

# Monitoring and Evaluation for Results

## The Practice of Impact Evaluation

## Sources

This presentation draws heavily on work done by several persons, including:

- Martin Ravallion, DEC
- Emmanuel Skoufias and Markus Goldstein, PREM
- Sebastian Martinez and Paul J. Gertler, HDN
- Howard White, IEG
- Robert Black, Johns Hopkins University
- Cesar Victora, University of Pelotas
- Jennifer Bryce, Johns Hopkins University
- Mickey Chopra, MRC Unit, South Africa
- Anuraj Shankar, WHO Geneva
- Development Impact Evaluation (DIME) Initiative

## Why monitor or evaluate?

**"Without data, you are just another person with an opinion"**

**–Anonymous**

## What is the purpose of M&E?

- **Improve the well being of the population!**
- **Improve quality of services**
- **Provide up-to-date estimates of status**

## Purpose of M&E (continued)

- **Ascertain changes and trends in indicators**
  - Changes over time
  - Differences in equity
- **Provide service or intervention coverage data**
- **Inform decision-making to improve:**
  - Services
  - Resource allocation
  - Policy

## M&E is an intervention

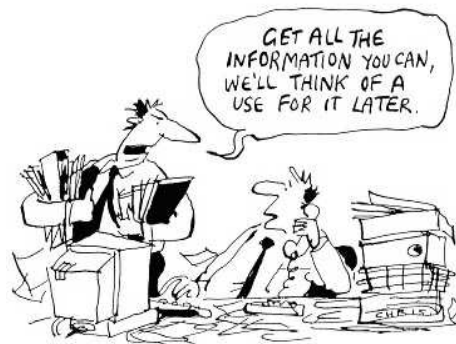
- Evidence transforms opinions into facts and can catalyze action....
- M&E can be applied anytime and anywhere:
  - Need to identify the purpose
  - Select the indicators
  - Develop the assessment methodology
  - Analyze the data
  - Use the results!

## What to measure?

- Characteristics of good indicators:
  - Can be measured with required accuracy and precision
  - Reflect the health status of the population
  - Sensitive to change
  - Easily interpreted
- Some common health indicators include:
  - Mortality
  - Vaccination coverage
  - Proportion of births with a skilled attendant

## Common Pitfalls in M&E

- Not action driven
- Collection and analysis is separate
- Poor reporting:
  - Not understandable or too complex
  - To the wrong persons
- Delay in feedback of results
- Lack of culture of use of data.  
M&E is equated with just filling in data, registers etc.,





## A sad state of affairs?

### • Data remains:

- Unprocessed
- Unanalyzed
- Not written-up
- Not read
- Not acted on

Source: Chambers, 1983

## desire for proxies of impact leads to epidemic of indicators and targets

### • By disease programmes (from M&E guides)

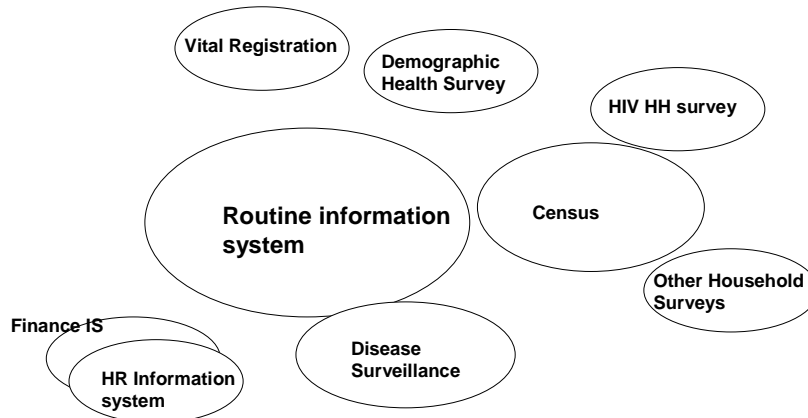
<u>Activity</u>	<u>Indicators</u>
– HIV/AIDS:	142
– TB:	57
– Reproductive health	148
– Adolescent reproductive health	292
– Child health:	102
– Essential drugs:	98
– Decentralization and health:	83

### • By global initiative

– Health for All by the Year 2000:	20
– MDGs	17 (30)
– UNICEF World Summit for Children:	40
– UNICEF World Fit for Children:	101

Chopra 2008

## Fragmentation of different data sources



Chopra 2008

## Where has this led?

- Too much data, not enough information
  - Poor data quality and use
  - Donor or project driven M&E approach
  - Confusion on indicators
  - Lack of use of data....

## Some steps for solutions

- Develop evaluation approach that is valid, as simple as possible, and implemented at the inception of the program
- use available data and tools as a starting point, and enhance as needed based on a clear rationale
- Promote use of data for decisions that can be implemented
- Make information available at all levels



## The Practice of Impact Evaluation

**Evaluation** of social programs, projects, or policies:

A systematic assessment to determine the relevance, efficiency, effectiveness, impact, and sustainability of a planned, ongoing, or completed interventions.

- Needs assessment
- Evaluability assessment and design assessment
- Process evaluation, implementation evaluation
- Tracer studies, mid-term evaluations, etc.
- Impact evaluation (e.g., “rigorous”) and meta-evaluation

## What is impact evaluation?

Impact evaluations assess the specific outcomes attributable to a particular intervention or program. They do so by comparing outcomes where the intervention is applied against outcomes where the intervention does not exist.

An appropriate comparison represents what would have happened in the absence of the intervention. By establishing a valid comparison of outcomes for these two groups, an impact evaluation seeks to provide direct evidence of the extent to which an intervention changes outcomes.

