Designing and Implementing Randomized Field Trials of Interventions in Schools

George Bohrnstedt
Jeanne Poduska
Mike Garet
Sheppard Kellam
David Myers
American Institutes for Research

Annual Meeting of the American Educational Research Association
April 8, 2007
Designing and Implementing Randomized Field Trials (RFTs) for Education Interventions

George Bohrnstedt, AIR
The Purpose of the Session

- To provide a way of thinking about randomized field trials. What do they entail? What are the major issues in mounting one?
- Done by applying a framework that includes a core set of inter-related parameters for thinking about, planning and implementing randomized field trials (RFTs).
Reasons Why You Might Be Here

- We recognize that the audience will be diverse and likely will include:
  - Those who are already doing RFTs but want to learn more
  - Those who have never done an RFT, but would like to learn how to do one
  - Consumers who may never do one, but want to understand how they are done
  - People with strong methodological skills who would like to learn more about experimental methods

- We will do our best to help address issues that each and every one of you wants answered, but realize that we cannot simultaneously satisfy all these various needs in a single session.
Much was learned historically about what was effective and what was not in large educational initiatives from previous evaluation work in education.

We are seeing efforts to integrate what we have learned from research in the public health domain into education research.

Most important, we are being challenged to move to the next stage of education research where we have more incontrovertible evidence about “what works” and “what doesn’t.”
In this new stage a heavy emphasis is being placed on the role of randomized field trials (RFTs)—the “gold standard” for research designs.

A clarification on RCTs versus RFTs

- Randomized control trial (RCT) is a more general term than RFT. Indeed, RFTs are subsets of RCTs.
- RFTs occur in naturally occurring field settings such as communities, districts or schools, and the interventions often use natural raters (e.g., teachers).
- But many controlled trials occur in far more laboratory-like settings, trials that would be RCTs but not RFTs.
Why Experiments?

- Randomization permits us to examine the effects of interventions by minimizing the effect of any confounding.
  - Guarantees that observed differences between the treatment and comparison groups are, with specifiable probability, due to the intervention or treatment (e.g., a professional development program).
  - Creates a comparison group that is not systematically different from the experimental group (differences are due only to chance).
Why Experiments? (cont.)

- Allows for establishing causal relationships of the $x$ causes $y$ sort by ruling out alternative explanations for the $x$-$y$ relationship, where $y$ is the outcome and $x$ is the intervention.

- This is accomplished by randomly assigning units (e.g., persons, teachers, schools) to either an experimental or a comparison group such that each unit has a specified probability of being assigned to one of the groups (e.g., 0.5).
Is There a Place for Other Kinds of Research Designs?

- Yes, definitely—both quantitative and qualitative designs are important in and of themselves.
- Also, other designs are important in helping to build a foundation for RFTs.
- Furthermore, RFTs can be strengthened by integrating other methods with them.
Types of Research Designs

- Research Aimed at Description
  - Used to characterize populations of students, teachers, classrooms, schools or school districts in a defined geographic area (e.g., state)
  - Allows an understanding of the scope or severity of problems or successes
  - Allows documenting change over time (e.g., longitudinal studies or time series studies)
  - Can describe associations between variables (e.g., between involvement in extra-curricular activities and staying in school)
  - Can be either quantitative or qualitative
Descriptive studies can be useful for generating hypotheses that can be tested with more rigorous designs (especially in areas of limited prior knowledge).

They are also useful for generating hypotheses about why interventions are not working as intended.

But descriptive designs cannot answer the question of whether x causes y.
Pre-Post Designs

- Units (students, teachers, classrooms, schools) are measured on outcomes of interest prior to and after an intervention.
- Supposition is that any difference between pre and post is due to the intervention.
- But without randomization there is no way to know whether the observed change would have occurred due to maturation or history.
Types of Research Designs

- Matched Comparison Designs
  - Comparison groups selected to be as similar to the intervention group as possible.
  - Either matched at the group or the individual level.
  - Differences are observed either pre-post or post-only.
  - Presumption is that observed intervention-comparison group differences are due to the intervention.
  - Nemeses include:
    - Inability to match on more than a few variables
    - Self-selection or other-selection into the intervention
Are Randomized Field Trials Feasible?

- Yes, in many circumstances it is not only feasible, but desirable to do RFTs. Indeed, IES has now launched approximately 30 of them.

