Obesity-Related Policy Evaluation Webinar Series

Session 1

Basics of Design to Evaluate Policy Interventions

Dr. Kathryn E. Newcomer

The Trachtenberg School of Public Policy and Public Administration

The George Washington University

February 27, 2009

Instructional Objectives

- Provide background for evaluators of obesity-related policies and programs to help them address likely questions and possible obstacles
- Offer advice about measurement strategies
- Identify key challenges to drawing credible conclusions from evaluations, and provide suggestions on how to address these challenges

Issues Addressed in this Session

1. The current context for evaluating policy interventions affecting diet and physical activity behavior
2. Measurement challenges
3. Basic study design objectives
4. Selecting appropriate evaluation designs

Section 1:
The Current Context for Evaluating Policy Interventions Affecting Diet and Physical Activity Behavior

What Do We Mean by Obesity-Related Policies?

- Public policies at federal, state, and local levels intended to improve diet, increase physical activity, and/or reduce obesity prevalence

  Examples:
  - School wellness policies
  - Calorie labeling in restaurants
  - Fruit and vegetable carts
  - Implementation of parks or bikes trails
  - Federal funding to create incentives for bicycle commuting

- Contrasts with individual-level, behavioral interventions

These Policies Operate in the Outer Two Rings of the Social Ecological Model

Source: Orleans, 2007

The Current Policy Evaluation Context

- Daunting standards espoused by proponents of "evidence-based policy making"
  - The Campbell Collaboration
  - Office of Management and Budget (OMB) guidance that randomized control trials (RCTs) are the "gold standard" for evaluation
- Difficulty of establishing a causal link between policy interventions and behavior change
  - Numerous factors affect individual behaviors
  - Trends in life style and choices offered to consumers are changing faster than ever before
- Ethical prohibitions and logistical impossibilities that do not allow random assignment of subjects in health-related evaluation

The Current Context: Diet and Physical Activity Policies

- There is not an abundance of theory to inform evaluations of policy interventions designed to change diet and physical activity behavior
- Consensus has not formed about how best to measure the variety of potential policy interventions in this relatively new field
- Physical activity and diet are among the most private of all human behaviors, thus challenging to monitor closely

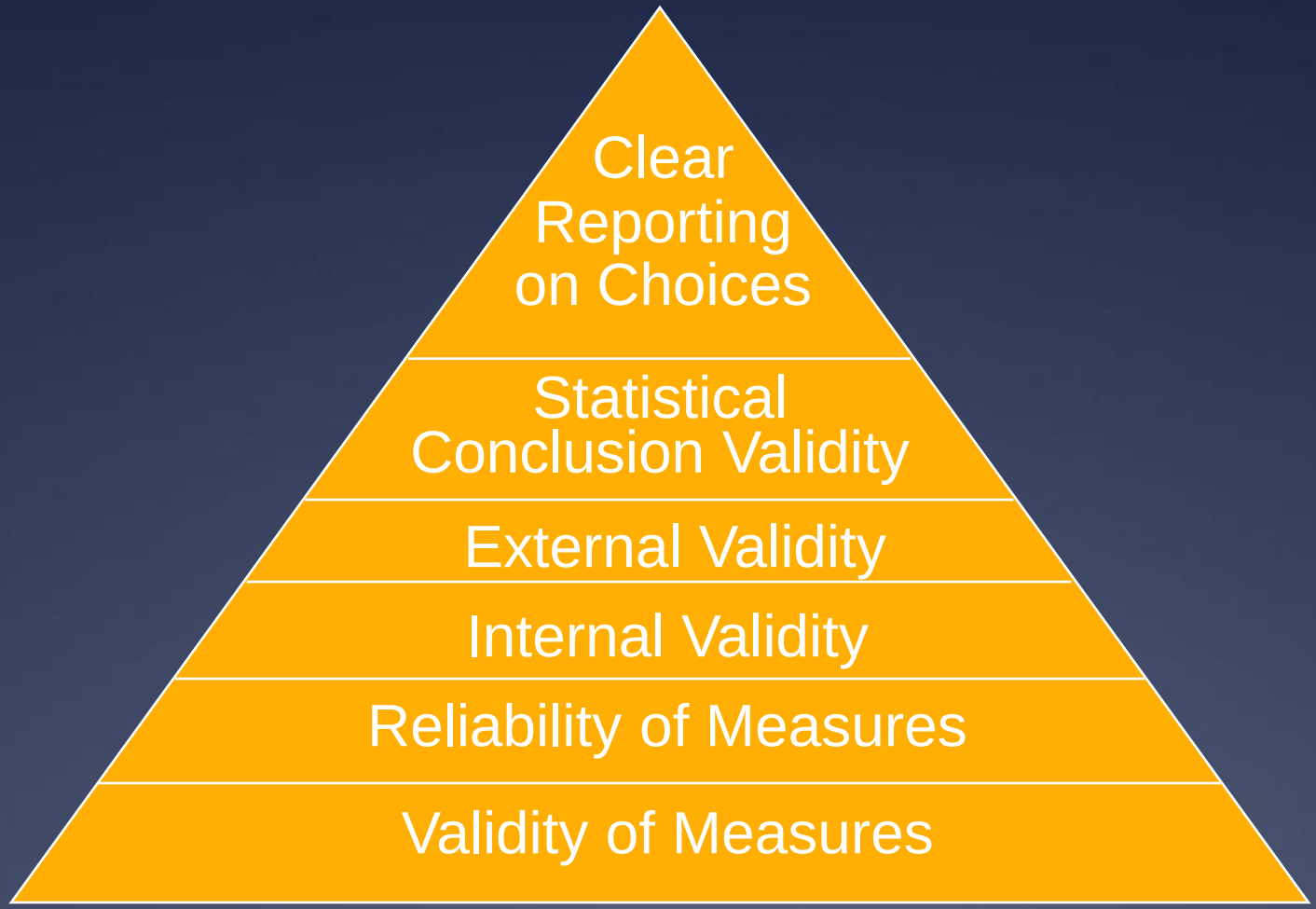
Given the Challenging Environment: Where To Start?

