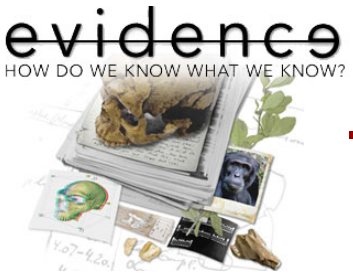
Cornell University



Evidence and Evaluation

William M.K. Trochim

Presentation to the
Australasian Evaluation Society
International Conference
Canberra, Australia

2 September, 2009

Overview – Some Questions About Evidence

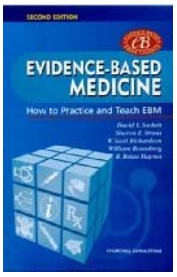
- Where did this evidence movement come from?
- Why is there so much emphasis on it today?
- What is the relationship among evidence-based practice, practice-driven evidence and research-practice integration?
- What constitutes evidence and how do we know it when we see it?
- What is the "unit" of evidence?
- How is evidence stored, retrieved and disseminated?

Overview – Some Questions About Evidence

- How do we determine the quality of evidence?
- What role do methods play in determining quality of evidence?
- How does the move to an evidence focus influence our thinking about evaluation?
- What role can evaluation play in generating or creating evidence and in influencing this movement?
- What role should evidence play in influencing evaluation?

Where did this evidence movement come from?

- Origins in in biomedical research – evidence-based medicine
- Sackett (2000) played a leading role in the development of evidence-based medicine. He describes four reasons for its development:
  - The need for clinicians to have immediate information based on evidence
  - The inadequacy of existing sources
    - Out of date textbooks
    - Frequently wrong experts
    - Ineffective medical education
    - Overwhelming research literature
  - The decline in up-to-date knowledge as clinicians move past medical school days
  - Time pressures of treating patients immediately



Where did this evidence movement come from?

Meta-analysis in medicine

Extension to Public Health

Systematic Reviews

Extension to Social Programs

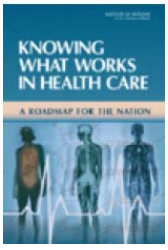
Guidelines

Extension to Education

Comparative Effectiveness Reviews (AHRQ)

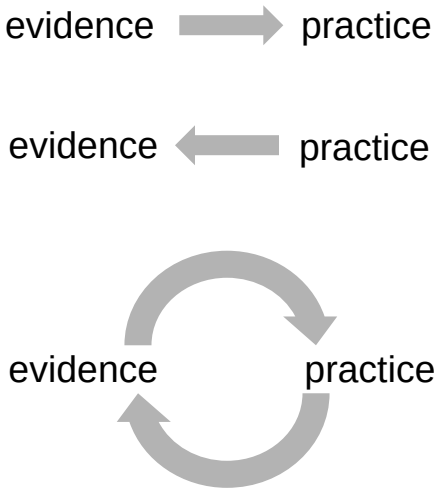
Why is there so much emphasis on it today?

"Before the move toward evidence-based practice, medical textbooks and articles were filled with thousands of statements and care recommendations that were based solely on the belief of the author or at best a consensus of experts."
(IOM, 2008, p123)



- Overwhelming evidence base
- Evidence that evidence was not being used in practice
- Tensions between research and practice models
- The ever-present pressure for accountability
- The need to control quality of treatment

What is the relationship among evidence-based practice, practice-driven evidence and research-practice integration?



