Assessing Implementation Fidelity and Achieved Relative Strength in RCTs: Concepts and Methods

David S. Cordray
Vanderbilt University
Presentation
The Nebraska Center for Research on Children, Youth, Families and School
University of Nebraska
Lincoln Nebraska
April 19, 2010
Overview

• Research Context and Definitions
• A 4-step approach to assessment and analysis of implementation fidelity (IF) and achieved relative strength (ARS):
  – Model(s)-based
  – Quality Measures of Core Causal Components
  – Creating Indices
  – Integrating implementation assessments with models of effects
Distinguishing Implementation Assessment from the Assessment of Implementation Fidelity

• Two ends on a continuum of intervention implementation/fidelity:

• A purely descriptive model:
  – Answering the question “What transpired as the intervention was put in place (implemented).

• Based on a priori intervention model, with explicit expectations about implementation of program components:
  – Fidelity is the extent to which the realized intervention (\(t^{Tx}\)) is faithful to the pre-stated intervention model (\(T^{Tx}\))
  – Infidelity = \(T^{Tx} - t^{Tx}\)

• Most implementation fidelity assessments involve descriptive and model-based approaches.
Dimensions Intervention Fidelity

• Aside from agreement at the extremes, little consensus on what is meant by the term “intervention fidelity”.
• Most frequent definitions:
  – True Fidelity = Adherence or compliance:
    • Program components are delivered/used/received, as prescribed
    • With a stated criteria for success or full adherence
    • The specification of these criteria is relatively rare
  – Intervention Exposure:
    • Amount of program content, processes, activities delivered/received by all participants (aka, receipt, responsiveness)
    • This notion is most prevalent
  – Intervention Differentiation:
    • The unique features of the intervention are distinguishable from other programs, including the control condition
    • A unique application within RCTs
Linking Intervention Fidelity Assessment to Contemporary Models of Causality

• **Rubin’s Causal Model:**
  – True causal effect of X is \( (Y_i^{Tx} - Y_i^C) \)
  – RCT methodology is the best approximation to this true effect
  – In RCTs, the difference between conditions, on average, is the causal effect

• **Fidelity assessment within RCTs entails examining the difference between causal components** in the intervention and control conditions.

• Differencing causal conditions can be characterized as *achieved relative strength* of the contrast.
  – Achieved Relative Strength (ARS) = \( t^{Tx} - t^C \)
  – ARS is a default index of fidelity
Achieved Relative Strength = 0.15

Infidelity

\[ \text{Expected Relative Strength} = (0.40 - 0.15) = 0.25 \]

\[ d = \frac{85 - 70}{30} = 0.50 \]

\[ d = \frac{\bar{Y}_t - \bar{Y}_c}{s d_{\text{pooled}}} \]

\[ d_{\text{with fidelity}} = \frac{90 - 65}{30} = 0.83 \]

\[ d_{\text{with fidelity}} = \frac{\bar{Y}_t - \bar{Y}_c}{s d_{\text{pooled}}} \]

\[ d = 0.50 \]
Why is this Important?

• **Statistical Conclusion validity**
  – **Unreliability of Treatment Implementation:** Variations across participants in the delivery receipt of the causal variable (e.g., treatment). Increases error and reduces the size of the effect; decreases chances of detecting covariation.

• Resulting in a reduction in statistical power or the need for a larger study....
The Effects Structural Infidelity on Power

Fidelity: .60 .80 1.0

α = 0.050

ρ = 0.13, k= 30, n=20, R^2_L = 0.78
Influence of Infidelity on Study-size

![Graph showing power analysis for different number of clusters with fidelity values 1.0, 0.80, 0.60.](image)

- **Power** vs **Number of clusters**
- **Fidelity** values: 1.0, 0.80, 0.60
- **α = 0.050, n = 20**
- Lines represent different scenarios with:
  - $\delta = 0.31, \rho = 0.13, R^2 = 0.78$
  - $\delta = 0.25, \rho = 0.13, R^2 = 0.76$
  - $\delta = 0.19, \rho = 0.13, R^2 = 0.78$
If That Isn’t Enough….