- Draw upon the social science methods that are used to bolster the credibility of research
- Take advantage of existing wisdom about the art of evaluating complex policies and programs
- Build knowledge in this relatively new field through careful sharing and testing of methods and measures and practicing transparency in the methodological decision-making in our work

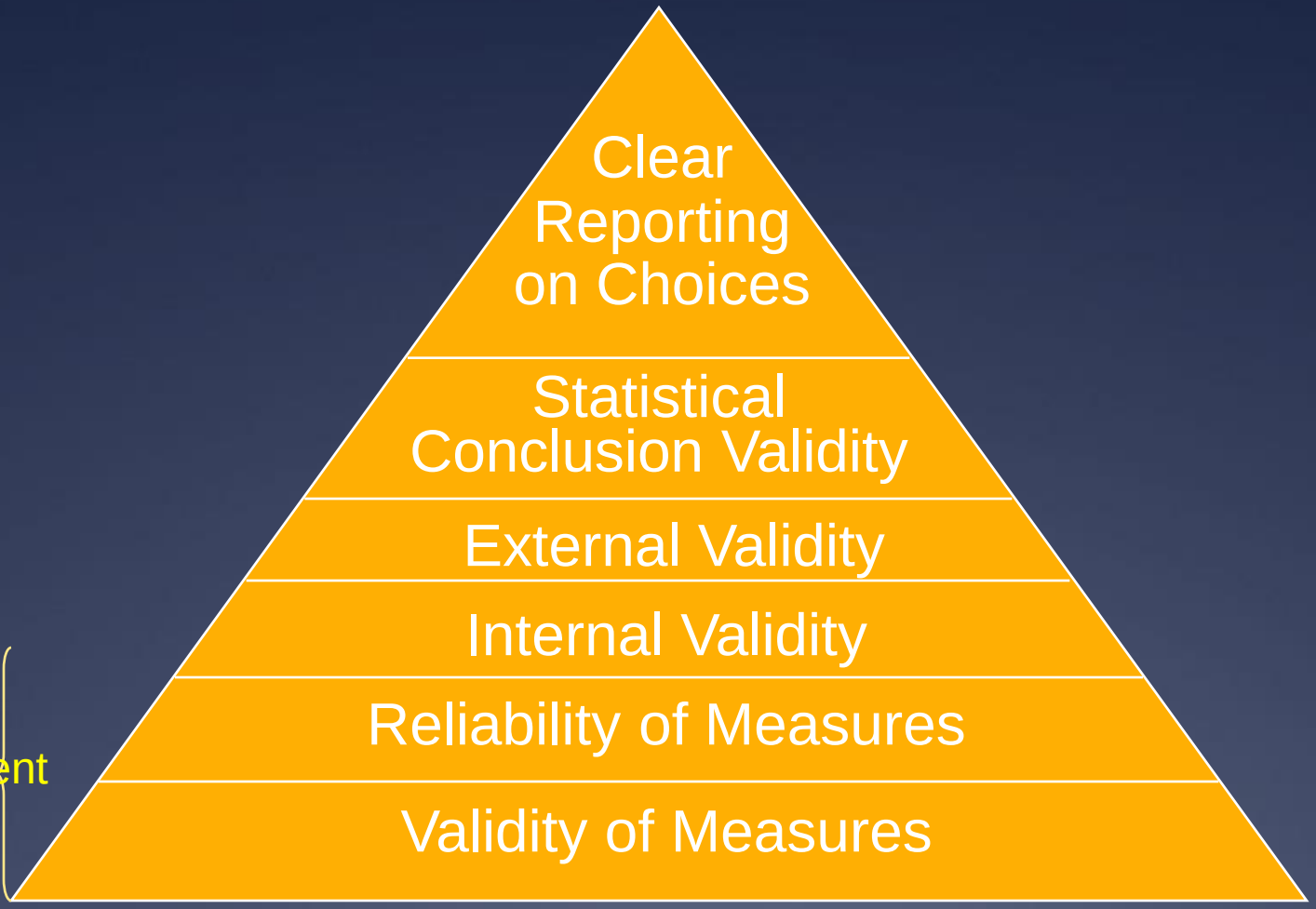
Credibility of our evaluation work is dependent upon…

