

1. Designs and methods for impact evaluation.
2. The quality of Sida evaluations.



EBA seminar 16/5 2014

Kim Forss, Andante – tools for thinking AB  
and EBA

# 1. Broadening the range of designs and methods for impact evaluations



- ❧ DFID Working Paper 38, 2012
- ❧ Elliot Stern (team leader), Nicoletta Stame, John Mayne, Kim Forss, Rick Davies, Barbara Befani.
- ❧ [www.gov.uk](http://www.gov.uk)
- ❧ Background - policy on evidence-based decision-making
- ❧ Does scientific rigour and evidence equal experimental methods?
- ❧ Challenges in evaluation practice

# Examples of evaluation tasks



- ❧ Strategic Climate Institutions Programme in Ethiopia
- ❧ District Development Programme in Afghanistan
- ❧ Violence against Women, Bihar, India
- ❧ Civil Society Governance Fund, Malawi
- ❧ Response to Haiti Earthquake
- ❧ Results Based Aid in the Education Sector in Ethiopia

# Nailing down the impact evaluation questions



- ❧ To what extent can a specific impact be attributed to the intervention?
- ❧ Did the intervention make a difference?
- ❧ How has the intervention made a difference?
- ❧ Will the intervention work elsewhere?



# Understanding causality – when can an event be said to cause an effect



- Both necessary and sufficient
  - Necessary but not sufficient
  - Sufficient but not necessary
  - Neither necessary nor sufficient but a contributory cause
- 
- The notion of causal packages

# Understanding how an intervention makes a difference



- How and why have the observed impacts come about?
- What causal factors or mechanisms in what combination have resulted in the observed impacts?
- Has the intervention resulted in any unintended impacts, and if so, how?
- For whom has the intervention made a difference?

# Will the intervention work elsewhere?



- ❧ Can this 'pilot' be transferred elsewhere and scaled up?
- ❧ Is the intervention sustainable?
- ❧ What generalisable lessons have we learned about impact?

Or, if the hoped for impacts are not being realized we want to know,

- ❧ Why have the impacts not been realized?
- ❧ What contribution has the intervention made?
- ❧ Were the impacts not realized because of programme failure or implementation failure? :

# Distinctions



- ❧ Distinguishing designs and methods
- ❧ 'Design' refers to the overarching logic of how research is conducted and consists of four elements: research questions, theory, data and the use of data.
- ❧ Different designs may share similar methods and techniques
- ❧ Distinguishing evaluation and research
- ❧ Impact evaluation is inevitably pushed towards the norms and principles of scientific research



# Design approaches (1)

Approaches	Specific variants	Basis for causal inference
Experimental	RCTs and other 'true' experiments Quasi experiments Natural experiments	Counterfactual logic
Statistical	Statistical modelling Longitudinal studies Econometrics	Correlation between cause and effect or between variables, influence of (usually) isolatable multiple causes on a single effect
Theory-based	<u>Causal process designs:</u> Theory of Change, Process tracing, Contribution Analysis, impact mapping <u>Causal mechanism designs:</u> Realist evaluation, Congruence analysis	Identification/confirmation of causal processes or 'chains'  Supporting factors and mechanisms at work in context

# Design approaches (2)

Design approaches	Specific variants	Basis for causal inference
Case based approaches	Interpretative: Naturalistic, Grounded theory, Ethnography Structured: Configurations, QCA, Within-Case- Analysis, Simulations and network analysis	Comparison across and within cases of combinations of causal factors. Analytic generalisation based on theory.
Participatory approaches	Participatory or democratic evaluation, Empowerment evaluation	Validation by participants that their actions and experienced effects are 'caused' by programme
Synthesis studies	Meta analysis, Narrative synthesis, Realist based synthesis	Accumulation and aggregation within a number of perspectives (statistical, theory based, ethnographic etc.)

# Which approach to choose?

	Requirements	Potential strengths	Potential weaknesses
Experimental	Two identical cases for comparison, ability to control the intervention	Avoiding bias	Generalisation Role of context
Statistical	Many/diverse cases, independent causes	Uncovering 'laws'	Difficult explaining 'how' and 'why'. Construct validity
Theory and/or case based/multiple causation	Sufficient number of cases. Availability of cases with comparable characteristics.	Discovery of typologies. Dealing with limited complexity.	Difficulties interpreting highly complex combinations.
Theory and/or case-based/generative mechanisms	One case with good access to multiple data sources. Theory to identify 'supporting factors'.	In-depth understanding. Focus on the role of contexts. 'Fine-grained' explanation.	Estimating extent. Risks of bias/ loss of evaluators' independence.

