Challenges to Finding a Number: Developing a System to Measure Fidelity of Implementation of an Adolescent Reading Intervention

Jill Lammert, Ph.D., Westat
JillLammert@westat.com

Sonia Jurich, Ph.D., Independent consultant
soniajurich@cox.net

Paper presented at the APPAM Annual Fall Research Conference, November 7-9, 2013, Washington, D.C.

Introduction

Since 2009, an increasing number of notices for funding opportunities in the U.S. Department of Education have included references to fidelity of implementation or a similar term (U.S. Government, 2013). Notably, the Obama Administration's signature education programs, Race to the Top (RTT) and Investing in Innovation (i3), require grantees to conduct external evaluations that include studies of fidelity of implementation in addition to impact studies.

This increased awareness at the highest levels of government of the need to measure fidelity of implementation of interventions has been accompanied, or perhaps even influenced, by greater interest among researchers and evaluators in the growing field of implementation research. In 2011, the National Implementation Research Network organized the first Global Implementation Conference (GIC), which was held in Washington, D.C. and attended by over 750 participants from 18 countries. The GIC served as "the first global forum for reporting implementation research and evaluation results, sharing implementation best practices, and advancing public policy to support implementation science and practice across human services" (GIC, 2013). Another GIC was held in August 2013 and additional conferences are scheduled every two years from now until 2020.

Notwithstanding this greater emphasis on advancing the field, currently "implementation science is not well organized, best implementation practices are not commonly used, and policymakers are slow to invest in new ways of thinking and doing" (Global Implementation Initiative, 2013). Researchers and evaluators tasked with developing systems to measure fidelity face a number of challenges when trying to generate a score that can accurately represent fidelity of implementation across a wide array of different types of interventions, contexts, programs and practices.

This paper describes our experience developing the system to measure fidelity of implementation of an adolescent reading intervention being implemented in two states.[1] Following a brief description of the intervention, we outline the process we went through to develop the fidelity measurement system. We then describe the benefits and limitations associated with our choices and conclude the paper with a discussion of the lessons learned from our experience and implications for future studies.

Description of the Intervention

In 2009, the U.S. Department of Education's (ED) Striving Readers program awarded discretionary grants to eight states to implement reading interventions designed to raise middle and high school students' literacy levels in Title I-eligible schools with significant numbers of students reading below grade-levels. The Striving Readers program purposely sought to build a strong, scientific research base for identifying and replicating strategies that improve adolescent literacy skills. To that end, grantees were required to contract an external evaluation of the intervention that included (a) a study of the impact of the intervention on students' performance on standardized assessments using a randomized control trial (RCT) design, with students randomly assigned to treatment or control groups within schools and (b) a study of fidelity of implementation of the intervention. We were awarded contracts to conduct the external evaluations for two of these state projects.

Although the target students were in different grades, both states chose to implement the same reading intervention, Cambium Learning Systems' *Passport Reading Journeys* (PRJ). PRJ is an

[1] This work was completed while both authors worked at RMC Research Corporation.

adolescent reading intervention that blends teacher-led targeted instruction with student-centered strategies, and uses information technology to engage students and reinforce instruction. The program is formatted as a series of lessons designed to be delivered over the course of one school year. Across grade levels, the intervention maintains the same structure but the content and reading level changes.

PRJ is designed to be delivered in daily, 50-minute lessons that provide explicit, systematic instruction in critical reading skills. The lessons are organized into Expeditions, with an average of 15 Expeditions per grade level. Each Expedition is organized in 10-lesson routines to facilitate teacher-led instruction and students' independent practice and lessons alternate between whole group and individual practice. In selected lessons teachers are allowed to choose from a range of activities and students are expected to spend some of their time in independent or paired practice working on Cambium's Strategic Online Learning Opportunities (SOLO) program. SOLO is an interactive, web-based reading resource component that gives students opportunities to engage in self-paced practice of vocabulary and comprehension skills. SOLO also includes assessment features to gauge student learning over time. The intervention materials for every PRJ classroom include teacher guidelines, student workbooks, DVDs, and a library of age-appropriate fiction and non-fiction books and magazines that are designed to engage the adolescent reader.