the **Methodological Integrity** of our work!

Methodological Integrity is affected by a variety of decisions about…

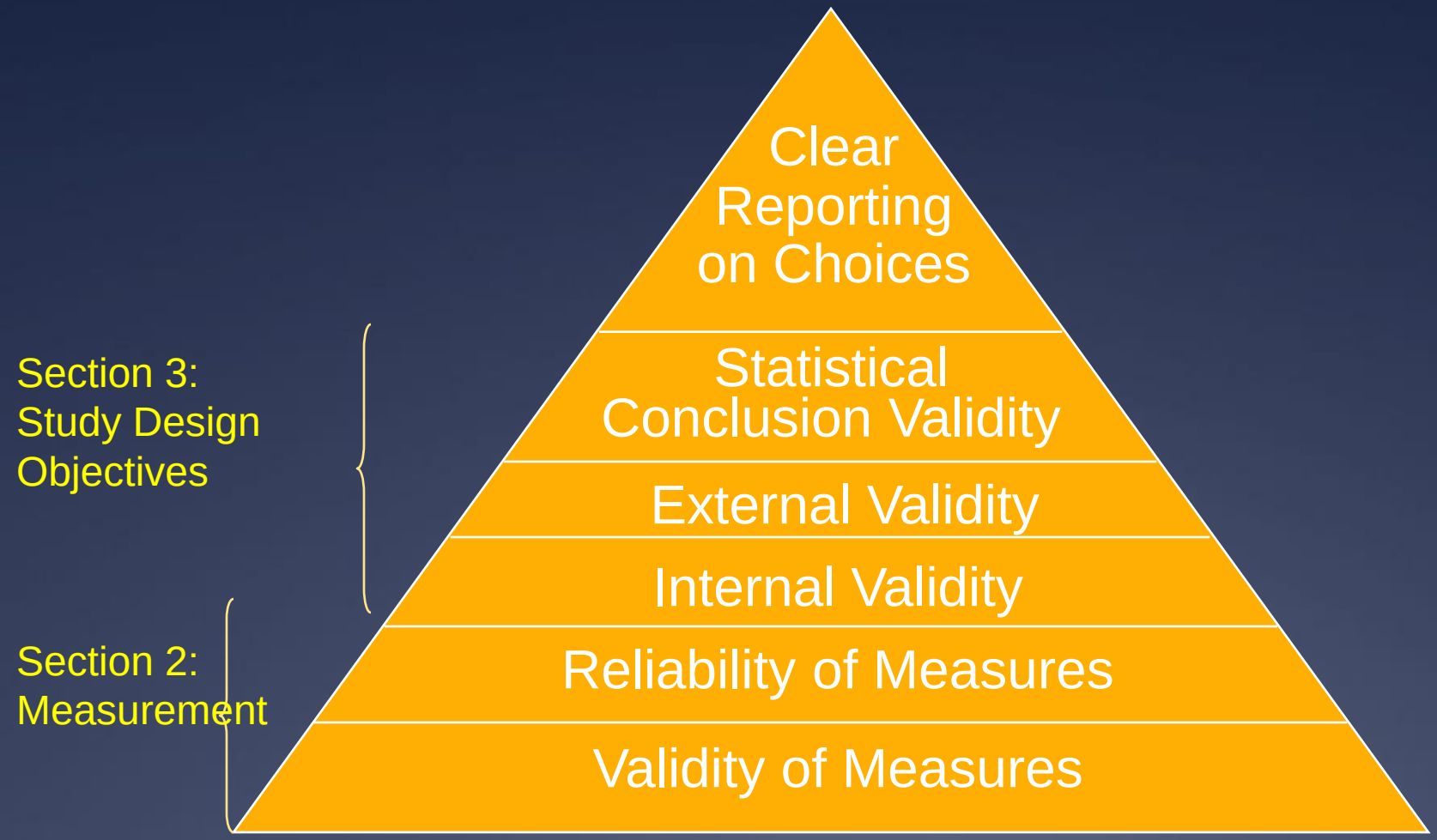


Methodological Integrity is affected by a variety of decisions about…

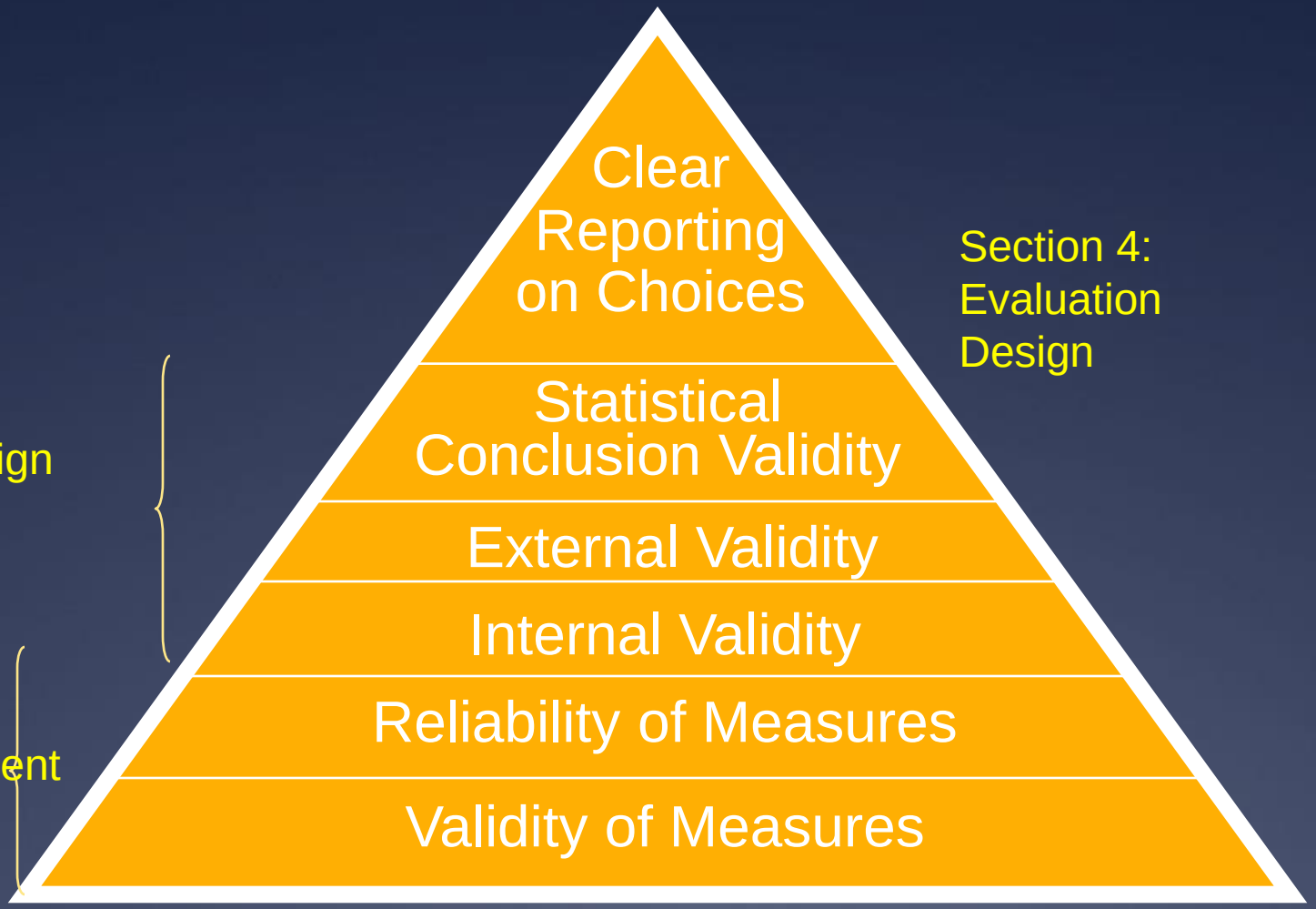


Section 2: Measurement

Methodological Integrity is affected by a variety of decisions about…



Methodological Integrity is affected by a variety of decisions about…



Section 2:
Measurement

Building Upon a Firm Foundation…

We must start with good, credible measures and strong procedures in place to insure we measure consistently across time and space.

Without credible measures, the rest of the considerations are inconsequential!

Examples of Measures

Policies	Food and Physical Activity Environments	Individual-level
Legislation language (e.g., "must" vs. "may")	Instruments: Observed; Self-reported	24-hour dietary recalls
Presence of funding vs. unfunded mandate	Geographic Information Systems	Food frequency questionnaires
Presence of enforcement strategies	Sales Analysis (food only)	Physical activity diaries
	Menu analysis (food only)	Pedometers
	Nutrient Analysis (food only)	Accelerometers

Measures

Measures need to be:

1. Valid
2. Reliable

Example: School Wellness Policies

1. Validity of Measures

Measurement validity is concerned with the accuracy of measurement

- Are we accurately measuring what we really intend to measure?

