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

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

3ie International Initiative for Impact Evaluation

Sense about Science

The Centre for Evidence and Social Innovation

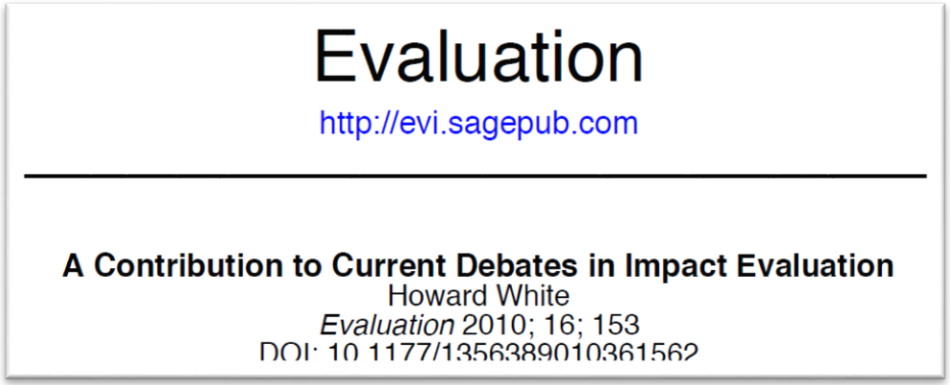
QUEEN'S UNIVERSITY IN BELFAST

A Contribution to Current Debates in Impact Evaluation

By Howard White

Evaluation 2010;

16; pp.153-164



10 key points from Howard White

1. of 10

• Outcomes do not tell us about success

*„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.

• Outcomes do not tell us about success

I am affraid 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 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“

• 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

• More rigorous evaluation

Using a word „counterfactual“ is not guarantee of rigour :-), and rigourous could be also in other methods – Howard White mentions QCA, think about rigorous process-tracing.

Without deep understanding the methods there is no rigorous evaluation.

3. of 10

• 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.

• Effects ≠ impact?

„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

• Effects ≠ impact?

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.

• Counterfactual to the rescue

„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

effect = impact = Y_1-Y_0

• Counterfactual to the rescue

- Most narratives on 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**.

• Counterfactual to the rescue

- Y_1 is what happens while the intervention is observed, but we don't know is 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 base-line 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~~

- **Counterfactuals is a wider notion than control group**

„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 counter factual, and 2) quantification allows analysis of the efficiency of the intervention in achieving welfare outcomes." (p. 157)

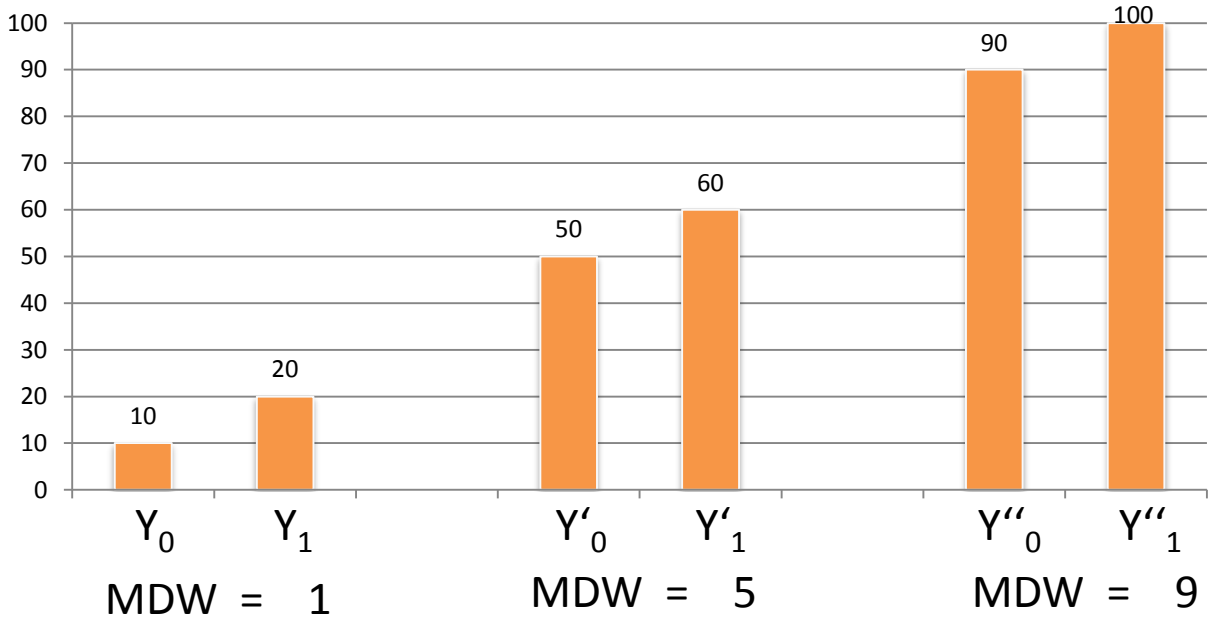
Here White is a bit sloppy with the word EFFICIENCY. He should have said **COST-EFFECTIVENESS, that is, the comparison of cost per unit of impact.**

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.