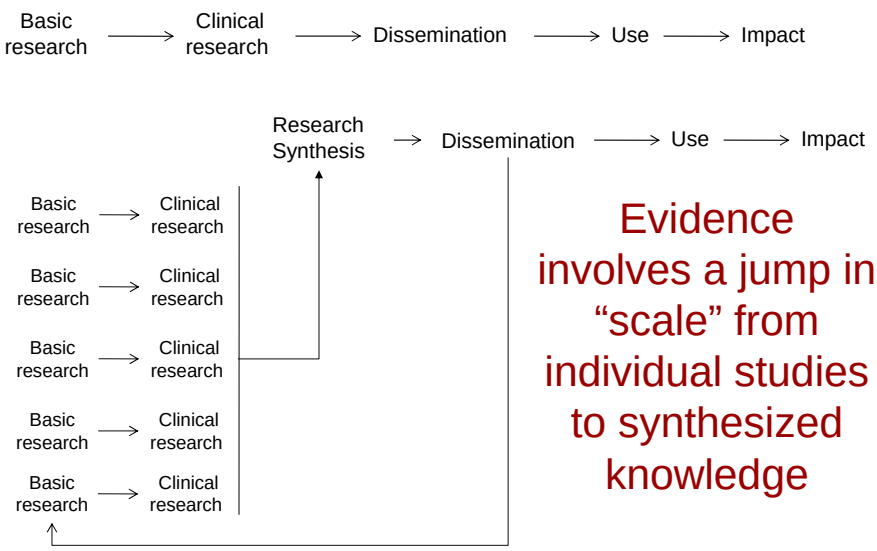
7

What constitutes evidence and how do we know it when we see it?

- Evidence is *synthesized empirical knowledge* that can be understood and used by practitioners
- Evidence differs in quality and most evidence-based systems use an *evidence hierarchy* – to help practitioners judge credibility
- Evidence differs in *strength of recommendation* for practice and most evidence-based systems rate this

8

Two models of research – practice process



Evidence involves a jump in "scale" from individual studies to synthesized knowledge

What is the "unit" of evidence and how is evidence stored, retrieved and disseminated?

"Units" of Evidence

Meta-analysis	Quantitative synthesis of the results of multiple research study results in order to arrive at an estimate of the average effect size across studies.
Systematic review	A synthesis that uses systematic methods to identify, select, assess and summarize findings across similar but separate studies. A systematic review may or may not include a meta-analysis.
Guideline	Practice recommendations developed systematically by a panel of experts who have access to the evidence, an understanding of the practice problem, knowledge of research methods, and time to absorb the information and make considered judgments.
Comparative effectiveness review	A type of systematic review that depicts how the relative benefits and harms of a range of practice options compare in the context of real-world practice. They answer more than the narrow question of whether a single therapy is safe and effective.

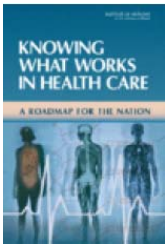
What is the "unit" of evidence and how is evidence stored, retrieved and disseminated?

Storage

11

How do we determine the quality of evidence?

- "Numerous hierarchies and typologies have proliferated – each with its own system of letters, codes and symbols (Schunemann, 2003). …the end result is greater confusion rather than clarification." (IOM, 2008, p. 102)
- "Evidence hierarchies have helped raise awareness that some study designs are less subject to bias than others (Glasziou et al, 2004). Hierarchies, however, consider just the type of research study (e.g., RCTs or prospective observational studies) and not the quality of the individual studies (Poolman et al, 2006). Findings from a poorly conducted trial should not necessarily trump evidence from a nonrandomized study." (IOM, 2008, p.102)



12

How do we determine the quality of evidence?

Quality of evidence	Numbers	Letters	Circles	Stars	Multiple
High	1	A	●	☆☆☆☆	⊕⊕⊕⊕
Moderate	2	B	◕	☆☆☆	⊕⊕⊕◆
Low	3	C	◒	☆☆	⊕⊕◆◆
Very low	4	D	◒	☆	⊕◆◆◆

Action based on balance between benefit and harm	Numbers	Letters	Traffic lights	Thumbs	Arrows
Do	1	A			↑↑
Probably do	2	B			↑?
Probably don't do	3	C			↓?
Don't do	4	D			↓↓

Fig. 1: Examples of possible symbols for representing quality of evidence and the balance between benefits and harm in health care recommendations. See Tables 1 and 2 on the *CMAJ* Web site for selection criteria (see www.cmaj.ca).

From Schunemann, H.D., Best, G. Vist, G. and A.D. Oxman. 2003. Letters, numbers, symbols and words: How to communicate grades of evidence and recommendations. Canadian Medical Association Journal, 169(7): 677-680.

What role do methods play in determining quality of evidence?

