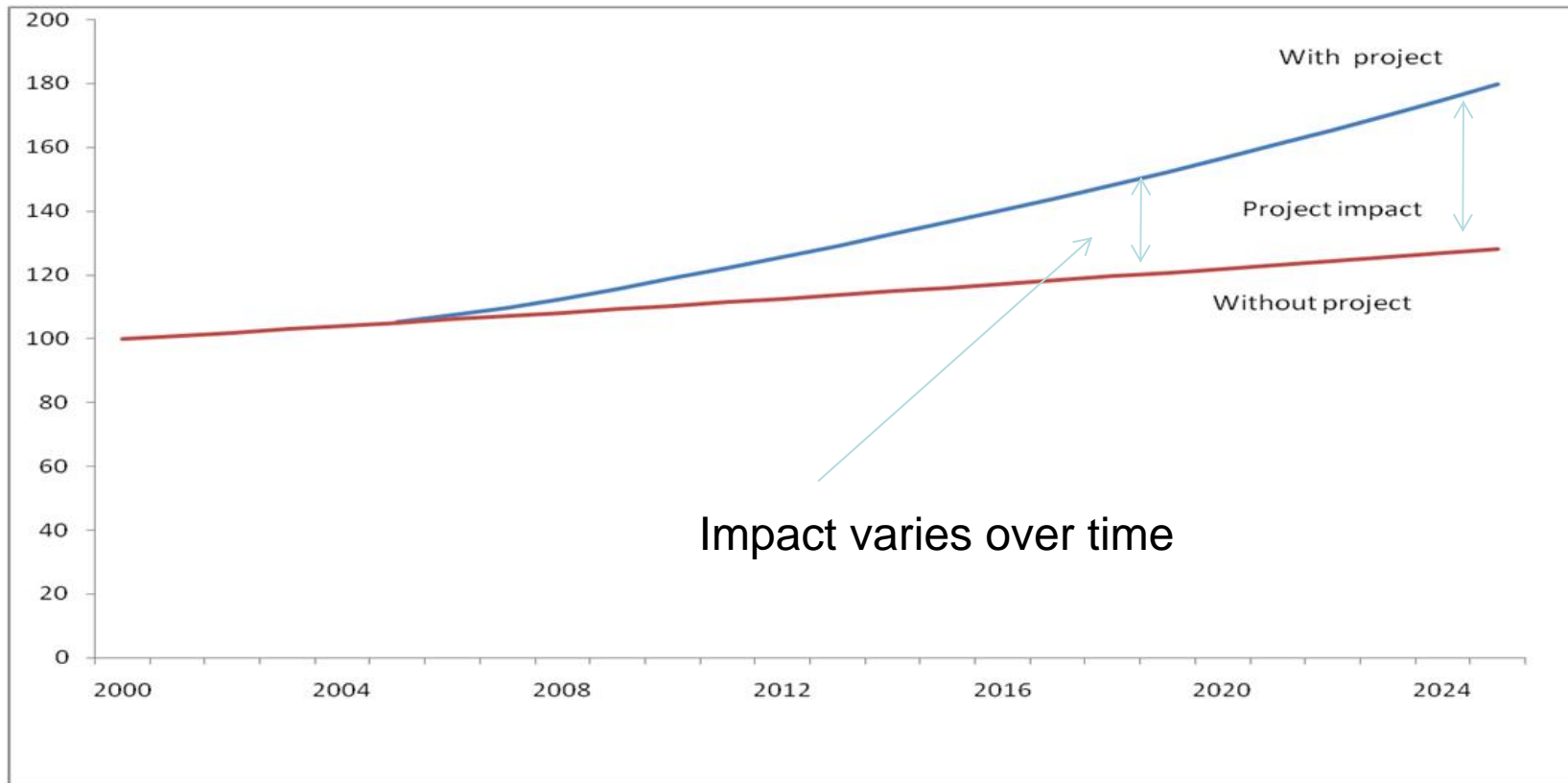
Howard White
International Initiative for Impact Evaluation

What is impact?

