THE "EVIDENCE" IN *EVIDENCE-BASED* PROGRAMS

Utilizing Blueprints Standards to Judge an Intervention's Effectiveness and Utility

Blueprints Preconference Session
April 30, 2018

Pamela Buckley, Ph.D.
Laura Michaelson, Ph.D.
Christine Steeger, Ph.D.

The Blueprints Team

- **Buckley, Pamela (PhD) – Co-Director**
- Elliott, Del (PhD) – Board chair
- Hill, Karl (PhD) – Principal Investigator
- **Ladika, Amanda – Professional research assistant**
- **Michaelson, Laura (PhD) – Research faculty**
- **Mihalic, Sharon – Co-Director**
- Pampel, Fred (PhD) – Research professor
- **Steeger, Christine (PhD) – Research faculty**
- 7 doctoral students from Sociology, Psychology and Education

History of Blueprints

- Began in 1996 with a focus on youth programs to prevent violence, crime, and drug use
- In 2012, expanded its scope to include mental and physical health, self-regulation, and educational achievement outcomes
- Further expanded in 2016 to include a focus on adult crime prevention programs

Plan For Today

Session 1: Overview of “Evidence-Based”

Session 2: Stages of The Blueprints Review Process

Session 3: Unpacking The Blueprints Standards

Session 4: After Blueprints Review

- Blueprints Certification
- Non-Certified Evidence

Summary and Closing Remarks

Blueprints: Overview

At Blueprints, we identify and review studies and reports that test effects of an intervention on positive youth development

The activity, program, policy, or practice intended to produce effects

Changes caused by an intervention

We then summarize our conclusions for those who seek to make evidence-based decisions

Discussion Question #1

What makes a program, practice, or policy "evidence-based"?

Discussion Question #1

What makes a program, practice, or policy "evidence-based"?

Defining "Evidence-Based"

Confusion exists around the term "evidence-based"

Evidence falls on a continuum

For today, two dimensions:

- Stages of evidence
- Types of evidence

Stages of Evidence

Anecdotal
Evidence from focus groups, surveys, opinions, and experiences

Correlational
Evidence of reliable relationships between variables

Causal
Evidence that changes in one variable can be directly attributed to another

Time

Types of Evidence

Anecdotal
Evidence from focus groups, surveys, opinions, and experiences

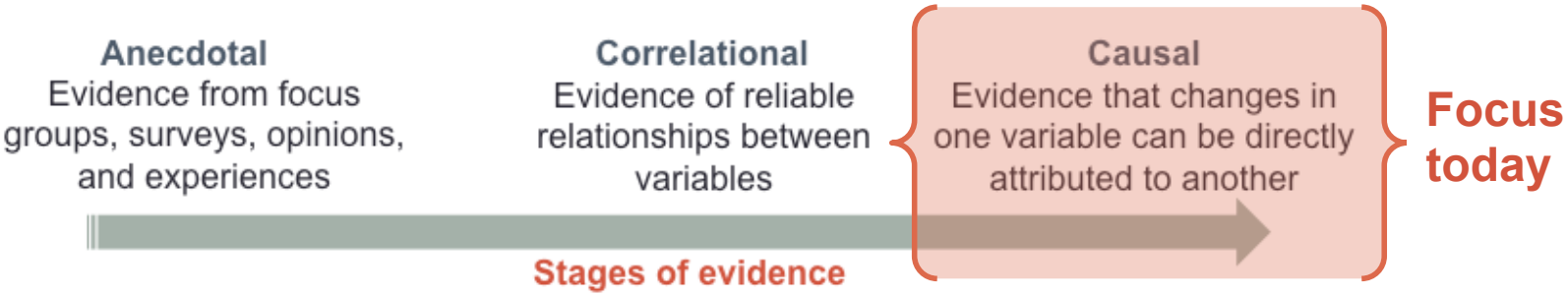
Correlational
Evidence of reliable relationships between variables

Causal
Evidence that changes in one variable can be directly attributed to another

Stages of evidence

National Center for Injury Prevention and Control, Division of Violence Prevention, Centers for Disease Control.

Types of Evidence



National Center for Injury Prevention and Control, Division of Violence Prevention, Centers for Disease Control.

Types of Evidence

Causal evidence is just one part of the larger evidence base

- May be most vulnerable to misunderstanding and misinterpretation

National Center for Injury Prevention and Control, Division of Violence Prevention, Centers for Disease Control.

Why We Prioritize This Evidence

Blueprints also considers factors such as dissemination readiness when determining certification

Today we will discuss our standards for causal evidence

- Internal validity as opposed to external validity

Whether observed changes can be attributed to the intervention

Whether the study can be generalized to other situations and to other people

Two reasons to focus on causal evidence:

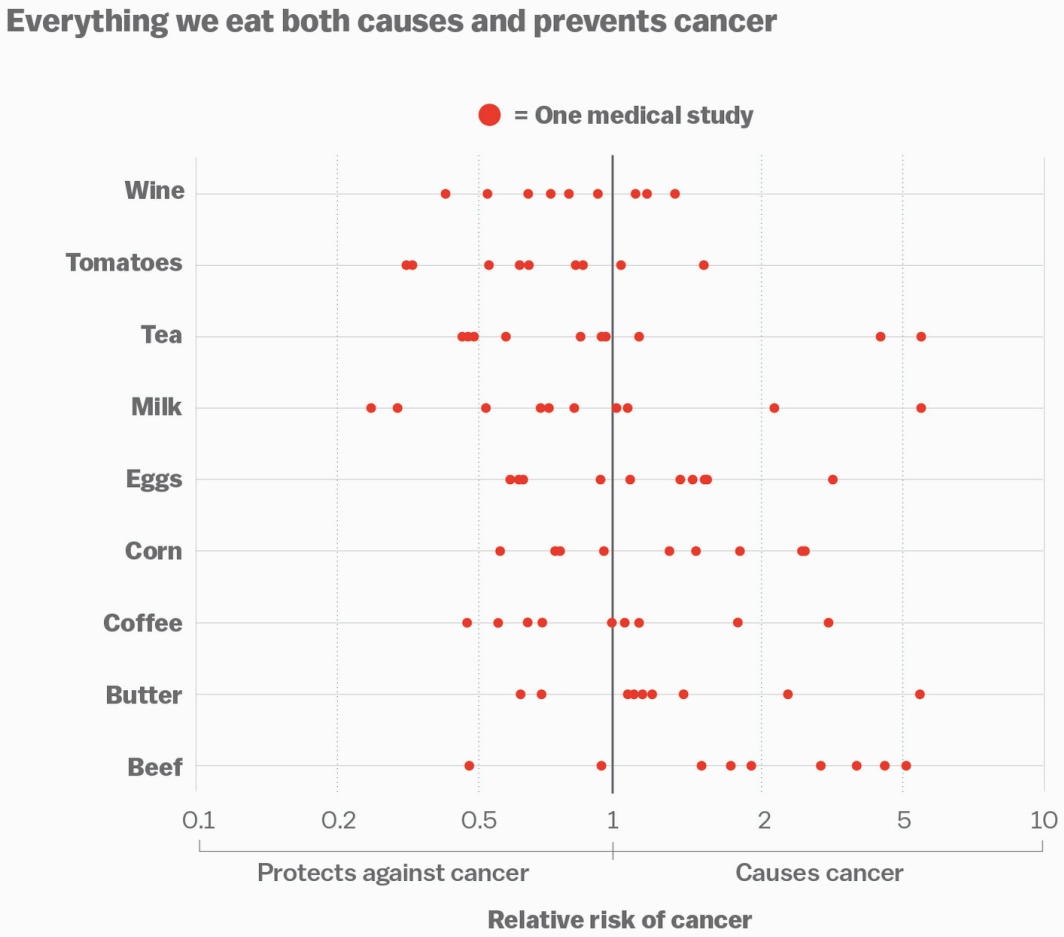
- Different studies produce different findings
- Evidence from a single study is often overblown

Different studies produce different findings

A study of other studies linking common foods to cancer

- Randomly selected 50 common ingredients from cookbook recipes
- Searched medical literature for studies linking those ingredients to cancer prevalence

Need to look at quality of individual studies to know what to believe



Evidence is often overblown

Press and media outlets often portray non-causal evidence as causal

Session 1 Summary

Session 1: Overview of “Evidence-Based”

When a study claims an intervention *caused* positive effects:

- Blueprints judges the ability of that study to produce causal evidence
- This is important because:
  - Different studies produce different findings
  - Evidence is often overblown

Next Up: Session 2

Session 1: Overview of “Evidence-Based”

