Defining "Evidence-Based": Developing a Standard for Judging the Quality of Evidence for Program Effectiveness and Utility

Delbert S. Elliott
Center for the Study and Prevention of Violence
University of Colorado, Boulder

Evidence: Something that furnishes or tends to furnish proof (Webster)

Nature of Evidence varies with Question Asked

- Is the intervention grounded in theory, practical and logical?
- How difficult is it to implement the intervention as designed?
- Does the program have the intended effect on the targeted outcome?
- What is the magnitude of change on the targeted outcome?
- Can the IV be replicated with fidelity; can it be integrated into existing service systems with fidelity?
- Is the IV valued sufficiently to be given a high social, economic and political priority for funding?

Program Evaluation is a process with each stage contributing to the overall evidence for a program's effectiveness, utility and acceptance by professionals

Each of the following types of evidence may be involved in the cumulative evidence for an IV

- Systematic reviews of findings
- Systematic reviews of records/documents
- Case studies, Qualitative Methods
- Surveys
- Non-experimental studies, e.g. pre-post studies of risk factors & outcomes
- Experimental/Quasi-Experimental studies

I. Evaluation of the Program's Theoretical Grounding

- Linking the targeted outcome to specific risk and protective factors
  - Review level of empirical support for this link
  - Are these factors relatively strong or weak causal variables in the theory?
  - Review of findings informs expected variability in dependent variable- effects sample size
  - Informs variables to be included as controls – statistical power needed
  - Informs causal gap – timing of measurement
  - Informs expected effect size –sample size

Evaluation of Theoretical Grounding – Cont'd

- Linking the intervention services/action to change in the targeted risk/protection factors
  - Reviews to determine if these risk/protection factors are manipulable
  - What evidence is there that these services will be effective in changing them?
  - Change time gap- how long to effect change ?– timing of measurement

II. Process Evaluation: Is the program delivering the intervention as designed?

- Is there an effective data collection, storage and retrieval system in place?
- Are all staff adequately trained to deliver the intervention?
- To what extent is the appropriate IV being delivered to the intended population, with the intended dosage, for the intended duration, with high quality? (fidelity)
- Are reliable and valid pre- and post-intervention assessments of client risk/protection and behavioral outcomes being collected and analyzed? (Pre-Post analysis to determine if the there is any change in risk/protection conditions and/or behavior)

III. Outcome Evaluation: Does the program work? Is it effective?

- Implies a standard for judging the quality and generalizability of the evidence
- There are multiple strategies for estimating effectiveness
- There is little consensus within the research community regarding the appropriate standard for certifying a program as “evidence-based”

The Blueprints Strategy

- A systematic review of individual program evaluations to identify violence, drug abuse and delinquency prevention programs that meet a high scientific standard of effectiveness
- Individual programs meeting this standard are certified as Model or Promising evidence-based programs
- Only Model programs are considered eligible for widespread dissemination

Blueprint Systematic Review

- Ideally: A Meta Analysis of multiple RCT's of a given program. Provides best estimates of expected effect-size and generalizability.

- In Practice: A review assessing the quality of each study (similar to TTIS* criteria), the consistency of findings across studies, effect sizes and external validity.

* Brown et al., 2000. Threats to Trial Integrity Score.

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 not independent.
- III. *Promising*: Q-E or RCT, no 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. www.ncjrs.gov/pdffiles1/nij/220889.pdf

Outcome Evaluation Components

- Designs: 1)RCT's; 2)Strong QE, e.g., interrupted time series, regression discontinuity;3)Minimum:QE with control group and strong internal validity
- Samples: 1) Random samples; 2) Purposive modal samples; 3) Purposive heterogeneous samples;4) theoretical directed sample
- Special Analyses that strengthen generalizability: Causal modeling and mediating effects
- Confirmatory rather than exploratory methods generally

Threats to RCT and QED internal and external validity *

- Selection bias
- Statistical power
- Assignment to condition
- Participation after assignment
- Diffusion/Receiving another intervention
- Implementation of intervention (fidelity)
- Inadequate measurement
- Clustering effects
- No mediating effects analysis
- Effect decay
- Attrition and tracking N's
- Improper analyses, e.g., wrong unit of analysis

*adapted from Brown et al., 2000, Threats to Trial Integrity Score.

