

DAY #2

# Vladimír Kváča

*vladimir.kvaca@gmail.com*

I was:

- Assistant professor of Development studies at University of Economics, Prague (5 years)
- 9 years in ESIF (till May 2016), including:
  - Head of ESF Evaluation Team,
  - Director General for ESF at CZ Ministry of Labour,
  - Director of Technical Assistance OP Managing Authority,
  - Director of Department of Partnership Agreement, Evaluation and Strategy (National Coordination Authority, CZ Min. of Regional Development)
- Community of Practice on Results Based Management regional coordinator

I am:

- Independent consultant on parental leave
- Expert in EC Public Administration and Governance Network
- Assistant professor at Charles Uni, Prague (Policy design, Evaluation)

Learning about ToRs from others



Howard White

worked at the World Bank

he left the Bank to create

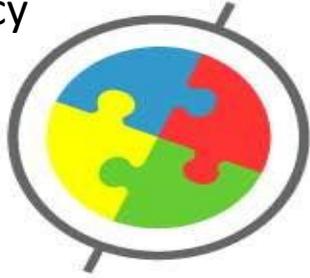
the

**International Initiative on  
Impact Evaluation (3ie)**

He was recently  
appointed Director of the  
**Campbell Collaboration**

I would define Howard White  
a key player in the  
international movement  
for **better evidence**  
**on what works**  
for whom and why  
(BE4W)

Putting evidence at the heart of policy and practice.



London,  
26-28 September  
2016

**WHAT WORKS**  
**GLOBAL SUMMIT 2016**



**International  
Initiative for  
Impact Evaluation**

**Sense about Science**

## **The Campbell Collaboration**

The Campbell Collaboration promotes positive social and economic change through the production and use of systematic reviews and other evidence synthesis for evidence-based policy and practice.

PROMOTED BY NUMBER  
OF ORGANIZATIONS  
INCLUDING  
THE USUAL  
SUSPECTS

**The Centre for Evidence and  
Social Innovation**

QUEEN'S UNIVERSITY IN BELFAST

**A Contribution to  
Current Debates in  
Impact Evaluation  
By Howard White  
*Evaluation* 2010;  
# 16; pp.153-164**

Evaluation

<http://evi.sagepub.com>

---

**A Contribution to Current Debates in Impact Evaluation**

Howard White

*Evaluation* 2010; 16; 153

DOI: 10.1177/1356389010361562



# 10 key points from Howard White

*„The results agenda has seen a welcome **shift in emphasis from inputs to outcomes**. However, it has been realized that **outcome monitoring does not tell us about the success**, or otherwise, of government programmes or the interventions supported by international development agencies.“ (p. 153)*

## Implications for ToR writing

What is an absorption rate?  
What is a take-up rate?  
When it is appropriate to use these concepts, when not to?

Btw. try to think about still so frequent „process evaluation“ as about impact evaluation of your technical assistance priority axis.

I am afraid that Howard White's despair from focus on outcome monitoring (as experienced by him in the development cooperation world) is still a kind of optimism in our ESIF world.

Here, the focus is still on absorption and error rate. Despite of all the effort...

“The most evident weaknesses which indicate the need for reform of cohesion policy are:

- A remarkable lack of political and policy debate on results in terms of the well-being of people, at both local and EU level, most of the attention being focused on financial absorption and irregularities.”

*April, 2009: Independent report “An Agenda for a Reformed Cohesion Policy” delivered at the request of Commissioner for Regional Policy, Ms Hübner*

*„Barca's Report“*

- “Results and result indicators
  - The intended **result** is the specific dimension of **well-being and progress for people** that motivates policy action, i.e. what is intended to be changed, with the contribution of the interventions designed.
  - **Result indicators** are variables that provide information on **some specific aspects** of results that lend themselves to be measured“

*March, 2014: Guidance Document on Monitoring and Evaluation (EC, DG Regio)*

⇒ Suggested further study: Amartya Sen’s Capability Approach.

