PROVE WORKSHOP:

*P5: Applying Rigorous Research*

---

**PROVE – Building Rigorous Evidence**

Today:

- Need for rigorous research and evaluation to build evidence base within GEAR UP
- How to move your project in this direction

Facilitator:

- Alina Martinez, Principal Associate, Abt Associates
Building Rigorous Evidence

Objectives:

- Gain an understanding of trends in research and evaluation as we move towards increased evidence-based outcomes in college access and success evaluation.
- Learn about techniques that can be applied to your overall evaluation or in a new grant application to increase evidence-based outcomes.
- Learn about the application techniques for increasing rigorous research and evaluation in GEAR UP programs.

Workshop Overview

- What came before
- Federal expectations regarding evidence
- ED’s evidence levels
- What Works Clearinghouse Standards and their relationship to ED’s evidence level
- Key components of rigorous evaluation
Evaluation in GEAR UP – Strong Plan (2014 RFP)

- A strong evaluation plan should:
  - Shape the development of the project from the beginning to the end of the grant period;
  - Include benchmarks to monitor progress toward meeting specific project objectives;
  - Include a summative evaluation approach for assessing the likely contribution of the project to improving student outcomes.
GEAR UP Evaluation – Plans (2014 RFP)

- Plan should clearly indicate:
  1. What types of data will be collected;
  2. When various types of data will be collected;
  3. For which students data will be collected – participants only, participants and non-participants, which cohort (i.e., in what year did the students start);
  4. What data collection methods will be used;
  5. How the data will be analyzed; and
  6. When reports and outcome data will be available.

GEAR UP Evaluation Plans (2014 RFP)

- Plans should maximize how much confidence the Project or the Program can have in the study results
- Strongest evaluation designs are experimental/randomized control studies
- Other evaluation designs may be easier to implement but do not provide such rigorous results.
Evaluation Independence

- The evaluation should be conducted by an evaluator that is independent of the entities developing or implementing the intervention.
- Independence is important in various aspects of the evaluation, including:
  - Collection of outcome data
  - Analysis of impacts
  - Reporting of the estimated effects of the intervention

Activity – Your Project’s Evaluation Plans

- Discuss in small group
  - What do/did you want to learn?
  - What are/were the key components of the evaluation design?
- Share with full group
  - Areas of investigation
  - Key components
    - Similarities
    - Differences
Activity – Reporting Out

- Areas of investigation
- Key components of evaluations
  - Similarities
  - Differences

Some Issues Important to Rigorous Evaluation

We will tackle a few
- Construction of comparison group
- Attrition
- Baseline equivalence
- Appropriate outcomes
- Confounds

- And point to resources for others
  - Logic models
  - Implementation fidelity
  - Power analyses
Why Does Rigor Matter?

- Increasing federal emphasis on rigorous evaluation
- Providing evidence of effectiveness
- Ensuring that federal dollars are invested wisely
- Informing continuous improvement of programs
  - Change what is not working
  - Do more of what is working

Federal Context

<table>
<thead>
<tr>
<th>Presidential Administrations</th>
<th>Federal Guidance</th>
</tr>
</thead>
<tbody>
<tr>
<td>President Obama (2009 – 2017)</td>
<td>• GPRA Modernization Act of 2010</td>
</tr>
<tr>
<td></td>
<td>• Office of Management and Budget Memoranda</td>
</tr>
</tbody>
</table>
Convergence of Efforts to Strengthen Federal Research and Evaluation

- Office of Management and Budget (OMB)
- Congress
- Focus on evidence across Agencies
- ED efforts
  - What Works Clearinghouse™
  - Evaluation technical assistance to grants

OMB Memoranda

- M-10-01 Increased Emphasis on Program Evaluation
- M-12-14 Use of Evidence and Evaluation in the 2014 Budget
- M-13-17 Next Steps in the Evidence and Innovation Agenda
- M-14-07 Fiscal Year 2016 Budget Guidance, Evidence and Evaluation
Focus on Evidence across Agencies

<table>
<thead>
<tr>
<th>Department or Agency</th>
<th>Example Initiative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>Investing in Innovation (i3)</td>
</tr>
<tr>
<td>Health and Human Services</td>
<td>Maternal, Infant, and Early Childhood Home Visiting Program</td>
</tr>
<tr>
<td>Labor/Education</td>
<td>Trade Adjustment Assistance Community College and Career Training (TAACCT)</td>
</tr>
<tr>
<td>Labor/Education/Health and Human Services</td>
<td>Workforce Innovation Fund</td>
</tr>
<tr>
<td>Corporation for National and Community Service</td>
<td>Social Innovation Fund</td>
</tr>
</tbody>
</table>

