Oregon Program Evaluators Network [OPEN]
Brown Bag Series
September 27, 2007

RealWorld Evaluation: Working under budget, time, data and political constraints.

Michael Bamberger

Freely adapted by the presenter from:

Michael Bamberger, Jim Rugh and Linda Mabry. 2006. *RealWorld Evaluation: Working under budget, time, data and political constraints*. Sage Publications

THE NEW CONTRACT

A Day in the Life of a Program Evaluator

Hi!

We would like to contract you to conduct an evaluation of the impacts of our Creating New Opportunities program. We are seeking new funding from the State and private donors so we want a really rigorous assessment of the program impacts.

Great!

We would be delighted to help you and we are familiar with all of the best evaluation designs for impact assessment.

Fantastic,

I should explain that the evaluation is starting a little late - the three year program ends in six months – and we are a little tight for money as no budget was approved for the evaluation when the program began. But we do need a really rigorous evaluation

I understand and we will do our best. Could you explain what type of baseline data was collected when the program began

Umm…….
Well we didn't actually collect any baseline data … but I am sure you can fix that.
So can I tell our funders that you will be able to deliver the final report in two months from today …… and don't forget we really need a rigorous assessment of how successful the program has been

Did you say we only have two months!!!……
….. and not much money !!…
….. and no baseline data!!!

Client Evaluator

RealWorld Evaluation scenarios

Scenario 1

The evaluation is commissioned at the start of the project BUT ….

The evaluation is commissioned at the start of the project or program but for budget, time, technical or political reasons:

- It is difficult to spend the required time on consultations and evaluation planning
- It is difficult to collect data on a comparison group
- The client is unwilling to authorize collection of the necessary baseline data on project participants
- There are political influences on who can be interviewed and what data can be collected

Scenario 2

The evaluation is not commissioned until late in the project ….

The evaluation is not commissioned until the project or program has been operating for some time (or is nearing completion or completed):

- No baseline data has been collected on a comparison group and often not on the project group either
- Secondary data is lacking or of poor quality
- There are time pressures to complete the report
- There are budget constraints
- The are political constraints on the evaluation methodology and pressures to ensure "objective but positive" findings"

Evaluators usually face one or more of the following constraints

Limited budget

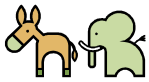
- Budget resources were not committed or released
- Budgets unpredictable and can be cut expectedly
- Pressures to reduce costs of data collection ["Do you really need all of that information?"]

Time pressures

- Heavy workload and other commitments
- Time pressures on evaluation clients/stakeholders
- Tight deadlines
- Pressures to reduce time spent on data collection and work in the field
- Pressures to start data collection without adequate preparation
- Pressures to evaluate impacts before the project has been underway for sufficient time to produce measurable results

Lack of data

- No baseline surveys
- Administrative data on the project is incomplete or of poor quality
- Secondary data is not adequate for the purposes of this evaluation
  - Wrong population
  - Wrong time period
  - Does not ask the right questions
  - Does not interview the right people
  - Unreliable
- Difficult to find data on qualitative outcomes and impacts



Political and institutional constraints

- Pressures from funders, implementing agencies and stakeholders to:
  - Only use certain methodologies
  - Only ask certain questions
  - Only interview certain groups
  - Only present positive findings and not "rock the boat"
- Multiple clients with different agendas
- Scarce local evaluation expertise
- Lack of an evaluation culture in many agencies
  - Fear of evaluation
- Lack of ownership of the evaluation by clients and stakeholders

The "Big Six" Challenges ……
Why evaluators can't sleep at night

1. What do the clients really want to know?
- What information is "essential" and what is only "interesting"
- What are they going to do with the information?
- Is there a hidden agenda?

→ **Stakeholder analysis** [Free resource page A]

2. What is the logic model underlying the program?
- What is the program trying to achieve?
- How will the intended outcomes and impacts be achieved?
- What are the critical assumptions on which success is based?
- How will outcomes be affected by the local contexts?

→ **Program theory evaluation** [page 6]

3. The dreaded counterfactual
- How do we know that the observed changes are due to the project intervention and not to: local contextual factors [economic, political, organizational, socio-cultural etc]; other programs or participant selection bias?