⇒ [https://en.wikipedia.org/wiki/Capability\\_approach](https://en.wikipedia.org/wiki/Capability_approach)

- What matters is the end-to-end process:
  - From Euro invested to positive change in the well-being of the citizens.



- You can spend billions of € in programmes that were well written respecting the partnership principle, have close-to-zero error rate, disburse payments within 15 days and reach all the target values of your indicators and still make no change in the well-being.
  - You can be effectively and efficiently irrelevant.
  - You can have no IMPACT.

## • More rigorous evaluation

*„Advocacy by a number of groups asked for more rigorous evaluation of development programs, most notably from the Poverty Action Lab at MIT, from the World Bank, the Inter-American Development Bank and more recently 3ie“ (p. 153) as well as most of the academic world.*

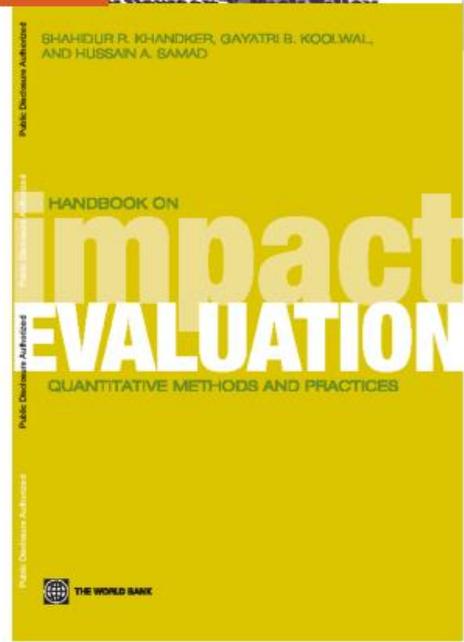
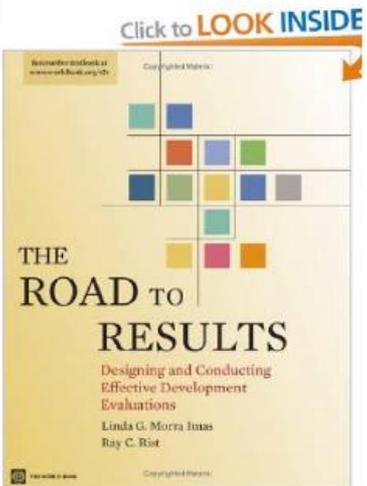
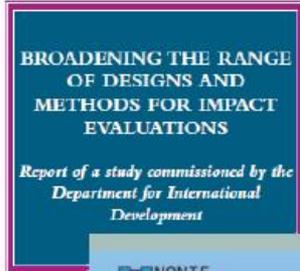
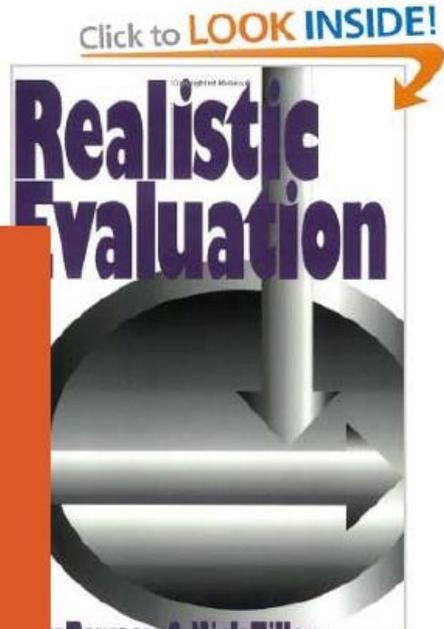
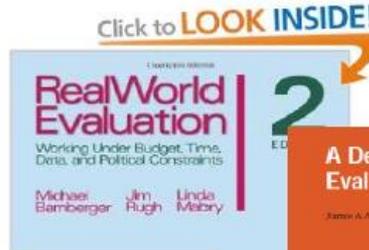
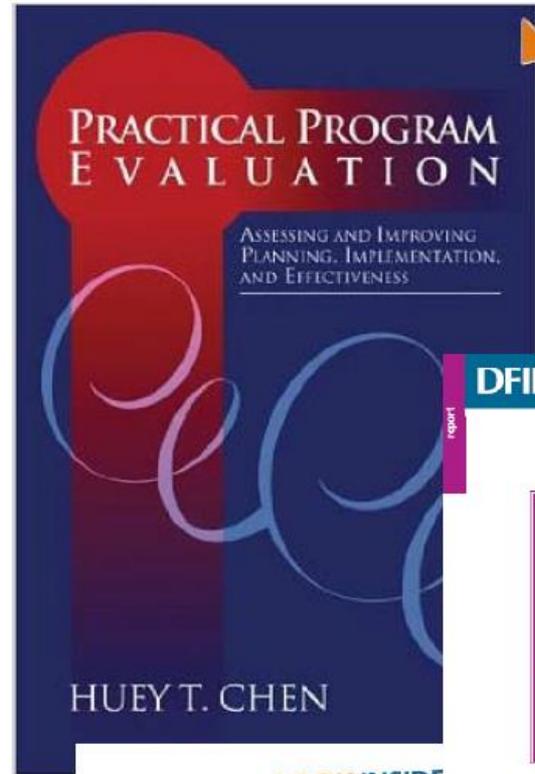
Beware: more rigorous evaluation has become in recent years a by-word for quantitative analysis using counterfactual methods.