*Development Impact Evaluation (DIME) Initiative, World Bank (2007)*

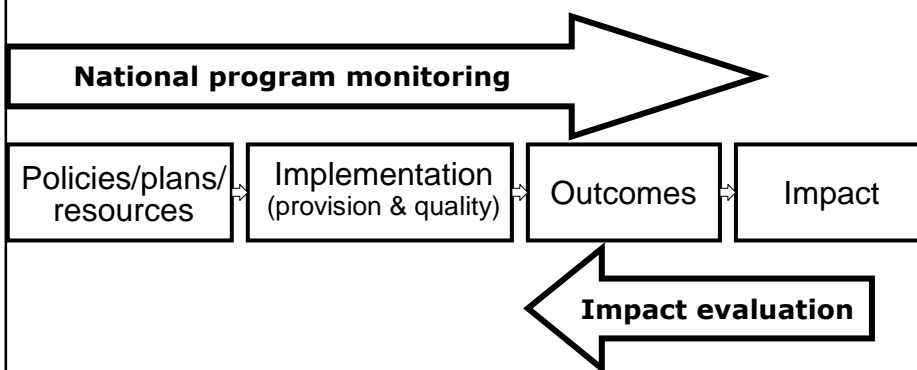
## What is results-based monitoring?

- Results monitoring is a continuous process of collecting and analyzing information to compare how well a project, program or policy is performing against expected results. (Sometimes referred to as results-based management, management for results, and performance measurement.)
- Typical features
  - No need for comparison groups.
  - No inherent causality (but predicated on a results framework or logic model)
  - Performance targets are clear and measurable.
  - Distinct from, but complementary to, an impact evaluation

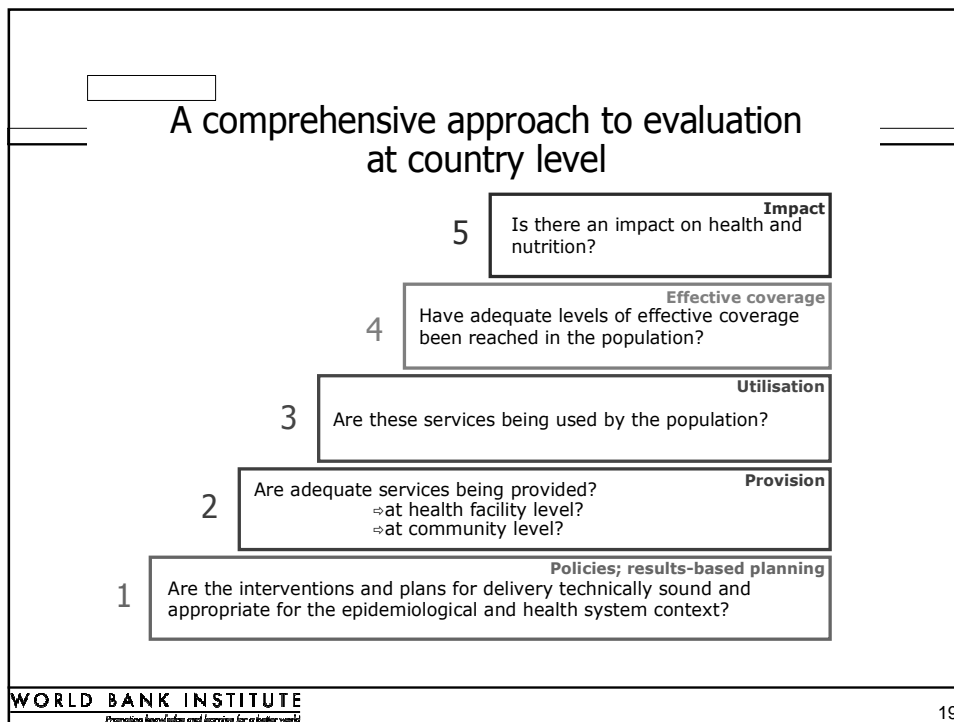
## When might it be appropriate to pursue impact evaluations?

- Similar projects, programs, or policies but with conflicting or divergent results
- **Allocating resources across** alternative investment options
- Identify issues around a problem (e.g., drug use, dropout rates)
- Pilot projects (however, if scientific basis already exists to justify cause-effect relationships, a “rigorous” impact evaluation may be unnecessary.)
  - If other evidence of effectiveness is good, and potential benefits are large, how much more evidence is necessary?
  - Difference between results monitoring and impact evaluation
- Increase transparency, credibility, or accountability

## National monitoring and impact evaluation are complementary



Black 2008



## The Basics of Program Evaluation

- Program evaluation is a set of techniques used to determine if an intervention ‘works’
- To determine the causal effect of the treatment we need knowledge of counterfactuals, that is, what would have happened in the absence of the intervention
- Observational studies, or associations, will not estimate causal effects

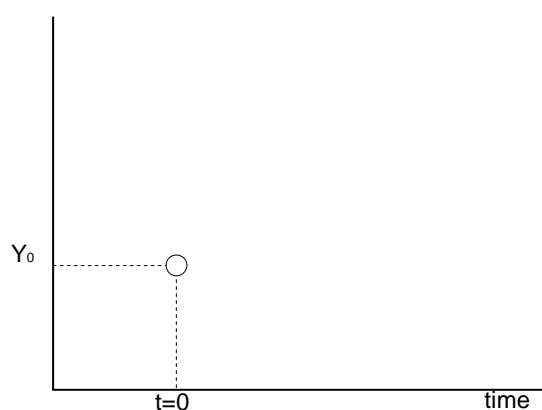
**WORLD BANK INSTITUTE**  
*Promoting knowledge and learning for a better world*

20

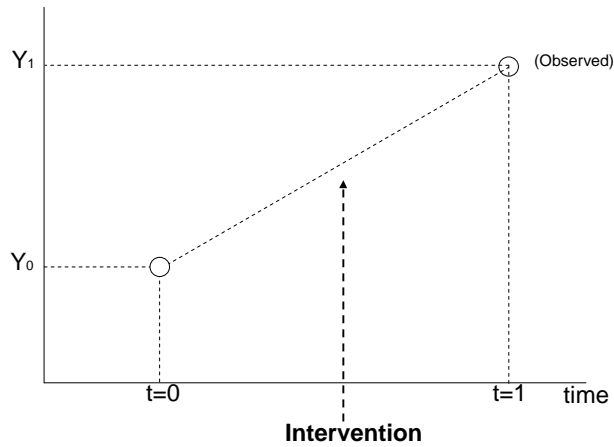
## Types of Impact Evaluations

- Note that almost all types of impact evaluations have a comparison group and baseline data
- The most widely used quantitative impact evaluation designs
- When baseline data do not exist, one can attempt to *reconstruct baseline data*

