The Debate About Randomized Controls in Evaluation:

The Gold Standard Question

Michael Quinn Patton
29 April, 2011
Copenhagen
What I will cover

• The issues and the debate about Randomized Controlled Trials (RCTs)
• Examples and applications of alternatives
• Methodological strengths and weaknesses of diverse approaches
• *Rigor Attribute Model*
• Contextually sensitive, contingency-based impact evaluation: Methodological appropriateness
My point of view

• Methods should be selected to answer evaluation questions and intended uses.
• No method should be designated as a “gold standard.”
• The international gold standard is methodological appropriateness.
• To label one method as the gold standard creates perverse incentives, violates basic evaluation design standards, reduces options, and does harm.
My point of view

• This is not an attack on experimental designs. I have conducted evaluations using experimental designs. RCTs are sometimes appropriate.

• This is an attack on treating RCTs as “the gold standard.” That is my focus here.

• RCTs are not inherently a superior method for any kind of evaluation, including impact evaluations.

Michael Quinn Patton
29 April, 2011
An example of an impact evaluation without an RCT: Public health campaign to reduce Sudden Infant Death Syndrome

The problem

Unexplained death of babies in North America, Europe and Australia – SIDS was the main cause of death among infants for most of the 1990s.
Multiple & Cumulative Methods

• Review of epidemiological data
• Review of individual cases
• Development of alternative hypotheses
• Investigation of evidence
• Development of guidelines for practice in the light of incomplete evidence
At least 19 retrospective case-control studies demonstrated a higher risk of SIDS when infants slept prone. Overall the studies showed a threefold or greater increased risk of SIDS when babies slept on their stomachs.

Public Health Association of Australia (1999/2005)
• An increased risk of SIDS when babies were exposed to cigarette smoke was found in over 30 case-control and cohort studies. This finding was consistent over time and place. Many studies reported a dose-response relationship.
FACT: In Australia, rates were higher in winter. 
POSSIBLE EXPLANATION: Linked to winter respiratory diseases.

ANOTHER FACT: Rates were not higher in winter in Canada or Sweden – which have colder winters. 

ANOTHER FACT: A review of cases found a large number who had been put to bed fully dressed or with additional bedding. 
POSSIBLE EXPLANATION: In under-heated houses, parents over-dressed their children. Over-heating seemed to be a risk factor. 

RECOMMENDATION FOR PRACTICE: Avoid overheating: The infant should be lightly clothed for sleep, and the bedroom temperature should be kept comfortable for a lightly clothed adult.
Guidelines for practice

- Avoid maternal smoking
- Lie the baby to sleep on its back
- Avoid over-heating
SIDS rates in Sweden

Risk reducing campaign, spring 1992

Alm et al 2001

Michael Quinn Patton
29 April, 2011
## Canada, 1990-2002

<table>
<thead>
<tr>
<th>YEAR</th>
<th>SIDS</th>
<th>Cause Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>327</td>
<td>98</td>
</tr>
<tr>
<td>1993</td>
<td>266</td>
<td>63</td>
</tr>
<tr>
<td>1994</td>
<td>267</td>
<td>61</td>
</tr>
<tr>
<td>1995</td>
<td>249</td>
<td>78</td>
</tr>
<tr>
<td>1996</td>
<td>166</td>
<td>88</td>
</tr>
<tr>
<td>1997</td>
<td>153</td>
<td>115</td>
</tr>
<tr>
<td>1998</td>
<td>158</td>
<td>56</td>
</tr>
<tr>
<td>1999</td>
<td>166</td>
<td>41</td>
</tr>
<tr>
<td>2000</td>
<td>121</td>
<td>0</td>
</tr>
<tr>
<td>2001</td>
<td>112</td>
<td>0</td>
</tr>
<tr>
<td>2002</td>
<td>111</td>
<td>0</td>
</tr>
<tr>
<td>2003</td>
<td>97</td>
<td>0</td>
</tr>
</tbody>
</table>

Michael Quinn Patton
29 April, 2011
Evidence Question

• How convincing is the evidence of impact?
• Can the drop in SIDS be attributed to the public health campaign with reasonable certainty based on the preponderance of evidence?
What does it mean for something to be the **GOLD STANDARD**?
The gold standard is a monetary system in which the standard economic unit of account is a fixed weight of GOLD.