• **Construct Validity:**
  – Which is the cause? \((T^{Tx} - T^C)\) or \((t^{Tx} - t^C)\)
    • **Poor implementation:** essential elements of the treatment are incompletely implemented.
    • **Contamination:** The essential elements of the treatment group are found in the control condition (to varying degrees).
    • Pre-existing similarities between T and C on intervention components.

• **External validity – generalization is about** \((t^{Tx} - t^C)\)
  – This difference needs to be known for proper generalization and future specification of the intervention components
So what is the cause? … The achieved relative difference in conditions across components

PD = Professional Development
Asmt = Formative Assessment
Diff Inst = Differentiated Instruction
Some Sources and Types of Infidelity

• If delivery or receipt could be dichotomized (yes or no):
  – Simple fidelity involves compliers;
  – Simple infidelity involves “No shows” and cross-overs.

• Structural flaws in implementing the intervention:
  – Missing or incomplete resources, processes
  – External constraints (e.g. snow days)

• Incomplete delivery of core intervention components
  – Implementer failures or incomplete delivery
A Tutoring Program: Variation in Exposure

4-5 tutoring sessions per week, 25 minutes each, 11 weeks

Expectations: 44-55 sessions

Random Assignment of Students

<table>
<thead>
<tr>
<th>Cycle</th>
<th>Average Sessions Delivered</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cycle 1</td>
<td>47.7</td>
<td>16-56</td>
</tr>
<tr>
<td>Cycle 2</td>
<td>33.1</td>
<td>12-42</td>
</tr>
<tr>
<td>Cycle 3</td>
<td>31.6</td>
<td>16-44</td>
</tr>
</tbody>
</table>
Variation in Exposure: Tutor Effects

The other fidelity question: How faithful to the tutoring model is each tutor?
In Practice….

• Identify core components in the intervention group
  – e.g., via a Model of Change
• Establish bench marks (if possible) for $T^{TX}$ and $T^C$
• Measure core components to derive $t^{TX}$ and $t^C$
  – e.g., via a “Logic model” based on Model of Change
• Measurement (deriving indicators)
• Converted to Achieved Relative Strength and implementation fidelity scales
• Incorporated into the analysis of effects
What do we measure?

What are the options?

1) **Essential** or **core** components (activities, processes);

2) **Necessary, but not unique**, activities, processes and structures (supporting the essential components of T); and

3) **Ordinary features** of the setting (shared with the control group)

- Focus on 1 and 2.
Specifying Intervention Models

• Simple version of the question: *What was intended?*

• Interventions are generally multi-component, sequences of actions

• Mature-enough interventions are specifiable as:
  – Conceptual model of change
  – Intervention-specific model
  – Context-specific model
An Illustrative Simple Model of Change

The Logic Model and Conceptual Model

Figure 1.2  Community Leadership Academy (CLA) Program Logic Model

The Generic Logic Model

From: W.T. Kellogg Foundation, 2004
The Other Half of the Picture

Fidelity assessment within RCTs should examine *the difference between causal components* in the intervention and control conditions.

- Differencing causal conditions can be characterized as *achieved relative strength* of the contrast.
  - Achieved Relative Strength (ARS) = $t^T_x - t^C$
  - ARS is a default index of fidelity
Quality Measures of Core Components

- Measures of resources, activities, outputs
- Range from simple counts to sophisticated scaling of constructs
- Generally involves multiple methods
- Multiple indicators for each major component/activity
- Reliable scales (3-4 items per sub-scale)
Core Reading Components for Local Reading First Programs

Design and Implementation of Research-Based Reading Programs

Use of research-based reading programs, instructional materials, and assessment, as articulated in the LEA/school application

Teacher professional development in the use of materials and instructional approaches

After Gamse et al. 2008

1) Teacher use of instructional strategies and content based on five essential components of reading instruction
2) Use of assessments to diagnose student needs and measure progress
3) Classroom organization and supplemental services and materials that support five essential components
### From Major Components to Indicators…