→ **RCTs, Quasi-experimental designs and qualitative approaches to causality** [pp. 6-7 and Table 2]

4. Combining depth with breadth
- The challenges of combining qualitative and quantitative methods
- How to generalize from in-depth qualitative data
- Putting flesh on the numbers

→ **Mixed method designs** [pp. 8-9]

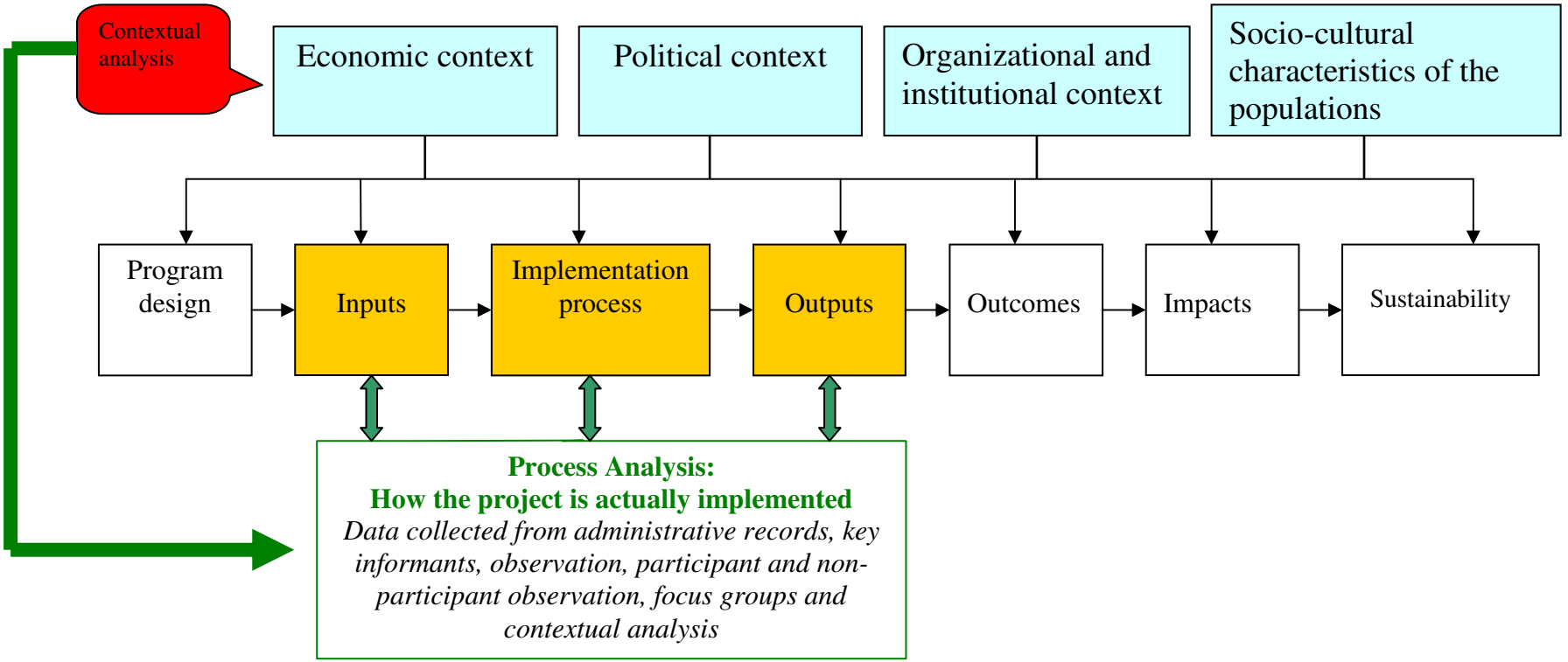
5. Why do the evaluation findings and recommendations not get used?
- Who "owns" the evaluation?
- Timing? Communication style?

