

# Evidence-Based Practice: Where do we really Stand?

Thomas D. Cook

Northwestern University

Orlando, October, 2007

# Some Recent History

- Tony Blair's Electoral Campaign
- George W. Bush's First Electoral Campaign
- Federal Agencies--e.g., FDA, IES, Justice and their lists of effective practices
- Professional Organizations - lists of evidence-based effective practices from which selections are to be made for improving local practice

# What is Evidence-based?

- My neighbor tells me that...is also evidence
- Best scientific evidence for a given type of knowledge - e.g. only experiments for cause
- Adequate and permissible for given type-which quasi-experiments should be added
- Organizationally certified as acceptable, but organizations vary in standards, even battling each other--e.g. school violence domain

# Some Realities

- Battles over aspects of standards of evidence are commonplace--in education today, for example
- Standards setting is a technical and political process, subject to historical change
- Ancillary standards always need to be applied -- e.g., measurement specifics, “exact” replication, transfer to other pops., settings, etc.
- Best and better standards for inferring generalization and cause relatively well settled
- Debates about how far down preference hierarchy to go; not about best or even second-best methods

# Four major Types of Evidence re:

- Entities - have I validly measured the constructs X and Y, and changes in them? Indicators of need, of causal levers and of outcomes of interest
- Co-variation--how certain am I that X and Y truly co-vary?
- Causation--how certain am I that X causes change in Y and that the size of the relationship is Z ?
- Generalization--does X generally causes change in Y; can I identify the conditions under which X causes change in Y; will X “work” in my context?

# Purposes of this Talk

- Analyze our capacity to draw strong conclusions about these four matters in education--how good are our relevant theories about each of the 4 tasks?
- Analyze our actual research practices to see how well we do in these four areas-to identify where we are doing well and where we need to improve
- Paint with a broad brush

# Conclusions

- We are generally doing well in education with measurement, assessments of covariation and with causation, though few areas we can improve on
- But theory and practice of causal generalization a real weakness
- In particular, efficacy study results are often over-generalized. Our struggle is; How to make evidence-based mean effectiveness-based rather than efficacy-based?

# Is it X or Y?: Construct Validity

- Need a substantive theory of X, specifying its “nature”, antecedents and consequences--we validate a theory and not a measure
- Need multiple measures for convergent validity, hopefully measured in different ways
- Need cognate measures for discriminant validity--anxiety is not depression
- Need validation against a criterion not subject to the same suspected biases. NCLB: respective role of state tests and NAEP.

# Practice Limitation: 1

- Measurement is context-dependent
- NCLB again--mismatch of state and NAEP trend results--why? How high the stakes for each?
- So which is valid if each is imperfect?
- Golden rule a: All indicators are corruptible.
- Golden rule b: Multi-measurement needed under conditions chosen to vary major sources of anticipated bias--every source of bias is impossible

# Practice Limitation: 2

- Measuring gaps and changes in gaps
- The Black-White ach gap revisited.
- As raw scores; in standard deviation form; as logs--what's the appropriate metric, especially over long time periods?
- Changing the cutoff point for sanctions

# Covariation: are X and Y related?

- Theory of stat sig testing well worked out
- Practice shortfall #1--capitalizing on chance with multiple tests, but improving
- Shortfall #2-- failure to control for clustering, but improving thanks to HLM
- Shortfall #3. Sample sizes and power- progress
- Shortfall #4--size of relationships by developer vs independent researcher, only recent sensitivity
- But our standards for inferring covariation are arbitrary anyway--Fisher and .05 level.

# Causation: Does X cause Y?

- Growing institutionalization of RCT--preferred for funding, publishing, training
- Upsurge of interest in cluster-based RCTs
- New interest in violated contamination assumption
- Huge interest in RDD in theory and practice
- Growing interest in abbreviated ITS as design
- Also in propensity scores for analysis
- Hierarchy at top getting accepted in general--RCT, RDD, matching--but not quite everywhere yet
- Growing acknowledgment of programs of research and meta-analysis --WWC

# Problems are:

- No consensus on what is good enough causal study
- None of these methods are necessary for causal knowledge--as in history
- RCT not a routine gold standard in practice
- Disaffected practitioners of old methods feel their identify is denied, but they are not well organized as a source of intellectual or political resistance
- Sad reality: Better studies needed because we are in the game of detecting modest effects

# Causal Dilemmas now most Evident in Educational Research

- The epistemological shift 30 years ago
- Made reputations since using other methods
- Enough limitations to RCT to make it reasonable to resist gold standard rhetoric
- BUT frequency of RCTs in school prevention work
- Frequency of RCTs in early education work with achievement as the dependent variable
- Role of funding agencies, and other institutions

# Within-Study Comparisons

- Nail in the coffin to the resisters
- RCTs vs RDD studies--3 of them
- RCTs vs a priori group matching studies--3
- RCTs versus workhorse design with or without pretest on same scale as outcome--disappointing unless selection process really well known and exquisitely measured

# How general is X-Y causal Link?

No general theory of causal generalization

- Two problems--of representation and of extrapolation
- What do most RCTs represent--one version of intervention, one version of outcome, one setting, one time, and one population
- Efficacy trials--contrast with effectiveness
- Developer presence; atypical fidelity?