Session 2: Stages of The Blueprints Review Process

Session 3: Unpacking The Blueprints Standards

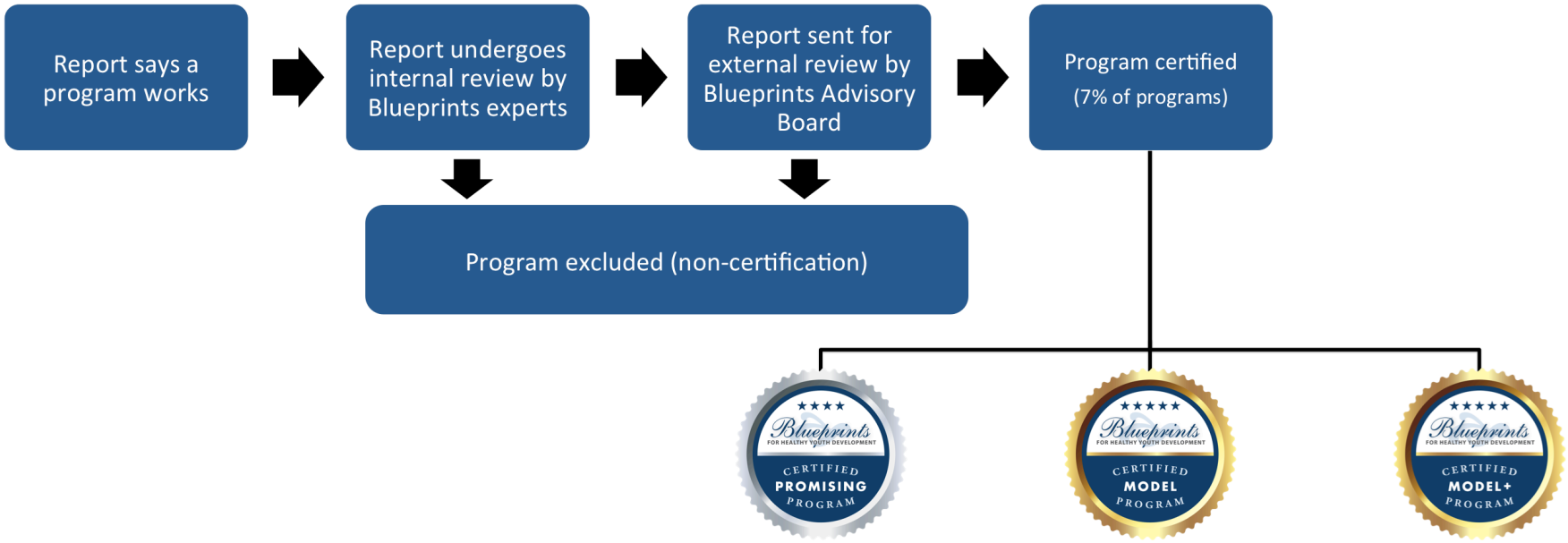
Session 4: After Blueprints Review

- Blueprints Certification
- Non-Certified Evidence

Summary and Closing Remarks

The Blueprints Review Process

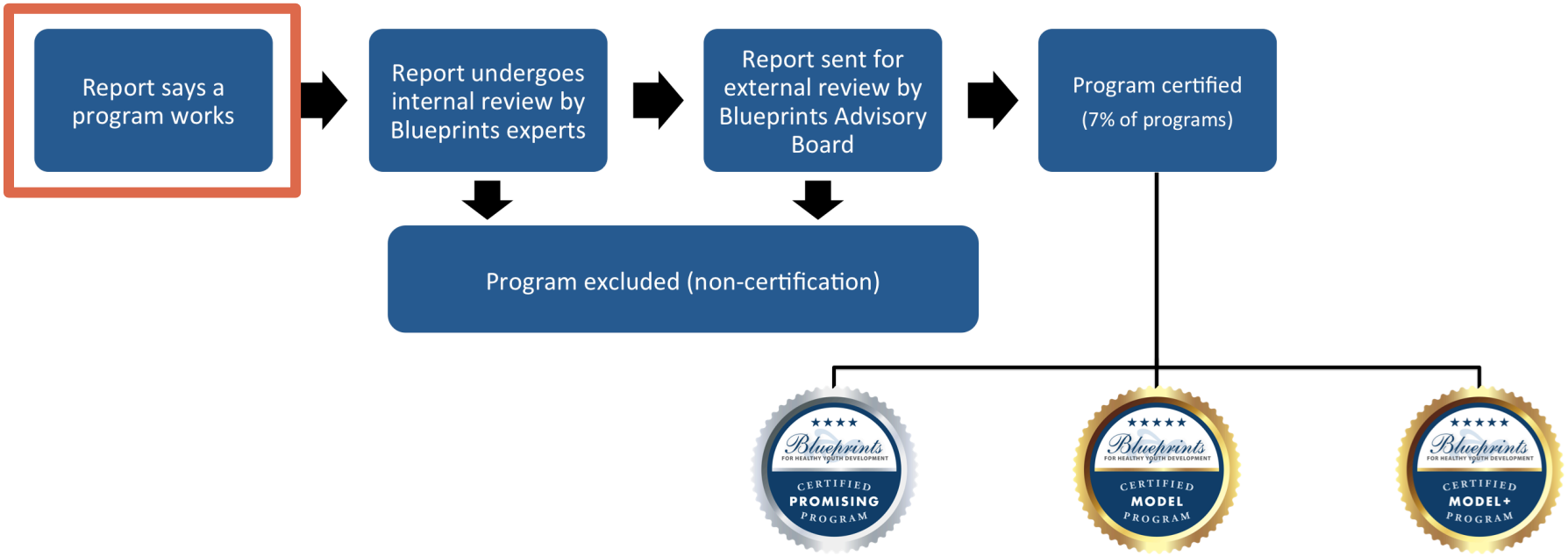
- A review is completed for each eligible study (“report”)
- Internal and external review stages
- Will go over each stage



The Blueprints Review Process

Report is identified for potential review

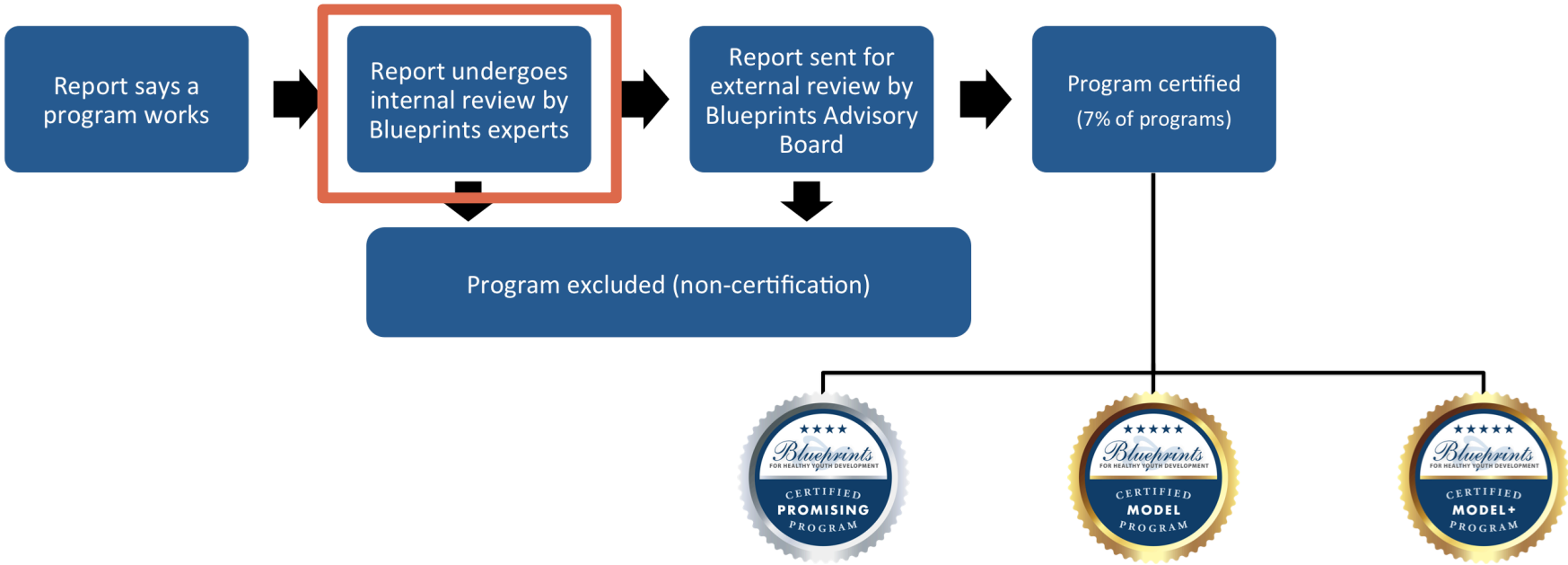
- Literature searches of electronic research databases
- Nominations from the field
- Eligible if group design (treatment, control), Blueprints outcome



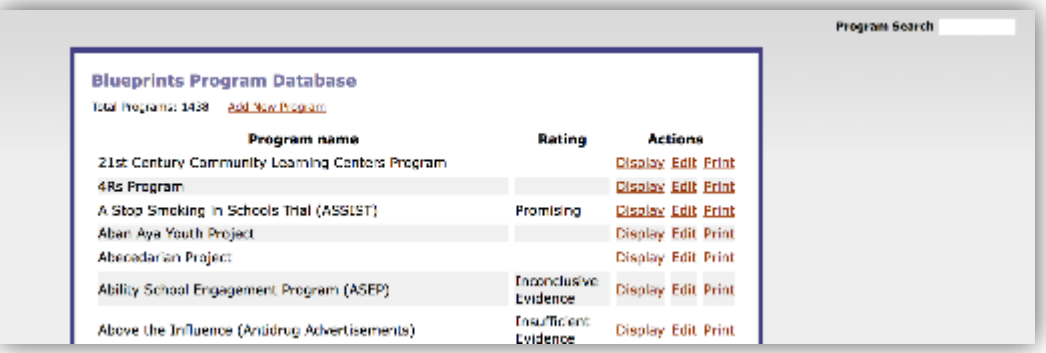
The Blueprints Review Process

Report undergoes internal review

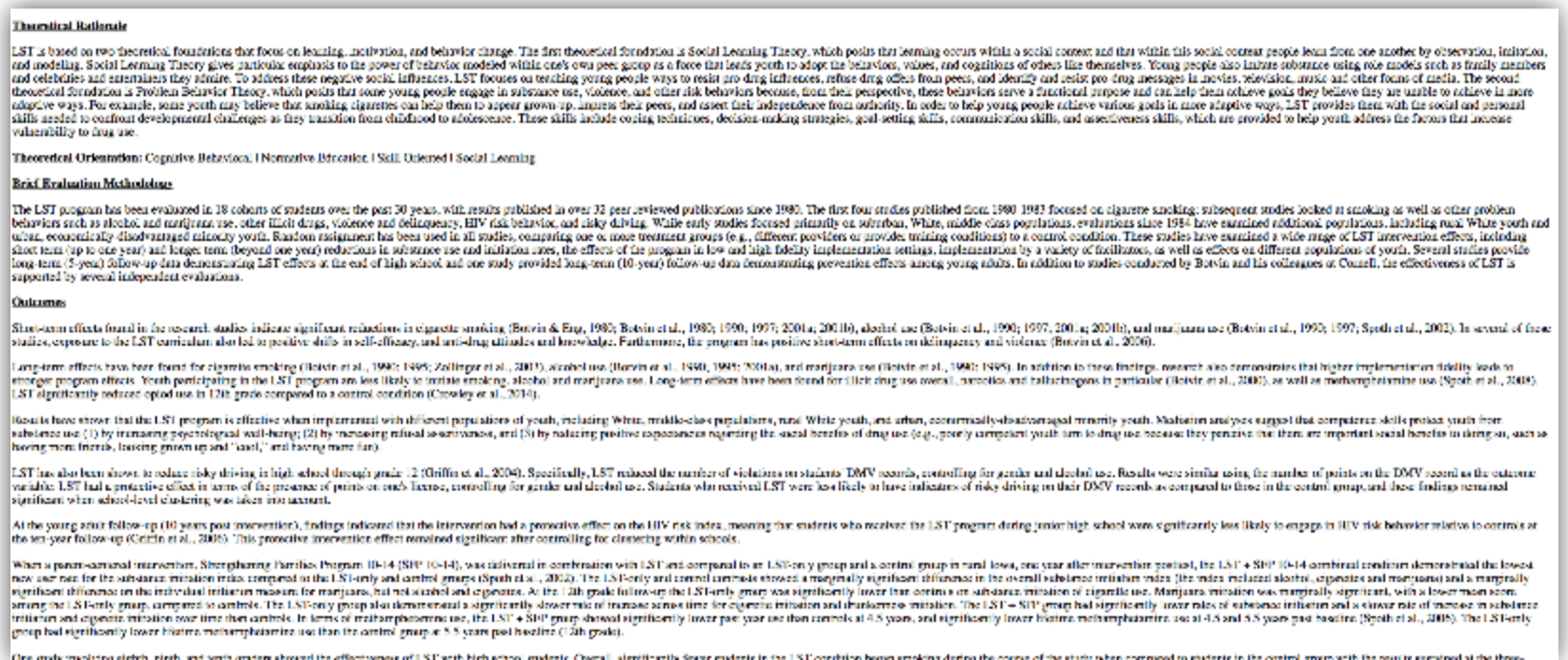
- Dyads of methodological experts trained in Blueprints standards
- Write-up (one for each program)
- Checklist (one for each report)
- Examples of write-up and checklist



Write-up and checklist examples



- Internal database (archive of write-ups for each program)
- One entry for each program; integrates across reports



Theoretical Rationale

LST is based on two theoretical foundations that focus on learning, motivation, and behavior change. The first theoretical foundation is Social Learning Theory, which posits that learning occurs within a social context and that within this social context people learn from one another by observation, imitation, and modeling. Social Learning Theory gives particular emphasis to the power of behavior modeled within one's own peer group as a force that leads youth to adopt the behaviors, values, and cognitions of others like themselves. Young people also imitate substance using role models such as family members and celebrities and entertainers they admire. To address these negative social influences, LST focuses on teaching young people ways to resist pro-drug influences, refuse drug offers from peers, and identify and resist pro-drug messages in movies, television, music and other forms of media. The second theoretical foundation is Problem Behavior Theory, which posits that some young people engage in substance use, violence, and other risk behaviors because, from their perspective, these behaviors serve a functional purpose and can help them achieve goals they believe they are unable to achieve in more adaptive ways. For example, some youth may believe that smoking cigarettes can help them to appear grown-up, impress their peers, and assert their independence from authority. In order to help young people achieve various goals in more adaptive ways, LST provides them with the social and personal skills needed to confront developmental challenges as they transition from childhood to adolescence. These skills include coping techniques, decision-making strategies, goal-setting skills, communication skills, and assertiveness skills, which are provided to help youth address the factors that increase vulnerability to drug use.

