ANZEA WORKSHOP

February 2011

HOW RIGOROUS IS RIGOROUS IMPACT EVALUATION ?

"Creating a culture in which rigorous randomised evaluations are promoted, encouraged and financed has the potential to revolutionize social policy during the 21st century just as randomised trials revolutionised medicine during the 20th century"
Esther Duflo

Robert Picciotto
European Evaluation Society

"Impact" means different things to different people

- **3iE:** the difference between indicators of interest with (Y1) and without the intervention (Y2), i.e. I = Y2 – Y1.

*This definition focuses on attribution, i.e. on assigning causality to effects observed by comparing them to the **counterfactual***

- **DAC:** "the positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended".

*This definition focuses on the **results chain** that connects the inputs, outputs, outcomes and the **impacts** of an intervention.*

Different definitions imply different evaluation questions

- The 3iE definition asks ***whether*** an intervention works (attribution)
- The DAC definition is concerned with ***whether*** an intervention worked over time- ***why*** and ***who*** was responsible
- The 3iE approach posits a ***gold standard***
- The DAC definition is pluralistic and implies more attention to ***context, accountability and sustainability***

"Rigour" in the physical sciences vs. "rigour" in the social sciences

- In the ***hard sciences*** rigour means cogency, exactness, accuracy and strict application of logical precepts and rules: evidence is only integrated into the knowledge base if researchers comply with replicable verification strategies
- But according to the dictionary to be rigorous is to be punctilious, severe, stern – even cruel, stiff and rigid (*rigor mortis*) reflecting dogmatic attitudes inconsistent with creativity and innovation in the ***social sciences***

Evaluation "rigour" is not limited to RCTs...

- Geology, astronomy and epidemiology demonstrate that causation ***can*** be established without RCTs
- Expert investigatory techniques, contestability protocols and rules of evidence are enough to fine, jail or even hang individuals convicted of a crime... Why are they not good enough for evaluation?
- In the social sciences quantitative measurements are not universally endorsed as the only valid proofs of rigour: rigour also applies to qualitative methods
- For Guba and Lincoln evaluation rigour = trustworthiness: (i) credibility (*accuracy)*; (ii) dependability (*reliability*) (iii) transferability (*external validity*) and (iv) conformability (*verifiability*)

But RCTs *can* be useful to ascertain attribution

- ***Before- after*** *comparisons* are the traditional way of assessing the impact of public interventions and they are appropriate where no other rival explanation for the changes exists
- But where influences other than the intervention could have affected the results the difference between the *ex-ante* and *ex-post* situation is ***not*** a reliable measure of the effects of an intervention
- Overcoming this indeterminacy is the basic rationale of ***with-without*** methods that compare observed outcomes with those of a ***counterfactual*** (what would have happened without the intervention)

Hence, RCTs are an integral part of the evaluation tool kit

- RCTs establish causality by comparing observed treatment effects with the counterfactual
- Random assignment ensures that the impact of the intervention is reliably ascertained since all the other contributing factors are identical except for stochastic errors
- RCTs eliminate the selection bias that arises from the programmatic choice of intervention groups or from participants' self selection
- RCTs enjoy the additional advantage of allowing evaluators to establish a measure of statistical significance of likely program impacts

Unfortunately, RCTs are rarely feasible or relevant...

- RCTs are not fit for complex, changing or adaptable programs
- RCTs are not feasible when no untreated target group can be identified
- RCTs “black boxes” do not distinguish between design and implementation issues
- RCTs cost a lot, require large studies and call on exceedingly scarce skills
- RCTs encourage simplistic interventions
- Ethical standards may prevent experimentation

… and it is very hard to achieve rigour in RCTs

Credibility (accuracy)

- "Indicators of interest" may leave out unintended, indirect and secondary effects
- Securing adequate group sizes can be problematic

Dependability (reliability)

- For programs with modest effects (high noise levels) RCTs can lead to the erroneous conclusion that "nothing works"

Transferability (external validity)

- RCT conclusions cannot be generalised since the groups affected and the circumstances in which an intervention takes place vary

Conformability (verifiability).

- Observed behaviour may be affected by the experiment itself

Alternatives to RCTs' estimations of the counterfactual do exist

- Regression and factor analysis
- Quasi-experimental designs
- Multivariate statistical modelling
- Qualitative approaches
- Surveys and sampling
- General elimination methodology
- Expert panels
- Benchmarking

Medical research is not as rigorous as commonly believed

- Administering a social program is not the same as administering a pill
- Research studies published in scientific journals regularly come up with different conclusions
- Dubious claims regularly slip through peer-reviews
- "***Most published medical research findings are false"*** (John P.A. Ioannidis, Director of the Prevention Research Centre, Stanford University)
- Re-testing of 45 of the most influential, highly cited medical research findings have led to refutation or convincing claims of significant exaggeration in over 40 percent of the cases

RCTs can be (and have been) subverted

- New drugs are often tested against placebos rather than drugs currently in use so that minor variations are amplified
- Comparisons are not always based on equivalent dosages
- Younger subjects who suffer less from side effects are selected for testing
- Testing is often of short duration even for drugs taken over a life time
- Private companies control data analysis and publication
- Findings from negative or inconclusive trials are usually suppressed and reports are written to show products in a favourable light

RCTs can be (and have been) captured by vested interests

Most drug trials in universities are funded by private companies

Private sponsors control trial designs, data analysis, interpretation and dissemination

They have often captured the regulatory framework

They employ hundreds of lobbyists and make large political contributions to major parties.