- Impact = the outcome with the intervention compared to what it would have been in the absence of the intervention
- Unpacking the definition
 - Can include unintended outcomes
 - Can include PAPs not just intended beneficiaries
 - No reference to time-frame, which is context-specific
 - At the heart of it is the idea of a attribution – and attribution implies a counterfactual (either implicit or explicit)

Defined in this way we have little evidence
on impact of development programs i.e.
we don't know the results of those
programs

The attribution problem: factual and counterfactual



What has been the impact of
the French revolution?

“It is too early to say”

Zhou Enlai

What do we need to measure impact? Girl's secondary enrolment

	Before	After
Project (treatment)		66
Control		

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

Before versus after single difference comparison

$$\text{Before versus after} = 66 - 40 = 26$$

	Before	After
Project (treatment)	40	66
Control		

Sometimes this can work e.g. Water supply and time use... but usually not

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

Outcome monitoring does not tell us about effectiveness

*Results... cannot as a rule be attributed specifically, **either wholly or in part**, to the Netherlands (“Results report 2005-06”)*

Before versus after: water supply

	Sri Lanka	Tanzania
Time taken to collect water (minutes)		
Before	24	176
After	14	13
Incidence child diarrhea (prevalence last 2 weeks)		
Before	1.9	12.6
After	1.8	10.4

Post-treatment control comparison

Single difference = $66 - 55 = 11$

	Before	After
Project (treatment)		66
Control		55

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

$$\text{Double difference} = (66-40)-(55-44) = 26-11 = 15$$

	Before	After
Project (treatment)	40	66
Control	44	55

Conclusion: Longitudinal (panel) data, with a control group, allow for the strongest impact evaluation design (though still need matching)

Main points so far

- Analysis of impact implies a counterfactual comparison
- Outcome monitoring is a factual analysis, and so cannot tell us about impact
- The counterfactual is most commonly determined by using a control group



If you are going to do impact evaluation you need a credible counterfactual using a control group

However....

- This is for ‘large n’ interventions
 - There are a large number of units of intervention, e.g. children, households, firms, schools.
 - Examples of small n are policy reform and many (but not all) capacity building projects.
 - Some reforms (e.g. health insurance) can be given large n designs
- ‘Small n’ interventions require either
 - Modelling (computable general equilibrium, CGE, models), e.g. trade and fiscal policy
 - Qualitative approaches, e.g. the impact of impact assessments
 - A theory-based large n study may have elements of small n analysis at some stages of the causal chain (this will be explained this afternoon)

But our focus is on learning why things work, not just what: **theory-based impact evaluation**
(*measurement is not evaluation*)

An example followed by principles

See 3ie working paper 3

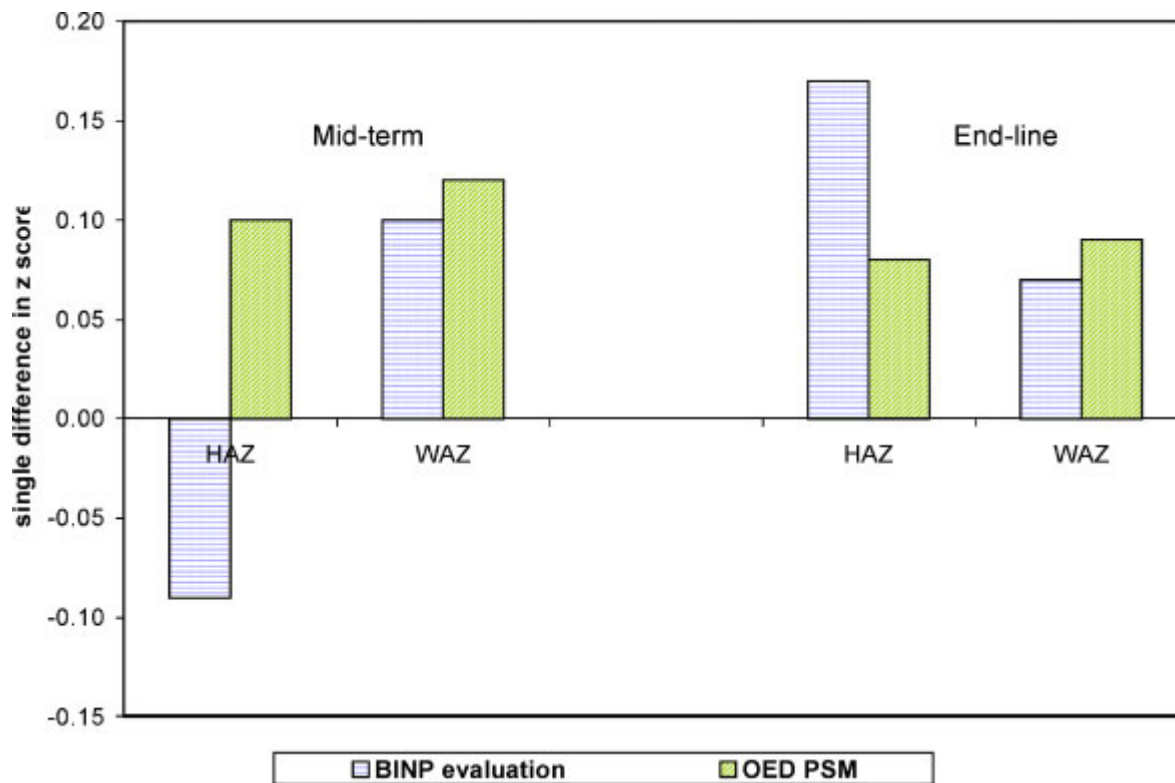
Theory-based impact evaluation: an example