- There are many opportunities for creative interventions in education with a potential for improving educational outcomes as well as other important developmental outcomes.
Format of the Session

- Remainder of the day will be broken into 6 sessions:
  - In the first session, we will review the framework, but before doing so we want to hear from you about why you are here and what you hope to take away from this session.
  - The second session will focus on determining what the research question is and its implications for designing the intervention.
  - In the third session we will focus on the importance of thinking about evaluating the impact of the intervention and about variation in impact.
  - Experimental designs and the analysis of data will be the focus of the fourth session.
  - The fifth session will stress the importance of partnerships for carrying out RFTs.
  - And the final session will be a review and wrap-up, hopefully with lots of input from the audience.
The Faculty

- **George Bohrnstedt**, Senior Vice-President for Research, American Institutes for Research

- **Michael Garet**, Chief Research Scientist, American Institutes for Research; Project Director on two ongoing RFTs delivering professional development to teachers

- **Sheppard Kellam**, Director, AIR’s Center for Integrating Education and Prevention Research in Schools, American Institutes for Research; Principal Investigator, Baltimore Education and Prevention Partnership; Senior Research Fellow, American Institutes for Research

- **David Myers**, Senior Vice-President for Education, American Institutes for Research; has directed several large RFTs, including one recently that evaluated reading programs for struggling elementary school readers

- **Jeanne Poduska**, Deputy Director, AIR’s Center for Integrating Education and Prevention Research in Schools, American Institutes for Research; Project Director, Baltimore Education and Prevention Partnership; Principal Research Scientist, American Institutes for Research
Some Details Before We Begin

- We will take breaks at 10:45 and 3:35, with a short 5 minute stretch break at 2:15.
- There will be a one-hour lunch break from 12:15-1:15 p.m. We have been told that there are “hundreds” of restaurants nearby.
- We want discussion during the sessions and again, just a reminder that the last session is designed for discussion as well.
- Any questions before we begin?
A Framework for Designing and Implementing RFTs for Interventions in Schools

Jeanne Poduska, AIR
Purpose of Session

- Learn about your interests

- Begin to build a framework for designing, implementing, and analyzing trials
What brings you to this workshop?
What are the salient issues?
Ten Parameters in the Framework for Doing Randomized Field Trials (RFTs)

1) Strategies and phases in research on interventions
2) Epidemiology as context for doing randomized field trials
3) Role of theory and previous research
4) Observable and measurable core elements of the intervention
5) Measurement of implementation, baseline status, mediating and moderating variables, and periodic outcomes
6) Design, sampling, and statistical power
7) Analyses of impact: Main effects and variation
8) Researcher/community and institutional partnerships
9) Economic analysis
10) Ethical considerations
Strategies and Phases of Research on Interventions
Four Strategies in Intervention Research

- **Developmental Epidemiology** — Interventions aimed at early risk factors to change long-term outcomes
  - Good Behavior Game done in 1st grade classrooms aims at decreasing aggressive behavior to prevent later teen and young adult drug abuse

- **More Immediate Risk** — Interventions aimed at risk factor that is close in time to outcome
  - Condom use to prevent infection during sex
  - Clean needles before injecting drugs
  - Curriculum for adolescents aimed at reducing initiation of drug and alcohol abuse by teens
Four Strategies in Intervention Research (cont.)

- **Community Strategy** — Community-level designs using interventions such as media, enforcing laws against tobacco sales to minors, organizing schools and parents to:
  - Prevent teen smoking
  - Promote use of seat belts
  - Prevent driving when drinking

- **Community/Societal** — Uses computer simulation modeling to assess impact of interventions such as policies and laws
  - Laws prohibiting billboards for alcohol within x feet of schools
  - Seat belt laws
  - Publishing names of DWI drivers
Phases of Research on Interventions

Efficacy

Effectiveness

Sustainability

Going-to-Scale

Sustaining Systemwide

Parameter 1  Kellam and Langevin, 2003
Epidemiology as a Context for RFTs
Levels of Scientific Inquiry

- **Level 1 — Descriptive Epidemiology/Demography**
  - Answers questions of who, what, where, when, how many

- **Level 2 — Correlation Studies**
  - Provides information about associations
  - Can provide base for generating hypotheses
  - Program evaluation

- **Level 3 — Randomized Field Trials**
  - Experiments require manipulation (e.g., drug, intervention, program)
  - Role of random assignment
    - Ensures comparability of two groups at beginning of experiment, before intervention or treatment begins
    - If design maintained, can attribute later impacts (effects) to the intervention (cause) because causes other than intervention have been ruled out
  - Alternative designs without randomization (pre-test/post-test, historical controls, comparison groups) cannot determine causality
Epidemiology Defined

- Study of the distribution and determinants of health-related states in a **defined population**.

- **Defined Population** — a population within a defined geographic space at a specific time; population has known characteristics
  - Controls selection bias
  - Allows study of variation within the population
  - Supports generalization of findings

- Can extend this concept into social, behavioral, educational, and community research
Risk and Risk Factors Defined

- **Risk** — the probability that an outcome will occur within a stated period of time

- **Risk factor** — an attribute or exposure associated with an increased probability of an outcome occurring
Multi-Stage Sampling Defined

- Begin with total defined population
- Choose smaller samples stratified on characteristics associated with risk—risk factors
- Always able to relate analyses back to the larger population
- Economic, efficient
Putting Epidemiology to Work

LEVEL 1
How many cases—prevalence? ➔ Reading failure—school, school district, state, national level

Where are cases located in the population? ➔ In which classrooms, in which schools, in which states

LEVEL 2
What is associated with some people being at risk? ➔ Teacher training, poverty
LEVEL 2
What are possible causes?

→ Poor reading skills, chaotic classrooms that disrupt teaching and learning

LEVEL 3
What can be done—prevention and treatment?