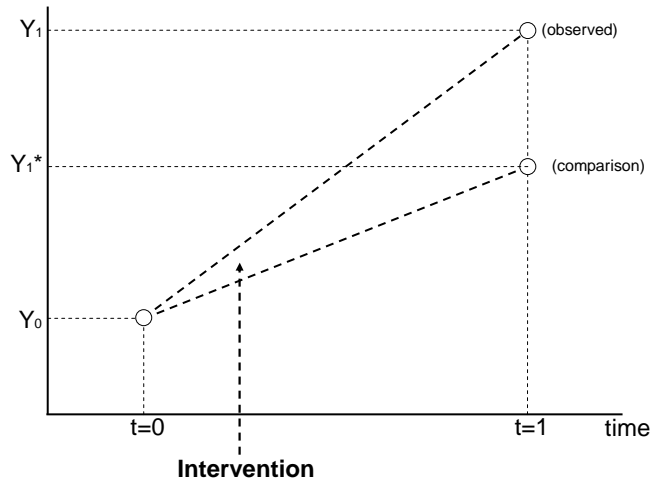
## We observe an outcome indicator



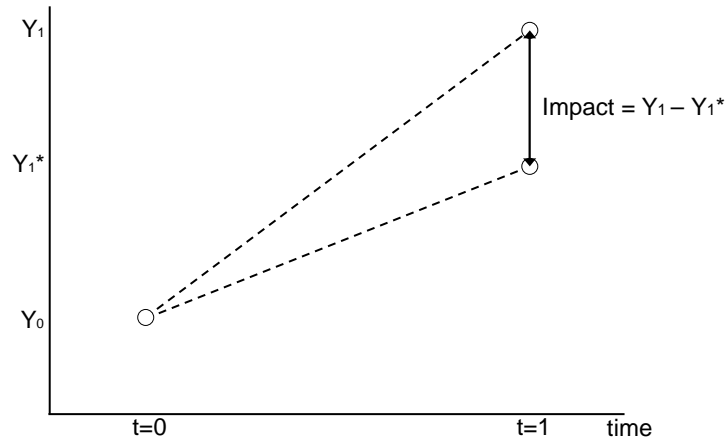
And its value rises after the intervention:



Having the “ideal” comparison...



## Allows us to estimate the true impact



## Continuation...

- The control is intended to estimate what would have happened to the project population if the intervention had not occurred
- The difference between the change in  $t_1$  for the project and  $t_1$  for the control group is an estimation of impact
- This is sometimes called the true impact or change score

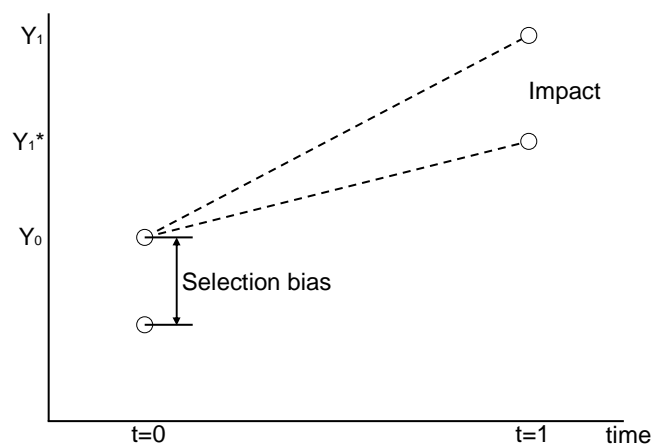
## Continuation...

- The validity of the measured impact depends in part on how closely the control group matches the treatment group on key characteristics relevant to the impacts being studied.

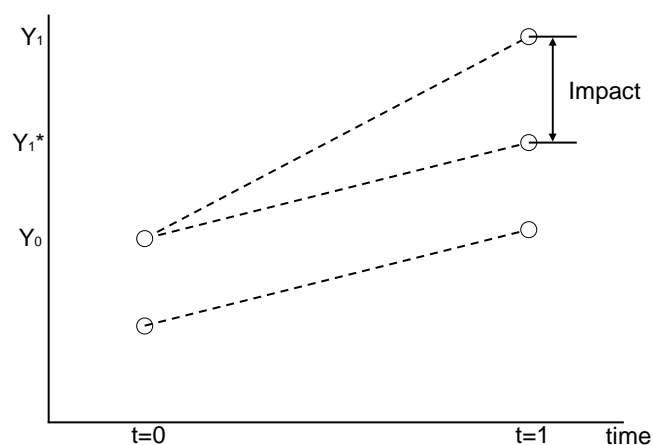
## Special challenges in constructing comparison groups

- Selection bias
  - Project communities or individuals are purposefully selected to target
    - Groups most likely to be successful
    - The poorest and most vulnerable groups (many of which may be less likely to be successful)
  - Participants are self-selected
    - Often people/communities who apply are most likely to be successful