- Bangladesh Integrated Nutrition Project (BINP)
- Growth monitoring, nutritional counselling and supplementary feeding (based on TINP)
- Implemented by NGOs at field level, using Community Nutrition Practitioners (CNPs)

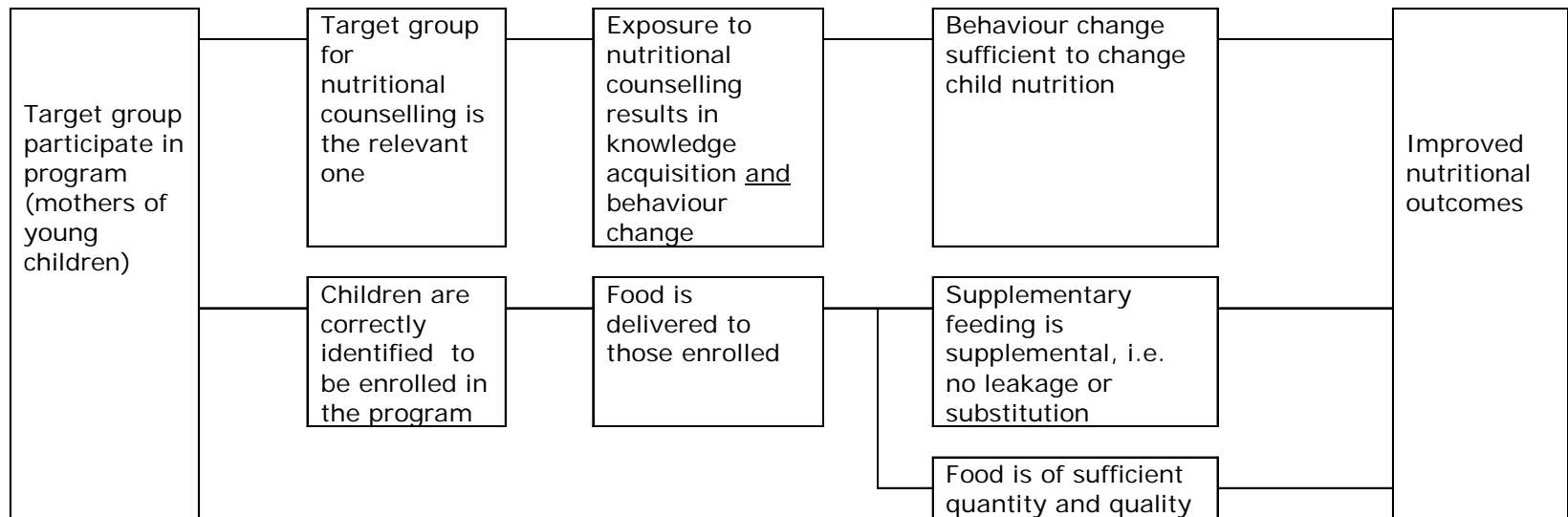
The evaluation story

- Looked like it was working – all bits in place and monitoring data showed sharp fall in severe malnutrition
- Bank agreed to scale up
- But Save the Children UK critical, though Bank's evaluation positive
- Bank's evaluation department (IEG) did evaluation – found little or no impact
- Theory-based approach explains why