→ **Utilization focused evaluation** [Free resource page C & D]

6. Will our methodology hold up under critical scrutiny?
- Have we assessed and addressed the main threats to validity?

→ **Threats to the validity of conclusions** [page 11 and Table 5]

A Program Theory Model Incorporating Contextual Analysis and Process Analysis



The Importance of the Counterfactual

Alternative explanations of the observed changes in the project population that the evaluation design must eliminate (control for)

1. Project selection bias
    - Self-selection
    - Project administrators select subjects most likely to succeed
2. Different experiences of participants and control groups during project implementation
    - Differential attrition
    - Demoralization of the control group
    - Response of other agencies to the project
3. The evaluation design
    - Sample selection bias
    - Data collection methods
        - Do not adequately capture impacts
        - Right people not interviewed
        - Cannot capture unexpected outcomes
4. External events during project implementation
    - Economic, political, organizational/institutional, socio-cultural characteristics of the target populations.
5. The influence of other programs
    - Providing similar services to sectors of the study populations

The dangers of not having a strong counterfactual

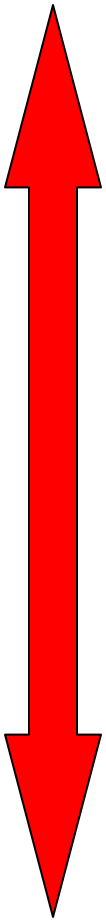
- Programs may be continued that are not producing any benefits
- Potentially good programs may be terminated
- Certain sectors of the target population may be excluded from the program or from receiving certain benefits

Stronger and weaker ways to define the counterfactual

Experimental designs

Strongest

- True experimental designs
- Randomized control trials (RCTs) in field settings



Quasi-experimental designs

Strong designs: [pre-test/post-test comparison of project and comparison groups

- Statistical selection of comparison group [e.g. propensity score matching]
- Judgmental selection of comparison group

Weaker quasi-experimental designs:

- No baseline data for comparison and/or project groups.

Non-experimental designs

Causality assessed through program theory models, case studies, focus groups or PRA techniques

- No comparison group
- No comparison group and no baseline data for project group

Weakest

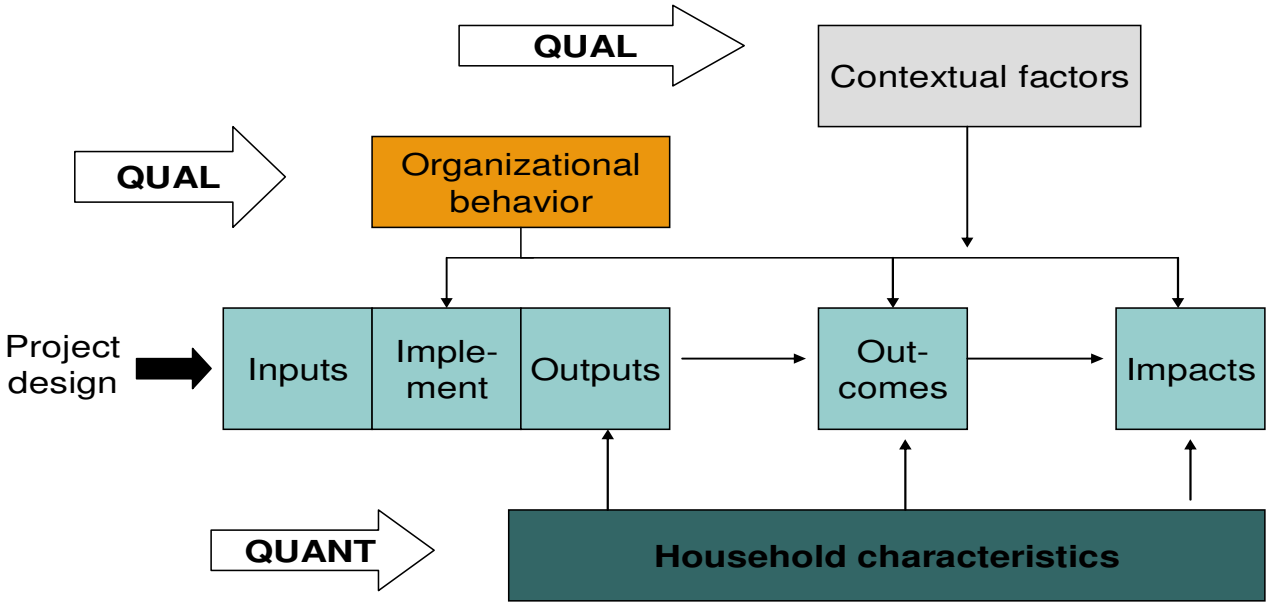
Mixed-Method approaches

Mixed methods "*research in which the investigator collects and analyzes data, integrates the findings, and draws inferences using both quanitative and qualitative approaches or methods in a single study or program of enquiry."*
[Tashakkori and Cresswell 2007]