→ Test intervention aimed at improving instructional practices around reading...classroom behavior management

POLICY
What are policy implications?

→ Can program be sustained?
  - Community acceptance
  - Training and staff
  - Cost

Parameter 2
Role of Theory
Role of Theory in RFTs

- **Theory of Cause** — set of inter-related hypotheses about underlying causes behind observed outcomes

- **Theory of Intervention** — hypotheses about how the intervention (cause) will effect or influence the observed outcomes
Observable and Measurable Core Elements of the Intervention
Observable and Measurable Core Elements of the Intervention

- Identify **core elements** of the intervention.

- Specify what you will **observe** to determine the **quality and quantity** of implementation as related to the core elements.

- Measure in the control condition as well as the treatment condition.
Parameter 5

Measurement of Implementation, Baseline Status, Mediating and Moderating Variables, and Periodic Outcomes
Developing a Measurement Framework

- The key constructs to be measured
- How the constructs are measured
- When and how often the constructs are measured
- The settings in which they are measured
- Multi-stage measurement

Parameter 5
Example of Measurement Objectives

- Fidelity of intervention
- Immediate effects of intervention on teacher practices
- Proximal effects of intervention on student outcomes
- Long term (distal) effects on student outcomes
- Hypothesized mediators and moderators
Objectives in Measurement

- Objective and unbiased
- Reliable
- Sensitive to change
- Multi-method
- Construct-specific
- Permits repeated assessment
Parameter 6

Design, Sampling, and Statistical Power
Things to Consider in Determining Design, Sampling, and Statistical Power

- What does theory tell you about the level at which to randomize?
- Determining number of groups (not the individuals in the groups)
- Effect size
- Is intervention impact expected to occur
  - Overall—main effect
  - Differentially between subgroups—variation
- When is intervention impact expected to occur?
- Maintaining the design
Maintaining the Design in the Abstract

Parameter 6

- Defined Population
- Located Eligible
  Consent Obtained
  Assessed at Baseline

- Random Assignment and Intervention begins

- Participation
  OR NOT

- Assessed at Follow up
  OR NOT

Selection Bias
Participation/Compliance Bias
Attrition Bias

Poduska
Maintaining the Population in Real Life

Schools 19

GBG 6 schools
14 classrooms

GBG Intervention
8 classrooms
238 students

GBG Control
6 classrooms
169 students

ML 7 schools
16 classrooms

ML Intervention
9 classrooms
274 students

ML Control
7 classrooms
205 students

External Control
6 schools
11 classrooms

Control
11 classrooms
310 students

Had Baseline Teacher Interview
Followed as Young Adult

Parameter 6

231 97%
179 77%
183 77%
258 94%
178 69%
188 77%
191 93%
153 75%
126 75%
111 74%
145 76%
147 74%
194 76%
255 82%
227 73%
Analyses of Impact:
Main Effects and Variation
Studying Variation in Intervention Impact

- Know that most drugs, interventions, and programs have differential effects
- What works for whom under what conditions
- Theory informs us as to known and possible mediation and moderating variables
- Implications for measurement, sampling and statistical power, and analytic plan
Parameter 8

Negotiating Community and Institution Partnerships
Theoretical Model & Technical Steps in Building Community and Institution Partnerships

- Analyze the social/political structure of the school district.
- Learn the vision and understand the challenges and priorities.
- Identify mutual self-interests within and across the leadership.
- Fit the prevention research/program interests under the visions of the leadership.
- Request ad-hoc oversight committee of leaders.
- Work through trust issues.
Economic Analysis
Types of Economic Analyses

- Cost effectiveness
- Cost benefit
- Cost utility
- Cost feasibility
Things to Consider in Economic Analysis

- Who is the audience for the analysis?
- How will results be used?
- Resources required to identify and measure costs
- Resources required to identify and quantify benefits
Ethical Considerations in the Design and Conduct of Field Trials
Ethical Considerations in the Design and Conduct of Field Trials

- Societal benefit
- Theoretical basis for experiment
- Inclusion/exclusion criteria
- Measurement burden
- Consent/assent processes and procedures
- Confidentiality
- Participant incentives
- Subgroup analyses
- Standards of reporting
Community and Institution Partnerships

- Define risk/benefit (mutual self-interest)
- Understand historical and cultural reasons for resistance to research
- Determine incentives/honoraria
- Review measures
- Determine crisis response procedures
Determining the Question and Designing the Intervention

Mike Garet, AIR
In this Session

We will focus on three key decisions in designing an RFT:

I. Identifying the outcomes and target population (parameters 1-3)
II. Specifying the intervention (parameter 4)
III. Determining what and when to measure (parameter 5)
To illustrate ideas, we will draw on the *Professional Development Impact Study* (PD Impact).

The study is supported by the Institute of Education Sciences (IES).

It involves 90 schools in 6 districts.