Hence RCTs have not protected the evaluation process from weak priority setting, misleading selection of comparators, “cherry picking” of data, biases in design and distorted reporting

Yet, RCTs have become all the rage in evaluation. Why?

- A public thirst for certainty?
- Budget stringency, i.e. tighter scrutiny of public policies, programs and projects?
- Poor evaluation quality: selection bias rarely addressed by evaluators?
- Traditional disciplinary bias (economists vs. others)?

Are we witnessing a rise in methodological fundamentalism?

Fundamentalists enjoy moral certitude; do not accept evidence that contradicts the revealed truth; cannot persuaded by logical arguments; exclude other perspectives; only associate with other believers; overcome resistance through exclusion and compulsion, etc.

"The problem is not randomised experimentation as such. It is the belief that there is one source of truth and one only. The error would be equally egregious if the government endorsed qualitative studies as the only source of truth." Ernest House

The roots of the philosophical debate run deep

- Experimentalism emerged as a revival of innocent religion in the enlightenment (Bacon)
- Positivism extended and experimentalism to human society (Comte; Durkheim; Weber)
- Since then historical materialists (Marx) and critical theorists (Adorno, Habermas) have denounced positivism now in retreat
- Post modern critics have even questioned the possibility of objectivity in the social realm

The history of evaluation methods reflects evolving worldviews

- ***Donald Campbell***, the visionary methodologist of the *Experimenting Society,* viewed RCTs as the methodological gold standard
- But he later recognized that *"to be truly scientific, we must re-establish the qualitative grounding of the quantitative" ,* part of a pluralistic trend
- ***Cook*** developed quasi-experimental techniques; ***Rossi*** and ***Weiss*** proposed theory-driven evaluations; ***Patton*** developed utilization based techniques; ***Scriven*** promoted valuing in the public interest as the central object of the discipline; etc.

Tools are just tools

- Since doctrinal differences cannot be reconciled it is best to focus on real world solutions to real world problems
- Understanding the limits of tools is key to evaluation quality
- Making them explicit is an ethical imperative
- All guidelines give equal weight and credence to qualitative and quantitative approaches.
- Overinvestment in a particular technique is a threat to quality

Conclusion: RCTs have an important but limited role in evaluation

Where adequate resources and skills are available and ethical dilemmas can be resolved, rigorously designed and independently implemented RCTs are the best way to assess attribution but only for relatively simple interventions the effects of which are realised in a short period of time and are large relative to other potential influences

"The only gold standard is appropriateness" (Patton)

1. In most real world situations mixed methods are the way to ascertain what works and doesn't work; why interventions succeed or fail; whether design or implementation problems need to be addressed and who among partners is responsible for particular outcomes
2. An explicit focus on the political and organizational constraints that affect evaluation processes is equally critical to evaluation trustworthiness and rigour under real-world constraints
3. Methodological fundamentalism has no place in the evaluation community and that a tolerant and pragmatic approach ought to prevail

THANK YOU FOR YOUR ATTENTION!

References

Michael Bamberger, Vijayendra Rao, Michael Woolcock, *Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development,* Brooks World Poverty Institute Working Paper 107, University of Manchester

Ernest R. House, 2004, *Democracy and Evaluation*, Paper presented to the European Evaluation Society, Berlin, October, IMA publications (http://www.informat.org/publications/ernest-r-house.html)

Ernest R. House, 2008, *Blowback: Consequences of evaluation for evaluation*, in American Journal for Evaluation, 29:416, December.

John P.A. Ioannidis, 2005a, *Why most published research findings are false,* PLoS Medicine 2 (8)

John P.A. Ioannidis, 2005b, *Contradicted and initially stronger effects in highly cited clinical research*, Journal of American Medical Association, 294 (2)

F. Leuuw and J. Vaessen, 2009, *Impact evaluations and development: NONIE guidance on impact evaluation*, World Bank, Washington D.C.

H. White, 2009a, *Some Reflections on Current Debates in Impact Evaluation*, International Initiative for Impact Evaluation, Working Paper 1, New Delhi

H. White, 2009b, *Theory based Impact Evaluation: Principles and Practice*, International Initiative for Impact Evaluation, Working Paper 3, New Delhi