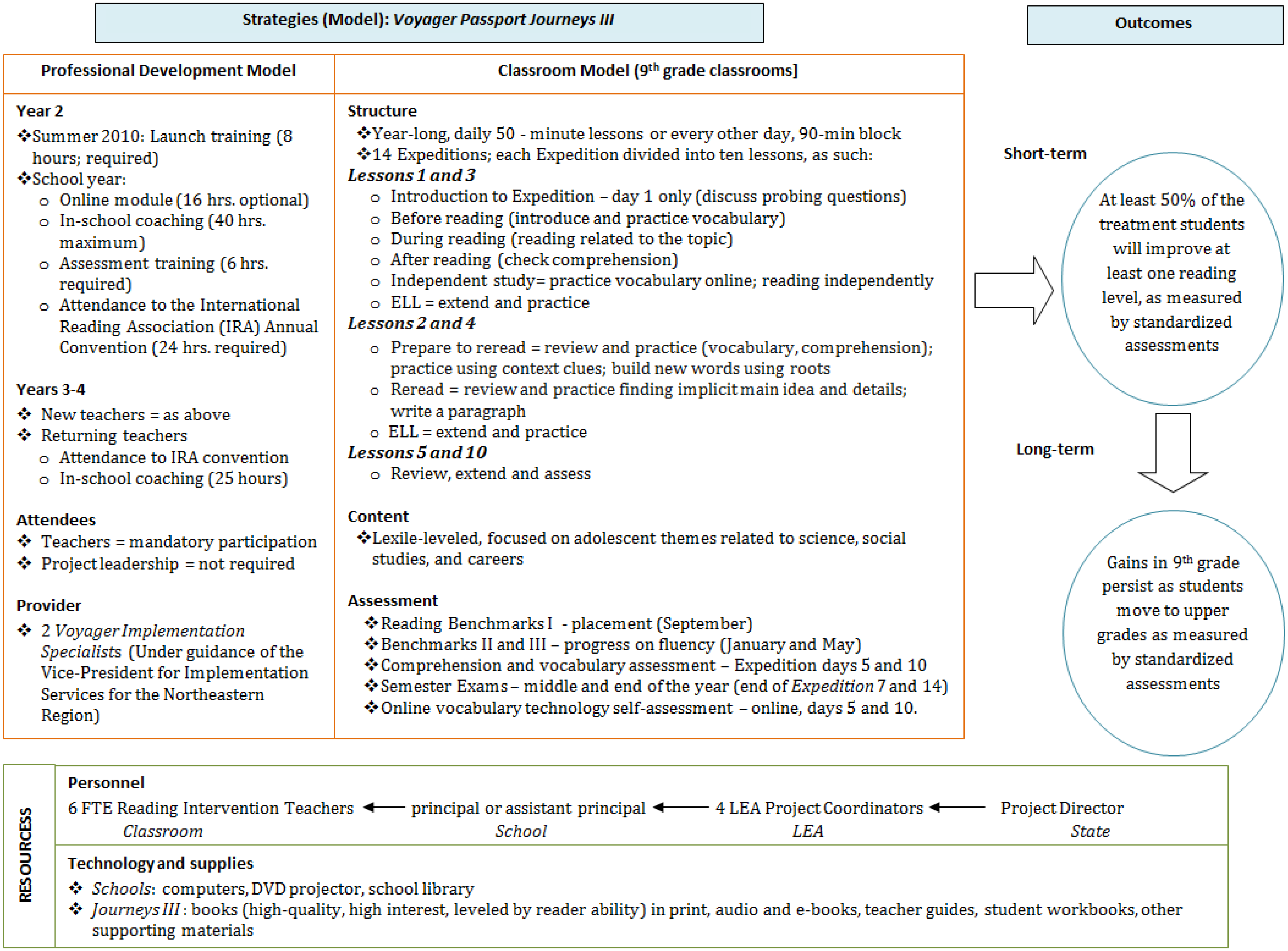
Cambium delivers the initial training prior to the start of the intervention, offers online professional development (PD) modules, and sells packages of coaching; the PD modules are included as part of the intervention, but customers purchase more or less coaching depending on their budget and their needs. All of the data collected through SOLO, along with data on teachers' use of PD modules, are stored in Cambium's online data system, VPORT. Teachers are expected to attend all of the required professional development activities (which may vary depending on the package that was purchased) and to follow a scripted guide that details what, how, and when they will teach. Both the PD and coaching are targeted at getting teachers to implement the intervention with high levels of fidelity to the scripted instructional materials. Diversions from the model generally are not expected or welcome, except for minor adaptations to adjust the expected pacing within the allotted classroom time or to differentiate instruction in suggested ways.