Theoretical Orientation: Cognitive Behavioral | Normative Education | Skill Oriented | Social Learning

Brief Evaluation Methodology

The LST program has been evaluated in 18 cohorts of students over the past 30 years, with results published in over 32 peer-reviewed publications since 1980. The first four studies published from 1980–1983 focused on cigarette smoking; subsequent studies looked at smoking as well as other problem behaviors such as alcohol and marijuana use, other illicit drugs, violence and delinquency, HIV risk behavior, and risky driving. While early studies focused primarily on suburban, White, middle-class populations, evaluations since 1984 have examined additional populations, including rural White youth and urban, economically disadvantaged minority youth. Random assignment has been used in all studies, comparing one or more treatment groups (e.g., different providers or provider training conditions) to a control condition. These studies have examined a wide range of LST intervention effects, including short-term (up to one year) and longer-term (beyond one year) reductions in substance use and initiation rates, the effects of the program in low and high fidelity implementation settings, implementation by a variety of facilitators, as well as effects on different populations of youth. Several studies provide long-term (5-year) follow-up data demonstrating LST effects at the end of high school and one study provided long-term (10-year) follow-up data demonstrating prevention effects among young adults. In addition to studies conducted by Botvin and his colleagues at Cornell, the effectiveness of LST is supported by several independent evaluations.

Outcomes

Short-term effects found in the research studies indicate significant reductions in cigarette smoking (Botvin & Eng, 1980; Botvin et al., 1980; 1990, 1997; 2001a; 2001b), alcohol use (Botvin et al., 1990; 1997, 2001a; 2001b), and marijuana use (Botvin et al., 1990; 1997; Spoth et al., 2002). In several of these studies, exposure to the LST curriculum also led to positive shifts in self-efficacy, and anti-drug attitudes and knowledge. Furthermore, the program has positive short-term effects on delinquency and violence (Botvin et al., 2006).

Long-term effects have been found for cigarette smoking (Botvin et al., 1990; 1995; Zollinger et al., 2001), alcohol use (Botvin et al., 1990; 1995; 2001a), and marijuana use (Botvin et al., 1990; 1995). In addition to these findings, research also demonstrates that higher implementation fidelity leads to stronger program effects. Youth participating in the LST program are less likely to initiate smoking, alcohol and marijuana use. Long-term effects have been found for illicit drug use overall, narcotics and hallucinogens in particular (Botvin et al., 2001), as well as methamphetamine use (Spoth et al., 2008). LST significantly reduces opioid use in 12th grade compared to a control condition (Crowley et al., 2014).

Research have shown that the LST program is effective when implemented with different populations of youth, including White, middle-class populations, rural White youth, and urban, economically-disadvantaged minority youth. Mediation analyses support that competence skills protect youth from substance use: (1) by increasing psychological well-being; (2) by increasing refusal self-assurance, and (3) by reducing positive expectancies regarding the social benefits of drug use (e.g., peer's competent youth turn to drug use because they perceive that there are important social benefits in doing so, such as having more friends, looking grown up and "cool," and having more fun).

LST has also been shown to reduce risky driving in high school through grade 12 (Griffin et al., 2004). Specifically, LST reduced the number of violations on students' DMV records, controlling for gender and alcohol use. Results were similar using the number of points on the DMV record as the outcome variable. LST had a protective effect in terms of the presence of points on each license, controlling for gender and alcohol use. Students who received LST were less likely to have indicators of risky driving in their DMV records as compared to those in the control group, and these findings remained significant when school-level clustering was taken into account.

At the young adult follow-up (10 years post intervention), findings indicated that the intervention had a protective effect on the HIV risk index, meaning that students who received the LST program during junior high school were significantly less likely to engage in HIV risk behavior relative to controls at the ten-year follow-up (Griffin et al., 2008). This protective intervention effect remained significant after controlling for clustering within schools.

When a parent-centered intervention, Strengthening Families Program: 10-14 (SFP 10-14), was delivered in combination with LST and compared to an LST-only group and a control group in rural Iowa, one year after intervention protocol, the LST + SFP 10-14 combined condition demonstrated the lowest new use rate for the substance initiation index compared to the LST-only and control groups (Spoth et al., 2002). The LST-only and control curricula showed a marginally significant difference in the overall substance initiation index (the index included alcohol, cigarettes and marijuana) and a marginally significant difference on the individual initiation measure for marijuana, but not alcohol and cigarettes. At the 12th grade follow-up the LST-only group was significantly lower than controls on substance initiation of cigarette use. Marijuana initiation was marginally significant, with a lower mean score among the LST-only group, compared to controls. The LST-only group also demonstrated a significantly slower rate of increase across time for cigarette initiation and drunkenness initiation. The LST + SFP group had significantly lower rates of substance initiation and a slower rate of increase in substance initiation and cigarette initiation over time than controls. In terms of methamphetamine use, the LST + SFP group showed significantly lower past year use than controls at 4.5 years, and significantly lower lifetime methamphetamine use at 4.5 and 5.5 years post baseline (Spoth et al., 2008). The LST-only group had significantly lower lifetime methamphetamine use than the control group at 5.5 years post baseline (12th grade).

Write-up and checklist examples

- One checklist is completed for each study
- #20: Program can be excluded or recommended for external review

Program Name: My Intervention Program (MIP)

Author(s): Michaelson et al. (2018)

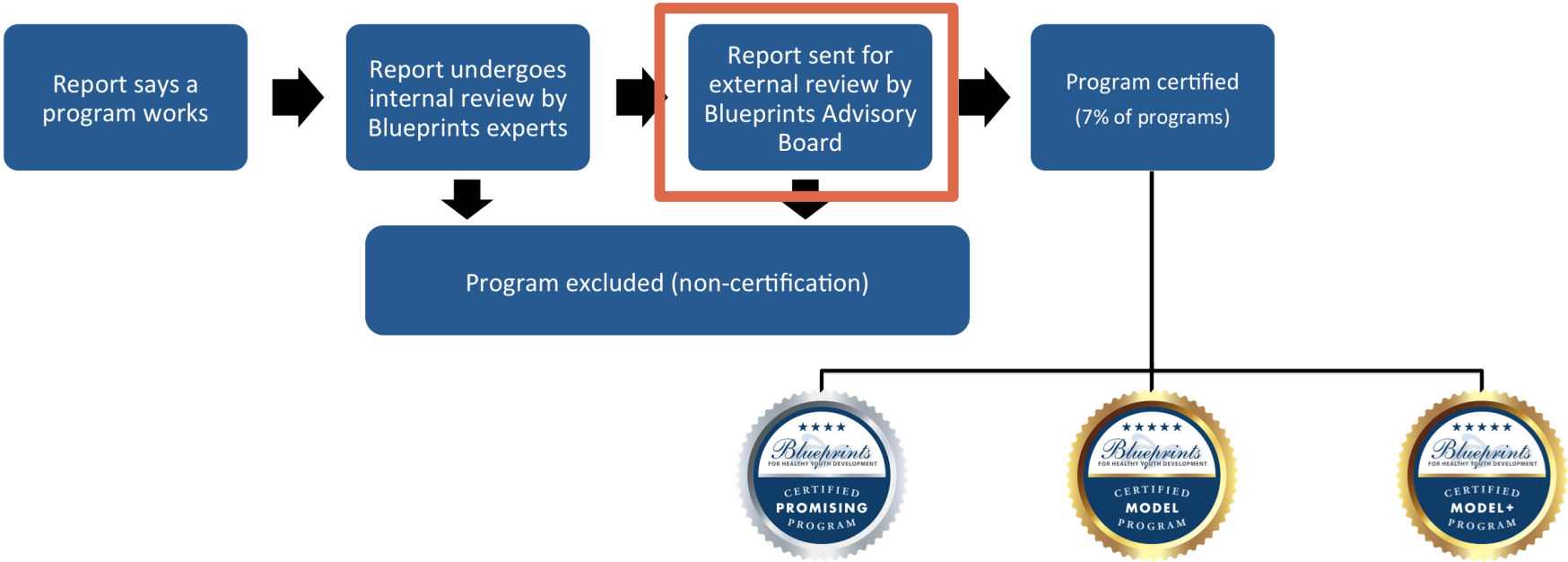
Primary Criteria

Yes	?	No	
☒	☐	☐	*1. High-Quality Design:* Classrooms randomly assigned to conditions
☒	☐	☐	*2. Sample Ns Tracked:*
☒	☐	☐	*3. Measures Independent:* Self-report questionnaires
☒	☐	☐	*4. Measures Valid/Reliable:* Validated measures with high reliability in present study sample
☒	☐	☐	*5. Behavioral Outcome Measure:* Tobacco use
☒	☐	☐	*6. Intent-to-Treat:* Used all participants with complete data
☐	☐	☒	*7. Proper Level:* Randomized classrooms but analyzed individuals
☒	☐	☐	*8. Baseline Outcome Controls:* Included baseline scores as covariate
☒	☐	☐	*9. Baseline Equivalence:* Though only tested baseline equivalence for analysis sample
☐	☐	☒	*10. Differential Attrition Minimal:* Not tested

The Blueprints Review Process

Report undergoes external review

- External Advisory Board (unique to Blueprints)
- Seven methodological experts with variety of content expertise
- Research and professional affiliations around the world



Blueprints Advisory Board

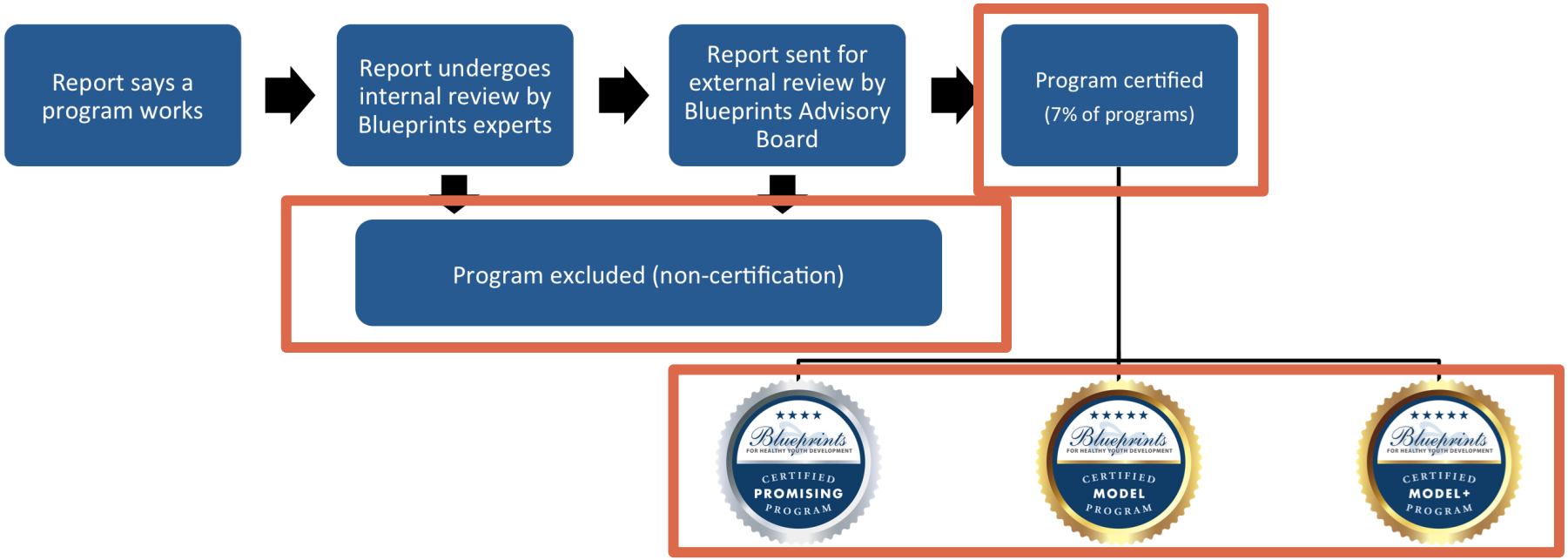
Distinguished board with expertise in research design and methodology from a variety of disciplines:

- **Cook, Thomas D. (PhD)**, Northwestern University
- **Elliott, Delbert (PhD)**, University of Colorado, Boulder
- **Gardner, Frances (PhD)**, University of Oxford
- **Gottfredson, Denise C. (PhD)**, University of Maryland
- **Hawkins, J. David (PhD)**, University of Washington
- **Hedges, Larry (PhD),** Northwestern University
- **Murry, Velma (PhD),** Vanderbilt University
- **Tolan, Patrick (PhD),** University of Virginia

The Blueprints Review Process

Report is certified or not

- Certified at one of three levels
- Excluded (not certified)
  - Classified according to reason for exclusion



Blueprints Review Process: Summary

- Each report scrutinized by multiple methodological experts
- Quality of evidence that program caused its intended effects
- Up to four stages for each report:
  1) A report is identified for potential review
  2) Internal review
     → Exclusion, or…
  3) External review
  4) Report is or is not certified
- Our high standards for making causal claims and the external review stage with distinguished Advisory Board is what makes us unique

Next Up: Session 3

Session 1: Overview of “Evidence-Based”

Session 2: Stages of the Blueprints Review Process

Session 3: Unpacking the Blueprints Standards

Session 4: After Blueprints Review

- Blueprints Certification
- Non-Certified Evidence

Summary and Closing Remarks

Blueprints Standards

If groups are the same at baseline, and nothing changes except the intervention, group differences at posttest can be attributed to the intervention.

