

# **Application of External Validity Criteria for Translation Research**

**Lisa M. Klesges  
Russell E. Glasgow  
Paul Estabrooks  
Lawrence W. Green**

*2<sup>nd</sup> Annual NIH Conference on the Science of  
Dissemination and Implementation  
Bethesda, Maryland  
January 28, 2009*

# Presenters

- **Russell E. Glasgow, Ph.D.**  
Senior Scientist and Director, Institute for Health Research, Kaiser Permanente Colorado, Penrose, CO
- **Lisa M. Klesges, PhD**  
Professor and Director, School of Public Health, University of Memphis, Memphis, TN
- **Paul Estabrooks, Ph.D.**  
Associate Professor, Center for Translational Obesity Research, Virginia Tech Riverside, Roanoke, VA



# Background

- **Field would benefit from ability to weigh evidence on external validity dimensions in addition to internal validity**
- **Can improve systematic reviews of the literature for program planning and policy decisions by including time, place and population characteristics**
- **Methods to calculate relative merits of programs would enhance comparisons and decisions about “real world” impact**

# Purpose of the Panel

- Establish why we need to be concerned about external validity (EV)
- Identify the salient EV elements to best support future translation of interventions
- Application of EV criteria in childhood obesity literature to highlight need for expanded EV reporting
- Demonstrate summary metric methods to compare programs on EV criteria for decision making

# **Presentations**

- **Importance of and Criteria for Evaluating External Validity (Glasgow)**
- **Evaluating External Validity Reporting in Childhood Obesity Studies (Klesges)**
- **Metrics for Comparing External Validity Dimensions of Interventions (Estabrooks)**
- **Discussion  
(All Participants)**

# IMPORTANCE OF AND CRITERIA FOR EVALUATING EXTERNAL VALIDITY

Russell E. Glasgow, PhD  
Lawrence W. Green, DrPH



# OVERVIEW

- **Rationale for and Importance of External Validity (EV) for Translation**
- **Proposed EV Reporting Criteria**
- **Editors' Meeting on EV Criteria**

# CURRENT SITUATION

- **CONSORT widely used for reporting trials (only 2 of 23 items address EV for non-pharmacologic trials)\***
- **Methodological quality ratings used in reviews almost solely internal validity focused**
- **No such widely used criteria for EV, which are critical for translation and dissemination**

\*1 of 22 criteria for pharmacologic RCTs

# RELATED EFFORTS

- **TREND** criteria for non-randomized trials
- **SQUIRE** quality improvement reporting guidelines include number of items on contextual issues

Des Jarlais DC, et al. *Am J Pub Health* 2004;94:361-366

Davidoff F, Batalden P. *Qual Saf Health Care* 2005;14(5):319-325

# DEFINITION AND DIMENSIONS OF EV

*External Validity* – “Inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments and outcomes.”  
(Shadish et al, 2002)

*External Validity* – “To what populations, settings, treatment variables and measurement variables can this effect be generalized?” (Campbell & Stanley, 1963)

Shadish WR, Cook TD, Campbell DT. 2002. Experimental and quasi-experimental design...Boston: Houghton Mifflin

Campbell DT, Stanley JC. Experimental and quasi-experimental designs for Research. Chicago, IL: Rand McNally. 1966.

## Perspective and Context

**“To different degrees, all causal relationships are context dependent, so the generalization of experimental effects is always at issue.”  
(Shadish et al, 2002)**



**“Everything is Contextual”  
(We made this up)**

# USES OF EXTERNAL VALIDITY INFORMATION

## Local Practitioners:

To determine if a program or study is relevant to their particular setting (patients, resources, staff, measures, etc.)

## Decision and Policy Makers:

To determine the range of conditions and settings across which a given program/policy/product will apply (generalizability)

***“Lack of consideration of external validity is the most frequent criticism by clinicians of RCTs, systematic reviews, and guidelines.”***

**Rothwell PM, *Lancet* 2005(365):82-93**

***“Where did the field get the idea that evidence of an intervention’s efficacy from carefully controlled trials could be generalized as THE best practice for widely varied populations and settings?”***

**L.W. Green**

**Green LW. From research to "best practices" in other settings and populations  
*Am J Health Behav* 2001; 25:165-78**

# DESIRED CHARACTERISTICS OF EV REPORTING CRITERIA

- **Address key conceptual dimensions of EV**
- **Address issues of concern to practitioners and policy makers**
- **Able to be reliably coded**
- **Feasible to report**

Criteria on next slides adapted from Green & Glasgow, *Evaluation and the Health Professions*, 2006;29(1):126-153

# EV REPORTING CRITERIA

## 1. Settings and Populations

- A. **Target Audience:** Are the intended “end users” identified for: 1) adoption (at the setting level, such as worksites, medical offices, etc.) and 2) application (at the individual level)?
- B. **Inclusion and Exclusion Criteria:** Are both 1) inclusion criteria and 2) exclusion criteria (e.g., run-in period, language, comorbid conditions, other treatments, demographic characteristics) reported?
- C. **Participation:** Are there analyses of the participation rate among potential a) settings, b) delivery staff, and c) patients (consumers).
- D. **Representativeness:** Are comparisons reported on the similarity of settings participating to the intended target audience of program settings--or to those settings that decline to participate?
- E. **Representativeness:** Are analyses reported on the similarity and differences between patients, consumers, or individuals who participate vs. either those who decline, or the intended target audience?