Benefits

- Broadening the conceptual framework
- Combining generalizability with depth and context
- Facilitating access to difficult-to-reach groups
- Strengthens understanding of the project implementation process
    - What really happens in the project and the affected communities?
- Control for underlying structural factors
- Permits multi-level analysis
- Enhances data reliability and validity through triangulation
- Strengthens the interpretation of findings
- Permits feedback to check on inconsistent findings.

Consecutive Mixed-Method Design
Integrating Multiple Levels of Analysis



Assessing and addressing threats to the validity of evaluation conclusions

Step 1: Applying the Standard Checklist for assessing the validity of QUANT, QUAL and Mixed-Method Evaluations

A. Confirmability: *Are the conclusions drawn from the available evidence and is the evaluation relatively free of researcher bias?*

B. **Reliability**: *Is the process of the evaluation consistent, reasonably stable over time and across researchers and methods?*

C. **Credibility**: Are the findings credible to the people studied and to clients and readers?

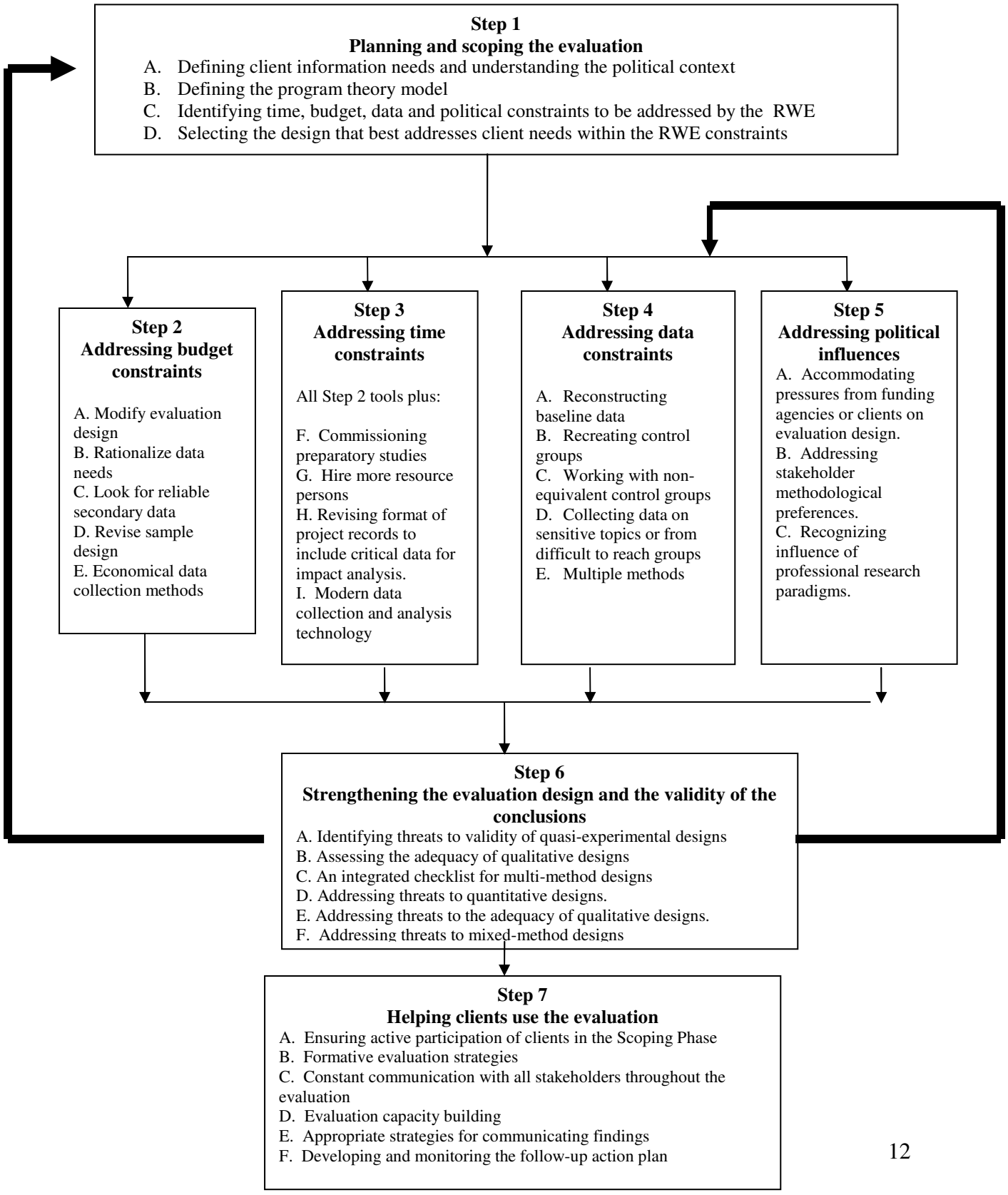
D. Transferability: *Do the conclusions fit other contexts and how widely can they be generalized?*