AMERICAN EVALUATION ASSOCIATION

The Coalition for Evidence-Based Policy

American Educational Research Association

OFFICE OF MANAGEMENT AND BUDGET

COMMUNITY Preventive Services — The Community Guide — What works to promote health

National Guideline Clearinghouse — AHRQ

ED.gov U. S. Department of Education *Promoting educational excellence for all Americans*

INSTITUTE OF MEDICINE OF THE NATIONAL ACADEMIES

THE CAMPBELL COLLABORATION What helps? What harms? Based on what evidence?

The Cochrane Collaboration

Methodological Pluralist Pragmatists

Views on Evidence Hierarchies

Experiment as "Gold Standard"

RCT "Gold Standard" View of Evidence Hierarchy

Prepared for the:
Institute of Education Sciences
Grover J. Whitehurst, Director
by the

COALITION FOR EVIDENCE-BASED POLICY

A Project Sponsored by

THE COUNCIL FOR EXCELLENCE IN GOVERNMENT

"Well-designed and implemented randomized controlled trials are considered the "*gold standard*" for evaluating an intervention's effectiveness, in fields such as *medicine*, welfare and employment policy, and psychology."

(U.S. D.O.E., 2003), p. 1. (emphasis added)

RCT "Gold Standard" View of Evidence Hierarchy

How to evaluate whether an educational intervention is supported by rigorous evidence: An overview

Step 1. Is the intervention backed by "strong" evidence of effectiveness?

Quality of studies needed to establish "strong" evidence:

- Randomized controlled trials (defined on page 1) that are well-designed and implemented (see pages 5-9).

+

Quantity of evidence needed:

Trials showing effectiveness in —

- Two or more typical school settings,
- Including a setting similar to that of your schools/ classrooms.

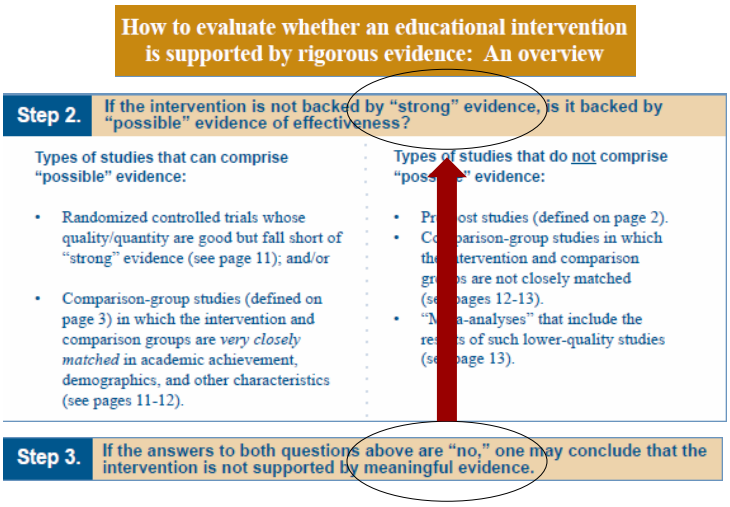
(see page 10)

=

"Strong" Evidence

U.S. Department of Education, (2003). Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide. Washington, D.C., p. v.

RCT "Gold Standard" View of Evidence Hierarchy



U.S. Department of Education, (2003). Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide. Washington, D.C., p. v.

RCT "Gold Standard" View of Evidence Hierarchy

the guidance points to the randomized controlled trial (RCT) as an example of the best type of evaluation to demonstrate actual program impact.

r agencies to
ate the
o the
e of
dded) Yet,
employed
en will need
to consider alternative evaluation methodologies. " (p. 1)

Well-de
(empha

RCTs are considered the *gold standard*

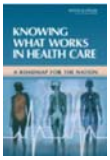
many diverse fields of human inquiry, such as medicine, welfare and employment, psychology, and education. (p. 4)."