Education Department General Administrative Regulations (EDGAR)

- Specifies
  - Requirements for all ED federal grants
  - Criteria used to evaluate GEAR UP applications are drawn from EDGAR 34 CFR part 75 sections 75.209(a) and 75.210.
Relevance of EDGAR (RFP)

- 7 Selection Criteria Stemming from EDGAR
  1. Need for the project (15 points)
  2. Quality of project design (15 points)
  3. Quality of project services (15 points)
  4. Quality of project personnel (10 points)
  5. Quality of the management plan (10 points)
  6. Quality of the project evaluation (20 points)
  7. Adequacy of project resources (15 points)

- In addition, there were competitive preference priorities (3 for States, 2 for Partnerships)

Simplified Overview of EDGAR Evidence Categories

- Strong theory
- Evidence of promise
- Moderate evidence of effectiveness
- Strong evidence of effectiveness
ED Efforts – Sources of Useful Resources

Common Guidelines

What Works Clearinghouse™

WWC: Who, What, and Why

Focus on causal validity of the study design and analyses
Identify standards for rigorous evidence
Developed by panels of national experts
Result in rating after study review of
- Meets WWC Standards without Reservations
- Meets WWC Standards with Reservations
- Does Not Meet Standards
Definitions of Summative Evaluation Design Approaches (RFP)

- Experimental study
- Quasi-experimental study
- Carefully matched comparison group design

A study that employs random assignment of, for example, students, teachers, classrooms, schools, or districts to participate in a project being evaluated (treatment group) or not to participate in the project (control group). The effect of the project is the average difference in outcomes between the treatment and control groups.
Definitions of Summative Evaluation Design Approaches (RFP)

- Experimental study
- Quasi-experimental study
- Carefully matched comparison group design

An evaluation design that attempts to approximate an experimental design and can support causal conclusions (i.e., minimizes threats to internal validity, such as selection bias, or allows them to be modeled). Well-designed quasi-experimental studies include carefully matched comparison group designs (as defined in this notice), interrupted time series designs (as defined in this notice), or regression discontinuity designs (as defined in this notice).

A type of quasi-experimental study that attempts to approximate an experimental study. More specifically, it is a design in which project participants are matched with non-participants based on key characteristics that are thought to be related to the outcome.
Activity – Constructing a Comparison Group

- How would you construct a comparison group to evaluate:
  - Your entire GEAR UP project?
  - A component of your project?

Activity – Reporting Out

- How many RCTs were proposed for:
  - Entire GEAR UP project
  - Components of projects
- What were some approaches to constructing a non RCT comparison group?
### Rigor of Design for Causal Inference

- Designs considered rigorous enough to provide strong evidence of effectiveness (All of these designs have a comparison group)
  - Randomized control trial (RCT)
  - Quasi-experimental design (QED)
  - Regression discontinuity design (not a topic for today–resource sheet)
  - Comparative interrupted time series (not a topic for today–resource sheet)

- Designs not rigorous enough to provide evidence of effectiveness (these designs do not have a comparison group)
  - Interrupted time series
  - Pre-post with no comparison group
  - Correlational studies
  - Post-test only

### Important Considerations/Components

- Logic model
- Fidelity of implementation
- Comparison group
  - Distinct groups
  - Baseline equivalence
- Attrition
- No confounding factors
- Appropriate outcome
- Sufficient power (consider level and units)
Logic Model
You need to know what ‘it’ is

- Implementing multiple strategies (i.e. interventions) concurrently poses challenges to measuring the impact of any of them.
- Need to clearly identify what is being implemented, for whom, how it’s different from “business as usual.”
- Use a logic model.

i3 evaluation TA resources [available on Abt website]

Implementation Fidelity
Evidence that ‘it’ was implemented.