The Critical Issue: Rejecting Plausible Alternative Hypotheses

- In some instances this may not be difficult, Pre-post studies maybe sufficient (Campbell, 1991)
- Most plausible alternative Hypothesis that invalidates QED's and Non-Experimental Designs- Confounding of selection and treatment

Case Study Evidence in Evaluation*

- Rich detail of context
- Allows important program variables/processes to emerge
- Primarily used in discovery role in evaluation
- Provides local credibility and *perceived* validity
- Empowers local stakeholders; disempowers more distant stakeholders
- Poor generalizability
- Requires confirmation from other observers
- Difficult to support abstractions
- Difficult to aggregate multiple case studies
- Weak evidence for validating hypotheses

* Shadish, Cook and Leviton, 1991

Defining “Evidence-Based”

- Programs classified as Model, Effective, or Promising on Federal Hierarchy
- Consistently positive effects from Meta Analyses
- Only Model programs should ever be taken to scale

Model and Effective Programs
Federal Working Group Standard*

- Model Programs
  - FFT, Incredible Years, MST, LST
- Effective Programs
  - BBBS, Midwestern Prevention Project, MTFC, NFP, TND, PATHS

*www.ncjrs.gov/pdffiles1/nij/220889.pdf

Promising Programs
Federal Working Group Standard

- Bullying Prevention, Guiding Good Choices, Raising Healthy Children
- CASA START, Strong African American Families Program
- Perry Preschool, I Can Problem Solve, Linking Families and Teachers
- Project Northland, Preventive Treatment Program
- Communities that Care, ATLAS, Strengthening Families (10-14)
- Triple P (Population level), Good Behavior Game
- Behavioral Monitoring and Reinforcement Program
- Brief Strategic Family Therapy, FAST TRACK
- Preventive Treatment Program

Federal Lists of Evidence-Based Programs: AS Behavior

- Blueprints (OJJDP): Model or Promising (100%)
- NIDA: Effective (60%)
- OJJDP Model Program Guide: Exemplary (52%)
- Office of Safe and Drug Free Schools (DOE): Exemplary (55%)
- Surgeon General (DHHS): Model or Promising (100%)

Best Alternative Strategy: Generic Program Meta-Analysis

- Good estimates of expected effect size for a given type of program
- Good estimates of generalizability
- Identifies general program characteristics associated with stronger effects
- Best practice guidelines for local program developers/implementers

Effective Strategies: Meta Analyses: AS Behavior

- Individual-Level Interventions
  - Self Control/Social Competency*
  - Individual counseling**
  - Behavioral Modeling/Modification
  - Multiple Services
  - Restitution with Probation/Parole
  - Wilderness/Adventure
  - Methadone Maintenance

*Only with cognitive-behavioral methods (Wilson et al., 2001)
**Only with non-institutionalized juvenile offenders (Lipsey and Wilson, 1998)

Effective Strategies: Meta Analyses, Cont'd

- Contextual (family, school and community)
  - School & Discipline Management
  - Normative Climate Change
  - Classroom/Instructional Management
  - Reorganization of Grades, Classes
  - Teaching Family Model
  - Community Residential*

* Effective only with institutionalized juvenile offenders

Meta-Analyses of Individual Model Blueprint Programs

- MST – 4 studies. 3 provide positive effects; 1 no significant marginal effects
- NFP – 1 study. Withdrawn due to major methodological problems

IV. Effect Size: the magnitude of change on the outcome

- Percentage change
- Odds Ratios
- Percentile change
- Standard deviations
- Effect size (Cohen): Recommended as the standard for BP Programs- for high and average fidelity; absolute and marginal effects

V. How valuable, important is the intervention in the real world of competing priorities for funding?

- Cost effectiveness-converts program input into monetary units; leaves effects in original metric
- Cost-Benefit Ratios – converts both inputs and effects into monetary units; calculates the ratio of benefits to costs. Recommended Standard for BP Programs

"…both benefit-cost analyses and meta analyses have proven quite appealing in public policy: they lead to simple , quantified results of general application, can be readily remembered, and are not hindered by multiple caveats". Shadish et al., 1991

The Ideal Evidence-Based Program*

- Addresses major risk/protection factors that are manipulatable with substantively significant effect sizes
- Relatively easy to implement with fidelity
- Causal and change rationales and services/treatments are consistent with the values of professionals who will use it
- Keyed to easily identified problems
- Inexpensive or positive cost-benefit ratios
- Can influence many lives or have life-saving types of effects on some lives

- *Adapted from Shadish, Cook and Leviton, 1991:445.

Thank You

Center for the Study and Prevention of Violence

colorado.edu/cspv/blueprints