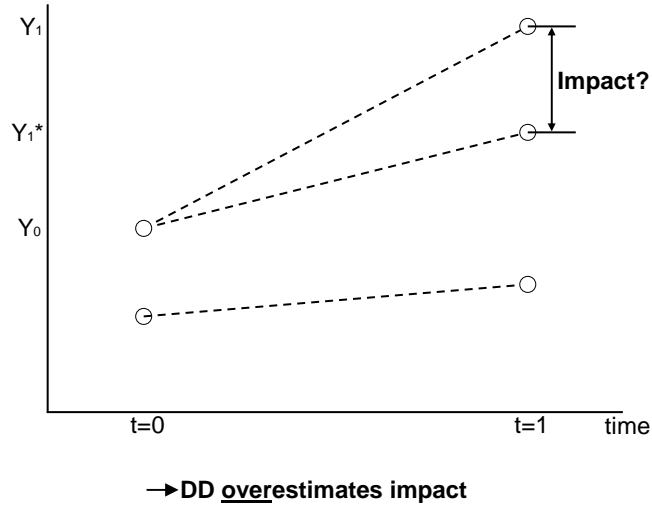
## Selection bias



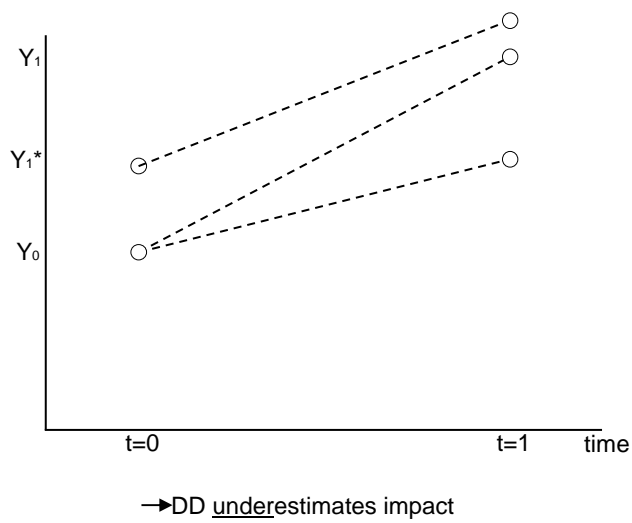
Difference-in-difference (“diff-in-diff”) requires that the bias is additive and constant over time



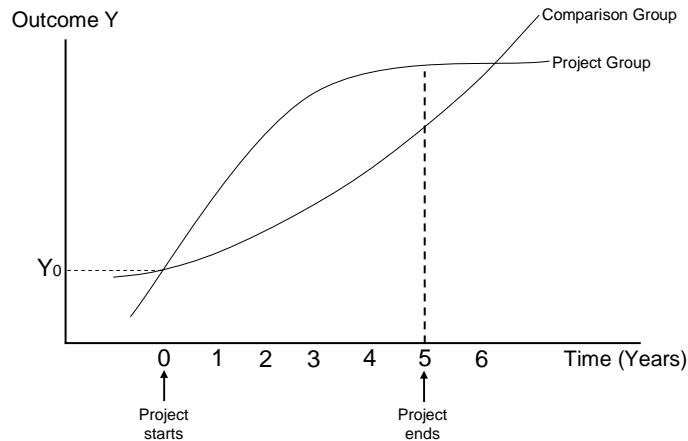
## The method fails if the comparison group is on a different trajectory



Or...



## When results are not additive and time invariant



## Evaluation purpose → design

Purpose	Primary question	Type of inference	Design implications
<b>Proof of program efficacy or effectiveness</b>	Is any measured effect on performance or health due to the implemented program?	Probability	Controlled trial usually randomizing clusters, eg health service areas, to program or not
<b>Demonstration of likely program effectiveness</b>	Is any measured effect on performance or health likely due to the program rather than other influences?	Plausibility	Concurrent, non randomized clusters with program or not; before-after or cross-sectional in program areas and non-program areas
<b>Demonstration of expected changes in performance or health</b>	Are behavioral or health indicators changing among program recipients	Adequacy	Before-after or time-series in program areas only

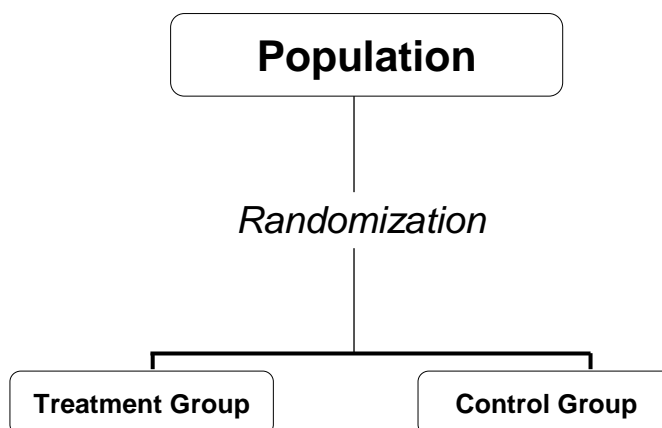
Black 2008

## EVALUATION OF EFFECTIVENESS OF HEALTH PROGRAMS

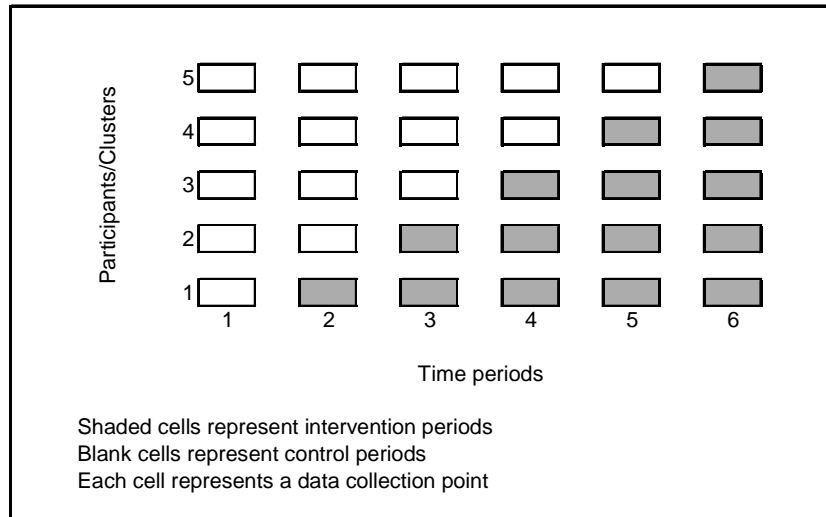
# Randomized Trials

- **Advantages**
  - Reduced selection bias
  - Controls for observable and unobservable factors
- **Threats to internal validity**
  - Lack of masking
  - Contamination
  - Residual confounding
  - Low power or imbalance if small number of clusters
- **Threats to external validity**
  - Non-representative area or program implementation (stronger) or utilization (“randomization bias”)
- **Limitations**
  - Provides average effect, not considering heterogeneity or effect modification, may not be possible for ethical or political reasons (staged or step-wedge designs may overcome some concerns)