- Four main elements considered
  1) Evaluation design
  2) Measurement
  3) Statistical analysis
  4) Group equivalence

Blueprints Standards for Designs

If groups are the same at baseline, and nothing changes except the intervention, group differences at posttest can be attributed to the intervention.

Four main elements considered

1) Evaluation design
2) Measurement
3) Statistical analysis
4) Group equivalence

Two main designs:
1.1) Randomized controlled trials (RCTs)
1.2) Quasi-experimental designs (QEDs)

1) Evaluation Designs

Two main evaluation designs

1.1) Randomized Controlled Trials (RCTs)

- Group assignment to treatment versus control is **random**

1.2) Quasi-Experimental Designs (QEDs)

- Group assignment to treatment versus control is **not random**

- There are also non-group designs (within-group pre/post comparison)
- Not reviewed by Blueprints, but important for building an evidence base

1.1) Randomized Controlled Trials

A random process is used to assign units to groups

- Coin toss, random number generator

Units can include:

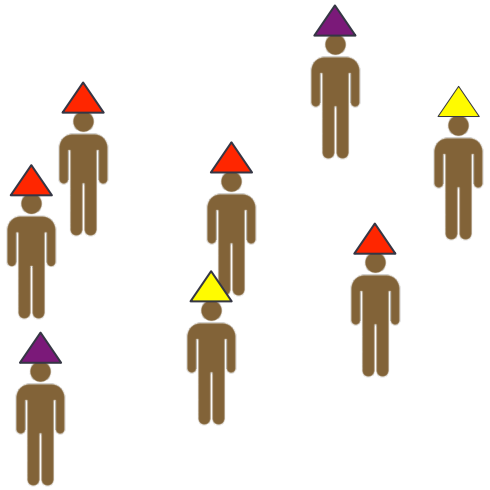
- Individuals (students, teachers)

- Clusters of individuals (classrooms, schools)

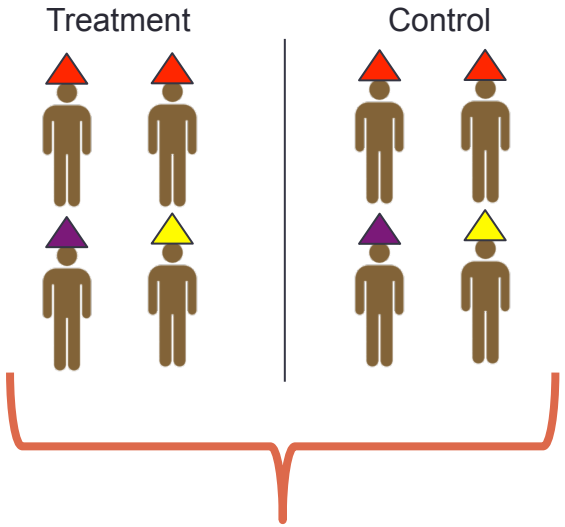
Randomization Creates Similar Groups

- If units in a study sample are randomly assigned, randomization should create similar groups

Study sample of individuals



Random assignment to conditions



Group differences here would be entirely random

If groups are the same at baseline, and nothing changes except the intervention, group differences at posttest can be attributed to the intervention.

1) Evaluation Designs

Two main evaluation designs

1.1) Randomized Controlled Trials (RCTs)

- Group assignment to treatment versus control is **random**

1.2) Quasi-Experimental Designs (QEDs)

- Group assignment to treatment versus control is **not random**

1.2) Quasi-Experimental Designs

Assignment to treatment versus control is **not** random

Researcher controls the assignment using some criterion other than random assignment (volunteering for a treatment, eligibility for a voucher, etc.)

Concerns regarding internal validity

- Treatment and control groups may not be comparable at baseline

QEDs and Internal Validity

To infer X (treatment) causes Y (outcome)

1. X must precede Y in time
2. X <-> Y must be related to each other
3. All other alternative explanations are eliminated through random assignment or experimental control

Here's an example of this concept

Example

Research Question:

Do students who take Advanced Placement (AP) courses in high school *(treatment group, or "X")* graduate from high school at higher rates *(outcome, or "Y")* than students who do not take AP courses (control group)?

If conducting a QED:

- Can "control" for baseline differences related to graduation
  - Achievement
  - Socio-demographic characteristics, etc.
- Cannot "control" for whether students who take AP are *more motivated* in school than students who do not take AP

Does "motivation" or "taking AP" improve high school graduation rates?

QEDs, continued

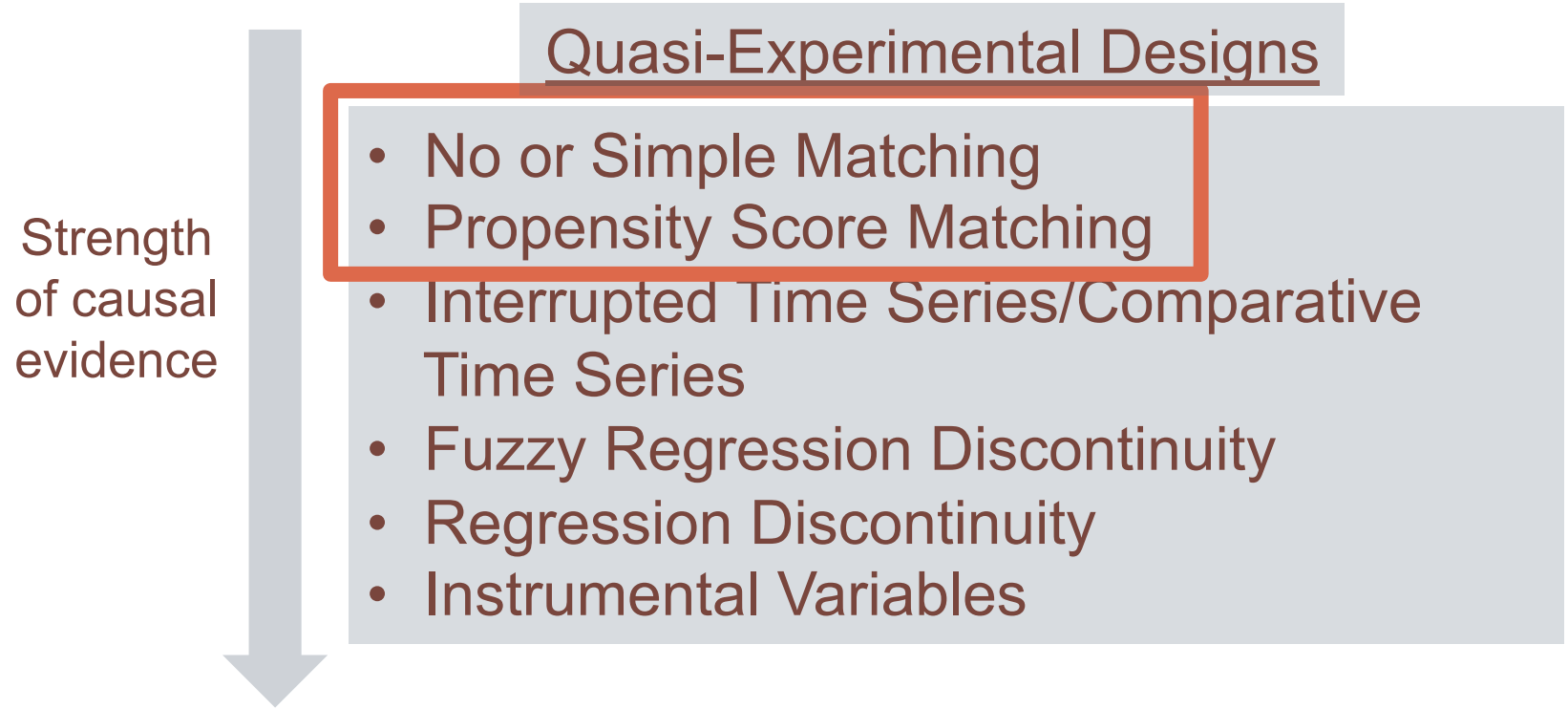
With QEDs, you can't rule out ALL alternative explanations, but you can try to minimize them

The extent to which a QED can eliminate possible threats to internal validity determines its usefulness

Continuum of QEDs: Limited to Better

Some QEDs are more internally valid than others

Vary in their credibility in providing causal evidence



Matching-Based QEDs

No matching (convenience sample)

- Example: first 20 participants who sign up will receive the treatment, everyone else will be waitlist controls

Simple matching

- Seeks to match each treatment unit to a comparison unit with similar characteristics

Statistical (“propensity score”) matching

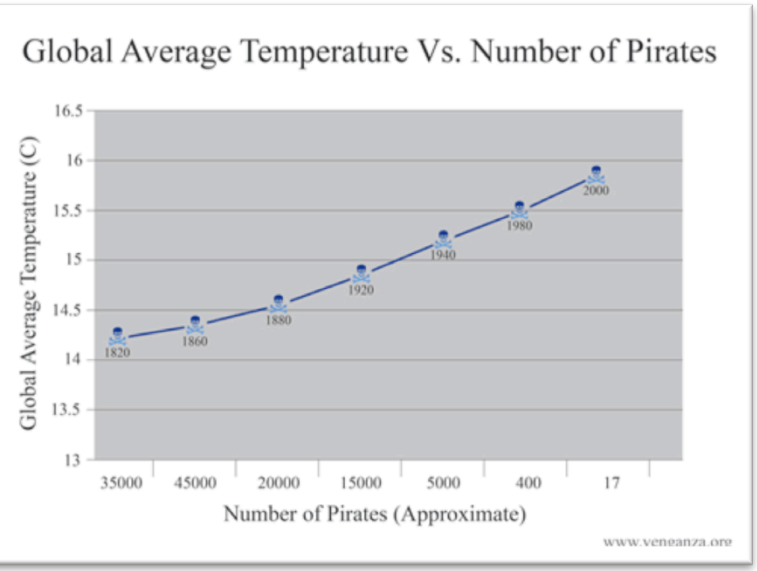
- Seeks to match each treatment with a “statistical twin” for comparison

QEDs and Causal Evidence

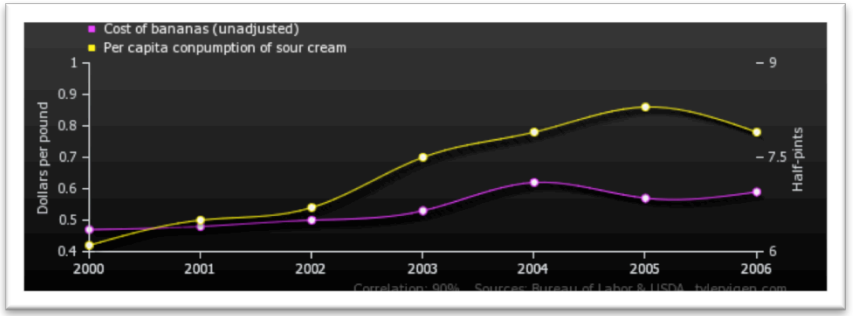
Sometimes, causal interpretations of correlational evidence are obviously absurd

- Example

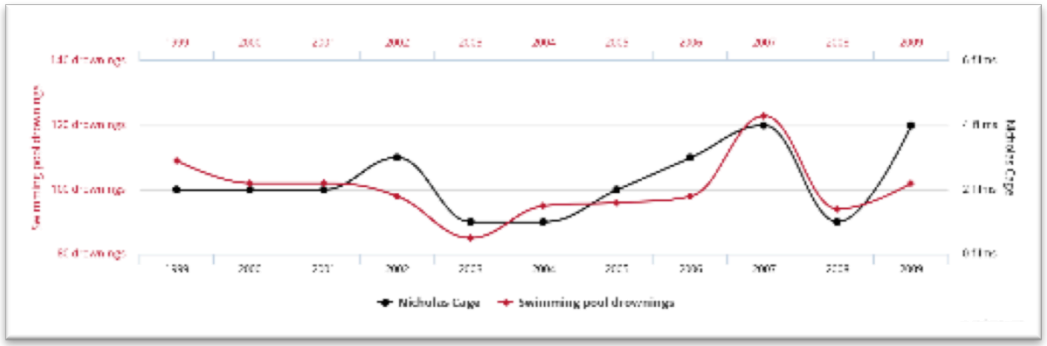
Absurd Causal Conclusions



Pirates cause global warming



When bananas are too expensive, people opt for sour cream



Watching Nicholas Cage movies makes people drown in their swimming pools

QEDs and Causal Evidence

Sometimes, causal interpretations of correlational evidence are obviously absurd

- Example

Other times, causal interpretations are more reasonable

Discussion Question #2

Teachers attending a social-emotional learning seminar were invited to test a school-based social-emotional curriculum in their classrooms. Researchers used sophisticated statistical techniques to identify matched comparison classrooms for each classroom in the treatment group.