When several nations are using such a fixed unit of account, the rates of exchange among national currencies effectively become fixed.
The United States stopped issuing promises to redeem dollars for gold in 1933 – part of a policy change for dealing with the Great Depression.
Preparing to rebuild global capitalism as World War II was still raging, 730 delegates from all 44 Allied nations gathered at the New Hampshire resort town of Bretton Woods, July 1944. Setting up a system of rules to regulate the international economy, they established the International Bank for Reconstruction and Development later divided into the World Bank and the International Monetary Fund.
The Bretton Woods system created an obligation for each country to maintain the exchange rate of its currency in terms of **GOLD**.

The system collapsed in 1971, following the United States’ suspension of convertibility from dollars to gold.
The international financial gold standard collapsed because of its RIGIDITY.
The Gold Standard issue in evaluation
First, in development, we’re not actually talking about the real RCT gold standard which is:

**Double and triple blind placebo control studies**

- What’s the placebo for a micro-loan?
Placebo Effect

• Huge literature on the placebo effect
• The MAJOR rival hypothesis in medical and pharmacological studies of all kinds
• Routinely ignored in evaluation impact studies using RCTs.

Michael Quinn Patton
29 April, 2011
GOLD STANDARD:

METHODOLOGICAL APPROPRIATENESS

not

methodological orthodoxy or rigidity
Political Context

- US Dept of Ed “Scientifically Based Evaluation Methods” policy (RCTs (Institute of Educational Sciences)
- Poverty Action Lab, MIT: RCTs as “the gold standard in evaluation…”
- Only 2% of World Bank programs “properly evaluated…” (NY Times, 28/7/04, p. A4; Chief economist of the WB)
- “When Will We Ever Learn” report and
- 3ie: International Institute for Impact Evaluation
OECD Handbook

Evaluation of Small & Medium Enterprises
(Professor David Storey,
Univ. of Warwick)

SIX STEPS TO HEAVEN
where heaven is an RCT
American Evaluation Association position and debate

• AEA dissent from RCTs as the gold standard
• Counter-statement from RCT advocates in AEA
AEA position (Nov 4, 2003):
(1) Studies capable of determining causality. Randomized control group trials (RCTs) are not the only studies capable of generating understandings of causality. In medicine, causality has been conclusively shown in some instances without RCTs, for example, in linking smoking to lung cancer and infested rats to bubonic plague. The secretary's proposal would elevate experimental over quasi-experimental, observational, single-subject, and other designs which are sometimes more feasible and equally valid.
RCTs are not always best for determining causality and can be misleading. RCTs examine a limited number of isolated factors that are neither limited nor isolated in natural settings. The complex nature of causality and the multitude of actual influences on outcomes render RCTs less capable of discovering causality than designs sensitive to local culture and conditions and open to unanticipated causal factors.
RCTs should sometimes be ruled out for reasons of ethics. For example, assigning experimental subjects to educationally inferior or medically unproven treatments, or denying control group subjects access to important instructional opportunities or critical medical intervention, is not ethically acceptable even when RCT results might be enlightening. Such studies would not be approved by Institutional Review Boards overseeing the protection of human subjects in accordance with federal statute.
In some cases, data sources are insufficient for RCTs. Pilot, experimental, and exploratory education, health, and social programs are often small enough in scale to preclude use of RCTs as an evaluation methodology, however important it may be to examine causality prior to wider implementation.
(2) Methods capable of demonstrating scientific rigor. For at least a decade, evaluators publicly debated whether newer inquiry methods were sufficiently rigorous. This issue was settled long ago. Actual practice and many published examples demonstrate that alternative and mixed methods are rigorous and scientific. To discourage a repertoire of methods would force evaluators backward. We strongly disagree that the methodological "benefits of the proposed priority justify the costs."
While we agree with the intent of ensuring that federally sponsored programs be "evaluated using scientifically based research . . . to determine the effectiveness of a project intervention," we do not agree that "evaluation methods using an experimental design are best for determining project effectiveness." We believe that the constraints in the proposed priority would deny use of other needed, proven, and scientifically credible evaluation methods, resulting in fruitless expenditures on some large contracts while leaving other public programs unevaluated entirely.