## Setting up a randomized evaluation



## Randomized step wedge design



## Challenges

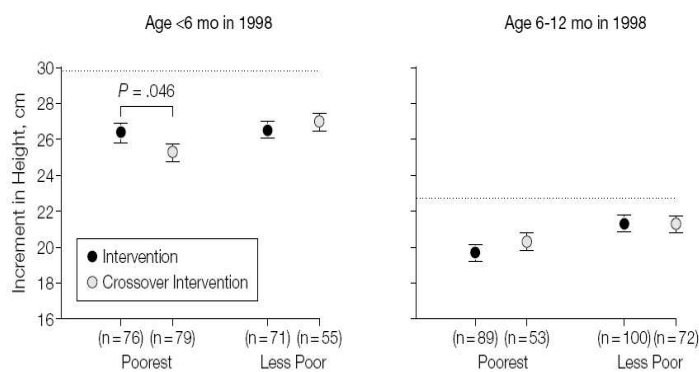
- Attrition (high dropout rates from control or treatment groups)
- Control group may be unwilling to cooperate or respond
- Hawthorne effects (threat to internal validity)
- Partial equilibrium analysis (threat to external validity). It may be difficult to extrapolate results from a pilot experiment to a national program
- Ethics and fairness of random assignment/selection
- Political feasibility of random assignment and selection for social programs

## Example: Progresa program in Mexico

- Pregnant and lactating women in participating households received fortified nutrition supplements, and the families received nutrition education, health care, and cash transfers.
- Participants were from low-income households in poor rural communities in 6 central Mexican states.
- A randomized effectiveness study of 347 communities randomly assigned to:
  - Begin in 1998 (intervention group; n=205)
  - Begin in 1999 (crossover intervention group; n=142).
- A random sample of children 12 months of age or younger in those communities was surveyed at baseline and at 1 and 2 years afterward.

## Progresa impact on child height

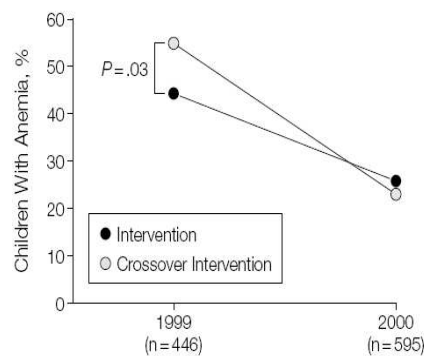
**Figure 2.** Incremental Growth in Height From Baseline in 1998 to 2000



Adjusted height increments by age and length in 1998 by using a random-intercept linear model. The expected growth from the World Health Organization reference standards is plotted for comparison (dotted lines).<sup>2</sup> Data are presented as mean (SE).

## Progresa impact on anemia

**Figure 3.** Prevalence of Anemia by Year of Survey and Intervention Group, 1999-2000



Data were adjusted by age using a generalized estimating equation model.

## Prospective, Structured Quasi-Experimental Designs

- **Advantages**
  - Allows matched or stratified comparisons
  - Before – after (difference in differences) analysis
  - Data on background and contextual factors
  - No perceived withholding of program
  - More likely to reflect routine implementation than CRT
- **Threats to internal validity**
  - As with CRT *plus*
  - More risk of selection bias
  - More influence of confounding factors
- **Threats to external validity**
  - As with CRT but without randomization bias
- **Limitations**
  - As with CRT but without ethical and political concerns, *plus*
  - Need to adjust for biases and confounding in analysis

## Other approaches

- **Retrospective “Natural experiments”**
  - Usually cross-sectional designs that compare outcomes for recipients vs. non-recipients or program and non-program areas
  - Severe problems with selection bias
- **Program adequacy evaluation (no comparison group)**
  - Program monitoring
  - Step-wise assessment of possibility of impact

## Other challenges to rigorous impact evaluation

- **Contagion** (e.g., spillover effects from “treatment” to “control” sites; introduction of another programs in either a “control” or a “treatment” site that may affect relevant outcomes there)
- **Budget, time and political constraints**, including administrative issues (e.g., timeliness of reporting results)
- There might be **different results depending on when the “before” or “after” data are collected**
  - For example, effects of interventions may not have had sufficient time to manifest themselves; or “before” data were collected too late.

## Other challenges to “rigorous” impact evaluation

- **Technical sophistication vs. relevance**
  - May be unable to address relevant issues (e.g., the change processes, sustainability of impacts, unintended consequences, and so on).
  - May be difficult to change research instruments once study commences; some rigidity of procedures
  - Little, if any, information on contextual factors (motivation, self-confidence, political/social factors, etc.) influencing outcomes
  - May not be applicable for all social interventions (e.g., no control group; difficulties with collecting data from target beneficiaries on sensitive topics (e.g., drug addicts) or from remote villagers)

## Evaluating System Strength

- **Need measures of functions of systems**
  - Stewardship
  - Mobilizing resources
  - Providing services
  - Strengthening capacity
  - Research and development
- **Implications for evaluations may include:**
  - Limited set of standardized indicators
  - Selected “fit to purpose” national indicators
  - Consider using index like “coverage gap” as aggregate score of system strength

## MEASUREMENT ISSUES AND CHALLENGES

# Evaluating Costs

- **Economic evaluations needed on costs of good practice and program implementation and scaling-up**
- **To what extent are costs, effects and cost-effectiveness assessed in one large-scale program generalizable?**
- **What accelerated strategies are financially sustainable?**
- **More focus on demand side re costs**
- **Implications for evaluations:**
  - **Common economic evaluation framework**
  - **Consideration of mix of domestic/external funding**
  - **Cost of demand creation**
  - **Role of private sector in increasing utilization**

# Evaluating Community Engagement

- **Communities not only as target of health services but as an integral part of implementation of health services**
  - **Possible roles include:**
    - **Selection, oversight and support of CHW**
    - **Oversight and accountability for health services**
    - **Analysis of health problems and related actions**
- **Implications for evaluations may include assessing roles and degree of engagement of community**

## Evaluating Coverage and Quality

- Indication that program interventions are reaching the intended recipients
- Critical for informing program implementation as well as predicting and understanding impact
- Quality of service usually assessed separately from coverage by program managers
- Can combine coverage and quality in a measure of “effective coverage” and express as the proportion of actual health gain delivered in comparison to the total potential gain