# Extrapolation

- Identify the set of conditions under which will work either thru meta-analysis or identifying causal mediating mechanism
- Meta-analysis of specific programs rare
- Definitively identifying necessary and sufficient conditions (crucial mediators) also very difficult.
- We have a practical problem of extrapolation from efficacy trials to conditions of application of general interest
- And to my local interests

# What we need to worry about

- Developer role
- Program specifics that limit implementation -- case of Success For All; SFA and gaps
- Fidelity vs adaptation dynamics
- Mismatch of outcome homogeneity in research and often broader in applications--early childhood
- To date we have evidence-based efficacy policy masquerading as effectiveness policy
- Example of class size reductions in Calif.

# New Frontier: Evidence-based Effectiveness Policy

- We are good at doing efficacy-based research, but is this the evidence we need?
- How good are we now?

# Result: Lists of Recommended Programs for Local Choice

- Growth of prescribing program selection from a limited list--smorgorsbord model to facilitate local tailoring
- Example of violence prevention in schools
- Multiple lists with some similar and some unique criteria
- Some programs on all lists; some on but one
- Agency pressure to include programs they funded

# Federal Program Lists

- Center for Mental Health Services (2000)
- National Registry (NREPP) (2002)
- Office of Safe & Drug Free Schools (2001)
- Blueprints for Violence Prevention (2007)
- National Institute of Drug Abuse (2003)
- Surgeon General Report (2001)
- (What Works Clearinghouse)

# Similarities and Differences Across Federal Lists

- All Include Violence, Drug, Delinquency and Antisocial Behavior Prevention Programs: Not Restricted to Agency Focus
- All but NIDA Include Multiple Categories of Effectiveness
- Scientific Standard for Certifying Programs Varies Widely

# Scientific Standards

- Ctr. for MH Services: *Effective and Promising*
  - RCT or Quasi-Experimental Design
  - Written Implementation Manual
  - Reduced Symptomology or Risk for Disorder
- NREPP\*: *Model, Promising and Effective*
  - Rationale
  - Fidelity
  - Good Internal Validity/Design
  - Replication
  - Utility
  - Dissemination Capability
  - Cultural and Age Appropriateness

\* As used in 2002. Currently undergoing a major revision

# Scientific Standards

## Cont'd

- NIDA: *Effective*
  - No Standard Specified
- OSDFS: *Exemplary and Promising*
  - Clear Rationale
  - Change Goals Appropriate for Population and Setting
  - Implementation Process Effectively Engages Intended Population
  - Sound Design and Internal Validity
- OJJDP-Title V: *Exemplary, Effective & Promising*
  - RCT (for Exemplary); Quasi-Exp. for Effective
  - Significant Association between program participation and reduction in criminal behavior for Promising

# Scientific Standards

## Cont'd

- Blueprints: *Model and Promising*
  - RCT or Quasi-Experimental Design
  - Good Internal Validity
  - Replication on one or more sites
  - Sustained Effects for at least one year post-intervention

# Consensus Across Lists

- Of 80 programs on any list
- Only One Program (LST) Appeared on 5 of the 6 Lists as a Model/Effective/Exemplary Program\*
- Three Programs were on 4 Lists: ATLAS, TND, Project Alert
- 12 Programs on half of the Lists: BBBS, Caring School Community, Early Risers for Success, Good Behavior Game, Incredible Years, MTFC, MST, NFP, Project Northland, TNT, PATHS, Strengthening Families

\* Top category on each list.

# Federal Working Group Standard for Certifying Programs as Effective\*

- Experimental Design/RCT
- Effect sustained for at least 1 year post-intervention
- At least 1 independent replication with RCT
- RCT's adequately address threats to internal validity
- No known health-compromising side effects

\*Adapted from *Hierarchical Classification Framework for Program Effectiveness*, Working Group for the Federal Collaboration on What Works, 2004.

# Hierarchical Program Classification\*

- I. *Model*: Meets all standards
- II. *Effective*: RCT replication(s) not independent
- III. *Promising*: Quasi-Experiment or RCT, no RCT replication
- IV. *Inconclusive*: Contradictory findings or non-sustainable effects
- V. *Ineffective*: Meets all standards but with no statistically significant effects
- VI. *Harmful*: Meets all standards but with negative main effects or serious side effects
- VII *Insufficient Evidence*: All others

\*Adapted from *Hierarchical Classification Framework for Program Effectiveness*, Working Group for the Federal Collaboration on What Works, 2004.

# What has happened to this great inter-agency collaboration?

- Agreed on criteria and standards
- Not one agency has implemented yet
- Why? Pressure of other things, and not so high; internal divisions within agency; fear of looking bad cos few meet; fear of conflict with already published lists

# Practice Dilemmas in Selecting from the Lists

- How do I know what best suits my project profile and how this will fit with what I already have?
- How do I navigate between fidelity to the program and adaptations to it to suit my circumstances?
- How do I navigate between what is on the list and what I prefer that is not on the list?
- Where's the booze?
- When can I retire?
- Can Tom Cook help?

# Short Answer: No

- These dilemmas are now salient again because of evidence-based rhetoric.
- We never solved them before in diffusion of innovations literature.
- They constitute the new frontier for research on evidence-based policy.
- But they could only become the frontier once we learned to measure better, to do analyses of association better, and to do causal studies better

# Longer Term Answer: Maybe Scholars can help--maybe!

We need new theories of causal generalization

We need to do qualitative studies of  
implementation of efficacy and especially  
effectiveness studies at scale

We need more attention, not on whether X causes  
Y, but on the conditions under which X causes Y.

Good news, though, is that our problems are the  
products of our progress in moving towards  
effectiveness-based practices.