Impact estimates (using propensity score matching)



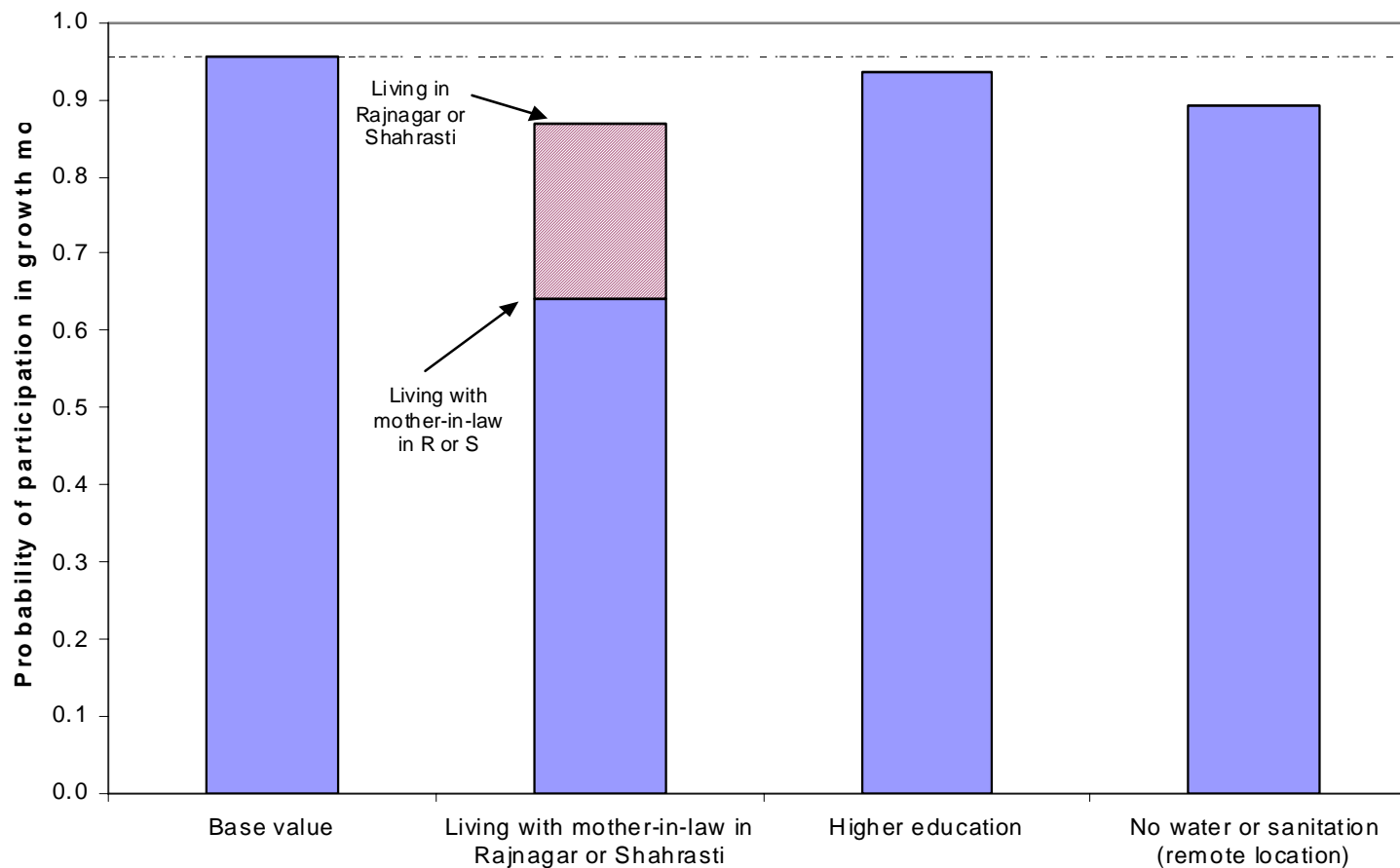
Program theory



Assumption	Findings
Provide nutritional counseling to care givers	Mothers are <u>not</u> decision makers, especially if they live with their mother-in-law
Women know about sessions and attend	90% participation, lower in more conservative areas
Malnourished and growth faltering children correctly identified	No – community nutrition practitioners cannot interpret growth charts
Women acquire knowledge	Those attending training do so
And knowledge is turned into practice	No there is a substantial knowledge-practice gap
Supplementary feeding is additional food for intended beneficiary	No, considerable evidence of substitution and leakage
Adopted changes are sufficient to improve intended outcomes	Only sometimes (not for pregnant women)