## Example: Child Mortality Changes in Shorter Periods

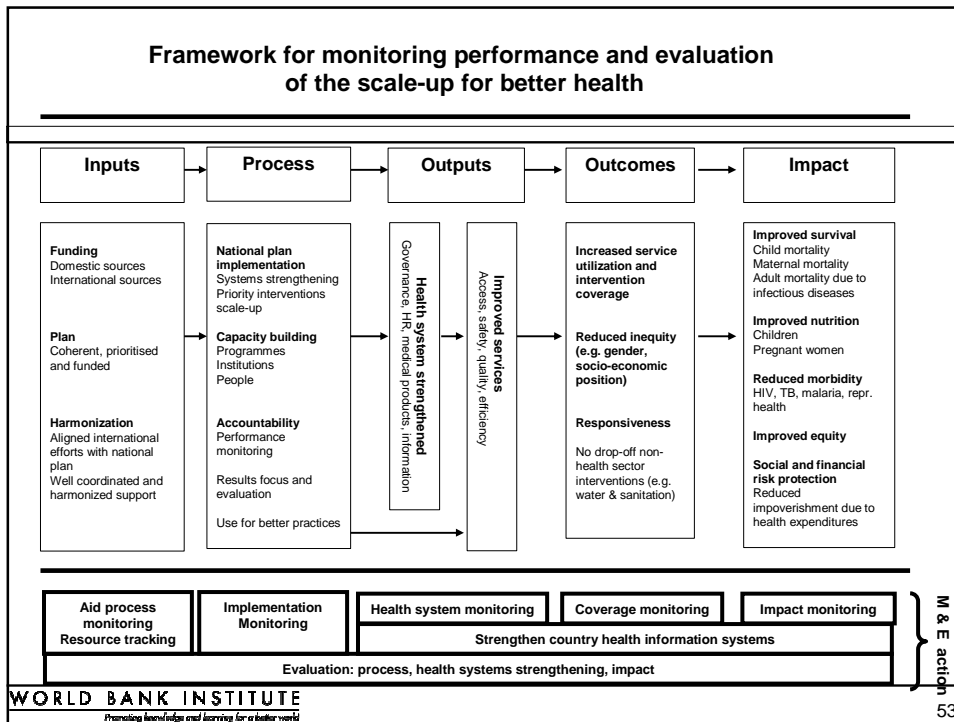
- Full or sample vital registration currently inadequate
- Survey methods provide mortality rate no more recent than several years ago
- Yet strong demand for more timely mortality data to assess accelerated programs
- Possible approaches to be tried:
  - Rolling surveys with birth history
  - Sample vital registration areas
  - Recording of deaths at community level
  - Deaths in health facilities adjusted by survey data

## Operational Considerations

- Need for team external to program with independent funding
- Leadership and involvement of national institutions throughout
- Close interaction with program implementers from the beginning eg impact framework, design, feedback, interpretation
- Coordination with national M&E plans and data collection efforts eg DHS, MICS if possible
- Dissemination in country and internationally

## Need for evaluation framework

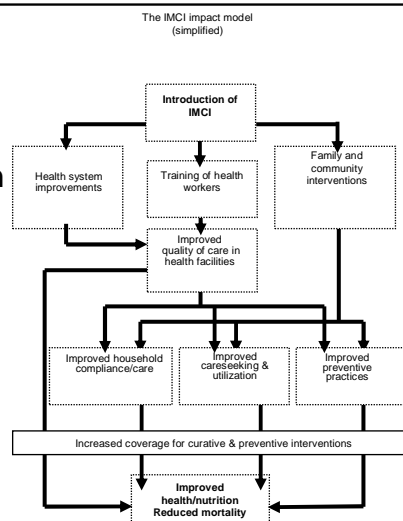
1. A strategic framework including general principles
2. A conceptual model specifying how activities will lead to outcomes and impact
3. A set of compatible designs for evaluation of country-level initiatives, to allow comparisons across places and time
4. A list of common indicators and other measures



53

## Why is a conceptual model needed?

- To clarify expectations of planners and developers
- To define the key evaluation questions
- To choose indicators
- To guide the design
- To estimate sample sizes