It’s in the 4th year, and we are in the process of analyzing the first wave of data.
Principal Staff

Michael S. Garet (project director) AIR
Fred Doolittle (co-project director) MDRC
Stephanie Cronen (deputy project director) AIR
  Meredith Ludwig, AIR
  Terry Salinger, AIR
Mary Dahlgren, Sopris West
  Marian Eaton, AIR
  Terry Anstrom, AIR
  Anja Kurki, AIR
Howard Bloom, MDRC
  Rob Ivry, MDRC
Partner Organizations

American Institutes for Research (AIR)
www.air.org
Conducts applied research and provides technical support in education

MDRC
www.mdrc.org
Evaluates the effectiveness of education reforms and other social policies and actively disseminates research lessons to inform policy and improve practice

REDA International, Inc.
www.redainternational.com
Specializes in social science research and program evaluation

Sopris West
www.sopriswest.com
Provider of staff development in reading instruction

CORE
www.corelearn.com
Provider of staff development and implementation support in reading instruction
I. Identifying the Outcomes and Target Population

- What outcomes are we interested in affecting and who is at risk?
- What is the hypothesized causal chain that leads to the desired or undesired outcomes?
- Where in the hypothesized causal chain should we intervene?
- What is the target population?
Many elementary students cannot read well enough to succeed in school. For example, 54% of disadvantaged 4th graders scored below the basic level on the 2005 National Assessment of Educational Progress (NAEP).

At the same time, there is evidence that specific instructional strategies can make a difference in early reading outcomes (National Reading Panel, 2000).
Successfully teaching all students to read requires a sophisticated understanding of the skills involved in reading and the typical errors children make.

Yet the pre-service preparation teachers receive to teach reading is often inadequate, especially for teachers in high-poverty urban schools (Moats, 1999).
The available evidence suggests that teachers’ instruction might be strengthened by participation in sustained in-service professional development (PD).

But survey evidence indicates that the PD teachers typically receive consists of brief workshops (Garet et al., 2002).

Thus, we set out to test the impact of sustained, content-focused PD.
The study is informed by an emerging conceptual model (small “t” theory) of features of PD that we hypothesize may support teachers in improving their knowledge, instruction, and student achievement (Garet et al., 2002).
Key Features of Effective Professional Development

- **Duration**: Long time-span, Many hours
- **Form of Activity**: Embedded in teaching
- **Collective Participation**: Department or grade level

Focus on Content
- Content students learn
- How students learn

Active Learning
- Practice
- Feedback
- Reflection

Coherence
- Aligned w/ standards
- Consistent w/ teacher goals

Teacher Characteristics
- Background, training, experience

Teacher Knowledge, Beliefs, and Attitudes

Classroom Instruction

Student Characteristics
- Background, prior achievement

Student Academic Outcomes
The study is being conducted in 6 districts, each using either Open Court or Houghton Mifflin’s Nation’s Choice or Legacy of Literacy.

We selected 90 high-poverty schools across the 6 districts (> 50% of students eligible for free lunch).

The study population includes all 2nd grade teachers in the 90 schools and their 2nd grade students.
II. Specifying the Intervention

- What are the core elements of the intervention?
- What elements must be in place to implement the intervention and ensure a meaningful service contrast?
- What is the counterfactual?
- What is the unit of intervention?
- How large an effect is the intervention likely to produce?
The theory suggests that 6 features of PD are important.

It is not possible to vary all 6 features in one study.

We decided to test two versions of the PD.

- PD treatment A: *Institutes and seminars* that incorporate 5 of the 6 features—all except PD embedded in teaching

- PD treatment B: *Institutes and seminars* plus *coaching*, an embedded component
PD Treatment A

3 institute days (summer 2005) and 5 follow-up seminar days (2005-06 school year)

- Based on *Language Essentials for Teachers of Reading and Spelling* (LETRS), by Louisa Moats
- Drawing on a detailed syllabus, with consistent materials and activities across sites
- Delivered by 4 experienced LETRS facilitators
- Supported by project staff to ensure high teacher attendance
Participation in institutes and seminars (treatment A), plus half-time reading coach in each school

- Based on *Consortium on Reading Excellence* (CORE) coach training model
- 20 coaches in total, working across 30 schools
- All received week-long training, coaching manual, and continuing support to ensure fidelity
Counterfactual

“In an experiment, we observe what did happen when people received a treatment. The counterfactual is knowledge of what would have happened to those same people if they simultaneously had not received the treatment. An effect is the difference between what did happen and what would have happened.” (Shadish, Cook, and Campbell, 2002, p 5.)
Control Condition (Group C)

The control condition (Group C) is “Business as usual”—that is, the PD each district currently provides

- The control condition is not the absence of PD.
- The control condition varies to some extent across districts (and in fact across teachers).
Unit of Intervention

- Our theory predicts that the effects of professional development will be maximized when teachers participate with their grade-level peers.

- Thus, we view the treatment as a school-level, not a teacher-level, intervention.

- We randomly assigned schools to treatment conditions separately within each of the 6 participating districts.
At each school, participants include:
- *All* 2nd grade teachers (not volunteers)
- School reading specialist who works with 2nd grade
- Special education teacher who works with 2nd grade
- Principal
Research Questions

- What is the impact of institutes and seminars on teacher knowledge, classroom instruction, and student achievement? (compare Group A to Group C)
- What is the added value of coaching in improving teacher knowledge, classroom instruction, and student achievement? (compare Group B to Group A)
- What is the impact of the combination of institutes, and seminars and coaching? (compare Group B to Group C)
Research Questions (cont.)