# EV REPORTING CRITERIA (cont.)

2. **Program or Policy Implementation and Adaptation**
  - A. **Consistent Implementation (“Fidelity” or well-delineated scope of adaptations):** Are data presented on the range of implementation variations of different program components during the evaluation/study?
  - B. **Staff Expertise:** Are data presented on 1) the level of training or experience required to deliver the program and 2) quality and extent of implementation by different staff?
  - C. **Program Customization or Adaptation:** Is information reported on the ways different settings modified or customized the program to fit their setting (or that no variation was observed)?

# EV REPORTING CRITERIA (cont.)

## 3. Outcomes for Decision Making

- A. **Significance:** Are the outcomes compared to either clinical guidelines (and their intended outcomes) or community preventive services guidelines or other standards of practice for best practices and their associated public health goals?
- B. **Adverse Consequences:** Do the outcomes reported potentially negative effects on quality of life or other outcomes?
- C. **Moderators:** Are there analyses of moderator effects-- including 1) different subgroups of participants and 2) types of intervention staff or settings--to assess robustness vs. specificity of effects?
- D. **Program Intensity:** Are data reported on either or both the total amount of staff time or patient/consumer contact time required?
- E. **Costs:** 1) Are data on the costs presented? If so, 2) are the assumptions made and perspective adopted (e.g., societal, health care payer, patient) and both physical and person costs reported?

## EV REPORTING CRITERIA (cont.)

### 4. **Time: Maintenance and Institutionalization**

- A. **Long-term Effects:** Are data reported on longer-term effects, at least 12 months following treatment / intervention?
- B. **Institutionalization:** Are data reported on the sustainability (or re-invention or evolution) of program implementation at least 12 months after the formal evaluation / study?
- C. **Attrition:** 1) Are data on attrition by condition reported, and 2) are analyses conducted of a) representativeness of those who drop-out or b) imputation?

# **EDITORS' MEETING ON EV**

- **Editors from 13 leading health and behavior journals**
- **Met in 2006 to discuss above issues and criteria**
- **Meeting sponsored by RWJF, OBSSR, CDC, AHRQ, and held at UNC**

# OUTCOMES OF EDITORS' MEETING

- Agreement all EV criteria were important
- Felt all EV criteria, except cost and institutionalization were feasible
- Majority felt that guidelines or principles, rather than mandatory reporting criteria, were best approach



# ACTIONS TAKEN BY JOURNALS

- Editorial and recommendations on EV reporting (n = ≥ 6)
- EV criteria for authors (n = ≥ 1)
- EV Criteria for reviewers (n = ≥ 1)

## OTHER IDEAS

Highlight article(s) with exemplary EV reporting

Special issue on EV topic

# WHAT IS NOT PROPOSED?

- ***NOT* saying all articles have to be strong on EV criteria**
- ***NOT* saying all articles have to report (but more should than is presently the case)**
- ***NOT* saying that internal validity (or RCTs) are not important (just that we need more of a balance)**

# FUTURE DIRECTIONS AND NEEDS

- Document reliability of EV coding criteria
- Consider “summary metrics”, composite, or overall EV quality scores
- Assistance to practitioners on how to combine with theory and local experience
- Evaluate which criteria most strongly related to long-term dissemination success
- Revise criteria based on lessons learned