# Concluding remarks



- ☞ IE must fit with contemporary development architecture that is decentralised, works through partnership and where developing countries are expected to be in the lead - practical implications for example, working through partners leads to multi-stage, indirect causal chains that IE has to analyse.
- ☞ Most interventions are 'contributory causes'. They 'work' as part of a causal package in combination with other 'helping factors' such as stakeholder behaviour, related programmes and policies, institutional capacities, cultural factors or socio-economic trends. Designs and methods for IE need to be able to unpick these causal packages.
- ☞ IEs should not be regarded as an everyday commission. Any IE that is thorough and rigorous will be costly in terms of time and money and will have to be justified.



## 2. The quality of Sida evaluations



- ❧ Mid Term Review of the Framework Agreement for Sida Reviews, Evaluations and Advisory Services on Results Frameworks. Transtec. LaGuardia, D. et al (2013)
- ❧ Are Sida Evaluations Good Enough? Sida studies in Evaluation 2008:1. Forss, K. et al.
- ❧ The Quest for Quality - or can evaluation findings be trusted? Evaluation, Vol (4), 481 - 502. Forss, K. and Carlsson, J. (1997)
- ❧ Några aspekter på kvaliteten inom utvärderingar av svenskt biståndssamarbete. Genomgång av databas upprättad vid Sekretariatet för Analys av Utvecklingssamarbete. En rapport beställd av Sekretariatet för Analys av Utvecklingssamarbete. Forss, K. (1994)

# Coverage



- 1994: 177 evaluations between 1969 and 1994
- 1997: 277 evaluations between 1969 and 1996
- 2008: 34 evaluations between 2002 and 2005
- 2013: 80 reports between 2011 and 2013(?)

# What defines evaluation quality?



☞ Utility

☞ Feasibility

☞ Propriety

☞ Accuracy

☞ The Program Evaluation Standards. Joint Committee on Standards. Sage (1994)

# Choice of evaluation design



Evaluation design	1994	1997	2008	2013
Experimental		-	-	
Statistical		-	-	
Theory-based		9%	33%	
Case-based		-	-	
Participatory		1%	-	
Synthesis		-	-	



# Conclusions on choice of design



- ❧ Choice of design does not immediately reflect poor quality of an individual evaluation
- ❧ In a system where many evaluations are commissioned you would expect a variety of designs
- ❧ Impact is often only one of many questions in ToR and hence design for impact analysis does not drive the evaluation effort
- ❧ Very low awareness of the choice of design. Evaluation methods are primarily discussed in terms of data collection (interviews, surveys, etc.)

# Aggregated rating 1994



Excellent	-	10	6%
Adequate	-	76	45%
Poor	-	31	18%
Not possible to assess	-	52	31%
Sum		169	100

# Aggregated rating 2008



Rating interval	Description	Number of evaluations	Percentage of total
26 – 30	Excellent in respect of several quality aspects and very good in the others (Very good)	0	0
21 – 25	Good in respect of many quality aspects, and satisfactory in respect of others (Good)	11	37%
16 - 20	Adequate, but not quite satisfactory, some shortcomings in essential aspects of quality ( Medium)	12	40%
11 – 15	Significant shortcomings in some respects and some shortcomings in most of the other (Poor)	6	20%
6 – 10	Significant shortcomings in all aspects of quality (Very poor)	1	3%

# Aggregated rating 2013

ALL REPORTS	Average (TOTAL)	Average (Last 12 Months)	Rankings
	ALL (80 Reports)	Aug '12 - Jul '13 (41 Reports)	
Terms of Reference	<b>3.96</b>	<b>4.08</b>	6 – Excellent
Evaluation Questions	<b>4.19</b>	<b>4.38</b>	5 – Adequate
Methodologies	<b>4.28</b>	<b>4.48</b>	4 – Minimally adequate
Data Collection	<b>4.39</b>	<b>4.45</b>	3 – not quite adequate
Data Analysis/Analytical Content	<b>4.18</b>	<b>4.26</b>	2 – significant problems
Conclusions & Recommendations	<b>4.48</b>	<b>4.57</b>	1 – very poor
Reporting	<b>4.75</b>	<b>4.80</b>	
<b>Data Analysis/Analytical Content, Conclusions &amp; Recommendations, and Reporting Average</b>	<b>4.41</b>	<b>4.54</b>	
<b>Total Components Average</b>	<b>4.31</b>	<b>4.41</b>	



# Concluding remarks



- ∞ The quality might be slowly increasing
- ∞ Fewer really poor evaluations
- ∞ Fewer really excellent evaluations (?)
- ∞ Still much confusion about what constitutes quality
- ∞ The assessments – though extensive – focus on some few aspects of quality
- ∞ Ethics and Feasibility are the aspects of quality that have been largely neglected
- ∞ The question of design for impact assessment has not been properly analysed in the quoted studies