E. Utilization: *Were the finding useful to clients, researchers and the communities studied?*

Step 2: Follow-up to address problems once they have been identified.

- Budgeting time and resources to return to the field to check up on inconsistent findings or to elaborate on interesting analysis.
- Applying the checklist at different points during the evaluation to allow time to identify and correct weaknesses in the evaluation design and analysis
- Rapid measures to address and correct threats to validity
- Identifying in the report the limitations of the analysis and recommendations.

The RealWorld Evaluation [RWE] Approach



Reference Table 1: Strategies for Reducing Costs of Data Collection and Analysis	
Quantitative Evaluations	**Qualitative Evaluations**
A. Simplifying the evaluation design	
• Simplify the evaluation design by eliminating one or more of the 4 observation points (pre-test/post-test, project and comparison groups). ***See Table 2 for the most common impact evaluation design options.***	• Prioritize and focus on critical issues. • Reduce the number of site visits or the time period over which observations are made. • Reduce the amount and cost of data collection. Reduce the number of persons or groups studied.
B. Clarifying client information needs	
Prioritize questions and data needs with the client to try to eliminate the collection of data not actually required for the evaluation objectives.	
C. Use existing data	
• Census or surveys covering project and comparison areas • Data from project records • Records from schools, health centers and other public service agencies	• Newspapers and other mass media • Records from community organizations • Dissertations and other university studies [for both QUAL and QUANT]
D. Reducing sample size	
• Using Power Analysis and Effect Size to determine the required sample size. • Lower the level of required precision • Reduce types of disaggregation required • Stratified sample designs (less interviews) • Use cluster sampling (lower travel costs)	• Consider critical or quota sampling rather than comprehensive or representative sampling • Reduce the number of persons or groups studied.
E. Reducing costs of data collection, input and analysis	
• Self-administered questionnaires (with literate populations) • Direct observation (instead of surveys) • Automatic counters and other non-obtrusive methods • Direct inputting of survey data through hand-held devices. • Optical scanning of survey forms	• Decrease the number or period of observations • Prioritize informants • Employ and train university students, student nurses, and community residents to collect data (for both QUAL and QUANT) • Data Input through hand-held devices.