Measurement Challenges Specific to Obesity-Related Policy

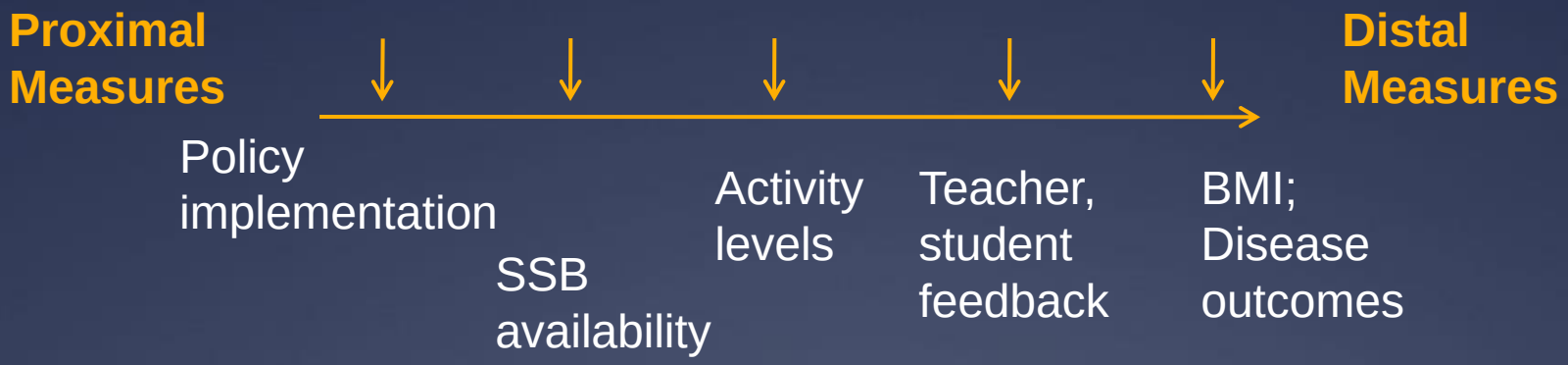
- Understand what you need/want to measure about:
  - 1) Policy intervention
  - 2) Target behavioral, and/or health outcomes

- Policy interventions may be complex sets of:
  - Changes to the environment

- Behavioral/health outcomes may include:
  - Individual behavior (e.g., diet, physical activity)
  - BMI, diabetes type 2, serum cholesterol, etc.
  - May be a continuum

For Example, School Wellness Policies:

- Two possible changes are:
  1. Increased physical education requirements
  2. Reduced access to sugar-sweetened beverages (SSBs)



- BOTTOM-LINE: There are many choices available. We need to consider the timeframe for our evaluation as well as the logical connection between the outcome measure and the policy intervention

Strengthening Measurement Validity: A Checklist

1. Face Validity:
   - ✓ Do other experts in the field use these measures?
   - ✓ Do experts find our newly constructed measures credible?

School Wellness Policies Example:

- ∗ Talk to principals, teachers – are we missing anything?
- ∗ Have other experts used similar measures?

Strengthening Measurement Validity: A Checklist

2. Content Validity:
   - ✓ Do the measures selected adequately represent the potential pool of similar measures?

School wellness policies example:

- ∗ Consult position papers, consensus statements, and reports from multiple organizations, e.g., federal and state-agencies, Institute of Medicine, the National Governor's Association

Strengthening Measurement Validity: A Checklist

3. Criterion Validity:
   - ✓ Do the measures correlate to a specific agreed-upon standard or criterion measure that is credible in the field?

School wellness policies example:

- ∗ You may be evaluating the perception of students to changes to school cafeteria options or access to physical activity equipment. How does that correlate to observed data?

Strengthening Measurement Validity: A Checklist

4. Construct Validity:
   - ✓ Do the measures behave (i.e. correlate) with other measures in ways consistent with existing theory and knowledge?

School wellness policies example:

- ∗ Examine surveillance systems for similar prevalence or trend data such as Behavioral Risk Factor Surveillance System (BRFSS), or the National Survey of Children's Health (NSCH), for youth obesity rates

Strengthening Measurement Validity: A Checklist

5. Predictive Validity:
   - ✓ Do the measures predict subsequent behaviors in ways consistent with existing theory and knowledge?

School wellness policies example:

- ∗ Does reducing access to SSBs in schools predict reduced intake of SSBs in children?

SSBs = sugar sweetened beverages

2. Reliability of Measures

Reliability is the extent to which a measure can be expected to produce similar results on repeated observations of the same condition or event.