### Implications for ToR writing

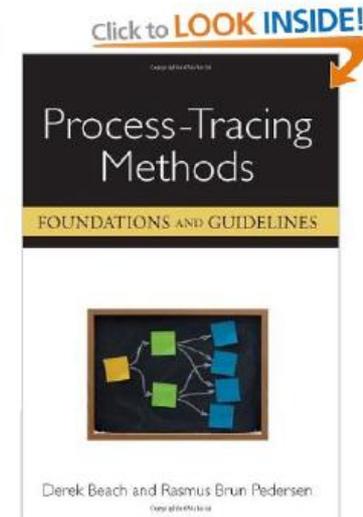
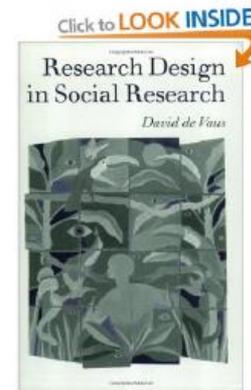
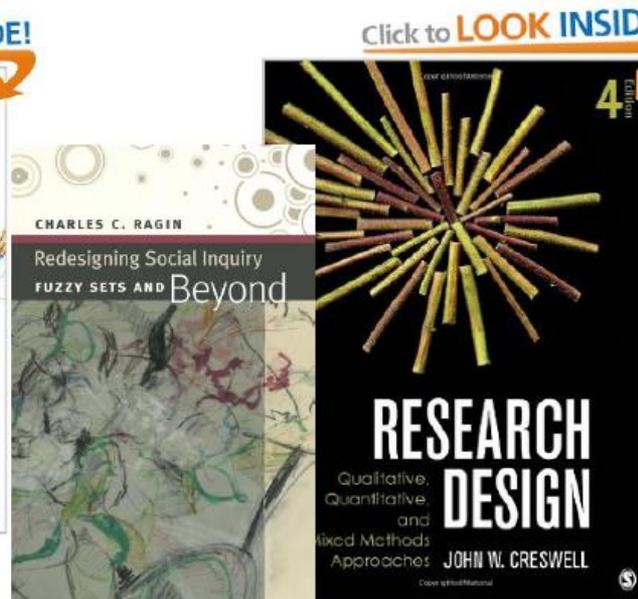
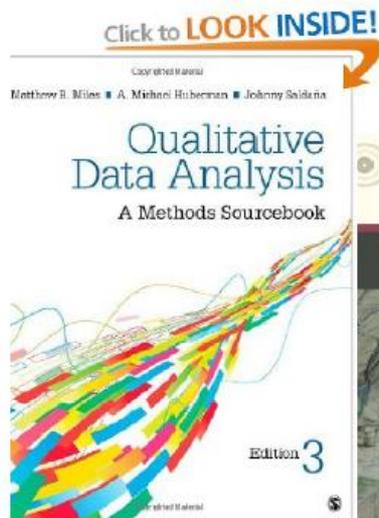
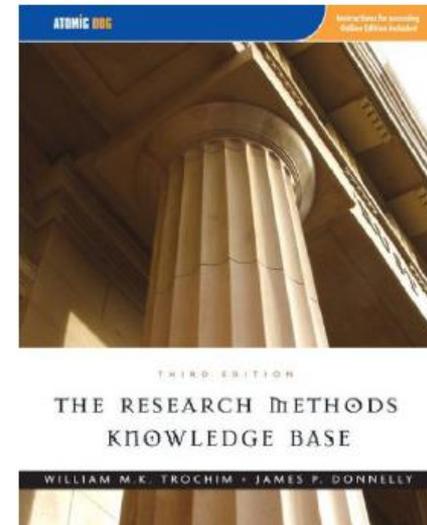
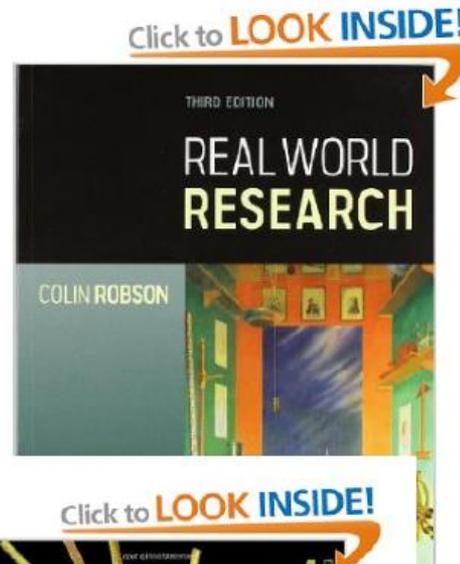
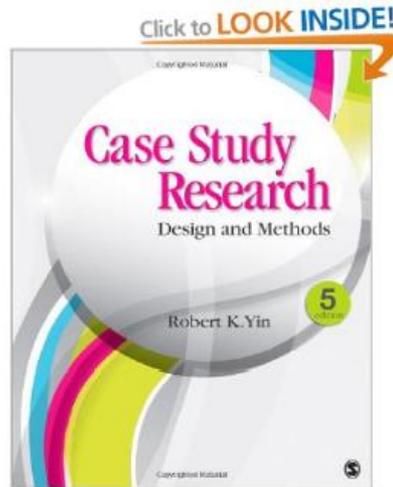
Be aware of this and keep in mind that deception, posturing and blind following of fashion trends are the rule around evaluation.