End of AEA position (Nov 4, 2003). Source: www.eval.org
In December 2007, the European Evaluation Society (EES) adopted a statement on “the importance of a methodologically diverse approach to impact evaluation—specifically with respect to development aid and development interventions.”

As context, the EES noted that …
“This statement was prepared in response to strong pressure from some interests advocating for ‘scientific’ and ‘rigorous’ impact of development aid, where this is defined as primarily involving RCTs. This debate has the potential to influence the future direction of evaluation—not only with respect to development but potentially in other areas as well.”
“EES however deplores one perspective currently being strongly advocated: that the best or only rigorous and scientific way of doing so is through randomised controlled trials (RCTs). In contrast, the EES supports multi-method approaches to IE [impact evaluation] and does not consider any single method such as RCTs as first choice or as the “gold standard.” (EES 2007:1)
GOLD STANDARD:

METHODOLOGICAL APPROPRIATENESS

not

Methodological orthodoxy or rigidity
Question should determine method

- Making RCTs the gold standard means that method determines question – and risks creating perverse incentives to design RCTs whether appropriate or not.
Perverse Incentives

When one method is treated as the gold standard, all other methods become inferior.

- Performance Bonuses
- EPA experimental study on effects of pesticides on young children
- AfrEA, Niger
Different Questions

• RCTs: Did it work?

• **Microcredit:** Does it reduce poverty?

• **Natural variation studies:** What works for whom, to what degree and in what ways, under what conditions.

• Who benefits from micro-loans, of what kinds, in what ways, with what results?
“A Proposal to Speed Translation of Healthcare Research into Practice: Dramatic Change Is Needed” Rodger Kessler, PhD, Russell E. Glasgow, PhD

“Randomized controlled efficacy trials using precisely defined interventions and highly selected participants have been the preferred and often exclusive design of choice. Designed for narrowly focused pharmacology trials, when applied to the other major issues facing health care today, such trials are limited in their ability to address the complex populations and problems we face. …”
“A moratorium is proposed on such research for the next decade, and pragmatic, transparent, contextual, and multilevel designs that include replication, rapid learning systems and networks, mixed methods, and simulation and economic analyses to produce actionable, generalizable findings that can be implemented in real-world settings is suggested.”

Michael Quinn Patton
29 April, 2011
“This shift would include greater focus on the needs of practitioners, patients, payers, and policymakers and generate more relevant evidence. Funding priorities would change to include greater focus on complex multimorbid patients in community settings. Changes would be made in grant review criteria and review sections would require reviewers with new methodologic skills and experience in pragmatic studies and contextual factors.”
“The current situation demands study of complex interventions that produce complex outcomes. Relying on an efficacy-based RCT research paradigm established to answer questions under decontextualized, optimal conditions will not produce the solutions needed.”
Dealing with Complexity

RCTs are about

• CONTROL
• STANDARDIZED, HIGH FIDELITY INTERVENTIONS
• RANDOMIZATION
• COUNTERFACTUAL LOGIC
COMPLEXITY

- ADAPTATION
- RAPID RESPONSE
- CONTEXTUAL SENSITIVITY
- EMERGENCE
- UNCERTAINTY & LACK OF CONTROL
- NATURAL VARIATION
The International Response to Conflict and Genocide: Lessons from the Rwanda Experience

Synthesis Report

Joint Evaluation of Emergency Assistance to Rwanda

Michael Quinn Patton
29 April, 2011
RTS and Shake Hands With the Devil
Classic RCT Problems