- Measuring implementation fidelity is important...
- ...but impossible to do without a clear understanding of how the intervention is intended to work!
- It is also important to understand whether participants in the comparison condition are having experiences that look at all similar to the treatment condition.

i3 evaluation TA resources [available on Abt website]
Comparison Group
Distinct Groups – (Two Groups, At Least)

- Compare treatment group to comparison group that does not receive the intervention
- Similar on everything except the intervention
- **Not** a comparison of treatment to itself (e.g. same students in previous grade)
- **Not** a previous or later cohort (e.g. freshmen form the previous year)

Baseline Equivalence

- Treatment and comparison groups are similar on observed pre-intervention measures.
- Established on (1) a pre-intervention measure of the outcome or (2) one or more variables that are highly correlated with this outcome.
- Difference in means is < 0.25 standard deviations.
  - Adjust statistically for differences > 0.05 standard deviations.
  - No adjustment required for differences <= 0.05 standard deviations.
Understanding Attrition

- Attrition occurs when the value of the outcome variable is missing for a unit (e.g., student) that was randomly assigned.
- The WWC has standards for the maximum acceptable attrition rate and the maximum acceptable difference in attrition rates between the treatment and control groups.
- WWC topic area protocols specify either the liberal attrition standards or the conservative attrition standards.

Confounds

- Confounds are design features that may yield differences in outcomes between the two groups that are not attributable to the intervention.
- Examples:
  - An “N=1 problem”, in which there is only one “unit”—teacher, student, school, or district—per condition.
  - Systematic differences in sample characteristics that are associated with outcomes (e.g., teacher education).
  - Bundling of the intervention with another intervention.
Appropriate Outcomes

- Follow from theory of intervention
- Used to determine impacts
- Include
  - Face validity – appears to measure intended outcome
  - Reliability – consistency
  - Not overly aligned with intervention
  - Collected in same manner for treatment and comparison

Outcomes:

1. **Face validity**

- An outcome has face validity if it appears to measure the domain into which it has been classified.
Outcomes:
2. Reliability

- Minimum WWC standards for reliability:
  - Test-retest reliability ≥ 0.40
  - Internal consistency ≥ 0.50
  - Inter-rater reliability ≥ 0.50
- Standardized tests (e.g., nationally normed tests and state assessments) are assumed to have adequate reliability.

Outcomes:
3. Alignment with the Intervention

- There should not be over-alignment.
- Over-alignment occurs when outcome measures are closely tailored to the intervention.
- Example: An outcome measure based on an assessment that relied on materials used in the intervention condition but not in the comparison condition (e.g., specific reading passages) likely would be judged to be overaligned.
Outcomes:
4. Collected in the Same Manner

- Should be **collected in the same manner** for the treatment and comparison groups.
- A measure would be collected in a **different** manner, for example, if:
  - Different modes, timing, or personnel were used to collect outcome data for the two groups.
  - Outcome measures were constructed differently for the two groups.

Power Analysis

- Find the needed sample size
  - For a given power and effect size
- Find the power
  - For a given sample size and effect size
- Find the minimum detectable effect size (MDE)
  - For given sample size and power

i3 evaluation TA resources [available on Abt website]
Activity – Common Issues

- What are some common issues encountered in RCT and QEDs?

Activity – Reporting Out

- Common issues for
  - RCTs
  - QEDs
  - Both
Common Challenges

- RCT
  - Perception of denied services/unfairness
  - Recruiting participants
  - Information needed to implement random assignment (*what do you need and how will you get it?*)
  - Attrition
- QED
  - Constructing a comparison group
  - Matching
  - Establishing equivalence at baseline
  - Extant data on the comparison group
- Both
  - Collecting outcome data in the same way across T and C groups
  - Accounting for clustering

Discussion – Moving toward More Rigor

- How do we collectively move to more rigorous accountability?
- What steps can you take to move toward more rigorous evaluation?
### Introducing More Rigor

<table>
<thead>
<tr>
<th>Instead of …</th>
<th>Consider producing stronger evidence …</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic model/theory</td>
<td>Correlational study</td>
</tr>
<tr>
<td>Correlational evidence</td>
<td>Quasi-experimental study</td>
</tr>
<tr>
<td>Quasi-experimental</td>
<td>Experimental study</td>
</tr>
<tr>
<td>Experimental study</td>
<td>A larger experimental study with varied populations</td>
</tr>
</tbody>
</table>

### Wrap Up – Resources Available

- **i3 resources for evaluators, selected resources recently uploaded**
  - Available on Abt website
- **Designing strong studies (webinar)**
- **Designing quasi-experiments (webinar, PPTs)**
  - [Link](http://education.ufl.edu/educational-research/files/2015/04/qed_presentation_slides_030315.pdf)
Thank you for attending this workshop session.