Non EC sources on evaluation



# Research methods sources



- **There are a number of misunderstandings**

*„The most important of these is that different people are using different definitions of ‘impact evaluation’. Since this is a purely semantic matter, neither side is right or wrong. The definitions are just different. It makes little sense to debate on the appropriate methodology when people are in fact talking about different things.“ (p. 154)*

## Implications for ToR writing

Before you even start try to agree on the fundamentals.

What are we measuring and why.

Your understanding of concepts is not the only possible and not the only correct. Don't blame others for not understanding you if you make no effort to explain your position.

*„The two sides of the impact evaluation debate are commonly talking about completely different things, but seem not to realize this.*

*Amongst evaluators, ‘impact’ typically refers to the final level of the causal chain (or log frame), with impact differing from outcomes as the former refers to long-term effects.“ (p. 154)*

Implications for ToR writing

~~IMPACT = long term EFFECTS?~~

**WE DO NOT THINK SO**

We will touch this in the coming soon discussion of Veronica Gaffey's paper.

Btw. taking impact aside, is there any good reason to distinguish outputs and outcomes? The border is blurred and definition wars can easily be both endless and useless.

⇒ *Suggested further study:*

⇒ *See the discussion on „Result“ in Wauters (2015), pp. 12-24*

[http://www.latitudeconsulting.eu/images/sourcebook\\_rbm.pdf](http://www.latitudeconsulting.eu/images/sourcebook_rbm.pdf)



„Impacts are ‘positive and negative, primary and secondary **long-term effects** produced by a development intervention, directly or indirectly, intended or unintended’. **Any evaluation which refers to impact (or often outcome) indicators is thus, by definition, an impact evaluation.** So even outcome monitoring can fall under the heading of impact evaluation.

But this definition is **not shared** by many working on impact evaluation, for example in the World Bank, or, indeed, *3ie*. Impact is defined as the difference in the indicator of interest ( $Y$ ) with the intervention ( $Y_1$ ) and without the intervention ( $Y_0$ ), known as **the counterfactual**.” (p. 154)

Implications for ToR writing

$$\text{effect} = \text{impact} = Y_1 - Y_0$$

- Most narratives of ESIF effects are based on *post hoc ergo propter hoc* fallacy. (After this, therefore because of this). „*OP ABC created 50,000 jobs in Farawayland*“ (based on monitoring indicators records).
- There is strong pressure from both national and European level decision makers to show the success, as the value of cohesion policy is being questioned.
- ***Post hoc ergo propter hoc* has been wrong for 2000 years. And still is!**
- Post Hoc fallacies are committed because leaping to a causal conclusion is always easier and faster than actually investigating the phenomenon. While it is true that causes precede effects (outside of Star Trek, of course), it is not true that precedence makes something a cause of something else. Thus, a causal investigation should begin with finding what occurs before the effect in question, **but it should not end there.**

- $Y_1$  is what happens while the intervention is observed, but we don't know what would have happened without the intervention ( $Y_0$ ). There are various ways of getting an estimate of  $Y_0$ . A common, though usually unreliable, one is the value of  $Y$  before the intervention. Still, very often we miss even baseline data.
- If one group pre-post evaluation design is usually too weak, one group post-only is then no evaluation design at all.

~~Post hoc ergo propter hoc~~

*„We might think that before versus after is adequate if there is no observed change in Y after the intervention. Y has remained unchanged, so clearly the intervention had no impact more unsettling the case in which Y has decreased. Once again, there is a counterfactual here, though there is no comparison group.“  
(p. 157)*

## Implications for ToR writing

In the evaluation of ESIF other confusion arises because of the mixing of baselines, targets, objectives, milestones, achievements. It could be all much simpler thanks to the unifying notion of counterfactual. But also more challenging because the **counterfactual is not observable by definition.**

- Even where benefits are obvious, quantification is still useful

*„It can be argued that ‘a program that gives food or shelter to those who need it immediately has beneficial consequences’. But this actually is not true for more complex outcomes, and even if it is, 1) there is still a counterfactual, and 2) quantification allows analysis of the efficiency of the intervention in achieving welfare outcomes.“ (p. 157)*

Here White is a bit sloppy with the word EFFICIENCY. He should have said **COST-EFFECTIVENESS**, that is, the comparison of cost per unit of impact.