OFFICE OF MANAGEMENT AND BUDGET

OMB PART Evaluation Guidance written by
The Coalition for Evidence-Based Policy

Pluralist Pragmatist View of Evidence Hierarchies

"RCTs can answer questions about the efficacy of screening, preventive, and therapeutic interventions… Observational studies are generally the most appropriate for answering questions related to prognosis, diagnostic accuracy, incidence, prevalence, and etiology (Chou and Helfand, 2005; Tatsioni et al., 2005). Cohort studies and case series are useful for examining long-term outcomes because RCTs may not monitor patients beyond the primary outcome of interest or for rare outcomes because they generally have small numbers of participants. Case series are often used, for example, to identify the potential long-term harms of new types of radiotherapy. Similarly, the best evidence on potential harms related to oral contraceptive use (e.g., an increased risk of thromboembolism) may be from nonrandomized cohort studies or casecontrol studies (Glasziou et al., 2004)." (IOM, 2008, p. 91)



Pluralist Pragmatist View of Evidence Hierarchies

AMERICAN EVALUATION ASSOCIATION

- Randomized control group trials (RCTs) are not the only studies capable of generating understandings of causality.
- RCTs are not always best for determining causality and can be misleading.
- RCTs should sometimes be ruled out for reasons of ethics.
- 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.
- Actual practice and many published examples demonstrate that alternative and mixed methods are rigorous and scientific.

American Evaluation Association Response To U. S. Department of Education Notice of proposed priority, *Federal Register* RIN 1890-ZA00, November 4, 2003 "Scientifically Based Evaluation Methods." http://www.eval.org/doestatement.htm

Pluralist Pragmatist View of Evidence Hierarchies

American Educational Research Association

While we appreciate the value of experimental designs as an evaluation method, we believe that a judgment of "best," as specified in the proposed language, does not adequately account for other methods of evaluation that might be as or more appropriate depending on the specific education program.

American Educational Research Association, (2003). Resolution on the Essential Elements of Scientifically-based Research. http://www.eval.org/doeaera.htm

Pluralist Pragmatist View of Evidence Hierarchies

AMERICAN EVALUATION ASSOCIATION

- RCTs are weak with respect to the goal of program improvement.
- RCTs do not by themselves explicitly address construct validity.
- RCTs are weak with respect to generalizability or external validity.
- Addressing RCTs' validity problems often entails investment in companion program evaluations that have methodological designs other than RCTs.
- The importance of mixed methods.
- The need to address feasibility and resource issues realistically.
- The need to address equity and human subjects concerns realistically.

Evaluation Policy Task Force (2008). Comments on What Constitutes Strong Evidence of a Program's Effectiveness? American Evaluation Association. http://www.eval.org/aea08.omb.guidance.responseF.pdf

Reformulating the Evidence Hierarchy Debate

- This struggle has been difficult
- It doesn't need to be this hard
- The danger right now is *overadvocacy* of RCTs as the basis of evidence
- The problem is not *whether* to use RCTs, it's *when* they should be used in the life of a program
- Or in other words a major problem is…

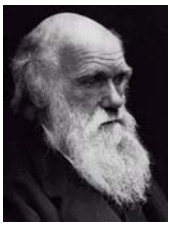
Premature Experimentation

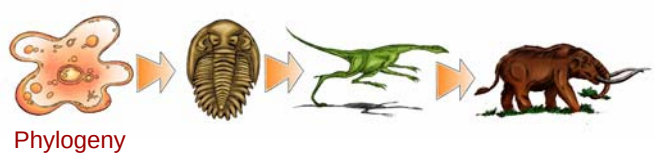
How does the move to an evidence focus influence our thinking about evaluation?

- Suggests that we need a *reframing* of evidence in evaluation
- Premise: a fundamental problem with the overadvocacy of RCTs is that the proponents have only *selectively adopted* what makes the biomedical research model so effective
  - They have adopted the emphasis on RCTs as a scientifically rigorous way to assess program effectiveness
- But they have not adopted the entire supporting system that has made that possible
  - The supporting system of evidence norms and phased trials that provide a necessary foundation for RCTs
- There is a scientific rigorous basis for adopting this broader system that preserves a central role for RCTs but *puts them in their appropriate place* in the larger evidence-generating endeavor…

A Potential Reframing

An evolutionary systems thinking approach





Phylogeny

Programs evolve just like species.
Blind variation and selective retention of those with "fitness to environment.

Symbiosis and Co-Evolution

Programs and their evaluation need to be linked appropriately. The right evaluation method for the right stage of development

Ontogeny

Programs change through a life-course. They grow through different stages.