- Does the impact of the PD differ for teachers with lower or higher knowledge as measured at baseline?
- Do teacher knowledge and classroom instruction mediate the impact of professional development on student achievement?
Required Sample Size

- We designed the study to be able to detect an effect on teacher knowledge and instruction of 0.4 SD, and an impact on achievement of 0.2 SD.

- Over second grade, students typically gain about one SD. An impact on achievement of 0.2 SD is roughly one-fifth of the typical annual gain.

- To achieve this objective, a sample of 90 schools is required, 30 per treatment condition.
## Participating Districts

<table>
<thead>
<tr>
<th>District</th>
<th>N of Group A schools</th>
<th>N of Group B schools</th>
<th>N of Group C schools</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>30</strong></td>
<td><strong>30</strong></td>
<td><strong>30</strong></td>
</tr>
</tbody>
</table>
III. Determining What and When to Measure

- What measures are needed to assess the fidelity of implementation of the intervention and the dosage received?
- What measures are needed to provide contextual information?
- Over what period are the hypothesized impacts expected to unfold?
- What measures are needed to assess mediating variables and outcomes?
Implementation
Fidelity and Dosage

- Observations of all institutes and seminars
- Sign-in sheets to monitor teacher attendance at institutes and seminars
- Logs of coaching activities completed daily
- Teacher surveys collecting data on all PD in which both intervention and control teachers participated
Contextual Information

- Teacher qualifications, experience, and prior participation in professional development at baseline
- These variables will serve as potential moderating variables in the analysis
Timing of Impacts

- The PD is delivered over a one-year period (summer of 2005 and 2005-06 school year)
- We are interested in:
  - The \textit{immediate impact} during the year of implementation of the PD
  - The \textit{sustained impact} in the year following the implementation of the PD
- The school (not the student or teacher) is the primary focus of longitudinal analysis
Outcomes

- Reading content and practices survey (measure of teacher knowledge)
  - 3 time points (1 pre and 2 post)

- Classroom observations of teacher practices during reading instruction
  - 4 time points (all post)

- Student achievement for consecutive cohorts of 2nd graders
  - 3 time points (1 pre and 2 post)
Timing of Outcome Data

<table>
<thead>
<tr>
<th></th>
<th>2005</th>
<th></th>
<th>2006</th>
<th></th>
<th>2007</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spr</td>
<td>Sum</td>
<td>Fall</td>
<td>Wtr</td>
<td>Spr</td>
<td>Sum</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

Institute, Seminars, and Coaching
Take-Away Points

- Designing an RFT requires articulating what outcomes one hopes to affect in what population(s).

- Specifying the treatment conditions depends on a theory of cause (hypotheses about factors leading to outcomes) and a theory of intervention (hypotheses about where and how to intervene).

- Achieving the desired service contrast rests on specifying the intervention with precision and fashioning routines to deliver it with fidelity.
Assessing whether the intended service contrast was achieved entails measuring the implementation of the intervention and the dosage received in both the treatment and control conditions.

Understanding the impact of the intervention entails assessing both the mediating variables and the outcomes that are part of the hypothesized causal chain.
Beyond Main Effects: Why Variation in Impact is Important in RFTs

Sheppard G. Kellam
NIH Support and Funding

- We are indebted to the NIMH and NIDA staffs and scientific peer reviewers.

- During the last 21 years, this research has been supported by NIMH Grants R01 MH 42968, P50 MH 38725, and R01 MH 40859, with supplements from NIDA for each of these grants.
Role of Theory in RFTs

- **Theory of Cause** — set of inter-related hypotheses about underlying “causes” and how they operate to influence outcomes

- **Theory of Intervention** — hypotheses about how the intervention will influence 1) the targeted cause and 2) the outcomes
Role of Theory in RFTs (cont.)

- Causal theory tells us that under certain conditions (risk factors, mediators, and/or moderators) an outcome is more or less probable.
- If our causal theory is correct, our intervention theory is correct, and the intervention is well implemented, the probability of the outcome will change in a predicted direction.
- The extent will be determined by the attributable risk—how much of the outcome is due to the targeted condition.
- And the “prevented fraction”—how much of the attributable risk is reduced/enhanced by the intervention.
Why Variation is an Issue

- Universal interventions—e.g., those directed at entire classrooms or schools—address children or contexts with differing levels of the targeted condition.
- Both causal theory and intervention theory are specifically directed at the targeted condition.
- Variation in the targeted condition must therefore be part of the analytic model—to test whether the targeted population is really benefited.
- Otherwise the rest of the population is lumped with the targeted population and hides the intended impact.
Inferences re Variation

The design must include:

- Defining the population epidemiologically—to understand the variation in one context vs. another.