- Weak external validity
- Controls rather than comparisons
- Poor treatment specification (Black Box)
- Missing important natural variations
- Greater within group variation than between group variation
- Experimental mortality
- Artificiality of randomization & control
Making Impact Evaluations More Useful

Should the randomistas rule?
Economists’ Voice (2009), 6 (2), 1-5

Martin Ravallion

Development Research Group (DECRG), World Bank

(next 3 slides)
Start with a policy-relevant question and be eclectic on methods

- Policy relevant evaluation must start with interesting and important questions.
- But instead many evaluators start with a preferred method and look for questions that can be addressed with that method.
- By constraining evaluative research to situations in which one favorite method is feasible, research may exclude many of the most important and pressing development questions.
Fully explore impact heterogeneity

• Impacts will vary with participant characteristics (including those not observed by the evaluator) and context.

• Participant heterogeneity

• Natural variation & social participation (beyond random individuals as the unit if analysis)

• Contextual heterogeneity
  – “In certain settings anything works, in others everything fails”
  – Local institutional factors in development impact
Examples of external invalidity: Problem of Scaling up from randomized pilots

- The people normally attracted to a program do not have the same characteristics as those randomly assigned + impacts vary with those characteristics
  **“randomization bias”** (Heckman & Smith)

- [With RCTs] we have evaluated a different program to the one that actually gets implemented nationally!
Counterfactual Logic

• **Counterfactual question**: What would have happened without the program

• Need to estimate counterfactual
  – i.e. find a control or comparison group

• **Counterfactual Criteria**
  – Treated & counterfactual groups have identical characteristics on average,
  – Only reason for the difference in outcomes is due to the intervention
The mystique and myth of the counterfactual

• What is the relevant counterfactual?
  Doing nothing is rarely a meaningful & relevant policy option.

• Philosophy of science literature on counterfactuals has fully debunked the meaningful of counterfactuals. Check it out.
Example of summative impact evaluation

Worm medicine and school attendance in LDCs

What constitutes quality evidence of impact?

What’s the value-added of the Counterfactual?
Conditions for which randomized control trials (RCTs) are reasonably appropriate

- A discrete, concrete intervention (treatment) – singular, well-specified
- Implementation can be standardized
- Implementation can be controlled
- Valid and reliable measures exist for the outcome to be tested
- Random assignment is possible and appropriate
- Random assignment is ethical
Examples of Appropriate Use of RCTs

- Drug studies
- Fertilizer and crop yields
- A single health practice:
  -- brushing teeth
  -- exercise regimen
  -- a standardized curriculum
Examples where RCTs are not appropriate in my view

• Complex, multi-dimensional and highly context-specific community interventions

• Ethical constraints (Example: Florida EPA test of effects of pesticides on infants)
What is possible and appropriate:

- Multiple sources of data about each case
- Triangulation of sources
- Modus operandi analysis
- Epidemiological field methods
Example where RCT is not needed in my view

• Face validity is high
• The observed changes are dramatic
• The link between treatment and outcome is direct
Vietnam:

December 15, 2007

Photos courtesy of Atlantic Philanthropies
Greig Craft of the Asia Injury Prevention Foundation:

December 15, 2007
“Nearly 100% of Vietnam’s motorbike users left home wearing a helmet. It was an unbelievable sight with a near instantaneous effect. Major hospitals reported the number of patients admitted for traumatic brain injuries in the two days after the law’s enactment was much lower than on previous weekends. In Ho Chi Minh City alone, serious traffic accident injuries fell by almost 50 percent compared with pre-helmet weekends.”
When Observational Data Suffices

A study in the *British Medical Journal* by Gordon Smith and Jill Pell (2003) found that....

http://bmj.bmjjournals.com/cgi/content/full/327/7429/1459?ck=nck
No randomized control trials of parachute use have been undertaken
Smith and Pell concluded:

“Only two options exist. The first is that we accept that, under exceptional circumstances, common sense might be applied when considering the potential risks and benefits of intervention.”
those who criticize interventions that lack an evidence base will not hesitate to demonstrate their commitment by volunteering for a double blind, randomized, placebo controlled, crossover trial.”
Examples of professional expertise and causal analysis