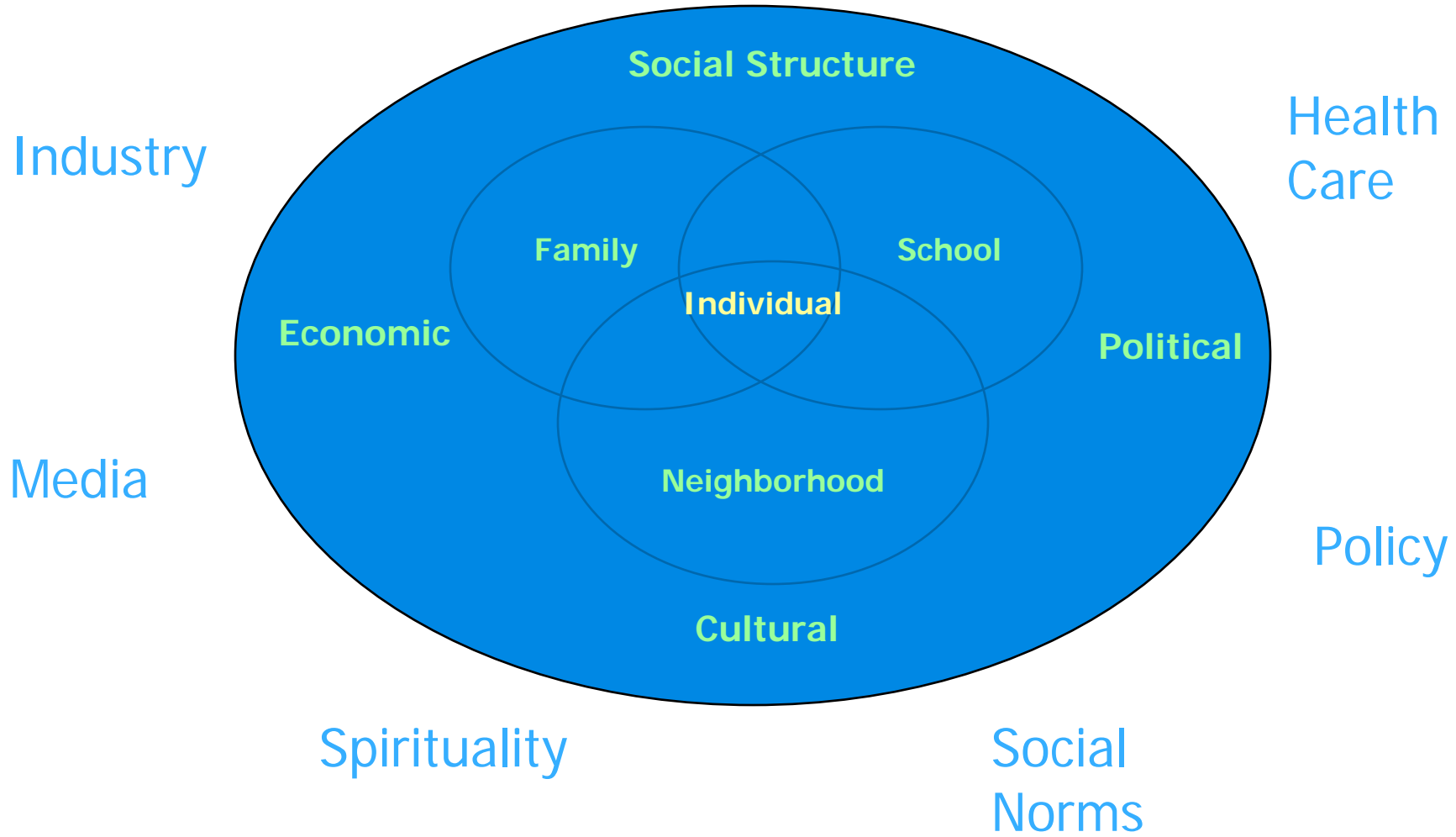
# **Evaluating External Validity Reporting in Childhood Obesity Studies**