- Having reliable measures means that operations consistently measure the same phenomena
- Reliable measurement entails consistently recording data with the same decision criteria

For Example:

- How do we word survey and interview questions in an evaluation of physical activity so that respondents will report their behaviors consistently? Across time? Across different socio-demographic subgroups?

- How do we ensure that observers of a school food environment record and assess exactly the same aspects of appearance and prominence in placement of healthy foods in a cafeteria?

Strengthening the Reliability of Our Measures and Measurement Procedures: A Checklist

1. Adequately pre-test instruments:

- ✓ Are we pre-testing instruments with representative samples of intended respondents before we go into the field?
- ✓ Are we implementing adequate quality control procedures to identify inconsistencies in interpretation of words by respondents in surveys and interviews?
- ✓ When we uncover problems with the clarity of our questions, do we revise them and then go back to re-survey or re-interview if the questions are vital?

Strengthening the Reliability of Our Measures and Measurement Procedures: A Checklist

2. Train observers and interviewers, and employ inter-coder/rater reliability checks:

   - ✓ Are we adequately training our observers and interviewers so that they consistently apply comparable criteria?
   - ✓ Are we implementing adequate and frequent quality control procedures to identify obstacles to consistent measurement in the field?
   - ✓ Are we testing observers or coders by asking them to all code a sample of the materials to test the levels of consistently, e.g., statistical checks?

Commonly Used Tests for Reliability

Test	Dimension Tested	Example	Commonly Used Statistics
Inter-rater reliability	How 2 coders score the same item/phenomenon	Evaluation of playground equipment safety by 2 observers	Kappa statistic; or simple percentage
Test-retest	The same measure performed at 2 points in time	Inventory the same vending machine contents 2 months apart	Pearson correlation coefficient
Internal consistency	Same construct within an instrument	Check that answers to survey questions designed to assess walking or biking to school correlate	Cronbach's alpha

Section 3:
Study Design Objectives

To Test the Effectiveness of Policy Interventions We Need to Ensure We Can Make Well-Founded Inferences About:

1. The relationship between the intervention and the observed effects (internal validity)

and

2. The generalizability of our results (external validity and statistical conclusion validity)

Internal Validity

Internal validity is concerned with our ability to determine whether an intervention (X) produced the intended outcome or result (Y) and in what magnitude.

- Are we able to definitely establish whether there is a causal relationship between a policy intervention and desired effect?
- Are we identifying unintended effects of a policy?

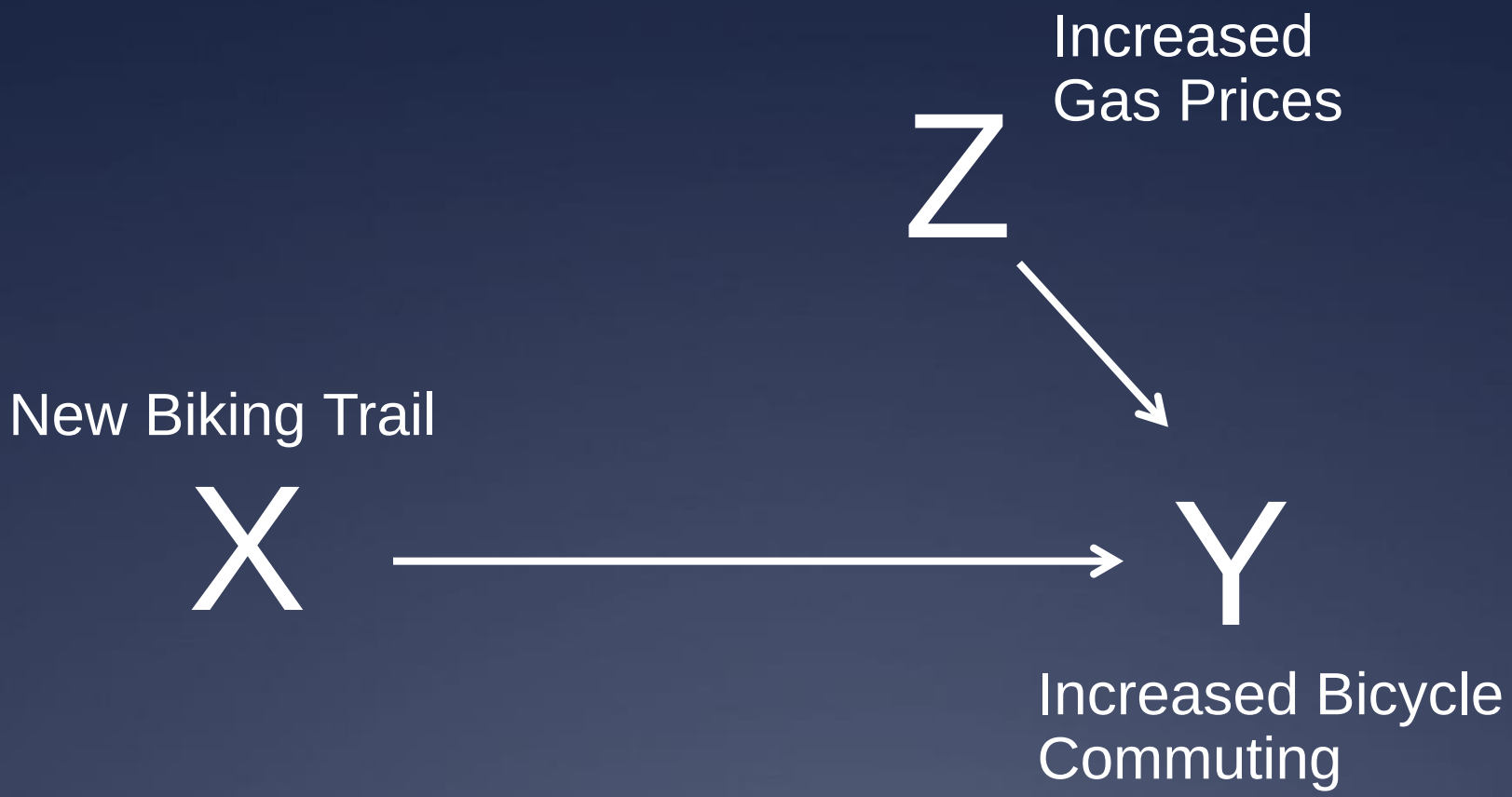
Causal Inference (Internal Validity)