At the end of the social-emotional learning program, treatment classrooms had fewer disciplinary referrals than control classrooms.

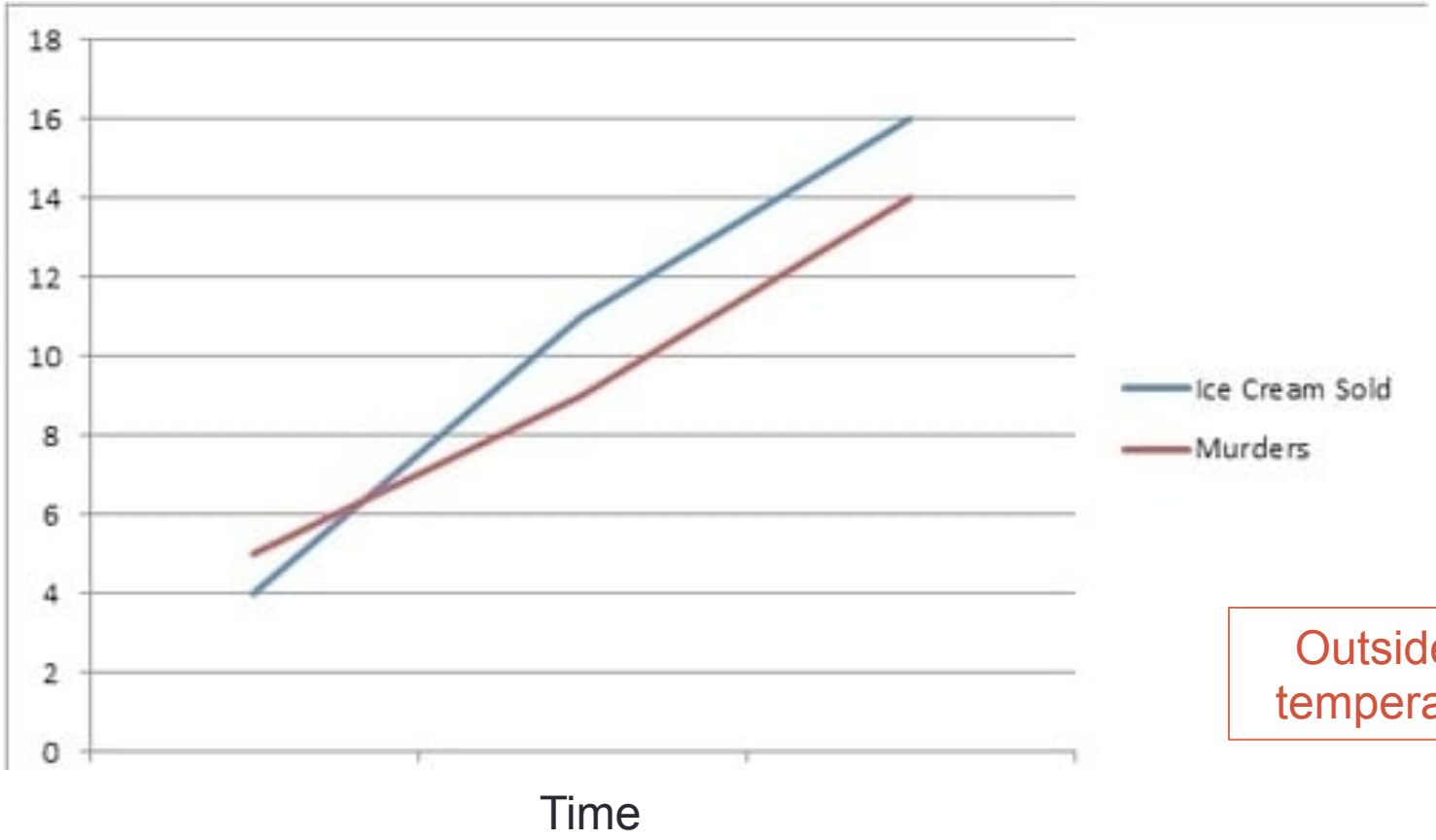
2a) Would it be reasonable to conclude that the social-emotional learning program caused disciplinary improvements?

2b) Other causal explanations?

Discussion Question #3

Positive correlation between ice cream and murder.

What third variable might be driving this correlation?



Outside air temperature!

Summary: Quasi-Experimental Designs

- Results from QEDs can be tricky to interpret
- So why do QEDs?
- Sometimes QEDs are necessary
  - Randomized trial is highly impractical or expensive
  - Unethical to assign to conditions
- QEDs are part of building an evidence base
  - Will touch on this more later

Exercise A

Question #1

At the beginning of the school year, 60 students are randomly assigned to receive a pull-out reading intervention, while 60 other students receive the normal curriculum. What evaluation design does this study employ?

A) A randomized controlled trial (RCT)
B) A quasi-experimental design (QED)
C) A within-study/no control group

Exercise A

Question #2

All second-grade classrooms in a school are participating in a study of a new math curriculum. Half of the teachers volunteer to use the new curriculum, while the other half use the standard curriculum. What evaluation design does this study employ?

A) A randomized controlled trial (RCT)
B) A quasi-experimental design (QED)
C) A within-study/no control group

Exercise A

Question # 3

Children were eligible to participate in a school-based reading program based on their standardized test scores. Among those who were eligible, children were assigned to receive the reading program if they did not have conflicts with other enrichment classes, and those who had conflicts made up the control group. What evaluation design does this study employ?

A) A randomized controlled trial (RCT)
B) A quasi-experimental design (QED)
C) A within-study/no control group

Exercise A

Question # 4

All clinics eligible for a health intervention were classified as either rural or urban based on their geographic location. Four urban and four rural clinics were randomly selected and agreed to participate in the evaluation. Within each type of geographic location (rural and urban), clinics were randomly assigned to treatment or control conditions. What evaluation design does this study employ?

A) A randomized controlled trial (RCT)
B) A quasi-experimental design (QED)
C) A within-study/no control group

Exercise A (Review)

Question #1

At the beginning of the school year, 60 students are **randomly assigned** to receive a pull-out reading intervention, while 60 other students receive the normal curriculum. What evaluation design does this study employ?

A) A randomized controlled trial (RCT)

B) A quasi-experimental design (QED)

C) A within-study/no control group

Exercise A (Review)

Question #2

All second-grade classrooms in a school are participating in a study of a new math curriculum. **Half** of the teachers **volunteer** to **use** the new **curriculum**, while the **other half use the standard curriculum.** What evaluation design does this study employ?

A) A randomized controlled trial (RCT)
B) A quasi-experimental design (QED)
C) A within-study/no control group

Exercise A (Review)

Question # 3

Children were eligible to participate in a school-based reading program based on their standardized test scores. Among those who were eligible, children were **assigned to** receive the reading **program** if they **did not have conflicts** with other enrichment classes, **and those who had conflicts made up the control group**. What evaluation design does this study employ?

A) A randomized controlled trial (RCT)
B) A quasi-experimental design (QED)
C) A within-study/no control group

Exercise A (Review)

Question # 4

All clinics eligible for a health intervention were classified as either rural or urban based on their geographic location. Four urban and four rural clinics were randomly selected and agreed to participate in the evaluation. Within each type of geographic location (rural and urban), **clinics were randomly assigned to treatment or control conditions**. What evaluation design does this study employ?

A) A randomized controlled trial (RCT)

B) A quasi-experimental design (QED)

C) A within-study/no control group

Bonus:
What kind of experimental design?

Blueprints Standards for Measurement

Four main elements considered

1) Evaluation design
2) Measurement
3) Statistical analysis
4) Group equivalence

Measures must be three things:
2.1) Well-established
2.2) Independent
2.3) Behavioral

2) Measurement

Measures must be:

2.1) Well-established

- Reliable and valid

Reliability:
Whether a measure gives similar results each time it is used

Validity:
Whether a measure reflects what it is intended to measure

2.2) Independent

2.3) Behavioral

2) Measurement

Measures must be:

2.1) Well-established

2.2) Independent

- Person delivering the program is not providing the assessment
    - Could be biased due to expectations, beliefs, social desirability

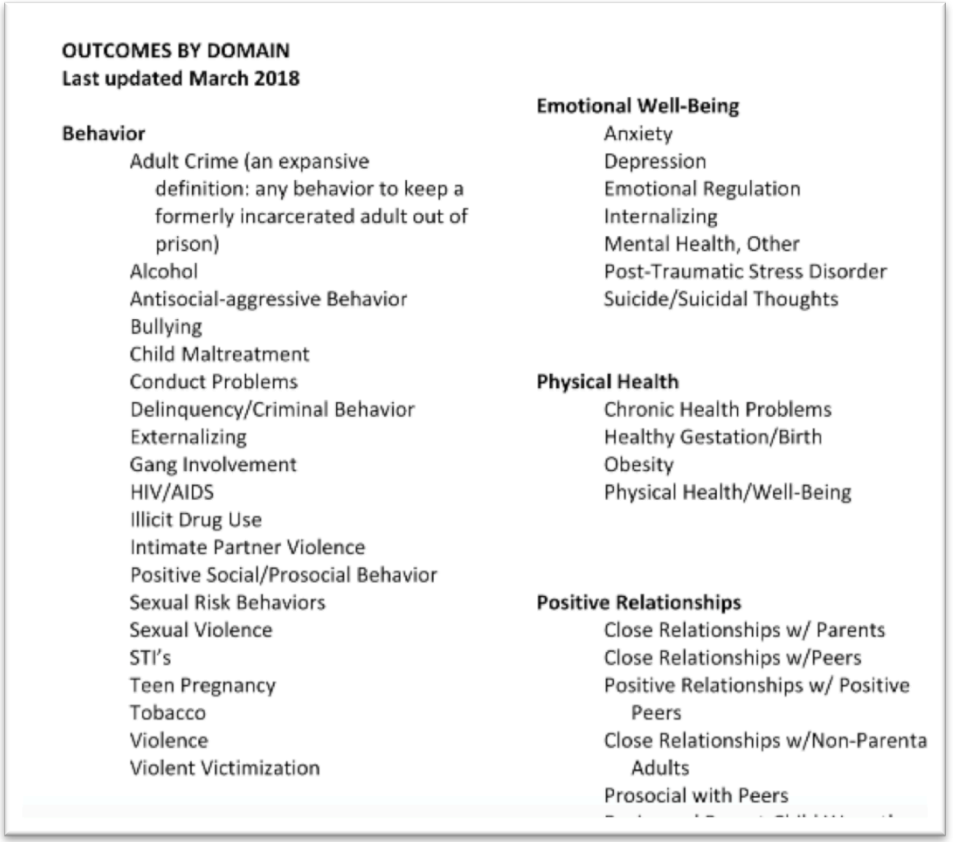
2.3) Behavioral

2) Measurement

Measures must be:

