WWC Standards for Regression Discontinuity and Single Case Study Designs

March 2011
Presentation to the Society for Research on Educational Effectiveness
John Deke    Shannon Monahan    Jill Constantine
What is the What Works Clearinghouse (WWC)?

- Initiative of the U.S. Department of Education's Institute of Education Sciences
- Central and trusted source of scientific evidence for what works in education
  - Develops and implements standards for reviewing and synthesizing education research
  - Assesses the rigor of research evidence on the effectiveness of interventions
What Works Clearinghouse

Producers of Information
- Universities
- Research Organizations
- Developers

Evidence That Meets Standards

Evidence

What Works Clearinghouse
- Evidence reports
- Practice guides
- Help desk

Effectiveness Information

Users of Information
- School boards
- Superintendents and Advisors
- State Education Departments
- Funders
- Intermediaries
- Consultants
- Technical Assistance Providers
- Principals
- Teachers
- Federal government
- State government
- National groups
- Public

Topics

Evidence-Based Interventions and Practices

Improved Student Outcomes
Current WWC Study Designs Standards

- Randomized Controlled Trials (RCTs)
  - Well executed RCTs can obtain the highest level of causal validity and *meet evidence standards*

- Quasi-experimental Designs (QEDs)
  - Some RCTs and matched comparison designs *meet evidence standards with reservations*

- Regression Discontinuity Designs (RDDs) — Version 1.0

- Single-Case Study Designs—Version 1.0
Goals and Procedures for Developing a Research Design Standard

- Establish design standards for causal inference studies
- Standard must be operational
  - Replicable and reliable
  - Transparent
- Operationalized through a protocol and a study review guide
- Development process
  - Convene an expert panel
  - Pilot test and revise
What is RDD?

- Treatment and control groups formed by design (like an RCT), not by unobserved self-selection
- Groups are formed purposefully, not randomly
- Groups formed using a cutoff on a continuous, “forcing” variable
Rationale for Adding RDD Standards

- RDD studies are becoming more common as large, longitudinal administrative data sets become more common in education.
- IES is conducting evaluation studies with RDDs.
- The literature on RDD methodological best practices has evolved to a point that supports the development of standards.
Process for Developing Standards

- Panel convenes
- Standards drafted
- Study Review Guide drafted
- Standards pilot tested
- Standards approved by IES
Regression Discontinuity Design (RDD)
Expert Panel Members

Peter Schochet, Co-Chair, Mathematica
John Deke, Co-Chair, Mathematica
Tom Cook, Northwestern
Guido Imbens, Harvard
J. R. Lockwood, RAND
Jack Porter, University of Wisconsin
Jeff Smith, University of Michigan
Key Considerations

- Should the WWC allow RDD studies to join RCTs as studies that “meet standards without reservations”?
- How can we distinguish between RDD studies that meet standards with versus without reservations?
Overview of RDD Standards

- Screener to identify studies as being RDD
- Three possible designations:
  1. Meets standards without reservations
  2. Meets standards with reservations
  3. Does not meet standards
- Four individual standards:
  1. Integrity of the forcing (assignment) variable
  2. Attrition
  3. Continuity of the outcome-forcing variable relationship
  4. Functional form and bandwidth
- Reporting requirement for standard errors
RDD Screener

- Treatment assignment is based on a “forcing variable”—units on one side of a cutoff value are in the treatment group, units on the other side are in the comparison group.
- The forcing variable must be ordinal with at least 4 values above and 4 values below the cutoff value.
- There must be no factor confounded with the cutoff value of the forcing variable.
1. Integrity of the Forcing Variable

- Primary concern is “manipulation”.
- Two criteria: institutional integrity and statistical integrity
- Both criteria must be met to pass this standard without reservations; one criterion must be met to pass with reservations.
2. Attrition

- Primary concern is nonresponse bias.
- Standard is almost the same as the WWC RCT standard on attrition.
- One difference: overall attrition and differential attrition can either be reported at the cutoff value of the forcing variable or for the full treatment and comparison samples.
3. Continuity of Outcome-Forcing Variable Relationship

- Primary concern is phantom impacts from lack of “smoothness” in the relationship between the outcome and the forcing variable.
- Two criteria: baseline equivalence on key covariates and no evidence of unexplained discontinuities away from the cutoff value
- Both criteria must be met to pass this standard without reservations.
4. Functional Form and Bandwidth

- Primary concern is misspecification bias.
- Five criteria:
  A. An adjustment must be made for the forcing variable.
  B. A graphical analysis must be included and should be consistent with bandwidth/functional form choice.
  C. Statistical evidence must be provided of an appropriate bandwidth or functional form.
  D. Model should “interact” forcing variable with treatment status or provide evidence that an interaction is not needed.
  E. All of the above must be done for every combination of outcome, forcing variable, and cutoff value.