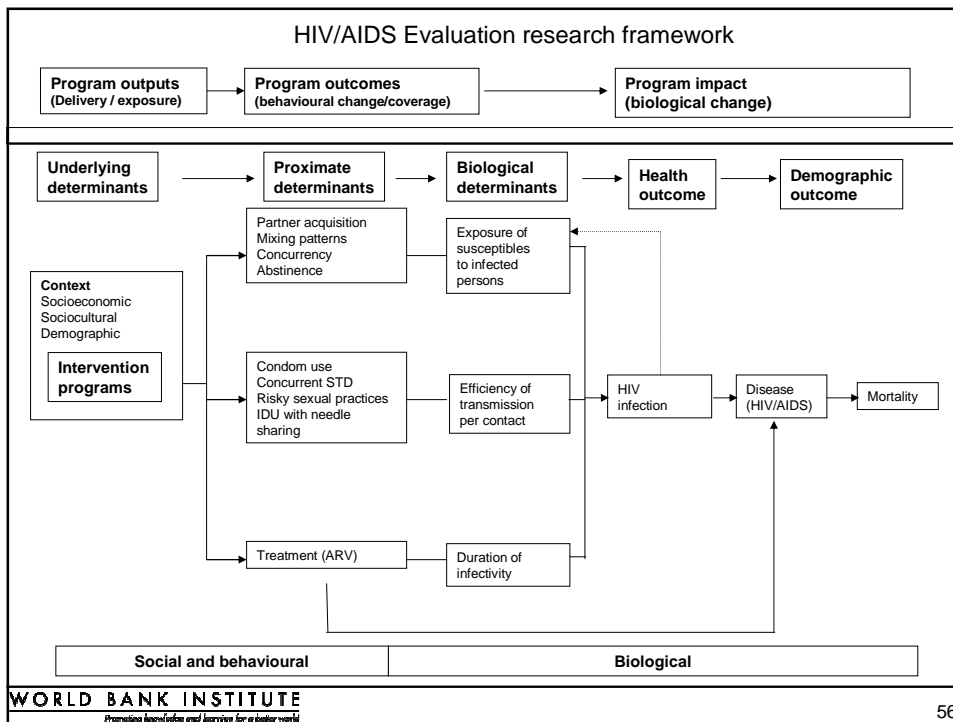


**WORLD BANK INSTITUTE**  
Promoting knowledge and learning for a better world

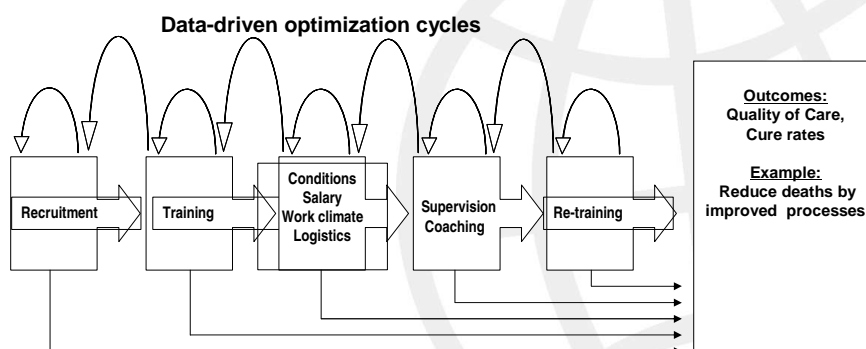
54

## Why is a conceptual model needed?

- To guide analysis and attribution of results
- To compare and interpret results across sites
- To track changes in assumptions as they evolve in response to evaluation findings
- To stay honest about what was expected



## Data driven local framework for human resources monitoring and enhancement



## Evaluation framework should address practical issues

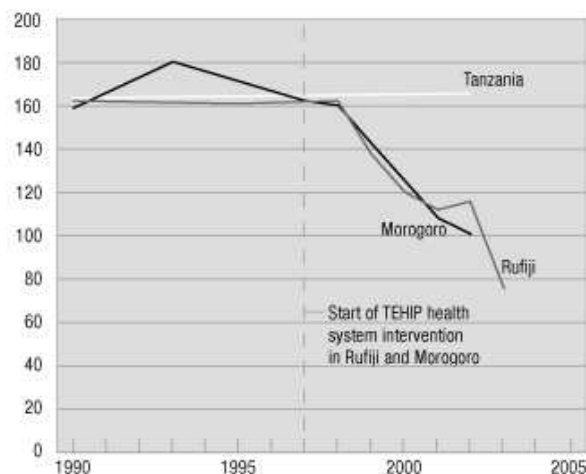
- Key points to address:
  - feasibility and acceptability of implementation
  - estimation of impact to make informed decisions about implementing at scale

## Example: Health Planning

- The Tanzania Essential Health Interventions Project (TEHIP): Goal to determine the feasibility of an "evidence-based" approach to health planning
  - Increase district level capacity to effectively monitor health interventions in terms of burden of disease and per capita cost
  - Strengthen capacity to plan and set funding priorities using locally obtained burden of disease and cost-effectiveness data

## Reversing the trend in child mortality after district-level health system interventions

Child <5 mortality  
per 1000 live births



## Example: Community intervention

### Effect of maternal multiple micronutrient supplementation on fetal loss and infant death in Indonesia: a double-blind cluster-randomised trial

*The Supplementation with Multiple Micronutrients Intervention Trial (SUMMIT) Study Group\**  
*Lancet 2008; 371: 215-27*

- Large scale (~40,000 women) RCT integrated into government prenatal care services indicates 18% reduction in early infant death due to maternal multiple micronutrient supplementation (*Lancet 2008; 371:215-27*)
- Data driven implementation framework used to promote effective processes

## Community Health Facilitators function: promotion, information, participation



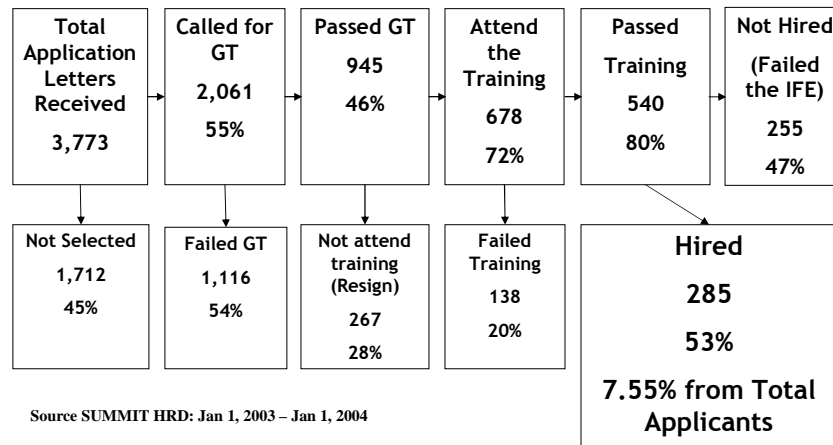
## Staff selection process

- Began by using traditional review of CVs and recommendations, and interview. Selected trainees based on:
    - Health backgrounds
    - Recommendations
    - Good impression in interview
- **BUT: trainees selected experienced high failure rate in training (i.e. >50%)**

## Staff selection process (cont'd)

- Based on review of recruitment data, traditional screening process was found to lack prediction of success for trainees. New more open process was developed.
- Criteria for selection
  - Motivation
  - Honesty
  - Intelligence
  - Compassion

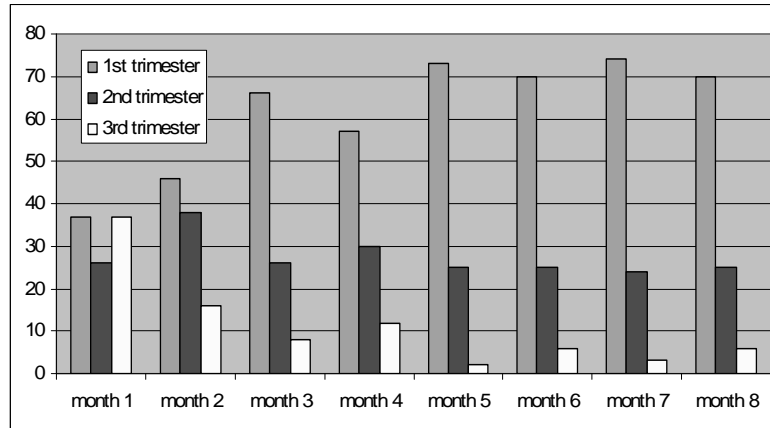
## Data driven process selects successful trainees