<table>
<thead>
<tr>
<th>Major Components</th>
<th>Sub-components</th>
<th>Facets</th>
<th>Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Instruction</td>
<td>Instructional Time</td>
<td>Block</td>
<td>Scheduled block?</td>
</tr>
<tr>
<td></td>
<td>Instructional Material</td>
<td>Actual Time</td>
<td>Reported time</td>
</tr>
<tr>
<td></td>
<td>Instructional Activities/Strategies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Support for Struggling Readers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professional Development</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Reading First Implementation: Specifying Components and Operationalization

<table>
<thead>
<tr>
<th>Components</th>
<th>Sub-components</th>
<th>Facets</th>
<th>Indicators (I/F)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reading Instruction</strong></td>
<td>Instructional Time</td>
<td>2</td>
<td>2 (1)</td>
</tr>
<tr>
<td></td>
<td>Instructional Materials</td>
<td>4</td>
<td>12 (3)</td>
</tr>
<tr>
<td></td>
<td>Instructional Activities /Strategies</td>
<td>8</td>
<td>28 (3.5)</td>
</tr>
<tr>
<td><strong>Support for Struggling Readers (SR)</strong></td>
<td>Intervention Services</td>
<td>3</td>
<td>12 (4)</td>
</tr>
<tr>
<td></td>
<td>Supports for Struggling Readers</td>
<td>2</td>
<td>16 (8)</td>
</tr>
<tr>
<td></td>
<td>Supports for ELL/SPED</td>
<td>2</td>
<td>5 (2.5)</td>
</tr>
<tr>
<td><strong>Assessment</strong></td>
<td>Selection/Interpretation</td>
<td>5</td>
<td>12 (2.4)</td>
</tr>
<tr>
<td></td>
<td>Types of Assessment</td>
<td>3</td>
<td>9 (3)</td>
</tr>
<tr>
<td></td>
<td>Use by Teachers</td>
<td>1</td>
<td>7 (7)</td>
</tr>
<tr>
<td><strong>Professional development</strong></td>
<td>Improved Reading Instruction</td>
<td>11</td>
<td>67 (6.1)</td>
</tr>
</tbody>
</table>

4 10 41 170 (4)

Adapted from Moss et al. 2008
# Reading First Implementation: Some Results

<table>
<thead>
<tr>
<th>Components</th>
<th>Sub-components</th>
<th>Performance Levels</th>
<th>ARSI (U3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Instruction</td>
<td>Instructional Time (minutes)</td>
<td>101</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>Support</td>
<td>79%</td>
<td>58%</td>
</tr>
<tr>
<td>Struggling Readers</td>
<td>More Tx, Time, Supplemental Service</td>
<td>83%</td>
<td>74%</td>
</tr>
<tr>
<td>Professional</td>
<td>Hours of PD</td>
<td>41.5</td>
<td>17.6</td>
</tr>
<tr>
<td>Development</td>
<td>Five reading dimensions</td>
<td>86%</td>
<td>62%</td>
</tr>
<tr>
<td>Assessment</td>
<td>Grouping, progress, needs</td>
<td>84%</td>
<td>71%</td>
</tr>
</tbody>
</table>

Adapted from Moss et al. 2008
Achieved Relative Strength = .15

Infidelity

“Infidelity”

Variability

Positive Infidelity
True Fidelity

Variability

Analyzing Variation in Implementation
Indexing Cause-Effect Linkage

• Analysis Type 1:
  – Congruity of Cause-Effect in ITT analyses
    • Effect = Average difference on outcomes \(\rightarrow\) ES
    • Cause = Average difference in causal components \(\rightarrow\) ARS
      (Achieved Relative Strength)
    • Descriptive reporting of each, separately

• Analysis Type 2:
  – Variation in implementation fidelity linked to variation in outcomes
    • Effect = outcomes
    • Cause = covariates (from ARSI)
## Common Cause-Effect Scenarios