• Auto mechanic
• Causes of fires
• Forensic science, autopsies
• Airline and auto crashes
• Accident investigations
The method of eliminating alternative possible causes (EAC):
This relies on expert knowledge, e.g., by using the diagnostic schemata of the coroner or the auto mechanic. The approach rests on three premises:
(i) the well-supported principle of macro-determinism;
(ii) well-supported checklists of possible causes of much-studied effects;
(iii) hard-earned knowledge about the distinctive pattern of operation (the modus operandi) of each possible cause.
The EAC approach can be used to establish causation in a particular case or in a family of cases.
(Michael Scriven EvalTalk, Jan. 4, 2004)
Problem of “Premature Experimentation”

In medical research, clinical trials are typically grouped into four phases. All four phases are considered essential to scientific assessment of the effects of treatments. Randomized experiments (and, when necessary, their quasi-experimental alternatives) are considered Phase III trials. (Next 3 slides from Trochim)
Phases of trials

- Phase I trials are essentially small sample exploratory studies that assess the nature of the intervention, how it is administered, potential side effects, patient reactions, etc.
- Phase II trials tend to be correlational studies of safety and effectiveness.
- Phase III trials are typically the randomized experiments designed for high internal validity.
• Phase IV trials explore dose responses, potential unintended effects, generalizability issues and new uses for the treatment.
All four phases are necessary to do "scientifically based" assessment in medicine. *Phase III trials -- randomized experiments - are not in themselves sufficient for scientific assessment.* And they would not typically be allowed even to commence without rigorous Phase I and II assessment, typically taking years to accomplish. The other phases (than Phase III) address issues of critical importance in scientific assessment including construct, conclusion and external validity. Most new treatments don't survive to the Phase III clinical trials. Typically well under half of all interventions are found safe and effective enough to proceed to Phase III research.

There is a very useful set of materials on clinical trials and their phases in cancer research at:

Even with strong RCT results, e.g., from medical or agricultural research, to actually apply findings, *the specific context and situation determine what to do*, that is, individual case analysis is essential.
Contribution Analysis

Preponderance of evidence
Contribution Analysis

Step 1: Set out the attribution problem to be addressed
Step 2: Develop the theory of change and risks to it
Step 3: Gather the existing evidence on the theory of change
Step 4: Assemble and assess the contribution story, and challenges to it
Step 5: Seek out additional evidence
Step 6: Revise and strengthen the contribution story
Ways of strengthening the contribution story

• Identifying other possible explanations and ruling them out.
• Identifying exceptions and seeking to explain them.
• Making predictions and checking them out.
• Use of expert opinion, beneficiaries or research evidence about the plausibility of program theory.
• Developing testable hypotheses not necessarily involving a group which has not received the intervention.
  – For example, if the intervention is working as expected, impacts might be predicted in favourable contexts (implementation environment, characteristics of participants) and not in others.
Rigor Attribute Model:

Measuring Attributes of Rigor in Information Analysis

Daniel J. Zelik, Emily S. Patterson, and David D. Woods

(next 10 slides)

Beyond mechanical methodology to critical thinking

- Beyond method as the focus of evaluation quality and rigor
- Focus on the quality of evaluative thinking
Rigor Attribute Model

• Rigor is not understood by assessing deviations of process, but rather, by assessing the contextual sufficiency of many different aspects of the analytic process.

• **Cognitive Systems Approach:** While professional intelligence analysts can make perceptive assessment about quality of an analysis based on product quality, these perceptions are apt to change with insight into the analytic process.

