Abstract Title Page

Title: Measuring Implementation Fidelity: Tradeoffs in Using Developer and Evaluator Measures

Authors and Affiliations:

Rekha Balu and Janet Quint: MDRC

Jose Blackorby: SRI International

Jill Lammert and Jill Feldman of Westat, and Sonia Jurich independent consultant

Beth Boulay of Abt Associates will serve as moderator
Abstract Body

Background / Context:

One of the essential elements of a well done evaluation is a careful analysis of the fidelity of implementation of the education intervention under study. Without this, it is impossible to know if the intervention was actually put in place or the impact study is studying the effect of something other than the intended treatment. Further, a fidelity analysis is part of understanding the “service contrast” between the program and control or comparison groups – the central difference that produces impacts. And the information on fidelity and service contrast is often important in efforts to understand why impacts vary across sites and how to improve program effectiveness.

Interventions vary in their complexity and the extent to which they are “prescribed or manualized,” so the definition of fidelity may vary from faithfulness to key principles of action to careful adherence to detailed procedures and organizational structure. In addition, some programs and evaluators feel their program can be dropped into most any setting “as is”, while others expect adaptations to occur.

Given this central role and varying analytic challenges, evaluators have tried many different techniques to measure fidelity. Many evaluators have created their own measures of program fidelity anchored on the theory of action of the intervention, but others have relied heavily on a measure of fidelity created by the program developer for operational purposes. Still others have adopted a hybrid approach. These choices have in part been driven by the nature of the intervention under study. If the program is relatively straightforward and key elements are visible in short, periodic observations, evaluators have tended to develop their own measures that are used in field visits. When the intervention is more complex and involves detailed procedures and many different elements, tracking fidelity through observation becomes costly and requires special knowledge and evaluators have sometimes shifted toward use of developer ratings of fidelity and/or surveys of those implementing the intervention.

Purpose / Objective / Research Question / Focus of Study:

This panel will explore the tradeoffs evaluators confront in designing a strategy to measure fidelity through a series of cases studies of recent evaluations of reading programs. These tradeoffs are often related to data collection cost, training and expertise needed to implement a method, measurement validity and reliability, and the extent to which the tool can also be used to assess the control condition well enough to contribute to measuring the service contrast or relative strength of implementation. The tradeoffs vary depending on whether one uses an instrument created by a program developer or a researcher or some hybrid.

Developer measures are often created for operational purposes to guide further technical assistance and are sometimes either highly detailed or involve measures of many different program features. They typically cannot be used in control schools to measure service contrast because they are program specific. However, developers may have highly trained staff quite familiar with the program who serve as the “raters” and these individuals often have frequent contact with evaluation sites. This may provide more ratings of implementation fidelity based on much more information about the program than researchers can glean in a few visits. In addition, data from a developer's instrument may be more easily compared to other rollouts of the program.

Researcher-created measurement tools can be more focused on the research questions in an evaluation and may contribute to measuring service contrast if they can be used in both program and control sites. And their properties as data can be more easily assessed through analysis such as inter-rater reliability checks. However, they may be difficult to compare to prior program iterations, and it may be costly to
train evaluators to assess complex multi-dimensional educational programs or have multiple visits to sites to conduct enough observations to represent the program actually.

This panel proposes to explore these tradeoffs through several evaluations that approached the fidelity measurement challenge in different ways. The case studies are:

1) an ongoing i3 scale up study of the Success for All (SFA) elementary school reading program in five school districts;

2) an ongoing i3 validation study of the Denver Public Schools (DPS) Collaborative Strategic Reading (known as Colorado CRS) program for middle school students; and

3) two studies of programs for struggling adolescent readers as part of the striving readers grant program.

Beth Boulay of Abt Associates, will moderate a discussion of the tradeoffs evaluators face in measuring fidelity using the cases as examples.

**Setting:**

The case study evaluations are of reading programs designed to address the needs of low performing students. They were all conducted in the United States in public schools and have either recently been completed or are still underway. All studies involve multiple schools and most involve multiple school districts.

**Population / Participants / Subjects:**

Students in the evaluations attend public schools in the United States and are in elementary or secondary schools.

**Intervention / Program / Practice:**

All studies are of reading instruction.