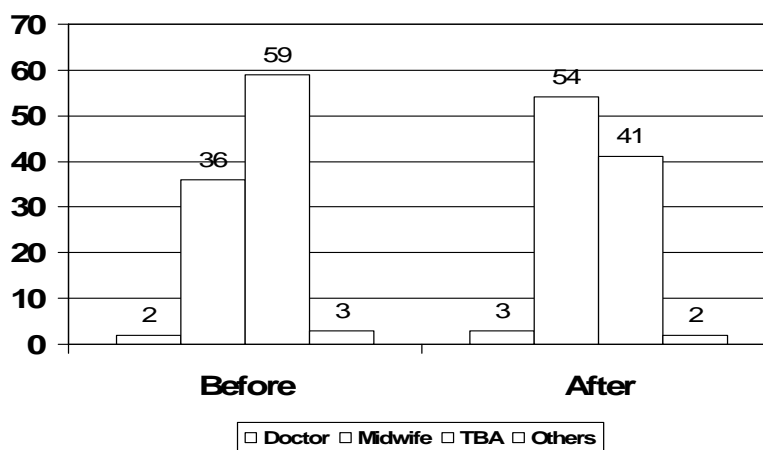
## Performance of staff: Head, Hand, Heart

- Monthly staff performance scores:
  - Head: Knowledge of job task
  - Hand: Task errors and on time & error in reporting
  - Heart: Feedback from clients and supervisor
- Performance score given to staff every month. Supervisors work as coaches in non-punitive system to help staff improve their performance

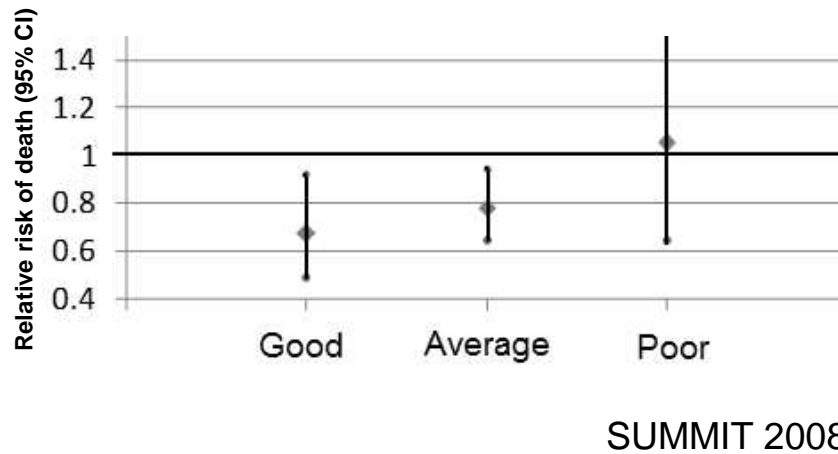
## Change in gestational age at ANC



## Change in delivery by skilled attendant



## Performance of community facilitator and impact of intervention on infant death



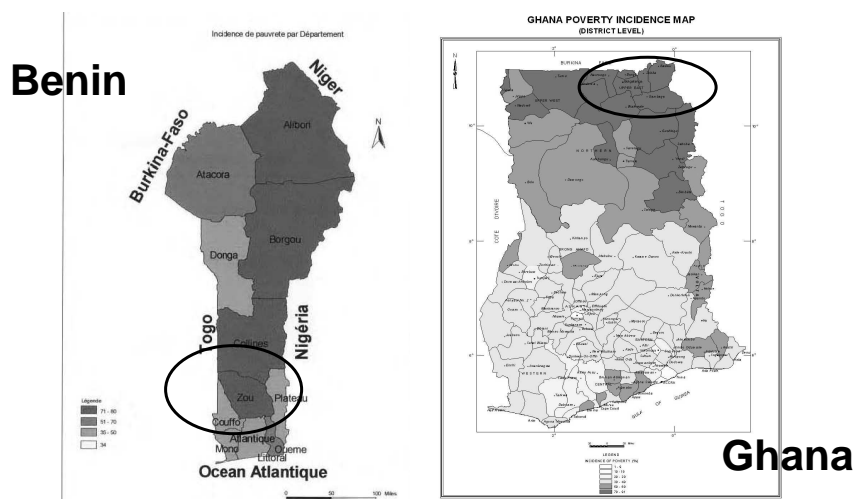
## Summary

- Data driven improvements in recruitment processes may enhance program impact
- Data driven adjustment to field procedure and supervision lead to improved staff performance
- Routine data-based Supervision to staff results in improved performance and better health for clients

## Does the evaluation framework incorporate equity considerations?

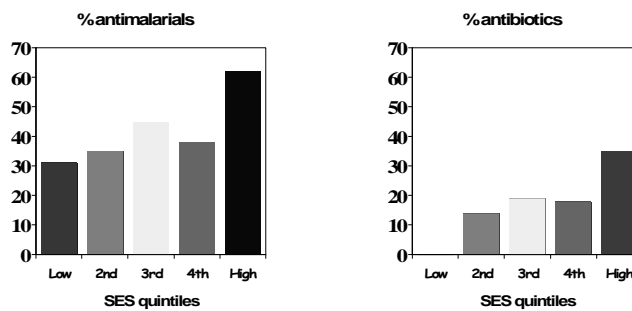
- Key issues:
  - Assess implementation by geographical area
  - Utilization, coverage, impact data broken down by SES, etc
  - Assessment of program effects on inequities over time
  - Implications for survey design and sample sizes

## Equity: poverty mapping (where is the program going?)



## Equity: Coverage by wealth (who is receiving the interventions?)

In rural Tanzania, poorer children are less likely to receive adequate care



## What is a healthy system?

**"To be in a good health means being able to fall sick and recover"**

**—Georges Canguilhem**

## Conclusion

1. Evaluation enables programs to be effective
2. A good evaluation is:
  - a. Designed at the outset of the program and is integrated with the conceptual framework
  - b. Aligned with national systems
  - c. Enhances routine monitoring and use of data
  - d. Provides reliable data that can be interpreted with actionable recommendations
3. Enhance data driven enhancement of local interventions for empowerment and results