Source: Howard White and Edoardo Masset (2007) 'The Bangladesh Integrated Nutrition Program: findings from an impact evaluation' *Journal of International Development* 19: 627-652

Participation rates



Illustrating the principles

- **Map out the causal chain** (programme theory): see figure
- **Understand context**: Bangladesh is not TN
- **Anticipate heterogeneity**: more malnourished children; different implementing agencies
- **Rigorous evaluation** of impact using an appropriate counterfactual: PSM versus simple control
- **Rigorous factual** analysis: targeting, KP gap, CNPs
- **Use mixed methods**: informed by anthropology, focus groups, own field visits

Problems in implementing rigorous impact evaluation: selecting a control group

- Contagion: other interventions
- Spill over effects: control affected by intervention
- Selection bias: beneficiaries are different
- Ethical and political considerations

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 program of school improvements is targeted at the poorest schools
- Since these schools are in poorer areas it is likely that students have home and parental characteristics associated with lower learning outcomes (e.g. illiteracy, no electricity, child labour)
- Hence learning outcomes in project schools will be lower than the average for other schools
- The comparison group has to be drawn from a group of schools in similarly deprived areas

Selection bias from self-selection

- A community fund is available for community-identified projects
- An intended outcome is to build social capital for future community development activities
- But those communities with higher degrees of cohesion and social organization (i.e. social capital) are more likely to be able to make proposals for financing
- Hence social capital is higher amongst beneficiary communities than non-beneficiaries regardless of the intervention, so a comparison between these two groups will overstate program impact

Examples of selection bias

- Infant mortality in Bangladesh:
 - Hospital delivery (0.115 vs 0.067)
 - Immunization status (0.062 vs 0.094)
 - Breastfeeding (0.03 vs. 0.77)
- Secondary education and teenage pregnancy in Zambia
- Male circumcision and HIV/AIDS in Africa

Main point

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

Dealing with selection bias

- Need to use experimental or quasi-experimental methods to cope with this; this is what has been meant by rigorous impact evaluation
- Experimental (randomized control trials = RCTs, commonly used in agricultural research and medical trials, but are more widely applicable)
- Quasi-experimental
 - Propensity score matching
 - Regression discontinuity
 - Pipeline approach
 - Regressions (including instrumental variables)

Randomization (RCTs)

- Randomization addresses the problem of selection bias by the random allocation of the treatment
- Randomization may not be at the same level as the unit of intervention
 - Randomize across schools but measure individual learning outcomes
 - Randomize across sub-districts but measure village-level outcomes
- The less units over which you randomize the higher your standard errors
- But you need to randomize across a ‘reasonable number’ of units
 - At least 30 for simple randomized design (though possible imbalance considered a problem for $n < 200$)
 - Can be as few as 10 for matched pair randomization

Issues in randomization

- Randomize across eligible population not whole population
- Can randomize across the pipeline
- Is no less unethical than any other method with a control group (perhaps more ethical), and any intervention which is not immediately universal in coverage has an untreated population to act as a potential control group
- No more costly than other survey-based approaches

Conducting 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)
- Adopt strict randomization protocol
- Maintain information on how randomization done, refusals and 'cross-overs'
- A, B and A+B designs (factorial designs)
- Collect baseline data to:
 - Test quality of the match
 - Conduct difference in difference analysis

Quasi-experimental approaches

- Possible methods
 - Propensity score matching
 - Regression discontinuity
 - Instrumental variables
- Examples are given this afternoon
- Advantage: can be done ex post, and when random assignment not possible
- Disadvantage: cannot be assured of absence of selection bias

Main points

Be issues-driven not methods driven

Find best available method for evaluation questions at hand

Randomization often *is* possible

But do ask, is this sufficiently credible to be worth doing?

So when to do an impact evaluation?



- Pilot programs
- Innovative programs
- Representative or important programs

Thank you

Visit www.3ieimpact.org