**Lisa M. Klesges  
School of Public Health  
University of Memphis**

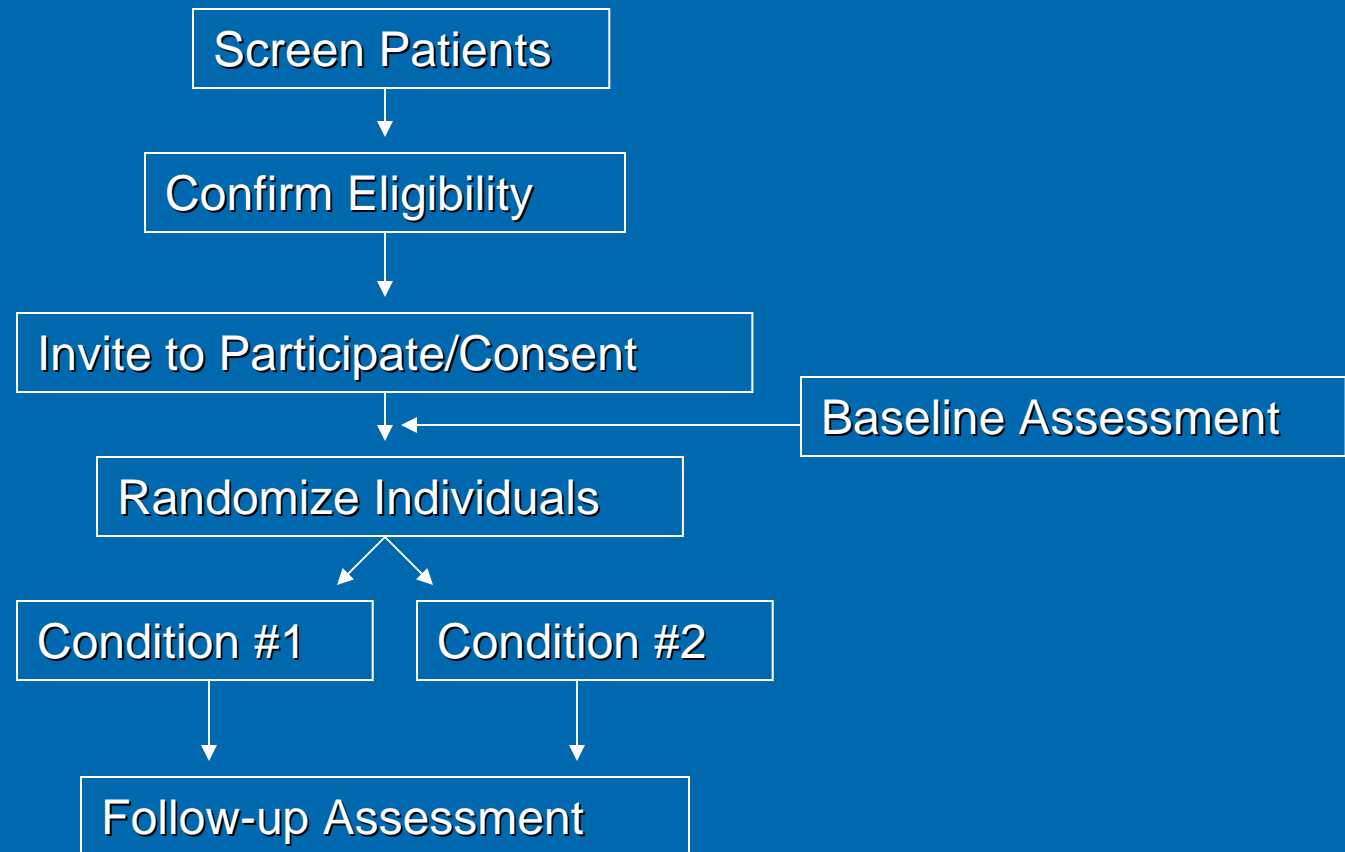
# Overview

- Improvements needed in how we design studies and report results
- Report results evaluating the extent that EV elements were reported in childhood obesity literature – prevention and treatment
- Recommendations to expand evidence based reporting of external validity and context elements

# Contextual Perspective



# RCT Diagram



# Contrasting Needs

	<b>Efficacy</b>	<b>Translation-focused</b>
<b>Goal:</b>	-Internal validity, isolate mechanisms and causes, evaluate theory decisions	-External validity, replication, understand context, evaluate practice & policy decisions
<b>Participants Reach</b>	-Homogeneous, motivated samples -Many exclusions and selection	-Broad, heterogeneous, representative samples
<b>Interventions Effects</b>	-Intensive, specialized interventions -Maximize effect size -Standardized protocols	-Brief, feasible interventions -Adaptable to settings
<b>Settings Adoption</b>	-Usually one setting to reduce variability and selected settings with resources and expert staff	-Appeals to and works in multiple settings
<b>Delivery Implementation</b>	-By “expert” research staff closely following specific protocol	-By variety of different staff with competing demands, using adapted protocol
<b>Long-term Change Maintenance</b>	-Often not evaluated (but expensive) -Often shorter-term outcome -Focus on individual level outcomes	-Major issues for decisions -Setting level maintenance important for investment

# Current Research Out of Context

---

- Translation decisions require evidence with high external validity and contextual relevance
  - Need more research specifically designed to address complexity and contextual variability
  - Need EV elements reported in efficacy-focused trials to support summary reviews

# Enhance Future Translation

---

- **Current issues specific to childhood obesity:**
  - **“Insufficient Evidence” available to make decisions and take action in many systematic reviews**
  - **Quantitative reviews weak in generalizability summaries to target interventions and translate into other settings**

# Enhance EV Reporting

**S** → **Settings & Populations**

**P** → **Program/Policy Implementation & Adaptation**

**O** → **Outcomes for Decision-Making**

**T** → **Time: Maintenance & Institutionalization**

# PURPOSES

- Evaluate extent that external validity dimensions are reported in behavioral interventions for childhood obesity
- Demonstrate feasibility of applying EV reporting elements
- Conducted 2 systematic reviews: 1) obesity prevention interventions and 2) treatment studies for already overweight or obese children



# Study Methods

- Controlled interventions published 1980-2004, with n=19 prevention studies and n=56 treatment studies

## Inclusion Criteria:

- Child or adolescent sample ( $\leq 18$  yrs)
- Anthropometric outcome (1° or 2°): e.g., BMI, Body Fat
- Control or Comparison condition (could be non-randomized)
- Health behavior target of nutrition, physical activity and/or lifestyle

## Exclusion Criteria:

- Non-English language publication
  - Designed as feasibility study
- 2 Trained, independent reviewers coded each study
  - Analyzed % of studies reporting criteria