- Randomization at the levels at which we want to make inferences.
Authors

Sheppard G. Kellam, M.D.\textsuperscript{1}
Jeanne Poduska, Sc.D.\textsuperscript{1}
C. Hendricks Brown, Ph.D.\textsuperscript{2}
Carla Ford, Ph.D.\textsuperscript{1}
Amy Windham, Ph.D.\textsuperscript{1}
Natalie L. Keegan\textsuperscript{1}
John Reid, Ph.D.\textsuperscript{3}
Nicholas Ialongo, Ph.D.\textsuperscript{4}
Hanno Petras, Ph.D.\textsuperscript{6}

Charlene Cooper-Boston, Ph.D.\textsuperscript{5}
Linda Chinnia\textsuperscript{5}

\textsuperscript{1}American Institutes for Research, Center for Integrating Education and Prevention Research in Schools
\textsuperscript{2}University of South Florida, Prevention Science Methodology Group
\textsuperscript{3}Oregon Social Learning Center
\textsuperscript{4}Johns Hopkins University Bloomberg School of Public Health
\textsuperscript{5}Baltimore City Public School System
\textsuperscript{6}University of Maryland
LIFE COURSE-SOCIAL FIELD CONCEPT

A Theory of Cause: Life Course/Social Field Theory

- **Natural Raters**: define **social task demands** and rate the individual’s performance

- **Social Adaptation**: interactive process of task demand and behavioral responses

- **Social Adaptational Status (SAS)**: rating of adequacy of individual’s performance on task demands

- **Psychological Well-Being (PWB)**: the individual’s internal state
Hypothesis 1: success in mastering social task demands specific to one stage of development and specific to one social field will lead to successes in the same and other social fields over the course of development.

Hypothesis 2: there is a potential reciprocal relationship between SAS and PWB—"feeling good" leads to better SAS performance and better SAS leads to better PWB.

Hypothesis 3: therefore, intervening into the social adaptational process between natural rater and individual can lead to better SAS and PWB.
Theory Supports Targeting First Grade Classroom SAS Process

- In the first grade classroom, sitting still, paying attention, learning, and participating as a student are social task demands set and rated by the natural rater—the teacher.

- Classroom aggressive, disruptive behaviors as early as first grade predict poor outcomes such as school failure, tobacco, drug and alcohol abuse, and violent/criminal behavior.
A Theory of Intervention: Good Behavior Game

- **Hypothesis:** improving teacher’s interaction in the classroom with children around the social task demands of behaving as a student will reduce aggressive, disruptive behaviors.

- **Hypothesis:** improving children’s student behavior in early classrooms will lead to better school SAS later.

- **Hypothesis:** These SAS successes early in life will lead later on to better SAS in school and other social fields and prevent substance abuse and behavioral and psychological problems and disorders where early aggressive, disruptive behavior is a risk factor.
The Baltimore Education and Prevention Partnership

- BCPSS has collaborated in three generations of education and prevention field trials.
- They were directed at helping children master key social task demands in 1st-grade classroom.
- Interventions were tested separately, then together.
- The first generation will be our main focus today, where the Good Behavior Game (GBG) was tested by itself and the children, now young adults, were recently followed to ages 19–21.
Baltimore (WD) Analytic Model

Classroom Behavior Management → Decreased Aggressive, Disruptive Classrooms → Decreased Individual Aggression → Decreased Later Conduct & Anti-social Personality Disorders

Teachers’ Effective Academic Instruction → Improved Reading Skills in Classroom → Increased Individual Achievement → Decreased Later Depressive Disorders

Effective Family-Classroom Partnerships → Decreased Individual Depressive Symptoms → School Success & Decreased Drop-Out

Whole-Day 1st-Grade Education and Prevention Program

Other mediating or moderating variables:
- Family structure and poverty
- Deviant peers
- School building
- Community economic health, resources, drugs, and violence
Impact of Poorly Managed Classrooms on Students

- Aggressive, disruptive behavior as early as 1st grade is a major risk factor for academic failure, later school drop-out, delinquency, drug abuse, depression, and other problem outcomes.

- Children with behavior problems in poorly managed 1st grade classrooms were up to 20 times more likely to exhibit severe aggressive problems in middle school compared to similar children in well-managed 1st-grade classrooms.
The number one reason for teacher burnout is the inability to manage classrooms.

Teachers need tested tools to manage classrooms, i.e., to teach children how to be students.

A large portion of 1st-grade teachers need such tools, e.g., ~50% in Baltimore.
The History of the Good Behavior Game (GBG)

- GBG was originally developed by Barrish, Saunders, & Wolfe at the University of Kansas, with the first report in 1969.

- At least 18 short-duration non-randomized trials followed and described positive results.

- These led to the developmental epidemiologically based randomized field trials in Baltimore to test GBG.
Goals of GBG

- Provide teachers a classroom-wide method to socialize children into the role of student.

- Reduce classroom aggressive, disruptive behavior to enhance classroom teaching and learning.

- Prevent school failure, drug abuse, delinquency, and other problem outcomes.
GBG Implementation

- In Baltimore, GBG consists of dividing the 1st-grade class into three heterogeneous teams.