Three Elements of Causal Inference

1. Temporal order
2. Co-Variation
3. Nonspuriousness (or lack of confounding)

(*and a fourth is sometimes given*
4. Grounded in Previous Research)

Causal Inference (Internal Validity)



Credible Causation versus Plausible Attribution versus Contribution?

When measuring outcomes, there are several challenges to capturing “net effects” or “net impacts” of the intervention since:

- There are other events and processes occurring in the neighborhoods or societies that affect the achievement of desired outcomes
- The time needed for the intervention to change the attitudes or behavior may be longer than the time given to measure outcomes
- There may be flaws in the evaluation design OR implementation of the intervention that reduce the ability of the intervention to even produce the desired outcomes

Examples

Credible Causation:

- Increased taxes on tobacco reduces consumption

Plausible Attribution:

- Increased Physical Education increases school test scores

Contribution:

- Providing calorie information on menus in restaurants improves dietary choices

Strengthening Our Ability To Attribute Effects to the Intervention: A Checklist

- ✓ Are we measuring the extent to which the intervention was implemented?

  For example: Are we asking about the availability of resources and guidance given to the principals and cafeteria staff?

- ✓ Depending on the intervention being evaluated, are we asking our study participants about other events or experiences they have had which also affected their decisions about diet or physical behaviors — before and during our study time frame?

  For example: Are we surveying samples of students in the school about their exposure to other programs or information campaigns of healthy diet and physical activity

Strengthening Our Ability To Attribute Effects to the Intervention: A Checklist

- ✓ Has enough time elapsed between implementation of the intervention and the measurement of the intended effects — given the existing knowledge about the likely time period needed to see effects?

  For example: Does the duration of our study permit collection of data over an appropriately extended period?

- ✓ Have we reviewed existing research to identify unintended effects, and then built in capacity to measure them?

  For example: Are we including open-ended questions when we talk to teachers and other school staff about unanticipated changes in student behaviors?

Generalizability

We should select sites and individuals that are truly representative of the populations to which we hope to generalize our results.

Generalizability includes:

1. External validity: the ability to generalize beyond the groups or context being studied
2. Statistical conclusion validity: the ability to generalize statistical findings beyond our sample; relevant only to quantitative data

Generalizability Cont'd

Why generalizability matters in policy evaluations:

- May be interested in assessing effects on a particular sub-population, e.g., youth, rural, racial/ethnic groups

For example, youth are far more sensitive to price increases in tobacco than adults

To Enhance Generalizability, We Make Sampling Choices

- Take care to identify sub-populations of interest, so that we have large enough sub-samples of the groups of interest to analyze
- Researchers should still examine a sample to ensure that it is **truly representative** of the population to which the researchers hope to generalize on demographic variables of interest to us, e.g., age, race/ethnicity

Generalizability – Cont'd

Estimating the sample size you will need in any evaluation to establish the statistical generalizability of the results depends on three general criteria:

1. The size of the population to which generalization is desired
2. The level of confidence you wish to have in your results – such as 95% or 99%?
3. The margin of error you are willing to provide – such as an error band of plus or minus 2%, or 3%, or 5%?