Adopting a Phased Approach to Evidence

AMERICAN EVALUATION ASSOCIATION

"Clinical trials involve four basic phases.

- **Phase I** trials are exploratory small sample studies that examine tolerance of the treatment and potential side effects.
- **Phase II** trials typically demonstrate whether the program is capable of achieving effects (efficacy) and examines its correlates.
- **Phase III** trials are typically controlled effectiveness studies.
- **Phase IV** trials typically examine generalizability and the fidelity of transfer of controlled interventions to field settings.

Randomized designs are usually not used until late in phase II or more likely in phase III studies when effectiveness is the focus. The FDA reports that the vast majority of interventions (approximately 70-75%) that begin clinical trials do not survive to controlled Phase III randomized trials because they do not meet the basic conditions that warrant subsequent efforts."

We need clear criteria that must be met before RCTs are mounted

Evaluation Policy Task Force (2008). Comments on What Constitutes Strong Evidence of a Program's Effectiveness? American Evaluation Association.
http://www.eval.org/aea08.omb.guidance.responseF.pdf

Some Potential Requirements for an RCT

- the program is well defined and has an articulated program model
- the program has been implemented consistently and with high fidelity
- there are high-quality (e.g., valid and reliable) outcome measures
- the program as implemented is capable of producing change
- there is sufficient statistical power to accomplish the study with high quality
- the participants can be kept unaware of the group (intervention or control) to which they have been assigned
- the random assignment can be implemented and maintained
- ethical and human subject protections have been approved and are in place

Evaluation Policy Task Force (2008). Comments on What Constitutes Strong Evidence of a Program's Effectiveness? American Evaluation Association. http://www.eval.org/aea08.omb.guidance.responseF.pdf

Need for an Evidence Generation ***Culture***

- Practitioners (program managers, deliverers and advocates) need to know that
  - Only a few programs *should* survive in the long run
  - It is our job to consciously evolve programs and variations (artificial selection as opposed to natural selection)
  - We need to consciously move programs through developmental stages (too much arrested development!)
  - Anticipate and plan for next stages
  - We succeed – we are doing our jobs – when we engage in this process, even if our program "fails"
- Institutions and organizations need to provide expectations and incentives to support these norms
- Decision makers and funders need to understand the game we're in

What role does evaluation play in generating or creating evidence and in influencing this movement?

- Intervention trials to generate evidence are *evaluation research*
- Evaluation for understanding the *process* of research-practice integration and translation (moving from research → synthesis → use)
- Evaluation can help assess *dissemination approaches*
- Evaluation as a profession can add a measured voice to the debates about evidence, especially with respect to methods. We can and should help shape *policies* about evidence generation.

What role should evidence play in influencing evaluation?

- Practice what we preach!
- Develop an *evaluation* evidence-base
- Emulate the full evolutionary model to study what works in evaluation
- Example – NSF project to develop and test a systems evaluation approach to planning, implementing and utilizing evaluations
  - Purpose – develop a systems thinking approach to evaluation
  - Program – partnerships, protocol, cyberinfrastructure
  - Phase I – development and formative/process evaluation (2006 – 2008)
  - Phase II – efficacy studies – is the approach correlated with key outcomes (2008 – 2013)
  - Phase III – effectiveness studies – a national or even international cross-site implementation (2013 – 2018)
  - Phase IV – dissemination studies

Conclusions

- Evaluation needs to understand the evidence movement and actively shape it based upon our experience
- Evaluation should embrace an evolutionary, ecological, systems-oriented perspective
  - It is based on a solid scientific foundation (there's almost no stronger theory out there than evolution!)
  - It provides a framework for viewing an evaluation as contributing to a broader evolution of knowledge
  - It provides a rationale for why multiple and different methods are needed throughout the life of an intervention
  - It provides an appropriate role for RCTs
- Evaluation needs to develop our own evidence base and model these values

Conclusions

- There are many other questions we need to address…
  - How can an emphasis on evidence be misused?
  - When have we gathered enough evidence?
  - How do we determine the balance of what types of evidence we need?
  - How do we allocate resources for evidence generation and use?
  - How do we train the next generation of researchers, evaluators, decision-makers and practitioners?
  - How do we educate the public about evidence?
- We have a lot of work to do!