<table>
<thead>
<tr>
<th>The Cause</th>
<th>The Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low/Low = Low/High</td>
</tr>
<tr>
<td>Low</td>
<td></td>
</tr>
<tr>
<td></td>
<td>= Low/High = ????</td>
</tr>
<tr>
<td>High</td>
<td></td>
</tr>
<tr>
<td></td>
<td>High/High = ????</td>
</tr>
</tbody>
</table>

- Low/Low = Cause-Effect Congruity
- High/Low = Dampening Process
- Low/High = ????
Cause-Effect Congruity: High/High Example

• Fantuzzo, King & Heller (1992) studied the effects of reciprocal peer tutoring on mathematics and school adjustment.
  – 2 X 2 factorial design crossing levels of structured peer tutoring and group reward
  – 45 min. 2-3 per week; 60-90 sessions

• Fidelity assessments:
  – Observations (via checklist) of students and staff, rated the adherence of group members to scripted features of each condition;
    • 50% random checks of sessions
  – Mid-year, knowledge tests to index the level of understanding of students about the intervention components in each of the four conditions.
Fantuzzo et al. Continued

- Fidelity results:
  - Adherence (via observations):
    - 90-100% across conditions,
    - 95% overall
  - Student understanding (via 15 item test):
    - 82% SD=11% (range 47-100%); ANOVA=ns
    - Reward+structure condition: 84% Control: 86%

- Effects on mathematics computation:
  - ES= (7.7-5.0)/1.71 = 1.58

- Congruity=High/High; no additional analyses needed
Exposure and Achieved Relative Strength

• Fantuzzo et al. example is:
  – Relatively rare;
  – Incorporates intervention differentiation, yielding fidelity indices for all conditions.

• More commonly, intervention exposure is assessed:
  – Yielding scales of the degree to which individuals experience the intervention components in both conditions
  – The achieved relative strength index is used for establishing the differences between conditions on causal components
Indexing Fidelity as Achieved Relative Strength

Intervention Strength = Treatment – Control

Achieved Relative Strength (ARS) Index

\[
ARS \text{ Index} = \frac{t^{Tx} - t^C}{S_T}
\]

- Standardized difference in fidelity index across Tx and C
- Based on Hedges’ \( g \) (Hedges, 2007)
- Corrected for clustering in the classroom
Average ARS Index

\[ g = ARS = \left( \frac{\bar{X}_1 - \bar{X}_2}{S_T} \right) \times \left( 1 - \frac{3}{4(n_{Tx} + n_C) - 9} \right) \times \sqrt{1 - \frac{2(n-1)p}{N-2}} \]

Where,
- \( \bar{X}_1 \) = mean for group 1 (t_{Tx})
- \( \bar{X}_2 \) = mean for group 2 (t_{C})
- \( S_T \) = pooled within groups standard deviation
- \( n_{Tx} \) = treatment sample size
- \( n_C \) = control sample size
- \( n \) = average cluster size
- \( p \) = Intra-class correlation (ICC)
- \( N \) = total sample size
A Partial Example of the Meaning of ARSI

Randomized Group Assignment → Professional Development → Differentiated Instruction → Improved Student Outcomes
### Very Large Group Difference, Limited Overlap Between Conditions

