Using Multiple Methods for Rigorous Evaluation of Voluntary After-school Programs

Silvia Swigert, Director of Evaluation Center for Educational Partnerships, UC Irvine
Presented at the 2012 CERA Annual Conference Monterey, CA
The Problem

“College interventions suffer from a serious lack of rigorous evaluation, in spite of the millions of dollars that are invested in them annually” (Gandara, 2002, p. 97)
Definition of Rigorous Program Evaluation-College Programs (Gandara)

• Specify the program goal – What is the expected outcome?
• Specify the selection criteria – who gets in and how?
• Collect baseline data – How are students performing when they start?
• Track students – how long have they been in the program and how often do they attend?
Rigorous Program Evaluation – College Programs (Con’t)

• Compare them to other students like them who are not in the program. Otherwise, how do we know they wouldn’t have done just as well without the program intervention?

• Collect some kind of process data to determine what about the program is effective?

• Collect information on program costs by component.
WWC Definition of Rigorous Evaluation

“An empirical study, using quantitative methods and inferential statistics, that includes a comparison group. Such designs include well-conducted randomized controlled trials (RCTs) and quasi-experiments with matching or equating of student samples on the pretest.”
Institute of Education Studies (IES) states:

“Well-matched comparison-group studies can be valuable in generating hypotheses about "what works," but their results need to be confirmed in randomized controlled trials. “
Typical Situations Needing Rigor (2010, USAID)

- Multiple approaches to achieving desired results have been suggested, and it is unclear which approach is the most effective or efficient;

- An approach is likely to be replicated if successful, and clear evidence of program effects are desired before scaling up;
Typical Situations Needing Rigor (2010, USAID)

• A program uses a large amount of resources or affects a large number of people; and

• In general, little is known about the effects of an important program or approach, as is often the case with new or innovative approaches.
Main Research Question

Does participation in an after school program for mathematics increase achievement for students who began below grade level in grade-level standards-based mathematics?
Intervention

- Replicable curriculum-based intervention using manipulatives aligned with recommendations of the WWC practice guide
- Duration of 42 to 48 hours
- Target populations included 7th Grade Pre-Algebra and 8th Grade Algebra students who scored in the lower proficiency levels of the California Standards Test (CST)
Professional Development
Active Learning
Evaluation Goals

• Conduct a rigorous evaluation of the program using an experimental design
• Study the impact of the program with low-achieving students
• Study correlates of program attendance
Evaluation Strategies

• Experimental Design
• Program Variation
• Fidelity Monitoring
• Stakeholder Surveys
• Attendance Monitoring
Experimental Design: Random Control Trial (RCT)

- 198 (60%) randomly assigned to treatment
- 130 (40%) randomly assigned to control
- No significant differences on pre-test (t-test, p < .05)
- No significant differences on key demographic variables including gender, ethnicity, English proficiency, and parent educational levels (chi-square, p < .05)
Random Control Trial (RCT) Fail

- 65 (33%) randomly assigned to treatment attended

- Treatment (65) and control (108) students were compared on the pre-test (post-attrition) with no significant differences ($t = 1.504$, $p=.134$)

- Treatment (65) and control (108) students were compared on the post-test with no significant differences ($t = 0.641$, $p=.522$)
Random Control Trial (RCT) Fail (Con’t)

• No significant differences on key demographic variables including gender, ethnicity, English proficiency, and parent educational levels (chi-square, p < .05)

• Attrition based on missing post-test realized for both treatment and control students roughly 15%

• Conclusion: Treatment effect on achievement scores appears to be null however matched study design may strengthen design because of treatment group attrition (non-attendance)
Traditional Matched Studies: Method 1

- Traditional method of matching one-to-one
- Sample stratified by gender than matched within gender on a combination of English language proficiency (yes/no) and parental high school completion (yes/no) classification
- 54 treatment and 54 control students matched (17% loss in treatment cases)
# Algebra Matched Study

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Treatment Group (54)</th>
<th>Control Group (54)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>38%</td>
<td>38%</td>
</tr>
<tr>
<td>Female</td>
<td>62%</td>
<td>62%</td>
</tr>
<tr>
<td>English Language Learner</td>
<td>40%</td>
<td>40%</td>
</tr>
<tr>
<td>Family Educational Level (Less than HS graduate)</td>
<td>40%</td>
<td>40%</td>
</tr>
</tbody>
</table>
Method 1 (Con’t)

• No significant difference on pre-test (post-attrition) ($t = 1.767, p = .08$)
• No significant difference on post-test (post-attrition) ($t = 1.575, p = .118$)
• **Conclusion**: Treatment effect on achievement scores appears to be null however students were not matched on quantitative pretest score - may strengthen design further
Method 2: Quantitative Covariate Matching with Stratification

<table>
<thead>
<tr>
<th>Quintile</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>258</td>
<td>319</td>
<td>301.89</td>
<td>15.324</td>
</tr>
<tr>
<td>@2011CSTMathScaledScores</td>
<td>35</td>
<td>322</td>
<td>335</td>
<td>329.31</td>
<td>4.562</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>35</td>
<td>339</td>
<td>348</td>
<td>344.56</td>
<td>3.775</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
<td>352</td>
<td>361</td>
<td>356.45</td>
<td>3.821</td>
</tr>
<tr>
<td>@2011CSTMathScaledScores</td>
<td>47</td>
<td>365</td>
<td>374</td>
<td>369.50</td>
<td>3.556</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>22</td>
<td>365</td>
<td>374</td>
<td>369.50</td>
<td>3.556</td>
</tr>
</tbody>
</table>

Band is **61** points

Band is **13** points

Band is **9** points

Band is **9** points

Band is **9** points
Method 2: Quantitative Covariate Matching with Stratification (Con’t)

• Method 2 with quantitative covariate of pretest score is aborted due to difficulty in obtaining adequate matching within quintiles

• Conclusion: More advanced matching methods on quantitative pretest scores might inform research – proceed to \textit{nmatch} (Abadie, Drukker, Herr, & Imbens, 2004)
Method 3: Quantitative Covariate Matching with *nmatch* in STATA

- Nearest Neighbor Method

**Depvar** = posttest achievement score

**Treatvar** = binary treatment membership

**ATE** = average treatment effect

**Varlist** = covariates that may bias results

Results: No significant difference between the treatment group (65) and the control group (108) – *confirms failed RCT analysis*
Summary

• Embrace rigorous evaluation methods for after school programs
• Begin with the most rigorous design such as RCTs
• Be prepared to incorporate additional matching designs to compensate for difficulties with RCTs
Contact

Silvia Swigert, Director of Evaluation
Center for Educational Partnerships (CFEP)
UC Irvine
Email: sswigert@uci.edu
Phone: (949) 824-0201
www.cfep.uci.edu