• Contribution versus attribution

„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.

9. of 10

• Impacts are likely to vary along many dimension

„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)

• Mixed methods

„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.

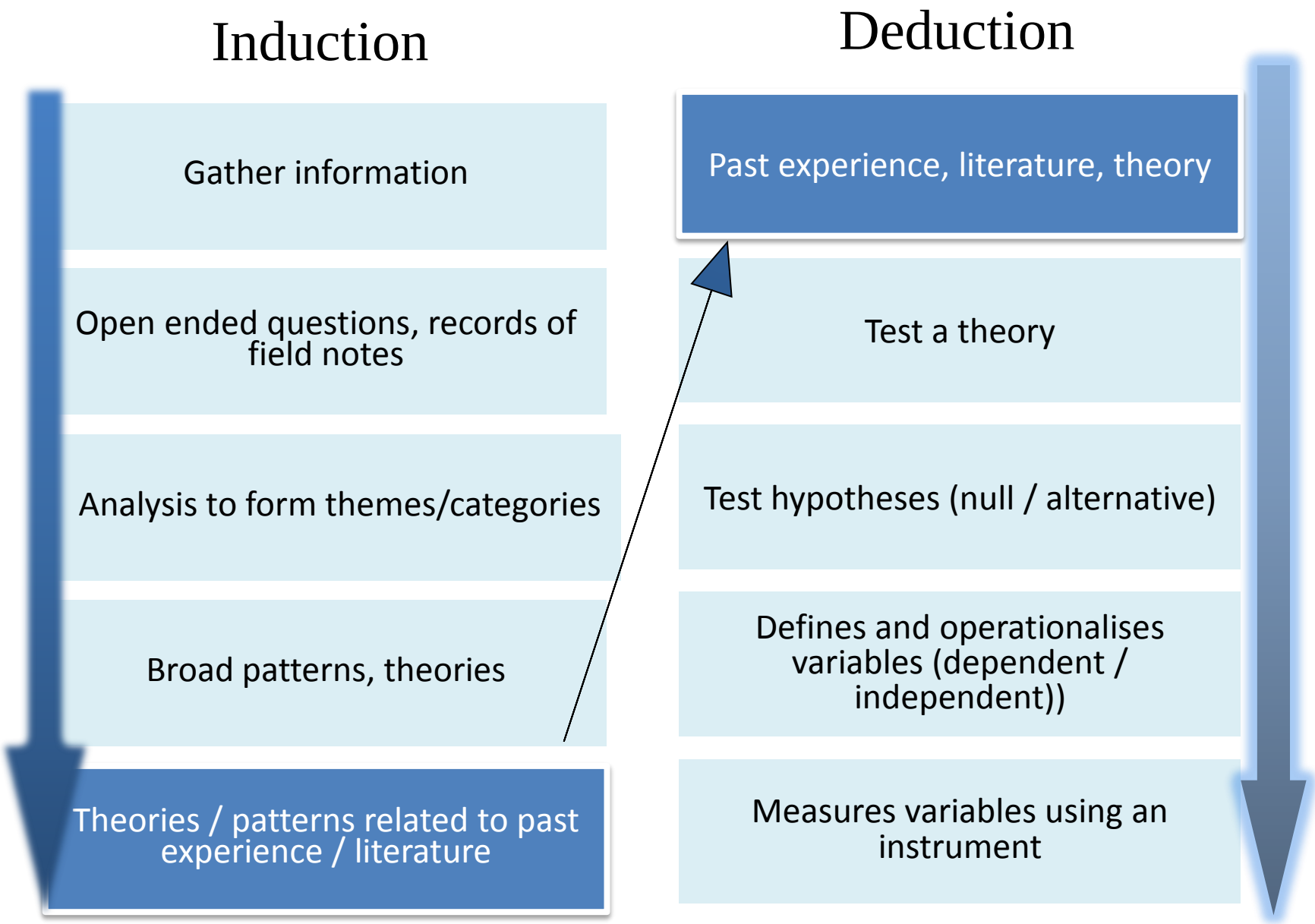
LONG DIGRESSION
to better illustrate

Mixed methods

as

A way to a good impact evaluation

Two sides of inferencing



Source: Creswell, 2003

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

Source: Creswell, 2003

Two sides of inferencing

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

=[illegible]uilding

Deduction used for
t[illegible]
-

Deduction = search for truth.
Validity is essential here.