2.1) Well-established

2.2) Independent



OUTCOMES BY DOMAIN
Last updated March 2018

Behavior
- Adult Crime (an expansive definition: any behavior to keep a formerly incarcerated adult out of prison)
- Alcohol
- Antisocial-aggressive Behavior
- Bullying
- Child Maltreatment
- Conduct Problems
- Delinquency/Criminal Behavior
- Externalizing
- Gang Involvement
- HIV/AIDS
- Illicit Drug Use
- Intimate Partner Violence
- Positive Social/Prosocial Behavior
- Sexual Risk Behaviors
- Sexual Violence
- STI's
- Teen Pregnancy
- Tobacco
- Violence
- Violent Victimization

Emotional Well-Being
- Anxiety
- Depression
- Emotional Regulation
- Internalizing
- Mental Health, Other
- Post-Traumatic Stress Disorder
- Suicide/Suicidal Thoughts

Physical Health
- Chronic Health Problems
- Healthy Gestation/Birth
- Obesity
- Physical Health/Well-Being

Positive Relationships
- Close Relationships w/ Parents
- Close Relationships w/Peers
- Positive Relationships w/ Positive Peers
- Close Relationships w/Non-Parenta Adults
- Prosocial with Peers

2.3) Behavioral

- Must be on the list of Blueprints behavioral outcomes
- Includes self-reports of behaviors

Exercise A

Question # 5

Researchers studying a sexual education program administered two outcome measures: a conventional risk aversion survey (commonly included in sex education studies), and a questionnaire on risky sexual behaviors created by the researchers. They did not report reliability or validity, but stated that the procedure minimized the potential for social desirability bias.

True or False:

Blueprints considers both of these to be established measures

Exercise A

Question # 6

Sessions for a six-month parenting intervention are delivered by clinically-licensed practitioners with expertise in education, social work, or counseling. Teachers and parents report on children's oppositional behavior at baseline and posttest.

True or False:

Blueprints considers this an independent measure

Exercise A

Question # 7

A substance use intervention program measured three aspects of alcohol use: peer prevalence of alcohol use, attitudes towards drinking, and intentions to use alcohol.

True or False:

If these were the only outcome measures included, this study would qualify for Blueprints certification.

Exercise A (Review)

Question # 5

Researchers studying a sexual education program administered two outcome measures: a conventional risk aversion survey (commonly included in sex education studies), and a questionnaire on risky sexual behaviors created by the researchers. **They did not report reliability or validity**, but stated that the procedure minimized the potential for social desirability bias.

True or False:

Blueprints considers this an established measure

Exercise A (Review)

Question # 6

Sessions for a six-month parenting intervention are **delivered by** clinically-licensed **practitioners** with expertise in education, social work, or counseling. **Teachers** and **parents report** on children's oppositional behavior at baseline and posttest.

True or False:

Blueprints considers this an independent measure

Exercise A (Review)

Question # 7

A substance use intervention program measured three aspects of alcohol use: **peer prevalence of alcohol use, attitudes towards drinking, and intentions to use alcohol**.

True or False:

If these were the only outcome measures included, this study would qualify for Blueprints certification.

3) Statistical Analysis

Four main elements considered

1) Evaluation design
2) Measurement
3) Statistical analysis
4) Group equivalence

3.1) Proper level
3.2) Intent-to-treat

3) Statistical Analysis

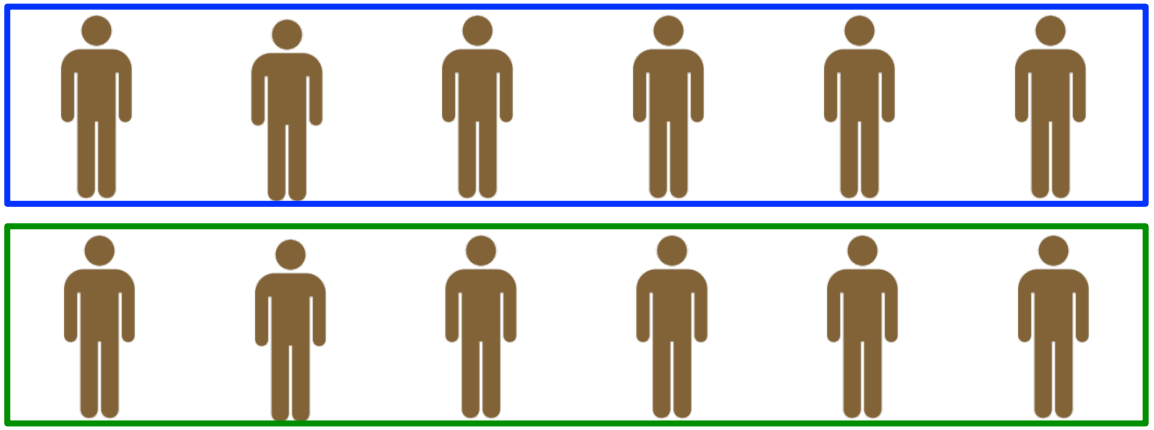
3.1) Proper level

- Must adjust statistically if there are clusters of individuals
- Usually with multilevel modeling
  - Example

3.2) Intent-to-treat

3.1) Proper Level of Analysis

Example



Treatment
N = 6

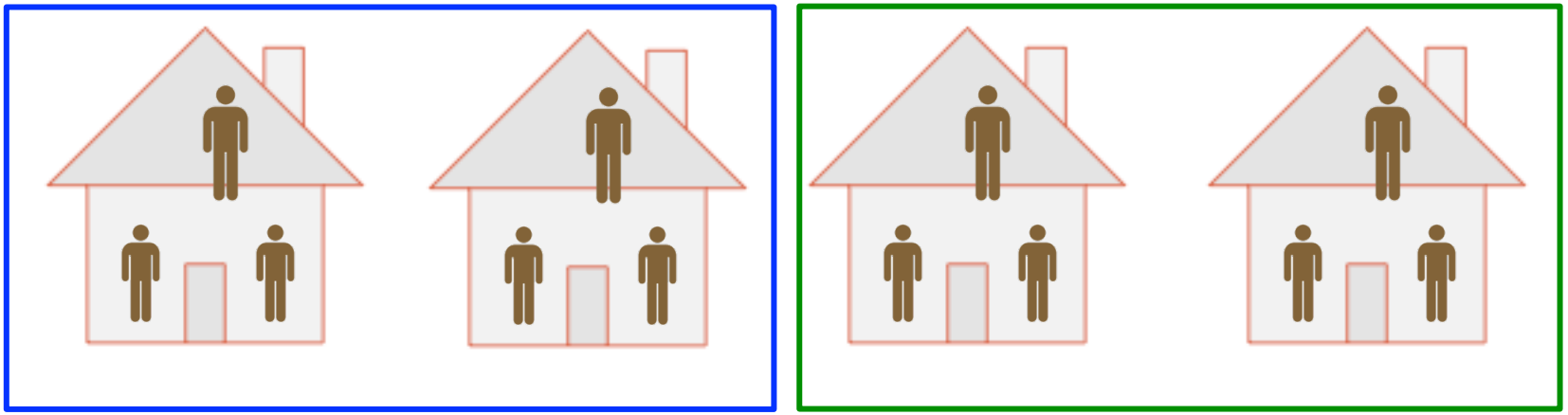
Control
N = 6

Units for analysis
N = 12

- 12 individuals
- No clusters
- Each **individual** assigned to one of two groups

3.1) Proper Level of Analysis

Example



Treatment
N = 2

Control
N = 2

Units for analysis
N = 4

- 12 individuals
- Clustered within 4 schools
- Each school assigned to one of two groups

3.1) Proper Level of Analysis

Multilevel models (AKA hierarchical linear models) are the statistical models that "adjust" for clustering

Units of analysis are usually individuals (at a lower level) who are nested within higher-level units such as classrooms or schools

Example

- Model of student performance that contains achievement measures for individual students as well as achievement measures for classrooms

3) Statistical Analysis

3.1) Proper level

- Must adjust statistically if there are clusters of individuals
- Names of statistical tests for cluster RCTs (groups assigned to condition)
  - Multilevel modeling
  - Hierarchical linear modeling (HLM)
- Names of statistical tests for RCTs (individuals assigned to condition)
  - ANCOVA
  - MANCOVA
  - Linear or Logistic Regression

3.2) Intent-to-treat

3.2) Intent-to-Treat (ITT)

Analyze all units according the group to which they were assigned, or the treatment they were *intended* to receive, no matter what happens

- What might happen?
  - Units might change conditions
  - Subjects could receive some, but not all, of the treatment
  - Subjects could show up for some assessments but not others

Example Test of ITT

- Classrooms randomly assigned to intervention or control conditions
- Justin is in a classroom assigned to the intervention condition
- The principal later moves Justin to a control classroom

If we were analyzing according to ITT, how should Justin be analyzed?

- According to his original assignment: the intervention condition
- Otherwise, randomization is compromised

Exercise A

Question # 8

A four-session group intervention designed to prevent the onset of eating disorders was evaluated in which a total of 148 female students were randomized to treatment (n=74) or waitlist control (n=74). Data were collected at baseline and post-intervention. An ANOVA was used to test differences between groups in outcomes from the pre to the post-test.

True or False:

According to Blueprints' standards, this analysis was conducted at the proper level

Exercise A

Question # 9

A school-wide anti-bullying program was evaluated by assigning 30 schools to either receive the program (n = 15 treatment schools) or to a control group (n = 15 schools) that did not receive the program. The analysis used multilevel models with students nested within schools to test whether behavior incidents, suspensions and expulsion rates from before the intervention to after the intervention were lower at the treatment schools compared to the control schools.

True or False:

According to Blueprints' standards, this analysis was conducted at the proper level

Exercise A

Question # 10

In a recidivism study, offenders in the treatment group were divided into two subgroups according to whether or not they completed the intervention. Each subgroup was compared to the control group to test for treatment effects.

True or False:

According to Blueprints' standards, this analysis violates intent-to-treat protocol

Exercise A (Review)

Question # 8

A four-session group intervention designed to prevent the onset of eating disorders was evaluated in which a total of **148 female students were randomized to treatment** (n=74) **or waitlist control** (n=74). Data were collected at baseline and post-intervention. An **ANOVA** was used to test differences between groups in outcomes from the pre to the post-test.

True or False:

According to Blueprints' standards, this analysis was conducted at the proper level

Exercise A (Review)

Question # 9

A school-wide anti-bullying program was evaluated by **assigning 30 schools** to either receive the program (n = 15 treatment schools) or to a control group (n = 15 schools) that did not receive the program. The analysis used **multilevel models** with students nested within schools to test whether behavior incidents, suspensions and expulsion rates from before the intervention to after the intervention were lower at the treatment schools compared to the control schools.

True or False:

According to Blueprints' standards, this analysis was conducted at the proper level

Exercise A (Review)

Question # 10

In a recidivism study, offenders in the treatment group were divided into two subgroups according to whether or not they completed the intervention. **Each subgroup was compared to the control group** to test for treatment effects.

True or False:

According to Blueprints' standards, this analysis violates intent-to-treat protocol

4) Group Equivalence

Four main elements considered

1) Evaluation design
2) Measurement
3) Statistical analysis
4) Group equivalence
    - 4.1) Baseline equivalence
    - 4.2) Attrition

4.1) Baseline Equivalence

"Baseline" refers to the pre-test (i.e., pre-treatment) assessments

Critical for causal conclusions

Reminder:

If groups are the same at baseline, and nothing changes except the intervention, group differences at posttest can be attributed to the intervention.