- All criteria must be met to pass standards without reservations; to meet with reservations, (A and D) plus (B or C).
Example of Graphical Analysis
Applying RDD Evidence Standards

- To *meet standards without reservations*, must meet each individual standard without reservations
- To *meet standards with reservations*, must meet individual standards 1, 3, and 4 with or without reservations
Reporting Requirement for Standard Errors

- Standard errors must reflect clustering at the unit of assignment.
- Lee and Card (2008)—clustering issue applies when forcing variable is not truly continuous (“random misspecification error”)
- This does not affect whether a study meets standards, but it does affect how its findings can be used.
Overview

- Defining Features of SCDs
- Common Designs
- Internal Validity and Replication
- WWC SCD Design Standards
- WWC SCD Evidence Standards (Visual Analysis)
Single Case Designs (SCDs)

- Involve repeated, systematic measurement of a dependent variable

- Measurement occurs before, during, and/or after the active manipulation of an independent variable (e.g., applying an intervention).

- Most often used in applied and clinical fields

- A systematic review that includes SCDs will speak to the effectiveness of an intervention in a variety of specific settings, but cannot speak to effectiveness in all settings.
Contributions of the SCD Standards

- Expands the body of rigorous research that can be reviewed by the WWC
- Offers guidelines to the research community at large of best practice in SCD
## Rationale for Adding SCD Standards

<table>
<thead>
<tr>
<th>WWC Topic Area</th>
<th>SCD Studies Planned for Review by June 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adolescent Literacy</td>
<td>20</td>
</tr>
<tr>
<td>Autism</td>
<td>190</td>
</tr>
<tr>
<td>Behavior Disorders</td>
<td>61</td>
</tr>
<tr>
<td>Intellectual Disabilities</td>
<td>159</td>
</tr>
<tr>
<td>Learning Disabilities</td>
<td>13</td>
</tr>
</tbody>
</table>
Single Case Design (SCD) Standards Panel

Tom Kratochwill, Chair, University of Wisconsin

John Hitchcock, Ohio University

Rob Horner, University of Oregon

Joel Levin, University of Arizona

Sam Odom, University of North Carolina

David Rindskopf, City University of New York

Will Shadish, University of California, Merced
Example SCD Graph

Source: Fraenkel and Wallen, 2006
Defining Features of an SCD

- An individual “case” is the unit of intervention and data analysis. A case may be a single participant or a cluster of participants (e.g., a classroom or community).

- Within the design, the case provides its own control for purposes of comparison.

- The outcome variable is measured *repeatedly* within and across different conditions or levels of the independent variable.

- The ratio of data points (measures) to the number of cases usually is large so as to distinguish SCDs from other longitudinal designs (e.g., traditional pretest-posttest and general repeated-measures designs).
Core Designs in the WWC SCD Standards

- Reversal/Withdrawal (ABAB designs)
- Multiple Baseline Designs
- Alternating Treatment Designs
Reversal/Withdrawal Design—ABAB

- Baseline (A₁ Phase)
- Intervention (B₁ Phase)
- Baseline (A₂)
- Intervention (B₂)

Days/Weeks/Months/Sessions
Multiple Baseline Design

- Participant 1
- Participant 2
- Participant 3
- Participant 4

- Dependent Variable
- Baseline
- Intervention

- Days/Weeks/Months/Sessions
Types of Questions an Individual SCD Might Answer

Which intervention is effective for this case?

- Does Intervention A reduce problem behavior for this case?
- Does adding B to Intervention A further reduce problem behavior for this case?
- Is Intervention B or Intervention C more effective in reducing problem behavior for this case?
The “5-3-20 Rule”

- A minimum of five SCD studies examining the intervention with a design that *Meets Evidence Standards* or *Meets Evidence Standards with Reservations*

- The SCD studies must be conducted by at least three research teams at three different institutions with no overlapping authorship.

- The combined number of cases (e.g., individual participants, classrooms) totals at least 20.
SCD Standards Are Designed to Mitigate Threats to Internal Validity

- Ambiguous Temporal Precedence
- Selection
- History
- Testing
- Instrumentation
- Additive and Interactive Effects of Threats
- Maturation
- Regression to the Mean
Importance of Replication

Methods of replication:

1. Introduction and withdrawal (i.e., reversal) of the independent variable (e.g., ABAB design)
2. Staggered introduction of the independent variable across different points in time (e.g., multiple baseline design)
3. Iterative manipulation of the independent variable across different observational phases (e.g., alternating treatments design)

Source: Horner and Spaulding, 2008
WWC SCD Rating Process

Evaluate the Design

- Meets Evidence Standards
- Meets Evidence Standards with Reservations
- Does Not Meet Evidence Standards

Conduct Visual Analysis for Outcome Variable

- Strong Evidence
- Moderate Evidence
- No Evidence

Effect-Size Estimation
Systematic Manipulation of the Independent Variable

- The independent variable (i.e., the intervention) must be systematically manipulated, with the researcher determining when and how the independent variable conditions change.