- Each teacher exhibits a large poster that states proper student behaviors, i.e., classroom rules.

- Teams are rewarded for each child’s pro-social behavior, and not rewarded when a child is disruptive. It is “group contingent.”
Early in the year, GBG is played systematically for ten minutes, three times a week, and announced to the class by the teacher.

The ten minutes are extended gradually over the year until the process is integrated into the entire day.

Rewards are more abstract as the year goes on.
Design of 1st Generation Trial

- 41 1st-grade classrooms in 19 schools.

- **ACROSS** schools: Schools were matched and randomly assigned to be an external control school, a GBG school, or an enhanced reading program school.

- **WITHIN** each intervention school: Children were balanced across all 1st-grade classrooms.

- Then 1st-grade classrooms and teachers were randomly assigned to standard program classrooms or to intervention classrooms.
Randomization at school level and classroom and teacher, and balancing children in classrooms allows testing theory-driven hypotheses:

1) Children overall improve with GBG.
2) Aggressive, disruptive children improve more with GBG.

Today we will focus on GBG children compared to standard program children within GBG schools combined with all control children in other schools.
Impact of GBG in 1st and 2nd Grades vs. All Controls on Drug Abuse or Dependence Disorder for Males Ages 19–21

Teacher Ratings of Aggression: Fall of 1st Grade

- GBG (n = 72)
- All Controls (n = 134)
Impact of GBG in 1st and 2nd Grades vs. All Controls on Alcohol Abuse or Dependence Disorder for Males Ages 19–21

Teacher Ratings of Aggression: Fall of 1st Grade

- GBG (n = 73)
- All Controls (n = 134)
Impact of GBG in 1st and 2nd Grades vs. All Controls on Regular Smoking for Males Ages 19–21

Teacher Ratings of Aggression: Fall of 1st Grade

GBG (n = 75)
All Controls (n = 136)
Impact of GBG in 1st and 2nd Grades vs. All Controls on Antisocial Personality Disorder for Males Ages 19–21

Teacher Ratings of Aggression: Fall of 1st Grade

Percentage of ASPD

GBG (n = 75)
All Controls (n = 138)
A relatively simple method of classroom behavior management aimed at improving the social adaptational process can have a long-term impact.

GBG, directed at 1st and 2nd grade teacher’s practices and children’s SAS around aggressive, disruptive behavior, resulted in improved risk for drug, alcohol, tobacco, and ASPD disorders—but mainly among higher risk males and less so females and lower risk males.

While Life Course/Social Field theory was supported, the results were not complete prevention. It left unanswered other sources (attributable risk or GBG prevented fraction).
Take-Away Points

- RFTs have 2 functions: 1) testing theory and 2) understanding what works for which children, under what conditions.
- Randomized field trials in epidemiologically defined populations, with multiple levels of randomization, can allow testing what works, for which children, in what conditions.
- Defining the population and the variation in the targeted conditions are possible and necessary.
- Nothing works for everybody and we need to know what works for whom and under what conditions.
Design and Analysis Considerations in Educational Experiments

David Myers, AIR
Design and Analysis

- Identifying key issues and parameters
- Role of a power analysis
  - Intuition
  - Formal analysis
  - Examples
- Real-world example
Critical Design Issues

- Prioritize research and policy questions
- Identify relevant subgroups \( (I(s)=0 \text{ or } I(1)=I(2)) \)
- Identify data collection issues
  - Sample attrition
  - Mobility of units assigned to T and C—not the same as attrition (mobility isn’t necessarily related to treatment)
- Establish relevant impact — ITT v TOT \( \rightarrow \) fraction of compliers
- Determine number of time points after completion of intervention
- Establish relevant effect size (P4K example)
- Determine relevant point of randomization—fit to program operations
Critical Design Issues (cont.)

- Fit randomization to school operations for effectiveness studies
  - Theory testing may require alternative approaches to sort out confounding mechanism (e.g., teachers and instruction)
- Develop analysis plan—be explicit (DOM, regression adjusted, multilevel, measurement strategy—use of multiple outcomes within domains, latent variable models)
- Estimate relevant design parameters
  - Variation “accounted for by covariates in model”
  - Between-cluster variance in multilevel models
  - Number of students per classroom
  - Variability of key outcome measures
- Establish relevant counterfactual—for example, classroom observations (use the P4K example on amount of reading by the controls versus treatments)
Establish Basic Design

- Number of treatment arms
- Within school or separate control group
  - Are you concerned about contamination/spill over?
- Unit of assignment
  - School level intervention?
  - Classroom intervention?
  - Student directed intervention?
- Need for external validity?
  - How many schools and how dispersed?
  - How many school districts and how dispersed?
Role of the Power Analysis

- Power analysis is a tool that helps us design our experiments
  - Power analyses are based on many assumptions; however, they do allow us to compare designs with common assumptions
- Question addressed in power analysis:
  - How large a sample and what allocation of units is needed to detect a statistically significant impact with a high probability?
Intuition for Power Analyses

- Power analyses are built around the following intuition
  - Imagine starting with a pool of students
  - Randomly allocate half to T and half to C
  - Compute the impact, which is defined as the difference in the means for T and C, and the standard error of the estimate
  - Compute a t-statistic and then compare to a critical value for the statistical test
  - With the power analysis, we establish how large a sample we need to ensure a high probability of detecting a difference if one exists
  - Think of repeating the “experiment many times” and each time recording whether you found a statistically significant impact—generally, we want 80% or more of the impact estimates to be statistically significant
    - Power of 80% is arbitrary, but an accepted standard
More Intuition

Critical value
Null
Alternative

How far do you move the distribution on the right so that 80% of the estimates are above the yellow line?