4.1) Baseline Equivalence

Blueprints requires that even in randomized designs, baseline equivalence must be tested and reported

Ideally, two sets of tests for baseline equivalence:

- Assigned sample (original sample of units that were assigned to conditions)
- Analysis sample (final sample of units available for analysis from each condition, after missing data and attrition)

Next up: Attrition

4.2) Attrition

Attrition

- The loss of participants from the beginning to the end of the study, resulting in a reduced sample size.

Differential attrition – Attrition that is selective (2 levels)
Characteristics systematically differ between

- "Attritors" (drop out) and "completers" (retained)
- "Completers" in treatment vs. control

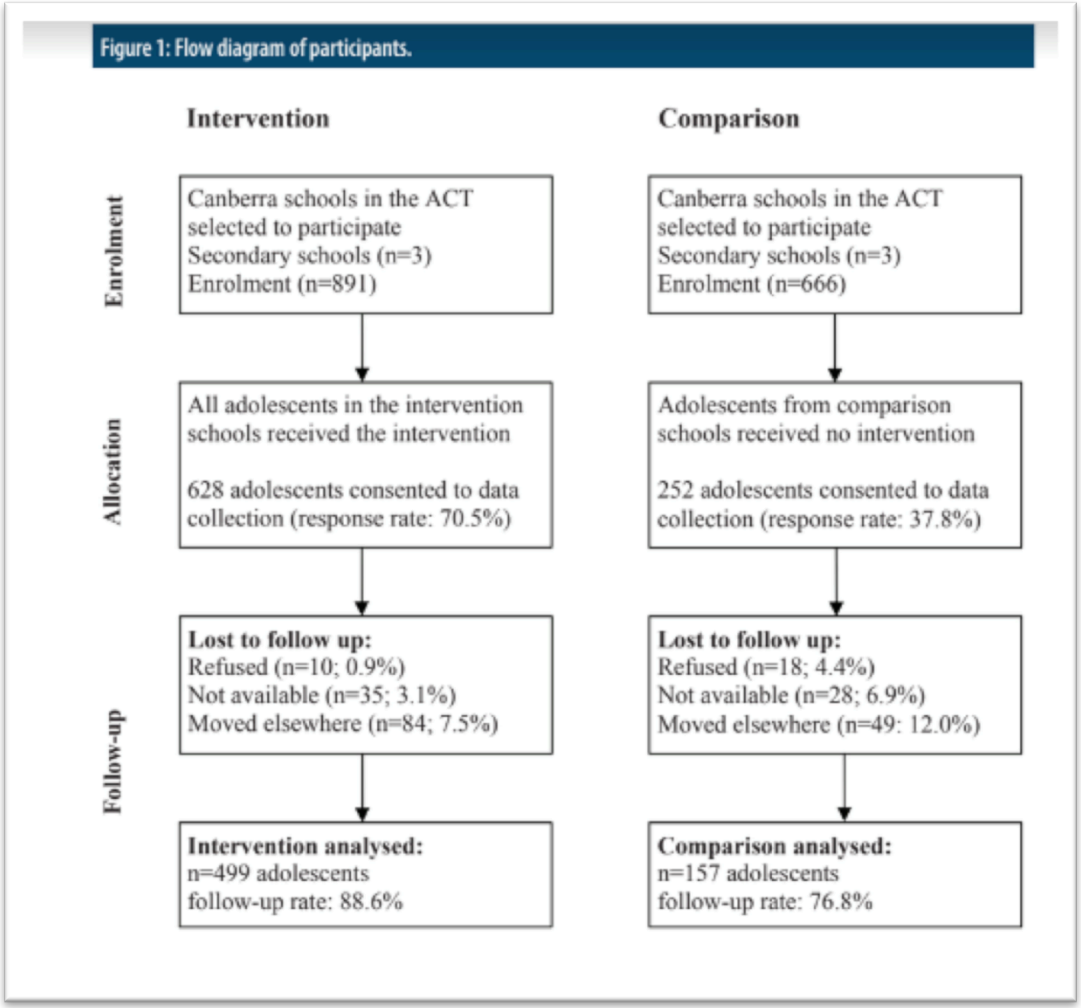
Blueprints threshold: 5%

- If overall attrition is less than 5%, not concerned about differential attrition
- Otherwise, must report tests

4.2) Attrition

Where we commonly look for the information needed to evaluate attrition:

- Example
- Flow chart or CONSORT diagram



Source: Figure 1, Malakellis et al., 2017

Exercise A

Question # 11

Researchers evaluating a cognitive training program found no significant differences on the pretest measures between participants who were randomly assigned to treatment versus control groups. Additionally, the treatment group was equivalent to the control group on all demographic variables, except maternal education, with the treatment group having lower levels than the control group. However, mother's education had no significant relationship with any of the outcome measures.

True or False:

This description satisfies the Blueprints standard for baseline equivalence

Exercise A

Question # 12

Researchers assigned 1,606 participants to conditions. By the end of the intervention, they had complete pre- and post-test assessment from 1,002 participants. The retained sample (n=1,002) was significantly different from the non-retained sample (n=604) on one variable: they had lower baseline gateway drug use scores.

This is an example of:

A) Differential attrition (attritors vs. completers)

B) Differential attrition-by-condition (completers in the treatment group vs. completers in the control group)

Exercise A

Question # 13

In a study on the long-term effects of a drug prevention program, complete data were available for 1,105 students (69% of the originally assigned sample). Attrition rates, drug use, and socio-demographic characteristics among students lost to follow-up did not differ between treatment and control schools.

True or False: Blueprints would require tests of differential attrition in this study

Exercise A (Review)

Question # 11

Researchers evaluating a cognitive training program found **no significant differences on the pretest measures** between participants who were randomly assigned to treatment versus control groups. Additionally, the treatment group was equivalent to the control group on all demographic variables, **except maternal education**, with the treatment group having lower levels than the control group. However, **mother's education had no significant relationship with any of the outcome measures**.

True or False:

This description satisfies the Blueprints standard for baseline equivalence

Exercise A

Question # 12

Researchers assigned 1,606 participants to conditions. By the end of the intervention, they had complete pre- and post-test assessment from 1,002 participants. **The retained sample** (n=1,002) **was significantly different from the non-retained sample** (n=604) on one variable: they had lower baseline gateway drug use scores.

This is an example of:

A) Differential attrition (attritors vs. completers)

B) Differential attrition-by-condition (attritors in the treatment group vs. attritors in the control group; completers in the treatment group vs. completers in the control group)

Exercise A

Question # 13

In a study on the long-term effects of a drug prevention program, complete data were available for 1,105 students (**69% of the originally assigned sample**). Attrition rates, drug use, and socio-demographic characteristics among students lost to follow-up did not differ between treatment and control schools.

True or False: Blueprints would require tests of differential attrition in this study

Session 3: Summary

Session 1: Overview of “Evidence-Based”

Session 2: Stages of The Blueprints Review Process

Session 3: Unpacking the Blueprints Standards

- Our core standard
- Four main elements considered
  1) Evaluation design
  2) Measurement
  3) Statistical analysis
  4) Group equivalence

If groups are the same at baseline, and nothing changes except the intervention, group differences at posttest can be attributed to the intervention.

Next Up: Session 4

Session 1: Overview of “Evidence-Based”

Session 2: Stages of the Blueprints Review Process

Session 3: Unpacking the Blueprints Standards

Session 4: After Blueprints Review

- Blueprints Certification
- Non-Certified Evidence

Summary and Closing Remarks

Plan for Afternoon

- What happens after Blueprints review:
  - Certifications
  - Non-certified studies and reasons for exclusion

Blueprints Certification

Only 81 of the 1400+ programs reviewed

4 certification standards:

1. Intervention specificity
   - Outcome(s)
   - R&P factors targeted to produce outcome change (if relevant)
   - Population
   - Program components
2. Evaluation quality (see next slide)
3. Intervention impact
   - Positive change
   - No harmful effects
4. Dissemination readiness
   - Capacity and materials
   - Implementation with fidelity

Blueprints Certification – Evaluation Quality

Promising

At least 1 high-quality RCT or 2 high-quality QEDs suggest the program is effective

Model

2 high quality RCTs, or 1 high quality RCT and 1 high quality QED, with effects sustained for 12+ months after the intervention ended

Model Plus

Meets all criteria for Model and includes at least one independent evaluation

Plan for Afternoon

- What happens after Blueprints review:
  - Certifications
  - Non-certified studies and reasons for exclusion

Non-certified studies

In 2016, we received funding from the Laura & John Arnold Foundation to

- Classify non-certified programs
- Code reasons for exclusion

Four classifications for non-certified programs

- Inconclusive
- Insufficient
- Ineffective
- Harmful

Inconclusive

Missing information or incomplete analyses:

- Attrition not reported
- No info on reliability/validity of outcome measures
- No tests for baseline equivalence
- Attrition is >5% and no tests for differential attrition are reported
- No controls for baseline outcomes

Request more information if all other standards met

Some concerns, however, we cannot follow-up on:

- Only 1 high-quality QED
- Problems with reliability or validity of outcome measures
- Some differences between conditions at baseline
- Evidence of differential attrition

“Inconclusive” typically = 2 or more of these limitations

Insufficient

"Fatal" flaws

- QED with limited or no matching
- No control group
- No intent-to-treat analysis
- No measures of behavioral outcomes
- No independently measured behavioral outcomes
- No effects on behavioral outcomes
- Effects but not for independently measured outcomes

These design limitations cannot be corrected

"Insufficient" can include 1 or more limitation(s) listed above, and can also include limitations from the "*inconclusive*" rating

Other

No design or analysis flaws that would render the evidence *insufficient* or *inconclusive*

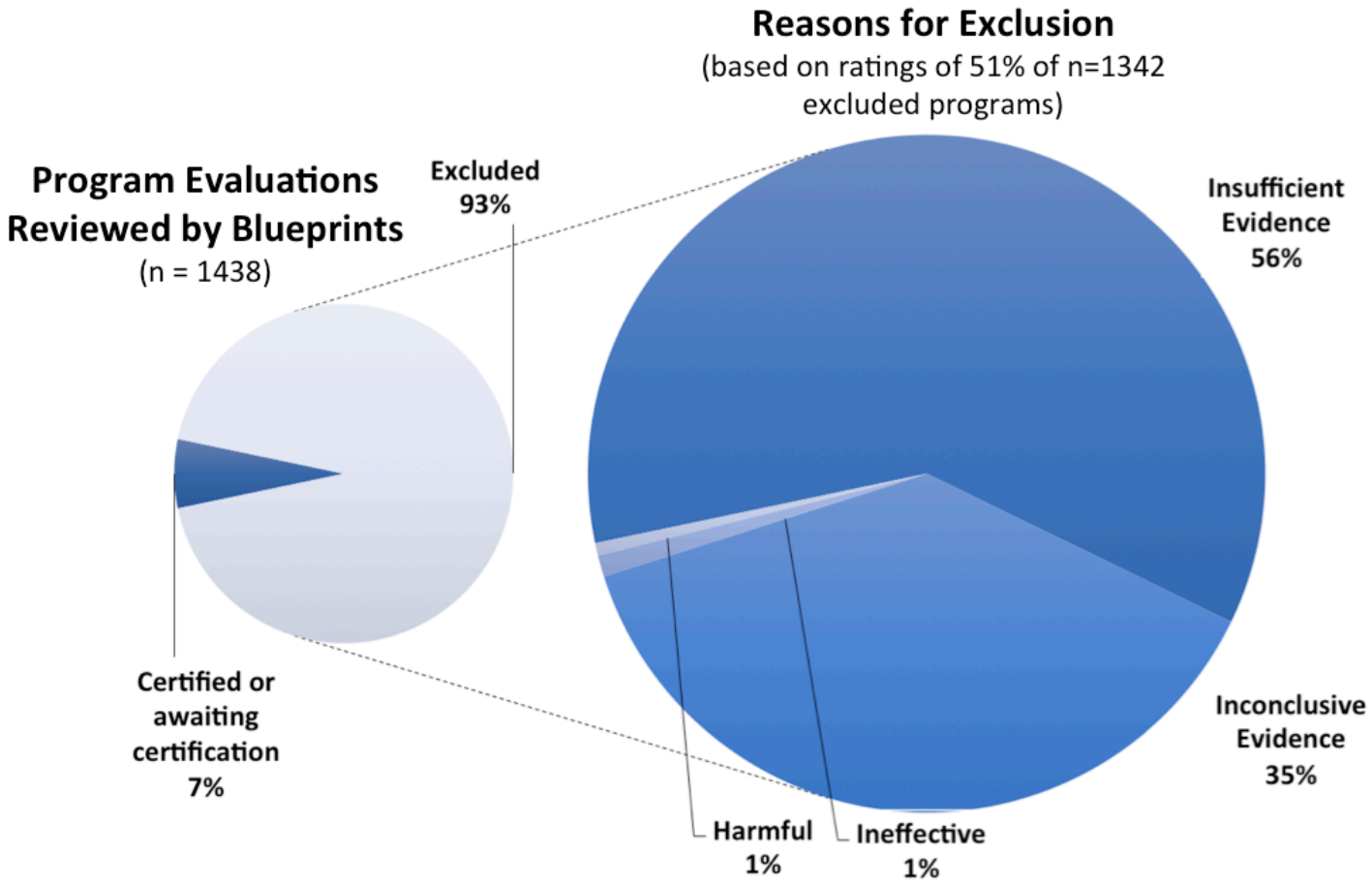
Harmful

- Results suggest the program caused worse outcomes than would otherwise be expected