- If this standard is not met, the study Does Not Meet Evidence Standards.
Inter-Assessor Agreement

- Each outcome variable must be measured systematically over time by more than one assessor.

- The study needs to collect inter-assessor agreement:
  - in each phase;
  - on at least 20% of the data points in each condition (e.g., baseline, intervention); and
  - and the inter-assessor agreement must meet minimal thresholds (e.g., 80% agreement and .60 for Cohen’s kappa).

- Regardless of the statistic, inter-assessor agreement must be assessed for each case on each outcome variable.

- If no outcomes meet these criteria, the study
  *Does Not Meet Evidence Standards.*
At Least Three Attempts to Demonstrate an Effect

- The study must include at least three attempts to demonstrate an intervention effect at three different points in time or with three different phase repetitions.

- If this standard is not met, the study *Does Not Meet Evidence Standards*.

- Designs that could meet this standard include:
  - ABAB Design
  - Multiple Baseline Design with at least 3 baseline conditions
  - Alternating Treatment Design

- Designs not meeting this standard include:
  - AB Design
  - ABA Design
  - BAB Design
Minimum Data Points for a Reversal Withdrawal Design

For a phase to qualify as an attempt to demonstrate an effect, it must have a minimum of 3 data points.

- To *Meet Standards*, a reversal/withdrawal design must have a minimum of 4 phases per case with at least 5 data points per phase.

- To *Meet Standards with Reservations*, a reversal/withdrawal design must have a minimum of 4 phases per case with at least 3 data points per phase. Any phases based on fewer than 3 data points cannot be used to demonstrate existence or lack of an effect.
Minimum Data Points for Multiple Baseline

For a phase to qualify as an attempt to demonstrate an effect, it must have a minimum of 3 data points.

- To *Meet Standards*, a multiple baseline design must have a minimum of 6 phases with at least 5 data points per phase.

- To *Meet Standards with Reservations*, a multiple baseline design must have a minimum of 6 phases with at least 3 data points per phase. Any phases based on fewer than 3 data points cannot be used to demonstrate existence or lack of an effect.
Meets Standards

Figure 2. Percentage of intervals with on-task behavior across the three classes.

Source: Lee Kern et al., 1994
Meets Standards with Reservations as ABABs

Source: Stahr et al., 2006
Applying Visual Analysis

- **Goal:** Determine whether the evidence demonstrates a “basic effect” at three different points in time
  - A basic effect is a change in the dependent variable when the independent variable is manipulated.

- **Assessment of basic effects involves simultaneous evaluation of several features**
  - Level, trend, variability, overlap, immediacy, consistency across similar phases, baseline stability in non-intervened series
Level
Variability
Immediacy of Effect
SCD Evidence Ratings

- Applied to SCDs that *Meet Evidence Standards* (with or without reservations)

**Rating Algorithm**
- Strong: Three demonstrations of a basic effect and no noneffects
- Moderate: Three demonstrations of a basic effect and at least one noneffect
- No Evidence: Fewer than three demonstrations of a basic effect

**Additional requirement for Multiple Baseline designs**
- Stability in baseline patterns when intervention is introduced in an earlier series (described later)
4 Steps in WWC Visual Analysis

- Evaluate baseline
- Compare data within other phases
- Compare data in similar phases
- Compare data in adjacent phases
Evaluate Baseline

- Do baseline data document...
  - A pattern of behavior that demonstrates the problem the intervention is designed to address; and
  - A pattern that would be expected to continue if no intervention were implemented?

- If yes, proceed. Otherwise, outcome receives a No Evidence rating.
Baseline and Within Phases

Example Reversal Design

Number of Tantrums

Baseline

Intervention X

Baseline

Intervention X

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
Similar Phases and Adjacent Phases

Example Reversal Design

First Demonstration of Effect

Second Demonstration of Effect

Third Demonstration of Effect

Number of Tantrums
Example No Evidence and Strong Evidence

Source: Stahr et al., 2006
Additional Consideration for Multiple Baselines

- **Stability in non-intervened series**
  - Looking for carry-over effects

- **Highest rating is moderate if evidence of instability**

Source: Lohrmann and Talerico, 2004