Reference Table 2. The nine most widely used quantitative impact evaluation designs					
Key T = Time P = Project participants; C = Control group P_1, P_2, C_1, C_2 First and second observations X = Project intervention (a process rather than a discrete event)	Start of project [pre-test]	Project intervention [Process not discrete event]	Mid-term evaluation	End of project [Post-test]	The stage of the project cycle at which each evaluation design can to be used.
Quantitative Impact Evaluation Design	T_1		T_2	T_3	
RELATIVELY ROBUST DESIGNS					
1. ***Randomized control trials [RCT]***. Subjects are randomly assigned to the project (treatment) and control groups.	P_1 C_1	X		P_2 C_2	Start
2. ***Pre-test post-test non-equivalent control group design with statistical matching of the two groups.*** Participants are either self-selected or are selected by the project implementing agency. Statistical techniques (such as propensity score matching), drawing on high quality secondary data used to match the two groups on a number of relevant variables.	P_1 C_1	X		P_2 C_2	Start
3. ***Pre-test post-test non-equivalent control group design with judgmental matching of the two groups.*** Participants are either self-selected or are selected by the project implementing agency Control areas usually selected judgmentally and subjects are randomly selected from within these areas.	P_1 C_1	X		P_2 C_2	Start
LESS ROBUST QUASI-EXPERIMENTAL DESIGNS					
4. ***Pre-test/post-test comparison where the baseline study is not conducted until the project has been underway for some time*** (most commonly this is around the mid-term review).		X	P_1 C_1	P_2 C_2	During project implementation (often at mid-term)
5. ***Pipeline control group design***. When a project is implemented in phases, subjects in Phase 2 (i.e. who will not receive benefits until some later point in time) can be used as the control group for Phase 1 subjects.	P_1 $Ph[2]_1$	X		P_2 $Ph[2]_2$	Start
6. ***Pre-test post-test comparison of project group combined with post-test comparison of project and control group.***	P_1	X		P_2 C_1	Start
7. ***Post-test comparison of project and control groups***		X		P_1 C_1	End
NON-EXPERIMENTAL DESIGNS (THE LEAST ROBUST)					
8. ***Pre-test post-test comparison of project group***	P_1	X		P_2	Start
9. ***Post-test analysis of project group.***		X		P_1	End

Reference Table 3. Rapid Data Collection Methods	
	Ways to reduce time requirements
A. Mainly qualitative methods	
Key informant interviews	Key informants can save time either by providing data (agricultural prices, people leaving and joining the community, school attendance and absenteeism) or by helping researchers focus on key issues or pointing out faster ways to obtain information.
Focus groups and community interviews	Focus groups can save time by collecting information from meetings rather than surveys. Information on topics such as access to and use of water and sanitation; agricultural practices and gender division of labor in farming can be obtained in group interviews possibly combined with the distribution of self-administered surveys.
Structured observation	Observation can sometimes be faster than surveys. For example: observation of the gender division of labor in different kinds of agricultural production, who attends meetings and participates in discussions, types of conflict observed in public places in the community.
Use of preexisting documents and artifacts	Many kinds of pre-existing data can be collected and reviewed more rapidly than new data can be collected. For example, school attendance records, newspapers and other mass media, minutes of community meetings, health center records, surveys in target communities conducted by research institutions.
Using community groups to collect information	Organization of rapid community studies (QUAL and QUANT) using community interviewers (local school teachers often cooperate with this)
Photos and videos	Giving disposable cameras or camcorders to community informants to take photos (or make videos) illustrating, for example, community problems.
B. Mainly quantitative methods	
Rapid surveys with short questionnaires and small samples	Reducing the number of questions and the size of the sample can significantly reduce the time required to conduct a survey.
Reduce sample sizes	Sample sizes reduce costs but also reduce statistical power of the test
Triangulation	Obtaining independent estimates from different sources (e.g., survey and observation) sometimes makes it possible to obtain estimates from smaller samples hence saving both elapsed time and effort.
Rapid exit surveys	People leaving a meeting or exiting a service facility can be asked to write their views on the meeting or service on an index card which is put on the wall. Often only one key question will be asked. For example: "Would you recommend a neighbor to come to the next meeting or use this center)?"
Use of preexisting data	• Previous surveys or other data sources may eliminate the need to collect certain data • Previous survey findings can reduce the time required for sample design or by providing information on the *standard deviation* of key variables may make it possible to reduce sample size or to save time through more efficient *stratification* or *cluster sampling*.
Observation checklists	Observation checklists can often eliminate the need for certain surveys (for example pedestrian and vehicular traffic flows, use of community facilities, time required to collect water and fuel).
Automatic counters	Recording people entering buildings or using services such as water.