# Summary of Prevention Studies

## Study Sample Characteristics (n=19 studies)

- 79% - 15 long-term ( $\geq$  1-yr or school year in duration)
- 95% - 18 interventions school-based delivery
- 63% - 12 weight gain as primary outcome
- 37% - 7 cardiovascular risk reduction as primary
- 74% - 14 targeted diet and physical activity
- 16% - 3 targeted physical activity or sedentary behavior
- 10% - 2 targeted diet/nutrition education only

# Summary of Treatment Studies

- **Study Sample Characteristics (n = 56 studies)\***
  - 34% - 19 long-term ( $\geq$  1-yr in duration)
    - 16% - 9 clinic or medical setting
    - 14% - 8 research setting
    - 11% - 6 school-based
    - 5% - 3 other (church., home, mixture)
    - 54% - 30 unknown
  - 84% - 47 targeted diet and physical activity
  - 9% - 5 targeted physical activity or sedentary behavior
  - 7% - 4 targeted diet/nutrition education only

\*All studies targeted primary weight outcome

Klesges, et al., in progress

# Percent of Studies Reporting on External Validity Dimensions

<u>I. Settings and Populations</u>	<u>% Prev</u>	<u>%Tx</u>
<u>Individual Level</u>		
Individual inclusion/exclusion	90	86
Participation rate	63	43
Representativeness of participants	10	9
<u>Setting Level</u>		
Setting inclusion/exclusion	11	5
Participation rate: settings	22	2
Representativeness of settings	0	0
<u>Delivery Staff</u>		
Participation rate	5	2

# Percent of Studies Reporting on External Validity Dimensions

<u>II. Program Implementation &amp; Adaptation</u>	<u>%Prev</u>	<u>%Tx</u>
Consistent implementation of program	26	11
Staff expertise or training	89	54
Implementation differed by staff	5	0
Program adaptation	42	57

# Percent of Studies Reporting on External Validity Dimensions

<u>III. Outcomes for Decision Making</u>	<u>%Prev</u>	<u>%Tx</u>
Outcomes compared to standard goal	37	7
Adverse Consequences	32	16
Effect moderator by participant characteristic	53	45
Effect moderator by staff/setting	10	5
Program intensity	68	100
Costs	0	2

# Percent of Studies Reporting on External Validity Dimensions

<u>IV. Time: Maintenance &amp; Institutionalization</u>	<u>% Prev</u>	<u>%Tx</u>
Long-term effects (at least 12 mo.)	74	34
Program Sustainability (12 mo. after last eval.)	0	0
Attrition Rate	100	79
Differential attrition by condition	21	11
Drop-out representativeness	42	9

# Implications

---

- **Applicability and reliability of EV coding criteria were demonstrated**
- **Feasibility supported by exemplar articles that reported most elements**
- **Particular weaknesses in EV reporting:**
  - **Representativeness of participants, settings**
  - **Program implementation elements**
  - **Attrition – differences, representativeness**
  - **Costs & sustainability**

# Recommendations

- **It's not as difficult as you think...**
- **Reformation in reporting**
  - Large potential benefit for systematic and quantitative review
  - Provide EV evidence for translation and decision-making – “evidence to action”
- **Redesign in our approach to design**
  - Relevance of research conducted to address “wicked” problems
  - Greater focus on context and external validity

# Contributors

- **David A. Dzewaltowski**
- **Russell E. Glasgow**
- **Katherine Kitzmann**
- **Paul Estabrooks**
- **Kara Davis**
- **Joanna Buscemi**
- **Julka Almquist**

# **METRICS FOR COMPARING EXTERNAL VALIDITY DIMENSIONS OF INTERVENTIONS**

**Paul Estabrooks, George Davis, Ranju Baral  
September, 2008**

# RE-AIM

- **Standard metrics that accurately summarize complex and multidimensional outcomes are needed.**
- **The RE-AIM framework offers a comprehensive approach.**
  - **Reach: the number, percent, and representativeness of participants**
  - **Effectiveness: the intervention impact on outcomes and QOL**
  - **Adoption: the number, percent, and representativeness of settings and intervention staff**
  - **Implementation: the consistency of delivery by various staff**
  - **Maintenance: the extent to which individual participants maintain behavior change long term and, at the setting level, the degree to which the program is sustained within the delivery organizations**

# RE-AIM

- To date, RE-AIM has predominately been applied to a single dimension at a time.
- Combining two or more RE-AIM dimensions may be useful for making policy comparisons and decisions.
  - Individual Level Impact
    - RE: Reach X Effectiveness
    - RE2: Problem Prevalence X RE (Attributable Individual Level Impact)
    - RE3: Incremental cost of treatment-control/Incremental RE of Treatment-control (Efficiency)
  - Setting Level Impact
    - AI: Setting Adoption X Staff Adoption X Implementation
    - AI2: AI X number of target settings X Average number served per setting
  - RE-AIM Average