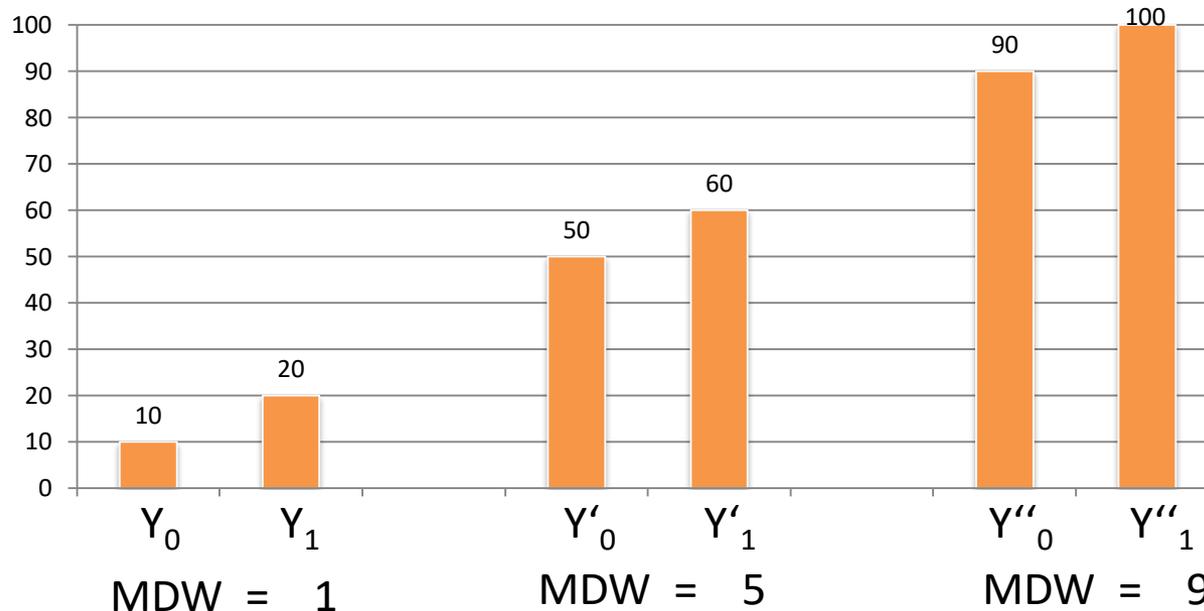
## Implications for ToR writing

Once you identify the impact ( $Y_1 - Y_0$ ) ask how much it costs and if the price is reasonable.

## SHORT DIGRESSION

to better illustrate **Cost effectiveness**

Good example of the equivalence of the terms  
(cost effectiveness – deadweight – efficiency)



The impact is positive and significant in all scenarios. However, it takes a lot more resources to produce the same result (10 %points) with the higher take up. So deadweight is much higher. A useful measure of deadweight, MDW' is the ratio on the cost paid to serve the controls/divided by the net impact.

*„By attribution, I mean attributing observed changes to the intervention being studied. A good study would (...) say that since Y changed by P percent over the period of the intervention, say, a quarter ( $p/P = 0.25$ ) of the overall change can be attributed to the intervention. In this sense, attribution analysis also addresses contribution.“ (p. 159)*

## Implications for ToR writing

Attribution is just another nickname of the counterfactual and it also addresses contribution.

*„Impact varies by intervention, characteristics of the treated unit and context. **Context is one aspect of impact heterogeneity.**“ (p. 160)*

Context is a weapon used to delegitimize any quantitative analysis. So we should welcome Howard White assertion that context is simply a group of potentially covariate factors.

## Implications for ToR writing

*„A study which presents a single impact estimate (the average treatment effect) is likely to be of less use to policy-makers than one examining in which context interventions are more effective, which target groups benefit most and what environmental settings are useful or detrimental to achieving impact.“ (p. 160)*

*„I believe the argument between proponents of theory-based evaluation and RCTs is overstated. A theory-based approach provides a framework for an evaluation. It still needs an analytical approach to determine if outcomes have changed as a result of the intervention, so experimental or quasi-experimental approaches are embedded in a mixed-methods approach.“ (p. 162)*

## Implications for ToR writing

Understand the role of theory, ask for mixed methods approach.

LONG DIGRESSION  
to better illustrate

## **Mixed methods**

as

A way to a good impact  
evaluation

# Two sides of inferencing

## Induction

Gather information

Open ended questions, records of field notes

Analysis to form themes/categories

Broad patterns, theories

Theories / patterns related to past experience / literature

## Deduction

Past experience, literature, theory

Test a theory

Test hypotheses (null / alternative)

Defines and operationalises variables (dependent / independent))

Measures variables using an instrument

# Two sides of inferencing

Induction used for  
developing a theory  
(exploratory  
research)

= Theory building

Deduction used for  
testing, validating:  
- Hypothesised causal  
relations (explanatory  
research)  
- Descriptive  
hypothesis  
(descriptive research)

= Theory testing

# Two sides of inferencing

Induction = is not search  
for truth,  
but for  
interesting concepts,  
plausible explanation:  
Validity has no sense here.