- Your friendly statistician can help you apply the formula once you answer these three questions!

You need also to consider sub-group populations to which you wish to extend your results and ensure you have sufficient sub-sampling of these groups

Statistical Significance versus Importance of the Measured Effects?

When reporting results we must report both:

1. Whether or not the sample size allows us to conclude that the intervention is effective based on the sample size

and

2. We need to apply credible criteria to discuss the importance and relevance of the size of the effect of the intervention

Strengthening Our Ability To Generalize the Effectiveness of the Intervention: A Checklist

- ✓ Did we determine the key demographic (or other) variables that we want to be represented in our sample?

  For example: Do we want to be able to generalize results to girls versus boys, or by age group?

- ✓ If we want to make statistical generalizations about the effectiveness of the intervention, did we produce a sample of sufficient size, with adequate sampling of important subgroups that we want to analyze?

  For example: Did we send enough surveys so that the respondent pool will permit statistical generalizations? (Check relevant literature for guides on response rates.) Did we over-sample for specific sub-groups?

Strengthening Our Ability To Generalize the Effectiveness of the Intervention: A Checklist

- ✓ Have we considered other possible groups, sites, or situations to which we will try to generalize?

  For example: Did we record potentially important aspects of the school or neighborhood so that we can describe the context of the policy intervention in enough detail so others many replicate the intervention?

- ✓ Are there aspects of the intervention that our evaluation identified that merit further study?

  For example: Were there any surprises or interesting mediating variables that appear to influence the effectiveness of the intervention that warrant further investigation?

Section 4:
Selecting Evaluation Designs

Your Evaluation Design Should Permit Relevant Comparisons

- Choices about selection of sites and subjects and the timing of data collection should be made so that your findings will be credible
- Evaluation practice offers guidance on strengthening the inferences made from policy intervention studies

Solomon Four Group — The "Gold Standard"

R	O_1	X	O_2
R	O_3		O_4
R		X	O_5
R			O_6

All four groups are drawn **randomly from the same population**

R = Randomly assigned; X = Policy Intervention; O = Observation

"Real World" Policy Evaluation Design

One-Group Posttest-Only

$X \quad O_1$

One Group Pretest-Posttest Design

$O_1 \quad X \quad O_2$

Multiple Time Series

$O_1\ O_2\ O_3\ldots \quad X \quad O_{25}\ O_{26}\ O_{27}\ O_{28}\ldots$

$O_1\ O_2\ O_3\ldots \qquad O_{25}\ O_{26}\ O_{27}\ O_{28}\ldots$

Strengthening Inferences

X = Policy Intervention; O = Observation

Design Your Evaluation to Permit Useful Comparisons!

- When possible, try to collect data on participants or the food or physical activity environment, as well as aspects of the intervention itself, at more than one point in time, such as pre-test post-test evaluations, or over a longer period, to view the apparent outcomes of interventions

or

- Try to find a comparison site or jurisdiction where an intervention was not implemented to permit useful comparisons

- You may use statistical analyses to attempt to control for other potential mediating variables when analyzing the data (as long you remember to measure them!)

Final Thoughts…

The ability to produce credible conclusions about the effectiveness of policy interventions will be strengthened by careful planning and reasoned decisions about evaluation methods:

1. Identify or develop valid and reliable measures
2. Identify the population and sub-populations to which you want to generalize
3. Design sampling strategies that allow you to generalize appropriately

Final Thoughts Cont'd…

4. Select a research design that will permit needed comparisons to assess the impact of the interventions and is feasible to apply
    - Obtain baseline (pre-test) data where possible
    - Identify comparison groups or jurisdictions where feasible
    - Measure the implementation of the intervention
    - Measure target behaviors/outcomes as many times as possible

Resources

Measures – Food and Physical Activity Environments

- Measures of the Food Environment searchable online database: www.riskfactor.cancer.gov/mfe
- *American Journal of Preventive Medicine* supplement: April, 2009 "Measures of the Food and Physical Activity Environments"
- Ohri-Vachaspati & Leviton: "Measuring Food Environments: A Guide to Available Instruments", forthcoming in *American Journal of Health Promotion*
- Active Living Research, physical activity measures: http://www.activelivingresearch.org/resourcesearch/toolsandmeasures
- Standardized Surveys of Walking and Biking Database: http://appliedresearch.cancer.gov/tools/paq/

Resources

Evaluation Texts

- Pawson & Tilley (1997): *Realistic Evaluation*, Sage Publications
- Rossi, Peter & Freeman (2003): *Evaluation: A Systematic Approach*, Sage Publications
- Weiss (1997): *Evaluation,* 2nd Ed. Prentice Hall
- Wholey, Hatry & Newcomer Eds. (2004): *The Handbook of Practical Program Evaluation*, Jossey-Bass