# Objectives

- **Examine individual level calculations for a commercial internet and incentives-based worksite weight loss program.**
- **Compare these calculations to fictitious, but realistic data from a worksite weight loss small group intervention and individual counseling**

# Data used in analyses

	Internet	Small Group	Individual
Total Employees	10523 (68% Women)	10523 (68% Women)	10523 (68% Women)
% Eligible	64%	64%	64%
# Informed	10523	10523	10523
# Participate (men/women)	1450 2748	350 450	100 150
Mean age participants	42.9	42.9	42.9
Mean age Total	42.4	42.4	42.4
Mean weight $\Delta$ intervention	-2.5 pounds	-12 pounds	-24 pounds
Mean weight $\Delta$ control	+1 pound	+1 pound	+1 pound
Cost PP	\$54	\$240	\$480

# Reach

**Definition:** It is the absolute number, proportion and representativeness of people willing to participate in a program

*1. What percent of the target population are you reaching?*

## 1.1 Participation Rate

$$= \frac{\text{No. of people willing to participate}}{\text{No. of people eligible}} * 100$$

$$= \frac{4198}{6728} * 100 = 62.4$$

# Reach

**1.2 Representativeness: It is the similarity or difference between those who participate and those who are eligible but do not participate.**

*Measured as Median Effect Size (MES differential characteristics)*

**1.2.1  $MES_{age} = (\text{Mean age of individuals who participated} - \text{Mean age of individuals who were eligible but did not participate}) / \text{Common S.D.}$**   
 **$= (42.9 - 42.4) / 12 = 0.04$**

**1.2.2  $MES_{gender} = (\text{Proportion of male who participated} - \text{Proportion of male who did not participate}) / \text{Common S.D.}$**   
 **$= (0.38 - 0.48) / 0.2 = -0.5$**

# Reach

$$\begin{aligned}\text{Reach} &= \text{Participation rate} - \text{MES}_{\text{age\&gender}} \\ &= 0.62 - (-0.27) \\ &= 0.89\end{aligned}$$

**As you see in this example,  
representativeness could increase reach  
score....**

# Effectiveness

**Definition:** It is the measure of impact of the intervention on targeted outcomes and quality of life and economic outcomes... and potential moderator effects.

**Composite Intervention Effectiveness (E) =**  
**(MES<sub>key outcomes</sub> - MES<sub>negative outcomes</sub> - MES<sub>differential impact</sub>)**

**Mean change from baseline (weight loss in lbs) among participants = 2.5 lbs**

**Mean change from baseline among non participants = 1 lbs gain**

**Common SD = 7.8 lbs**

**2.1 MES<sub>key outcomes</sub> = (2.5 - -1) / 7.8**  
**= 0.45**

# Effectiveness: Differential Effects

**MES<sub>differential impact</sub> ( weight loss in lbs)**

**Mean weight decrease among males =2.58 lbs**

**Mean weight decrease among females = 2.40 lbs**

**Common SD = 7.8**

**2.3 MES<sub>gender</sub> = (Mean of outcome among male – Mean outcome among female) / common SD**

$$= (2.58-2.40 ) / 0.78 = 0.02$$

**2.4 MES<sub>age</sub> = (Mean outcome among 50 & under– Mean outcome among >50 years) / common SD**

$$= (2.8- 1.7) / 7.7 = 0.14$$

# Effectiveness

Effectiveness =

$$\text{MES}_{\text{key outcomes}} - \text{MES}_{\text{negative outcome}} - \text{MES}_{\text{gender\&age}}$$

$$= 0.45 - (0) - .08$$

$$= 0.37$$

*{ Question: Is differential impact always negative? }*

# Reach and Effectiveness

**Definition:** RE is a composite measure for assessing the reach and effectiveness of an intervention at the individual level.

$$\text{RE1} = \text{Reach} * \text{Effectiveness}$$

$$= 0.89 * .37$$

$$= 0.33$$

# Reach and Effectiveness

**Definition: RE2 (Attributable Individual Level Impact)**

**It measures the total impact on the target population that can be attributed to the given intervention**

**Prevalence (P) = number of overweight and obese employees at a given time divided by the total number of employees in the same time period**

$$= \frac{6728}{10513} = .64$$

$$\begin{aligned} \text{RE 2} &= P * \text{RE 1} \\ &= 0.64 * 0.33 \\ &= 0.21 \end{aligned}$$

# Reach and Effectiveness

**Definition: RE efficiency (RE 3) measures the efficiency and reach in terms of money.**

**Per participant cost of treatment = \$108**

**Incremental impact of RE 1 = 0.33**

$$\begin{aligned} \text{RE 3} &= \frac{\text{Cost of treatment}}{\text{Incremental RE1 of treatment}} \\ &= 54 / 0.33 \\ &= 163.64 \end{aligned}$$