<table>
<thead>
<tr>
<th>Hours of Professional Development</th>
<th>Control</th>
<th>Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>1 1 1</td>
<td>70 1 1 1</td>
</tr>
<tr>
<td>60</td>
<td>0 0 0 1 1 1 1 1 2 5 6 7 7 7 7 8 9 9 9 9 9</td>
<td>60 0 0 0 1 1 1 1 2 5 6 7 7 7 7 8 9 9 9 9</td>
</tr>
<tr>
<td>50</td>
<td>0 0 0 1 1 1 1 2 5 6 7 7 7 7 8 9 9 9 9</td>
<td>50 4 5 6 7 7 7 7 8 8 8 8 9</td>
</tr>
<tr>
<td>40 0 1</td>
<td>0 0 0 1 2 2 2 4 7 8 8 8 9</td>
<td>40 7 9</td>
</tr>
<tr>
<td>30 0 0 0 1 2 2 2 4 7 8 8 8 9</td>
<td>20 0 1 4 4 4 5 6 6 6 6 7 7 7 7 7 7 9</td>
<td>30</td>
</tr>
<tr>
<td>20 0 1 4 4 4 5 6 6 6 6 6 7 7 7 7 9</td>
<td>10 2 3 8 9</td>
<td>20</td>
</tr>
<tr>
<td>10 2 3 8 9</td>
<td>0 0</td>
<td>10</td>
</tr>
<tr>
<td>0 0</td>
<td></td>
<td>0 0</td>
</tr>
</tbody>
</table>

Mean (Control) = 28.2
Mean (Intervention) = 61.8

\[
\text{ARSI: } \frac{61.8 - 28.2}{6.61} = 5.08
\]

U3 = 99%
Cohen’s U3 Index: Very Large Group Separation

- Control Mean
- Intervention Mean

50th percentile

U3=99th Percentile
ARSI=5.08
## Small Group Differences, Substantial Overlap

<table>
<thead>
<tr>
<th>Hours of Professional Development</th>
<th>Control</th>
<th>Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>60</td>
<td>50</td>
<td>60</td>
</tr>
<tr>
<td>50</td>
<td>40 0 1</td>
<td>50</td>
</tr>
<tr>
<td>40 0 1</td>
<td>30 0 0 1</td>
<td>40 0 0 0</td>
</tr>
<tr>
<td>30 0 0 1</td>
<td>20 0 1 4</td>
<td>30 0 0 0 0 1</td>
</tr>
<tr>
<td>20 0 1 4</td>
<td>10 2 3 8</td>
<td>20 3 4 5 6 6</td>
</tr>
<tr>
<td>10 2 3 8</td>
<td>0 0</td>
<td>10 6 8</td>
</tr>
<tr>
<td>0 0</td>
<td></td>
<td>0 0</td>
</tr>
</tbody>
</table>

### Mean and SD

- **Control**
  - Mean: 28.2
  - SD: 7.04

- **Intervention**
  - Mean: 30.8
  - SD: 6.14

### ARSI

- **ARSI:** $\frac{(30.8 - 28.2)}{6.61} = 0.39 \times 100 = 39\%$

### U3

- **U3:** 66%
Cohen’s U3: Little Group Separation

Control Mean

Intervention Mean

50th

66th Percentile

ARSI=0.39
High/High and Low/Low Congruity

Hulleman & Cordray (2009) examined the results of a motivation intervention in the lab and in classrooms, not surprisingly…..

<table>
<thead>
<tr>
<th>Measure</th>
<th>Lab</th>
<th>Classroom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived Utility Value</td>
<td>g = 0.45</td>
<td>g = 0.05</td>
</tr>
<tr>
<td></td>
<td>p = 0.03</td>
<td>p = 0.67</td>
</tr>
<tr>
<td>Achieved Relative Strength:</td>
<td>0.65</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Calculating ARSI When There Are Multiple Components

Infidelity

Augmentation of Control

PD = Professional Development
Asmt = Formative Assessment
Diff Inst = Differentiated Instruction
Weighted Achieved Relative Strength

\[ ARSI_{PD} = \frac{\bar{X}^E - \bar{X}^C}{Sd} = \frac{3 - 2.5}{2} = 0.25 \]

\[ ARSI_{Assess} = \frac{6 - 3.5}{3} = 0.83 \]

\[ ARSI_{DI} = \frac{7 - 4}{3.5} = 0.86 \]

\[ ARSI_{Weighted} = \sum w_j ARSI_j = .25(.25) + .33(.83) + .42(.86) = 0.69 \]

\[ U3 = 76\% \]
Converting ARS into a Composite Fidelity Index

Composite Fidelity = \frac{\text{ARSI}}{\text{RSI wgt}} = \frac{0.69}{1.94} = .36

Where:

\[ RSI = \sum w_j (RSI_j) \]

\[ RSI_{PD} = \frac{\overline{X}^T - \overline{X}^C}{Sd} = \frac{6 - 2}{2} = 2.0 \]

\[ RSI_{Assess} = \frac{8 - 2.5}{3} = 1.83 \]

\[ RSI_{DI} = \frac{10 - 3}{3.5} = 2.0 \]

\[ RSI_{Weighted} = .25(2) + .33(1.83) + .42(2) = 1.94 \quad U3 = 97\% \]
Main points….