Assumes $\alpha = 0.05$ two-tailed, and power $=0.80$
Implementation of Power Analyses

- **Method of simulation**
  - May be most practical for complex, non-text book problems (Gelman and Hill, 2007)

- **Analytic methods**
  - Useful for learning the basics and understanding key components and relationships (Schochet, 2005)
Analytic Approach

- Key expression
  \[ \delta_N + Z_\alpha SE(\delta) = \delta_A - Z_{(1-\beta)} SE(\delta) \]

- Left side shows how far we need to be from the “null value” to be treated as statistically significant

- Right side shows how far we need to be below the “alternative” value so that some percent of the parameter estimates are statistically significant

- In most evaluations, we set the null value to 0 and the alternative value \( \delta_A \) is equivalent to the size of impact we want to detect, at a minimum

- \( Z_\alpha, Z_{(1-\beta)}, \) and \( SE(\delta) \) correspond to the critical values for the probability of a Type I error, power, and the standard error of the impact estimate, respectively
Useful Expressions

Minimum detectable effect: a statistically significant impact of this size or larger will be detected with a high probability (generally 0.80)

\[ \delta_A = (Z_\alpha + Z_{(1-\beta)})SE(\delta) \]

Under a simple model (difference-of-means), where we assume equal variances across T and C and equal numbers of units in T and C, and we drop the subscript for “A”

\[ N_T = (Z_\alpha + Z_{(1-\beta)})^2 \left( \frac{(2\sigma^2)}{\delta^2} \right) \]

And, with regression adjusted model

\[ N_T = (Z_\alpha + Z_{(1-\beta)})^2 \left( \frac{(2\sigma^2 (1-R^2))}{\delta^2} \right) \]
Treatment Sample Sizes with Regression Adjusted Model

ES = 0.20: sample size for T is 392, 294, and 196

ES = 0.60: sample size for T is 44, 33, and 22
Implications of Using “Clusters” as the Unit of Assignment

- **Examples**
  - Random assignment of families
  - Random assignment of classrooms/instructional groups
  - Random assignment of schools

- **Multiple sources of variability introduced into the estimates**
  - Within “unit”
  - Between “units”
  - Unlike a fixed effects model which “accounts” for the additional variation, random effects model introduces more variability and reduces precision
Random Assignment of Classrooms: Number of Schools

Number of classrooms = 4
Number of students per classroom = 20
School is a fixed effect
Between classroom random effect

- ICC=0
- ICC=.05
- ICC=.10
- ICC=.15
Overview of P4K

- Questions
  - Could well-implemented, intensive pull-out interventions substantially improve the reading skills of struggling readers in grades 3 and 5?
  - For whom did the interventions work most effectively?

- Setting
  - 27 school districts around Pittsburgh (AIU)

- Funding
  - The evaluation was funded by IES and the Smith Richardson Foundation.
  - Foundations that funded the intervention component included the Monell, Barskdale, Grable, Haan, Kellogg, Raymond, and Rockefeller foundations, the Heinz Endowments, and IES
Estimating Impacts and Standard Errors

- Most design work is predicated on simple models
  - Difference-of-means
  - Regression adjusted (increasing precision)
  - Multi-level (better approximation to correct standard errors)
- Use the most relevant models for “powering the study”
- Think about issues of multiple comparisons when determining sample sizes
Overview (cont.)

- Interventions
  - Failure Free, Wilson, Corrective Reading, Spell Read
  - PD for pull-out teachers
    - One week of PD, plus 4-5 weeks of practice with 4th graders
  - 3 students per teacher, 5 days per week, 1 hour per day (goal was 100 hours)

- Outcomes
  - Word level skills (e.g., phonemic awareness and decoding)
  - Fluency
  - Comprehension
Overview of the Idealized P4K Design

40 School Units

School Level
- FF(8)
  - 12 T
  - 8 C
- SR(8)
  - 12 T
  - 8 C
- RA

Student Level
- WL(8)
  - 12 T
  - 8 C
- CR(8)
  - 12 T
  - 8 C
P4K: Single Intervention and Multiple Interventions

Goal for single intervention

6 treatment students
4 control students
R_sq = 0.50
ICC between schools
Take-Away Points

- Keep designs simple
- Stay focused on critical research and policy questions
- Don’t build “Christmas Trees” with lots of ornaments—they may collapse
- Understand program/school operations
- Make design fit the intervention
- When designing the study, consider the range of statistical analyses that will be used