# What do these numbers mean?

	Internet	Small Group	Individual
Participation Rate	62.4%	7.6%	2.4%
Representative	-.27	-.27	-.27
REACH	.89	.35	.26
Effect Size	.45	1.7	3.2
Diff. Effects	.08	.08	.08
EFFECTIVENESS	.37	.9	2.6
REACH*EFFECT	.33	.32	.68
RE2 (Attributable Individual Level Impact)	.21	.2	.43
RE3 (efficiency)	\$163.64	\$750	\$705.88

# Comments

- **All aspects of individual assessment will influence individual level composite scores**
- **The need for comparative data to determine normative values is clear**
- **Some issues with combining effect sizes and prevalence scores (the issue of negativity)**
  - **Closer adherence to the rules of probability may allow for a more interpretable range of scores (i.e., 0 to 1).**

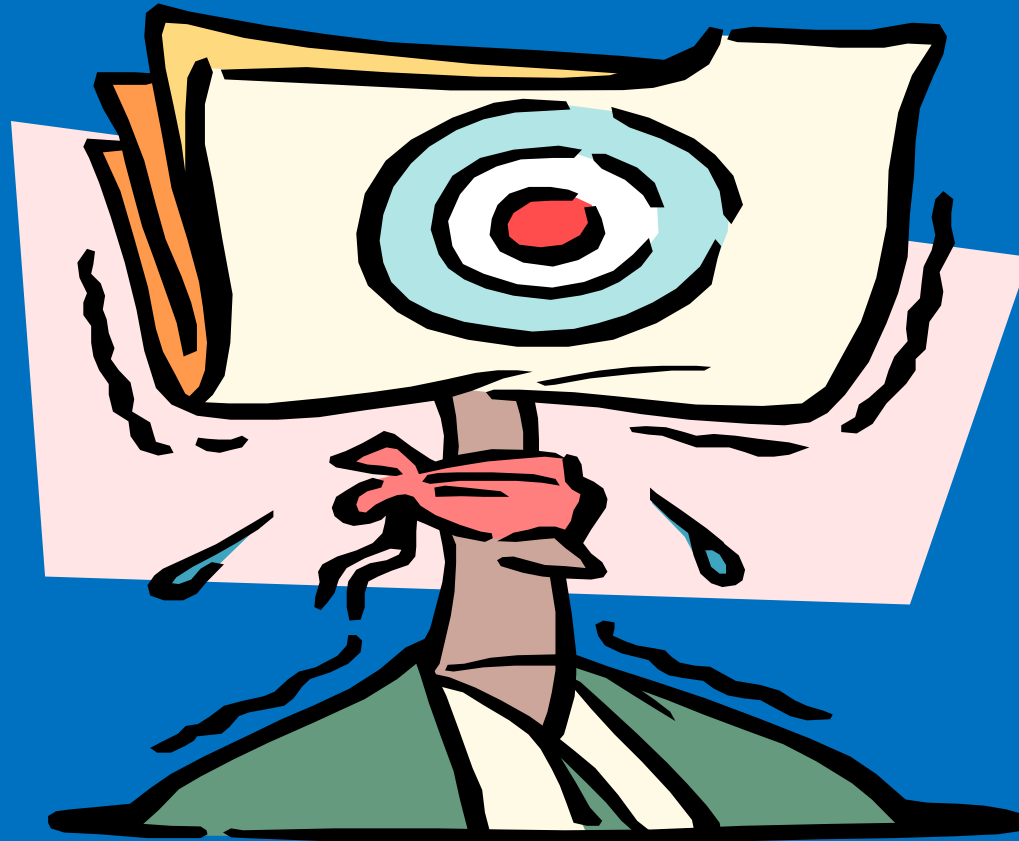
# Comments and work in progress

- **Will these metrics be compelling to policy makers?**
  - A need for more 'grounded' information.
- **Reach by effectiveness example using the same data:**
  - Approximately 10% of the employee population will benefit
  - Those who benefit will lose 9.5 pounds on average.
  - Women will be more likely to be successful
  - The cost is \$171 per successful employee
  - These data come from worksites where employees use computers for job related tasks.

# Next steps?

- **Data is nearly impossible to come by across the individual level factors**
- **The organization level metrics are equally (or more lacking in information in the extant literature.**
  - **Consistency in reported is needed (Editors workgroup)**
  - **Methods to present this data in a relatively brief way**
- **Integrate conceptual metrics with descript language that will be compelling for policy makers**