• Analysis of intervention fidelity and achieve relative strength is a natural counterpart to estimating ESs in ITT studies.
• They provide an interpretive framework for explaining outcome effects.
• When ES and ARSI are discordant, serve as the basis for additional analysis.
• Next section focuses on analysis of variation
Analysis II

Linking Variation in Treatment Receipt/Delivery to Outcomes
Analyzing Variation in Treatment Receipt/Delivery Within Groups: Fidelity Indicators

• Rather than relying on the 0,1 coding of groups, fidelity indicators replace the group variable.
• New question being answered: What is the effect of treatment on those receiving treatment or TOT.
• Value of fidelity indices will depend on their strength of the relationship with the outcome;
• The greater the group difference, on average, the less informative fidelity indicators will be; and
• High predictability requires reliable indices
Using Group, Fidelity Indicators, or Both: A Simple Example

Randomized Group Assignment

Fidelity Indicator = Hours of Professional Development
Outcome = Differentiated Instruction
Improved Student Outcomes
The “Value Added” of Implementation Fidelity/ARS Data

<table>
<thead>
<tr>
<th>Group Separation</th>
<th>U3</th>
<th>Predicting Level of Differentiated Instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$R^2_{\text{Group}}$</td>
</tr>
<tr>
<td>Small</td>
<td>0.39</td>
<td>0.01</td>
</tr>
<tr>
<td>Large</td>
<td>2.36</td>
<td>0.215*</td>
</tr>
<tr>
<td>Very Large</td>
<td>5.08</td>
<td>0.401*</td>
</tr>
</tbody>
</table>
EXAMPLE: Intent-to-treat (ITT) and Treatment-on-Treated (TOT): An Example

• Justice, Mashburn, Pence, & Wiggins (2008) examined:
  – Language-Focused Curriculum (LFC) in 14 classes;
  – Classes randomly assigned to LFC and control;
  – Core component of LFC is the use of language stimulation techniques (e.g., open questions, recasts, models); and
  – Outcome → Growth in expressive language examined (fall to spring)
Justice et al. Continued

• Implementation fidelity assessed:
  – 3 times using 2 hour observation (45 item check list) 50 min. video sample; and 40 weekly lesson plans.

• Fidelity score =
  – weighted sum of frequency of the use of 7 language stimulation techniques (range 0-21);

• Fidelity = score/21; averaged over observations

• Results:
  – LST teachers average fidelity = 0.57 (range 0.17-0.79)
  – Control teachers average fidelity = 0.32 (range 0.17-0.56)
  – ANOVA F=11.83, p = .005; d = ARSI = 1.71
Justice et al. Continued

Level 1

\[ Y_{ij} = \beta_{00} + \beta_{01} \text{(Fall Score)} + \beta_{02} \text{(Gender)} + \beta_{03} \text{(SES)} + \beta_{04} \text{(Attendance)} + r_{ij} \]

Level 2 $\rightarrow$ ITT

\[ \beta_{00} = \gamma_{00} + \gamma_{01} \text{(LFC)} + u_{0j} \]

Level 2 $\rightarrow$ TOT

\[ \beta_{00} = \gamma_{00} + \gamma_{01} \text{(LST)} + u_{0j} \]