(desc[illegible]esearch)

=Theory testing

Source: Creswell, 2003

Good impact evaluation

Inductive part

Work to be done:	Control questions:	
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)?	In non-existent perfect world, this is a part of policy design.
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?	**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?**	
Pre-research in the field. (Stakeholders interviews, focus groups…).	**Is the evaluator confronting own assumptions and ideas with the reality?**	
Output: theoretical explanation of why the intervention should (not) work.		

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 text 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 affraid here the indicators were born.
Road to hell is paved with good intentions.

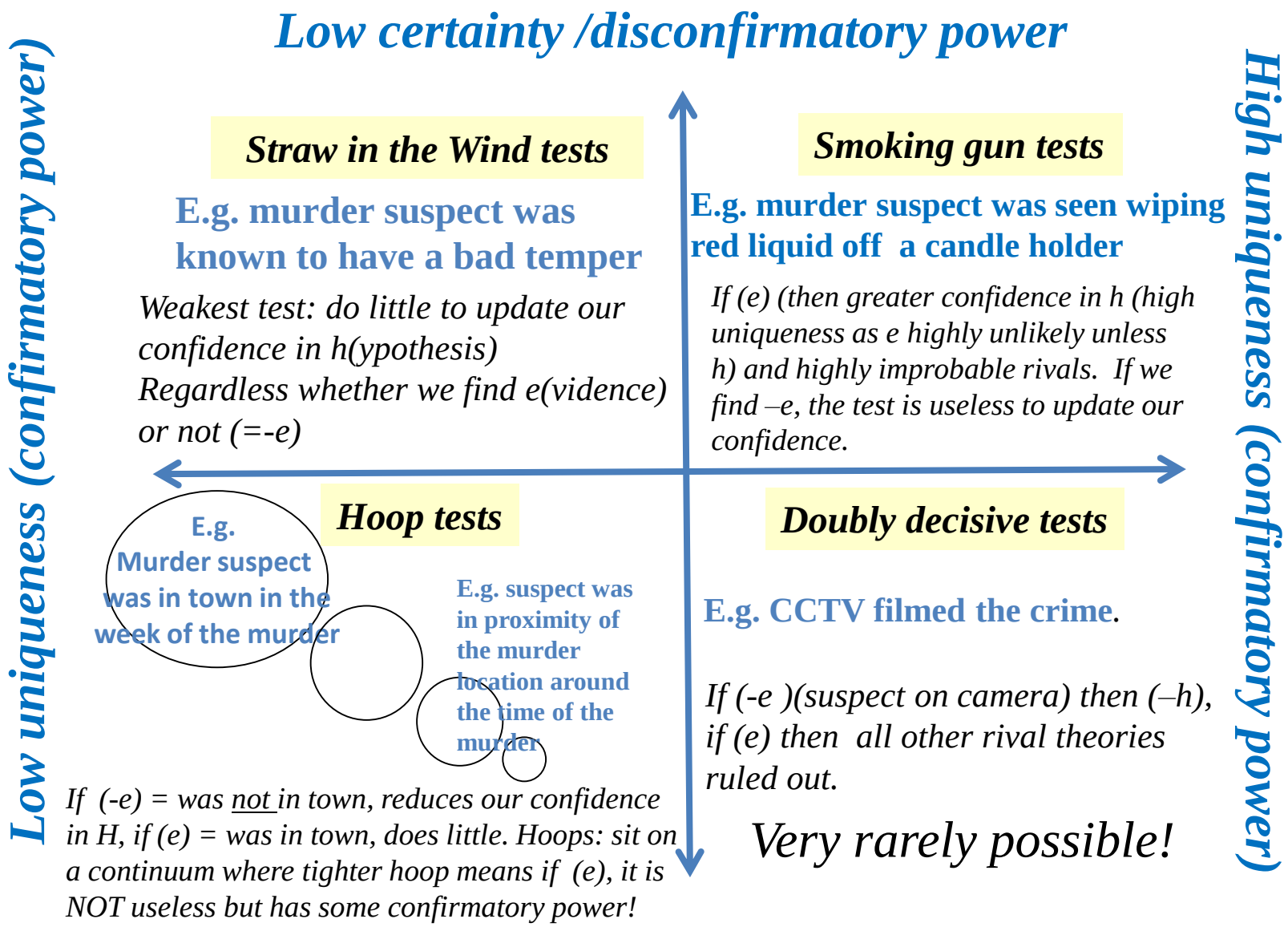
Good impact evaluation

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



Beach and Pedersen

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

What about
a break now?