# Questions, Crossfire, Discussion



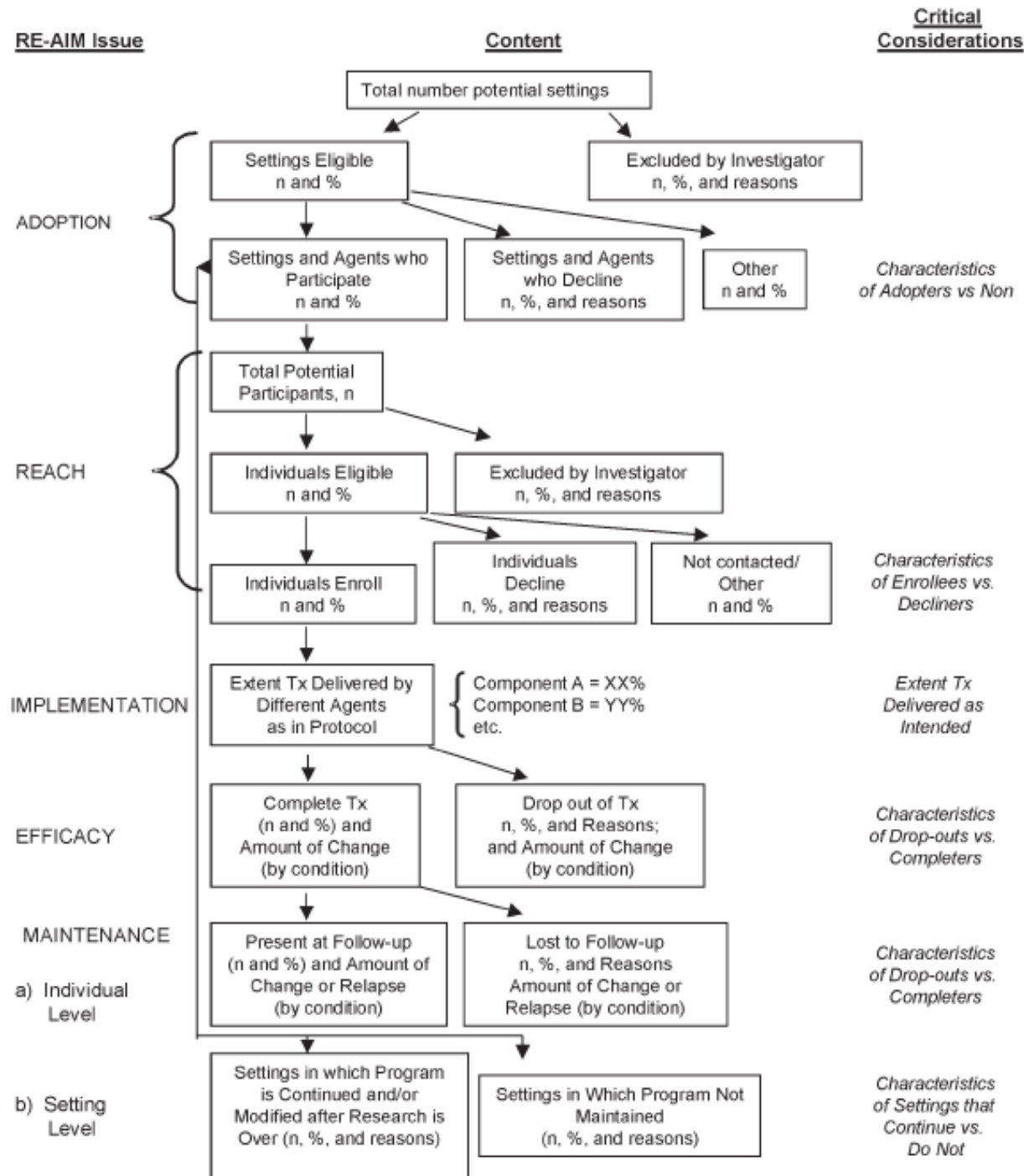
# Summary

- **Field would benefit from ability to weigh evidence on external validity dimensions in addition to internal validity**
- **Can improve systematic reviews of the literature for program planning and policy decisions by including time, place and population characteristics**
- **Methods to calculate relative merits of programs would enhance comparisons and decisions about “real world” impact**

# References

- [www.re-aim.org](http://www.re-aim.org) -- Website
- Green, L.W., Glasgow, R.E. Evaluating the relevance, generalization, and applicability of research: Issues in translation methodology and external validity. *Eval Health Prof* 2006;29(1):126-153.
- Klesges LM, Dzewaltowski DA, Glasgow RE. Review of external validity reporting in childhood obesity prevention research. *American Journal of Preventive Medicine*, 34: 216-223, 2008.
- Glasgow RE, Klesges LM, Dzewaltowski DA, Estabrooks PA, Vogt TM. Evaluating the impact of health promotion programs: using the RE-AIM framework to form summary measures for decision making involving complex issues. *Health Education Research*, 21:688-694, 2006.
- Glasgow RE, Emmons KM. How can we increase translation of research into practice? Types of evidence needed. *Annu Rev Public Health*. 2007;28:413-33.

# Proposed Reporting Standard to Enhance EV Criteria Inclusion



Glasgow RE, Bull SS, Gillette C, Klesges LM, Dzewaltowski DA. Behavior change intervention research in healthcare settings: a review of recent reports with emphasis on external validity. *Am J Prev Med.* 2002 Jul;23(1):62-9.