0,1 Group

Fidelity Score
## Justice et al. Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Reading Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td>Level 1</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>.139</td>
</tr>
<tr>
<td>Fall Language scores</td>
<td>0.29**</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.13</td>
</tr>
<tr>
<td>SES</td>
<td>0.10**</td>
</tr>
<tr>
<td>Attendance</td>
<td>0.19</td>
</tr>
<tr>
<td>Level 2 (ITT)</td>
<td></td>
</tr>
<tr>
<td>Treatment (1)/Control (0)</td>
<td>0.64</td>
</tr>
<tr>
<td>Level 2 (TOT)</td>
<td></td>
</tr>
<tr>
<td>Average observation</td>
<td>-0.03</td>
</tr>
</tbody>
</table>
What can we conclude about the ITT and TOT analyses?

• Few teachers exhibited high levels of LST use (core component of LFC)
• Fidelity overall = 0.45
• They argue, the large group difference (ARSI=1.71 for fidelity = 0.57 vs. 0.32) may not have been sufficient because the dosage (0.57) was so far below what is needed to affect language development.
• Other possibilities include:
  – Reliability of the scaling?
  – Use of average when trend in observations showed improvement?
  – Coverage of central constructs?
  – Functional form of fidelity-outcome linkage?
Hierarchy of Approaches
ITT and LATE

• ITT (Intent-to-treat) estimates (e.g., ES) plus:
  – an index of true fidelity:
    • ES=.50 Fidelity = 96%
  – an index of Achieved Relative Strength (ARS)
    • The Assign→ Hours of Professional Development example

• LATE (Local Average Treatment Effect):
  – If treatment receipt/delivery can be meaningfully dichotomized
    and there is experimentally induced receipt or non-receipt of
    treatment:
    – adjust ITT estimate by T and C treatment receipt rates.
      • Simple model can be extended to an Instrumental Variable Analysis
        (see Bloom’s 2005 book).

• ITT retains causal status; LATE can approximate causal
  statements.
Treatment-on-Treated

• TOT (Treatment-on-Treated).
  – Two-level linear production function, modeling the effects of implementation factors in Tx and modeling factors affecting C in separate Level 2 equations.
  – Regression-based model, exchanging implementation fidelity scales for treatment exposure variable.
  – Simple: ITT estimate adjusted for compliance rate in Tx, no randomization

• Subject to mis-specification

• Useful in identifying potential differentiated effects and basis for new studies.
Descriptive Analyses

• Descriptive analyses:
  – Dose-response relationship
  – Partition intervention sites into “high” and “low” implementation fidelity:
    • ATOD prevention studies, the
      \[ ES^{\text{HIGH}} = 0.13 \text{ to } 0.18 \]
      \[ ES^{\text{LOW}} = 0.00 \text{ to } 0.03 \]
Key Points and Issues

- Fidelity assessment serves two roles:
  - Average causal difference between conditions; and
  - Using fidelity measures to assess the effects of variation in implementation on outcomes.
- Degree of fidelity and Achieved Relative Strength provide fuller picture of the results
- Modeling fidelity depends on the assignment model
- Most applications, fidelity is just another Level 2 or 3 variable.
- Uncertainty and the need for alternative specifications:
  - Measure of fidelity
  - Index of achieved relative strength
  - Fidelity-outcome model specification (linear, non-linear)
- Adaptation-fidelity tension
Additional Examples
EXAMPLE 2: An Elaborated Model: The Welfare to Work Experiments

- Howard Bloom and his colleagues (2005) assessed the effects of employment training on earnings in a classic set of welfare to work experiments.
- They modeled the effects of site-level implementation and program variations, controlling for client characteristics and unique aspects of site-level control conditions.
- This approach is commonly referred to as a production function: unfortunately these types of examples are very rare (but a great model for the future).
Bloom et al. Model Specification

Factors affecting control group conditional mean earnings

\[ Y_{ji} = \alpha_j + \beta_j P_{ji} + \sum \delta_k CC_{kj} + \sum \gamma_k CC_{kj} + \kappa RA_j + \epsilon \]

Total Earnings
Assignment
Control
Client Characteristics
Conditional program impact on earnings, in each office
Level 2 models
Random Differences in the Control
### Some Bloom et al. Results