= **Building**

Deduction used for  
theory testing

-  
Deduction = search for truth.  
Validity is essential here.

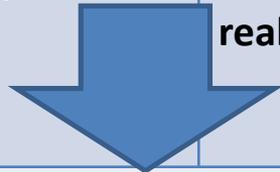
(descriptive research)

= **Theory testing**

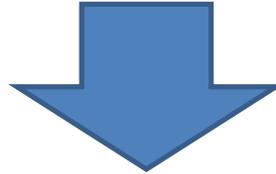
# Good impact evaluation

Inductive part

<b>Work to be done:</b>	<b>Control questions:</b>	In non-existent perfect world, this is a part of policy design.
Desk research. What is already known about the topic?	Is the evaluator pretending to be the first one on the planet Earth who is dealing with the topic of intervention (policy, programme, project)?	
<b>What are the relevant theories available in the scientific literature? Is there any theory explicitly or implicitly expected by the policymaker to be applicable in the intervention?</b>	<b>Is the evaluation using theories? Is he/she critical to assumption of the policymakers? Is there a specialist in the field of intervention involved in the evaluation team?</b>	
Pre-research in the field. (Stakeholders interviews, focus groups...).	Is the evaluator confronting own assumptions and ideas with the reality?	
<b>Output: theoretical explanation of why the intervention should (not) work.</b>		



# Good impact evaluation



Deduction: Test preparation

Explicit description of „observable implications of the theory“ = empirical prediction.  
„If the theory is right, I should be able to observe XY, if it is wrong, I will detect ABC...“

Is the evaluator considering the power of particular tests with regard to their confirmatory and disconfirmatory properties?

Operationalization of empirical predictions (tests) to the level of particular variable / observations.

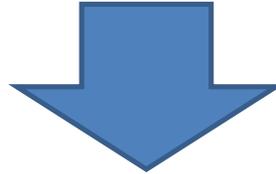
Is the evaluator working with existing indicators only?

**Output: Set of tests**

Again, in non-existent perfect world, this is a part of policy design. I am afraid here the indicators were born. Road to hell is paved with good intentions.



# Good impact evaluation



Deduction: Running the tests

**Empirical testing.**

No data used in **INDUCTIVE** part can be used in testing. You need different data for theory building and for theory testing. Otherwise tautological conclusions appear. (This one of the meanings of triangulation principle in evaluation)

**Output: Tests executed, results interpreted**

Try to ask for this approach  
in the ToR...

# Tests

Low uniqueness (confirmatory power)

High uniqueness (confirmatory power)

*Low certainty /disconfirmatory power*

## Straw in the Wind tests

E.g. murder suspect was known to have a bad temper

Weakest test: do little to update our confidence in  $h$  (ypothesis)  
Regardless whether we find  $e$  (vidence) or not ( $=-e$ )

## Smoking gun tests

E.g. murder suspect was seen wiping red liquid off a candle holder

If ( $e$ ) (then greater confidence in  $h$  (high uniqueness as  $e$  highly unlikely unless  $h$ ) and highly improbable rivals. If we find  $-e$ , the test is useless to update our confidence.

## Hoop tests

E.g. Murder suspect was in town in the week of the murder

E.g. suspect was in proximity of the murder location around the time of the murder

If ( $-e$ ) = was not in town, reduces our confidence in  $H$ , if ( $e$ ) = was in town, does little. Hoops: sit on a continuum where tighter hoop means if ( $e$ ), it is NOT useless but has some confirmatory power!

## Doubly decisive tests

E.g. CCTV filmed the crime.

If ( $-e$ ) (suspect on camera) then ( $-h$ ), if ( $e$ ) then all other rival theories ruled out.

*Very rarely possible!*

*High certainty /disconfirmatory power*

- In this view, CIE is nothing more (and nothing less!) than a high quality test and has to be embedded within a theory.

# Group Exercise



Go back to your ToRs and:

1. Formulate a theory why your intervention should (not) work.
2. Make at least one empirical prediction
3. Suggest a test of your theory

What about  
a break now?