Reference Table 4. Strategies for Reconstructing Baseline Data	
Reconstructing baseline data	
Approaches	**Sources/ Methods**
Identification and effective use of secondary data [All data sources must be assessed to determine reliability and validity]	• Surveys and census data • Project records • School enrollment and attendance records • Patient records in local health centers • Savings and loans cooperatives records of loans and repayment • Vehicle registrations (to estimate changes in the volume of traffic) • Records of local farmers markets (prices and volume of sales)
Using recall to obtain ***numerical*** (income, crop production, travel time, school fees) ***or qualitative*** (community violence, the level of consultation of government officials with the community) estimates of the situation at the time the project began	• Recall questions in surveys • Interviews with key informants • PRA and other participatory methods • Focus groups
Improving the reliability/validity of recall	• Conduct small pre-test post-test studies to compare recall with original information • Identify and try to control for potential bias (under-estimation of small expenditures, truncating large expenditures, distortion to conform to accepted behavior, intention to mislead) • Clarifying the context of required recall (time period, specific types of behavior, reasons for collecting the information) • Link recall to important reference points in community or personal history • Triangulation (key informants, secondary sources, PRA)
Key informants	• Community leaders • Religious leaders • Teachers • Doctors and nurses • Store owners • Police • Journalists

Reference Table 5

STANDARD CHECKLIST FOR ASSESSING THE ADEQUACY AND VALIDITY OF ALL EVALUATION DESIGNS[1]	
	√ [2]
A. Confirmability (and Objectivity)	
Are the conclusions drawn from the available evidence, and is the research relatively free of researcher bias?	
1. Are the study's methods and procedures adequately described? Are study data retained and available for re-analysis?	
2. Is data presented to support the conclusions?	
3. Has the researcher been as explicit and self-aware as possible about personal assumptions, values and biases?	
4. Were the methods used to control for bias adequate?	
5. Were competing hypotheses or rival conclusions considered?	
B. Reliability (and Dependability)	
Is the process of the study consistent, coherent and reasonably stable over time and across researchers and methods? If emergent designs are used are the processes through which the design evolves clearly documented?	
1. Are findings trustworthy, consistent and replicable across data sources and over time?	
2. Were data collected across the full range of appropriate settings, times, respondents, etc.?	
3. Did all fieldworkers have comparable data collection protocols?	
4. Were coding and quality checks made, and did they show adequate agreement?	
5. Do the accounts of different observers converge? If they do not (which is often the case in QL studies) is this recognized and addressed?	
6. Were peer or colleague reviews used?	
9. Are the conclusions subject to "Threats to Construct Validity" [See Section H]? If so were these adequately addressed?**	
10. Were the rules used for confirmation of propositions, hypotheses, etc. made explicit?	
C. Credibility (and Internal Validity or Authenticity)	
Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying?	
1. How context-rich and meaningful ("thick") are the descriptions? Is there sufficient information to provide a credible/valid description of the subjects or the situation being studied?**	
2. Does the account ring true, make sense, seem convincing? Does it reflect the local context?	
3. Did triangulation among complementary methods and data sources produce generally converging conclusions? If expansionist QL methods are used where interpretations do not necessarily converge, are the differences in interpretations and conclusions noted and discussed?**	
4. Are the presented data well linked to the categories of prior or emerging theory? Are the findings internally coherent, and are the concepts systematically related?	

[1] Source: Bamberger, Rugh and Mabry (2006) RealWorld Evaluation Appendix 1. Adapted by the authors from Miles and Huberman (1994) Chapter 10 Section C. See also: Guba and Lincoln (1989). Categories added, or modified by the current authors are indicated by **

[2] The check mark can be used to indicate if this question has been addressed. It is possible to develop this into two rating scales [Scale 1: The importance of this item to the validity of the evaluation; Scale 2: How well the issue is addressed in the evaluation]. However, the authors caution that rating systems (e.g. good, adequate, poor) are difficult to apply in an objective and consistent manner.