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Program Characteristic</th>
<th>B ($)</th>
<th>Adj B ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Implementation</strong></td>
<td>Emphasis on quick job entry</td>
<td>720***</td>
<td>720***</td>
</tr>
<tr>
<td></td>
<td>Emphasis on personal attention</td>
<td>428***</td>
<td>428***</td>
</tr>
<tr>
<td></td>
<td>Closeness of monitoring</td>
<td>-197</td>
<td>-197</td>
</tr>
<tr>
<td></td>
<td>Staff caseload size</td>
<td>-4***</td>
<td>-268***</td>
</tr>
<tr>
<td></td>
<td>Staff disagreement</td>
<td>124</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>Staff-supervisory disagreement</td>
<td>-159*</td>
<td>-159*</td>
</tr>
<tr>
<td><strong>Activities</strong></td>
<td>Basic education</td>
<td>-16**</td>
<td>-208**</td>
</tr>
<tr>
<td></td>
<td>Job-search assistance</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Vocational training</td>
<td>7</td>
<td>71</td>
</tr>
<tr>
<td><strong>Econ Environ</strong></td>
<td>Unemployment rate</td>
<td>-94***</td>
<td>-291***</td>
</tr>
</tbody>
</table>
EXAMPLE 3: Analyzing the Reasons for Implementation Failure

- Hulleman & Cordray (2009) examined the sources of implementation failure.
- Focused on the classroom results where there were no motivation effects.
- Student behaviors were nested within teachers:
  - Teacher dosage
  - Frequency of student exposure
- Student and teacher behaviors were used to predict treatment fidelity (i.e., quality of responsiveness/exposure).
Sources of Infidelity: Multi-level Analyses

Part I: Baseline Analyses

• Identified the amount of residual variability in fidelity due to students and teachers.
  – Due to missing data, we estimated a 2-level model (153 students, 6 teachers)

Student: \( Y_{ij} = b_{0j} + b_{1j}(\text{TREATMENT})_{ij} + r_{ij}, \)
Teacher: \( b_{0j} = \gamma_{00} + u_{0j}, \)
\( b_{1j} = \gamma_{10} + u_{10j} \)
Sources of Infidelity: Multi-level Analyses

Part II: Explanatory Analyses

- Predicted residual variability in fidelity (quality of responsiveness) with frequency of responsiveness and teacher dosage

Student: \[ Y_{ij} = b_{0j} + b_1(TREATMENT)_{ij} + b_2(RESPONSE FREQUENCY)_{ij} + r_{ij} \]

Teacher: \[ b_{0j} = \gamma_{00} + u_{0j} \]
\[ b_{1j} = \gamma_{10} + b_{10}(TEACHER DOSAGE)_{j} + u_{10j} \]
\[ b_{2j} = \gamma_{20} + b_{20}(TEACHER DOSAGE)_{j} + u_{20j} \]
## Sources of Infidelity: Multi-level Analyses

<table>
<thead>
<tr>
<th>Variance Component</th>
<th>Baseline Model</th>
<th>Explanatory Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Residual Variance</td>
<td>% of Total</td>
</tr>
<tr>
<td><strong>Level 1</strong> (Student)</td>
<td>0.15437*</td>
<td>52</td>
</tr>
<tr>
<td><strong>Level 2</strong> (Teacher)</td>
<td>0.13971*</td>
<td>48</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>0.29408</td>
<td></td>
</tr>
</tbody>
</table>

* p < .001.
Case Summary

• The motivational intervention was more effective in the lab ($g = 0.45$) than field ($g = 0.05$).

• Using 3 indices of fidelity and, in turn, achieved relative treatment strength, revealed that:
  – Classroom fidelity < Lab fidelity
  – Achieved relative strength was about 1 SD less in the classroom than the laboratory

• Differences in achieved relative strength = differences motivational outcome, especially in the lab.

• Sources of fidelity: teacher (not student) factors
And, finally....