Ineffective

- Results showed no effects

Reasons for Exclusion: Preliminary Results



Exercise B: Fatal Flaws

Examples of fatal design or analysis flaws for insufficient rating

Take about 10-15 minutes to work through 4 exercises

We will go over the answers together

Exercise B (Review)

Question #1

A tutoring program designed to help struggling readers was evaluated using data drawn from a stratified random sample of 230 participants who had attended the program's after-school tutoring sessions once a week for 35 consecutive weeks. The Gates-MacGinitie Reading Test was administered by researchers blind to condition at the beginning and end of the program. Findings showed students who received the program significantly improved in their reading scores.

"Fatal flaw," according to Blueprints standards:

A) No control group
B) No independently measured behavioral outcomes
C) No intent-to-treat analysis
D) No effects on behavioral outcomes

Exercise B (Review)

Question #1

A tutoring program designed to help struggling readers was evaluated using **data drawn from a stratified random sample of 230 participants who had attended the program's** after-school tutoring sessions once a week for 35 consecutive weeks. The Gates-MacGinitie Reading Test was administered by researchers blind to condition at the beginning and end of the program. Findings showed students who received the program significantly improved in their reading scores.

"Fatal flaw," according to Blueprints standards:

A) No control group

B) No independently measured behavioral outcomes

C) No intent-to-treat analysis

D) No effects on behavioral outcomes

Exercise B (Review)

Question #2

Nutritionists lead 1-hour sessions in a classroom once per month for 9 months to teach healthy food habits. Using a cluster randomized design, 40 classrooms were randomly assigned to the treatment (n=20) or control (n=20) group. At the posttest, no significant differences in BMI scores, body fat percentage or rates of overweight and obesity were found. However, student self-reports revealed those in the treatment group were found to consume significantly fewer cookies and sodas and eat more fruits compared to students in the control group.

"Fatal flaw," according to Blueprints standards:

A) No control group

B) No independently measured behavioral outcomes

C) No intent-to-treat analysis

D) No effects on behavioral outcomes

Exercise B (Review)

Question #2

Nutritionists lead 1-hour sessions in a classroom once per month for 9 months to teach healthy food habits. Using a cluster randomized design, 40 classrooms were randomly assigned to the treatment (n=20) or control (n=20) group. At the posttest, **no significant differences in BMI scores, body fat percentage or rates of overweight and obesity were found.** However, student self-reports revealed those in the treatment group were found to consume significantly fewer cookies and sodas and eat more fruits compared to students in the control group.

"Fatal flaw," according to Blueprints standards:

A) No control group

B) No independently measured behavioral outcomes

C) No intent-to-treat analysis

D) No effects on behavioral outcomes

Exercise B (Review)

Question #3

An intervention was designed to reduce child behavior problems by teaching parents positive discipline strategies. An evaluation was conducted involving 75 parents with children, aged 6 to 11, who were randomly assigned to a treatment (n=44) or control (n=31) group. A variety of parent-report standardized measures were used to assess child antisocial-aggressive behaviors. These data were collected at baseline (time 1), posttest (time 2) and at six-month follow-up (time 3). Findings showed at both the posttest and 6-month follow-up, compared to the control group, children in the treatment group showed significantly lower levels of aggressive behavior.

"Fatal flaw," according to Blueprints standards:

A) No control group
B) No independently measured behavioral outcomes
C) No intent-to-treat analysis
D) No effects on behavioral outcomes

Exercise B (Review)

Question #3

An intervention was designed to reduce child behavior problems by **teaching parents** positive discipline strategies. An evaluation was conducted involving 75 parents with children, aged 6 to 11, who were randomly assigned to a treatment (n=44) or control (n=31) group. A variety of **parent-report** standardized **measures** were used to assess child antisocial-aggressive behaviors. These data were collected at baseline (time 1), posttest (time 2) and at six-month follow-up (time 3). Findings showed at both the posttest and 6-month follow-up, compared to the control group, children in the treatment group showed significantly lower levels of aggressive behavior.

"Fatal flaw," according to Blueprints standards:

A) No control group

B) No independently measured behavioral outcomes

C) No intent-to-treat analysis

D) No effects on behavioral outcomes

Exercise B (Review)

Question #4

The family drug court (FDC) program aims to address parents' underlying substance abuse issues and give them the skills to become sober, functioning caregivers while also protecting the safety of the children involved. This study examined a total of 632 children involved in child welfare cases, 214 of which were adjudicated through the FDC program, and 418 matched control cases who received child welfare services-as-usual. Official child maltreatment reports 24 months' post-enrollment were assessed using administrative records. Results showed that participants who completed the program had significantly lower rates of child maltreatment allegations compared to the participants in the control group.

"Fatal flaw," according to Blueprints standards:

A) No control group
B) No independently measured behavioral outcomes
C) No intent-to-treat analysis
D) No effects on behavioral outcomes

Exercise B (Review)

Question #4

The family drug court (FDC) program aims to address parents' underlying substance abuse issues and give them the skills to become sober, functioning caregivers while also protecting the safety of the children involved. This study examined a total of 632 children involved in child welfare cases, 214 of which were adjudicated through the FDC program, and 418 matched control cases who received child welfare services-as-usual. Official child maltreatment reports 24 months' post-enrollment were assessed using administrative records. **Results showed that participants who completed the program** had significantly lower rates of child maltreatment allegations compared to the participants in the control group.

"Fatal flaw," according to Blueprints standards:

A) No control group

B) No independently measured behavioral outcomes

C) No intent-to-treat analysis

D) No effects on behavioral outcomes

Reasons for Exclusion: Next Steps

Complete classification of non-certified programs

Detect the patterns in fatal flaws to help people avoid them

Encourage evaluators to use designs that allow for stronger causal evidence

Disseminating Information About Certified Programs

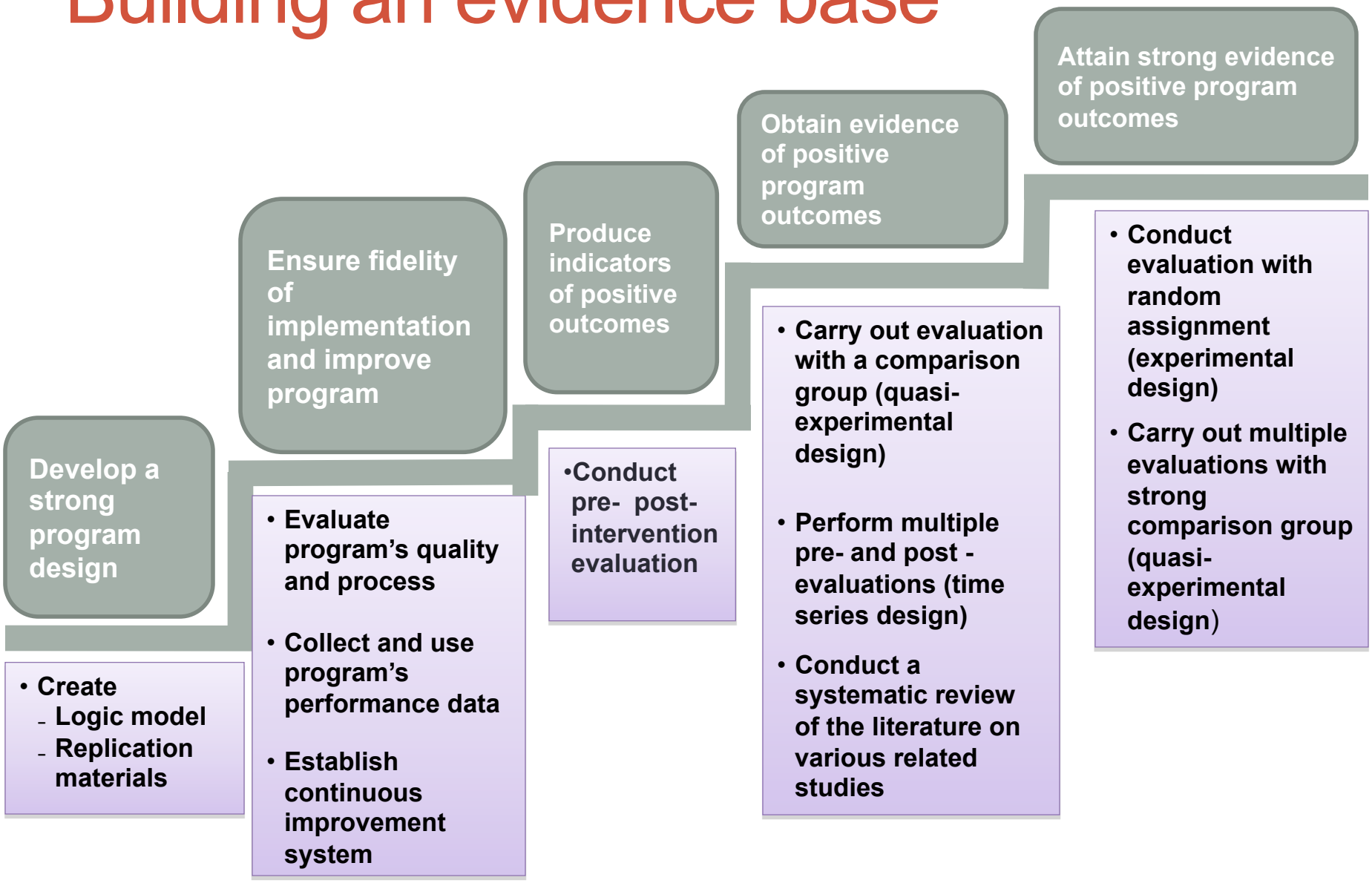
- Navigating the Blueprints website

Closing Remarks

Our goals:

- Building an evidence base
- Increase transparency
- Promote evaluations that yield strong causal evidence

Building an evidence base



Closing Remarks

Our goals:

- Building an evidence base
- Increase transparency
  - Blueprints review process
  - Blueprints standards

(Though we cannot fully standardize the process because the Advisory Board uses methodological expertise to ultimately certify Blueprints programs)

- Promote evaluations that yield strong causal evidence

Conclusions

Blueprints acts in a way similar to the FDA—evaluating evidence, data, and research on program effectiveness to determine those programs that actually work

Benefits of high scientific standards

- We can be confident that programs work
- Helps secure public and financial support for social programs
- Maximizes the efficient allocation of limited resources
  - Money
  - Time
  - Hope

Acknowledgements

THE ANNIE E. CASEY FOUNDATION

Blueprints
FOR HEALTHY YOUTH DEVELOPMENT

IBS
INSTITUTE
OF BEHAVIORAL SCIENCE