5. Are areas of uncertainty identified? Was negative evidence sought, found? How was it used? Have rival explanations been actively considered?	
6. Were conclusions considered accurate by the researchers responsible for data collection?	
7. Are the findings subject to "Threats to Internal Validity" [see Section G]? If so were these addressed?**	
8. Are the findings subject to "Threats to Statistical Validity"[see Section F]? If so were these adequately addressed?**	
D. Transferability (and External Validity or Fittingness)	
Do the conclusions fit other contexts and how widely can they be generalized?	
1. Are the characteristics of the sample of persons, settings, processes, etc. described in enough detail to permit comparisons with other samples?	
2. Does the sample design theoretically permit generalization to other populations?	
3. Does the researcher define the scope and boundaries of reasonable generalization from the study?	
4. Do the findings include enough "thick description" for readers to assess the potential transferability?	
5. Does a range of readers report the findings to be consistent with their own experience?	
6. Do the findings confirm or are they congruent with existing theory? Is the transferable theory made explicit?	
7. Are the processes and findings generic enough to be applicable in other settings?	
8. Have narrative sequences been preserved? Has a general cross-case theory using the sequences been developed?	
9. Does the report suggest settings where the findings could fruitfully be tested further?	
10. Have the findings been replicated in other studies to assess their robustness. If not, could replication efforts be mounted easily?	
11. Are the findings subject to Threats to External Validity [see Section I]? If so were these addressed?	
E. Utilization (and Application or Action Orientation)	
How useful were the findings to clients, researchers and the communities studied?	
• Are the findings intellectually and physically accessible to potential users?	
• Were any predictions made in the study and, if so, how accurate were they?	
• Do the findings provide guidance for future action?	
• Do the findings have a catalyzing effect leading to specific actions?	
• Do the actions taken actually help solve local problems?	
• Have users of the findings experienced any sense of empowerment or increased control over their lives? Have they developed new capacities?	
• Are value-based or ethical concerns raised explicitly in the report? If not do some exist that the researcher is not attending to?	

Free resources on RealWorld Evaluations

- RealWorld Evaluation: overview chapter including main tables from the book (currently available in English and French and Spanish version coming soon)
- Powerpoint presentations and reviews of the book also available

www.realworldevaluation.org

- Michael Bamberger (2006) Conducting quality impact evaluations under budget, time and data constraints. Independent Evaluation Group. The World Bank (available in English and Spanish).

www.worldbank.org/ieg/ecd

- Michael Bamberger 2004 and 2005 (editor) Influential Evaluations: Evaluations that improved performance and impacts of development programs. Summary volume and more detailed presentation of case studies. Summary available in six languages. Independent Evaluation Group. The World Bank.

www.worldbank.org/ieg/ecd

- Video of workshop presentation on Increasing the utilization of your evaluations.

http://www.worldbank.org/ieg/ecd/evaluation_utilization.html

Super-Evaluator Saves the Day

Discussion topic:

A pilot job training and placement program was started about 18 months ago in several low-income communities in the Portland area. The communities have a high unemployment rate and also high indices of crime and gang-related violence. The non-profit agency organizing the program believes that if more residents, particularly young people, can find jobs this will contribute to reducing crime and violence. The agency believes the initial results have been positive as quite a few people who have gone through the program have been able to find jobs and there are some reports from community residents that crime has decreased. "You can feel the change in the community every time you visit" says the Executive Director.

The agency is planning to meet with potential funders in December to seek support to expand the program, and they decide to commission an evaluation to assess project effectiveness and impacts – "and to show funders how successful the program has been." The evaluation has to be conducted on a relatively small budget and must be completed in 3 months. Even though the budget is limited the agency is anxious to present a "rigorous" assessment of project effectiveness.

No systematic baseline data was collected at the start of the project and although they know how many people were placed in jobs they do not have good records on how many people went through the training or were sent for job interviews.

Enter the Super-Evaluator

What kind of evaluation design would you propose?

How would you define the counterfactual to try to control for (eliminate) alternative explanations of the observed changes?

How could you reconstruct baseline data on the conditions of the communities and the project participants at the start of the project?

What strategies could you use to reduce the costs of the evaluation while ensuring maximum possible rigor?

Do you think the agency might try to influence the evaluation to produce a more favorable report? How might they do this and how should the evaluator respond?