- The SFA elementary school reading program involves a specific way to provide reading instruction to all students, use of data to measure progress and group students for instruction, and changes in school wide operations to support better reading achievement.
- Denver’s CRS program involves professional development for teachers in social studies and science to help students develop meta-cognitive awareness and learn four specific strategies associated with enhanced reading comprehension.
- The striving readers programs: One program provides professional development for teachers to support the integrated, school wide use of reading strategies to teacher math, English, science, and social studies plus use of the Read 180 program as a supplemental reading program. The second uses Voyager Passport Reading Journeys (VPRJ), a supplemental reading program for adolescent students reading below grade level. VPRJ is highly structured and blends teacher-led targeted instruction with student-centered technology in a series of lessons designed to be delivered over the course of one school year.

**Significance / Novelty of study:**

The topic is of great interest to evaluators and funders of evaluation. There has been a long history of efforts to study program fidelity in evaluations, so it is not so much that there are missing elements in
prior work. Rather, recent experience with large scale, multi-site studies of education interventions inside i3 and other programs has brought home the tradeoffs evaluators face in designing strategies to study fidelity. Intensive strategies that require extended observation of a program work well in small scale studies but can be prohibitively expensive in large scale projects. On the other hand, relying on developer created and completed measures of fidelity hampers the evaluator’s ability to craft a data collection tool that is tailored to the needs of the study and they are not really traditional research data collection instruments with known properties. And survey-based strategies for teacher reports of practice have their own limits, often related to concerns that teachers may not accurately report their instructional practice.

So the proposed panel is a chance for experienced evaluators who have faced these tradeoffs to share their experiences and set up a discussion with others about strategies to adopt. We think this case study-based discussion will be a unique contribution to the discussion.

**Statistical, Measurement, or Econometric Model:**

All of these evaluations used random assignment to estimate program impacts, some random assignment of schools and others random assignment of students. The aspect of methodology discussed in the proposal panel is how the evaluations assessed the fidelity of implementation of the intervention.

- The SFA evaluation largely relied on a snapshot of implementation fidelity completed by SFA coaches working with each treatment school that included 99 measures covering the many aspects of the program and then worked with the SFA staff to identify key features of SFA as rolled out in the first program year and subsequent years. In addition, the evaluation collected data on key aspects of the SFA program through surveys with principals and teachers in both the SFA treatment and control schools.

- The Colorado CRS evaluation measured fidelity through: teachers’ self-reports, through an online database, of the number of minutes spent each week implementing CSR; data from classroom observations of CSR implementation collected by CU-Boulder researchers and analyzed by SRI researchers; students’ response to questions designed to measure familiarity of terms unique to CSR collected at the end of the 2011-2012 school year by SRI researchers; and teachers’ response to questions about CSR implementation collected through their completion of an online survey at the end of the school year.

- The striving readers studies tackled implementation fidelity in two different ways. The innovation in the study of a school-wide reading program involved using an innovation configuration map to explicate targeted approaches and distinguish them from among a range of commonly used practices, designing an observation protocol aligned with the map, and providing evaluators with intensive training to collect observation data and assign fidelity ratings. The of the Voyager program developed a framework for research questions based on the program and included classroom observations and monthly surveys of each teacher and interviews with the project director and each Cambium VPRJ coach assigned to each school.

**Usefulness / Applicability of Method:**

Fidelity studies within evaluations typically rely on one or more of the methods covered in the case studies on the panel. Thus the challenges and tradeoffs faced in these studies should have wide applicability in education evaluations.
**Research Design:**

See discussion above. Random assignment is the method used for estimation of program impacts in all of the case study evaluations. Methods for assessing fidelity vary widely across the studies, reflecting the variety of the interventions and choices made by the evaluators.

**Data Collection and Analysis:**

See discussion above.

**Findings / Results:**

N/A

**Conclusions:**

The panel will illustrate the range of methods currently used to assess implementation fidelity in large scale evaluations of education initiatives and the tradeoffs in the design of data collection and analysis strategies. Key aspects of the tradeoffs include:

- budget costs driven by the need for repeated on-site time for program observation or the skill levels needed to conduct the observation;
- the complexity of the program and the resulting scope of needed data collection;
- the frequency of occurrence (daily versus over extended intervals) of key elements of the program;
- the accuracy and objectivity of assessments of fidelity by different types of observers;
- the ability to craft survey questions that draw out reliable reports from teachers and other staff of their actions; and
- the ability of researchers to distill detailed operational measures of fidelity into an understandable and accurate overall fidelity rating.