Michael Quinn Patton
29 April, 2011
Rigor Attribute Model

1. *Information Search* relates to the depth and breadth of the search process used in collecting data. A low-rigor analysis process does not go beyond routine and readily available data sources, whereas a high-rigor process attempts to exhaustively explore all data potentially available in the relevant sample space.
Rigor Attribute Model

2. Hypothesis Exploration -- extent to which multiple hypotheses are considered in explaining data. In a low-rigor process there is minimal weighing of alternatives. A high-rigor process, in contrast, involves broadening of the hypothesis set beyond an initial framing and incorporating multiple perspectives to identify the best, most probable explanations.
Rigor Attribute Model

3. *Information Validation* details the level at which information sources are corroborated and cross-validated. In a low-rigor process little effort is made to use converging evidence to verify source accuracy, while a high-rigor process includes a systematic approach for verifying information and, when possible, ensures the use of sources closest to the areas of interest.
Rigor Attribute Model

4. **Stance Analysis** is the evaluation of data with the goal of identifying the stance or perspective of the source and placing it into a broader context of understanding. At the low-rigor level an analyst may notice a clear bias in a source, while a high-rigor process involves research into source backgrounds with the intent of gaining a more subtle understanding of how their perspective might influence their stance toward analysis-relevant issues.

Michael Quinn Patton
29 April, 2011
Rigor Attribute Model

5. *Sensitivity Analysis* considers the extent to which the analyst considers and understands the assumptions and limitations of their analysis. In a low-rigor process, explanations seem appropriate and valid on a surface level. In a high-rigor process the analyst employs a strategy to consider the strength of explanations if individual supporting sources were to prove invalid.
Rigor Attribute Model

6. **Specialist Collaboration** describes the degree to which an analyst incorporates the perspectives of domain experts into their assessments. In a low-rigor process little effort is made to seek out such expertise, while in a high-rigor process the analyst has talked to, or may be, a leading expert in the key content areas of the analysis.
Rigor Attribute Model

7. *Information Synthesis* refers to how far beyond simply collecting and listing data analysts went in their process. In the low rigor process an analyst simply compiles the relevant information in a unified form, whereas a high-rigor process has extracted and integrated information with a thorough consideration of diverse interpretations of relevant data.
Rigor Attribute Model

8. **Explanation Critique** is a different form of corroboration that captures how many different perspectives were incorporated in examining the primary hypotheses. In a low-rigor process, there is little use of other analysts to give input on explanation quality. In a high-rigor process peers and experts have examined the chain of reasoning and explicitly identified which inferences are stronger and which are weaker. (Zelik, Patterson & Woods, 2007).

High quality evaluative thinking

• Follow the evidence where it leads you
• Bottoms up, contextual design & analysis
• High quality evaluative thinking throughout
• Matching the evaluation to the situation, primary intended uses, primary intended users, priority questions, and available resources
• Following DAC standards to determine evaluation quality – not methodological orthodoxy.
Alternative paradigm: Systems and complexity thinking

Website sample chapter:  
website for the book:  
http://www.guilford.com/cgi-bin/cartscript.cgi?page=pr/patton.htm&dir=research/res_eval&cart_id=824067.29797

See pages 288-293 for a discussion of RCTs and an example of methodological rigidity and the harm the gold standard dogma does.
The Challenge:

Matching the evaluation design to the evaluation’s purpose, resources, and timeline to optimize use.
• Contextually sensitive, contingency-based impact evaluation,

which means…
GOLD STANDARD:

METHODOLOGICAL APPROPRIATENESS

not

methodological orthodoxy or rigidity
References


   Print version: "Why have educational evaluators chosen not to do randomized experiments? Annals of American Academy of Political and Social Science, 589: 114-149


continued…/
4. Web reference:

"Determining Causality in Program Evaluation and Applied Research: Should Experimental Evidence Be the Gold Standard?"

Streaming video of a debate between Dr. Mark Lipsey (Vanderbilt University) and Dr. Michael Scriven (Western Michigan University) available on the Claremont Graduate University website: http://www.cgu.edu/sbos/pdw.html

This debate, held on July 15, 2004 was co-sponsored by the Southern California Evaluation Association following the Claremont Graduate University Professional Development Workshops. Selected excerpts from the debate are also available on the website in a text format.
Utilization-Focused Evaluation

Reference

• Michael Quinn Patton (2011)