A total of 15 schools were involved in the two projects. All schools were Title I-eligible schools that had not made, or were at-risk of not making, adequate yearly progress requirements under the *No Child Left Behind* Act of 2001. Students who scored at least two years below grade level were eligible for the study, but those students whose Individualized Education Plans precluded their participation in the study or whose parents requested that their children did not participate were excluded from the study. In total, 1,768 students were eligible to participate.

Each state was given a planning year before starting implementation of the classroom intervention. This allowed the state project directors and school staff to conduct important activities—such as purchasing and installing technology and conducting the PRJ training—prior to the start of the intervention. Although by the time we were awarded the evaluation contracts only approximately five months remained in the planning period, having that time allowed our evaluation team to meet with school staff and teachers to explain the requirements of the study and create buy-in, work with district and school staff to carry out the random assignment of students, and develop our fidelity measurement system.

The logic model presented in Figure 1 illustrates the intervention as implemented in one of the two states. While there were slight variations between the states (particularly related to the professional development component), the main features of the intervention and the general system developed to measure fidelity of implementation were the same in both states. For simplicity, we thus will focus our discussion on the process we went through to develop the system to measure fidelity of implementation in general, rather than on measuring specific aspects of the intervention in each state.

Figure 1. PRJ Logic Model



Strategies (Model): *Voyager Passport Journeys III*

Professional Development Model

Year 2

- Summer 2010: Launch training (8 hours; required)
- School year:
  - Online module (16 hrs. optional)
  - In-school coaching (40 hrs. maximum)
  - Assessment training (6 hrs. required)
  - Attendance to the International Reading Association (IRA) Annual Convention (24 hrs. required)

Years 3-4

- New teachers = as above
- Returning teachers
  - Attendance to IRA convention
  - In-school coaching (25 hours)

Attendees

- Teachers = mandatory participation
- Project leadership = not required

Provider

- 2 *Voyager Implementation Specialists* (Under guidance of the Vice-President for Implementation Services for the Northeastern Region)

Classroom Model (9th grade classrooms]

Structure

- Year-long, daily 50 - minute lessons or every other day, 90-min block
- 14 Expeditions; each Expedition divided into ten lessons, as such:

Lessons 1 and 3

- Introduction to Expedition – day 1 only (discuss probing questions)
- Before reading (introduce and practice vocabulary)
- During reading (reading related to the topic)
- After reading (check comprehension)
- Independent study= practice vocabulary online; reading independently
- ELL = extend and practice

Lessons 2 and 4

- Prepare to reread = review and practice (vocabulary, comprehension); practice using context clues; build new words using roots
- Reread = review and practice finding implicit main idea and details; write a paragraph
- ELL = extend and practice

Lessons 5 and 10

- Review, extend and assess

Content

- Lexile-leveled, focused on adolescent themes related to science, social studies, and careers

Assessment

- Reading Benchmarks I - placement (September)
- Benchmarks II and III – progress on fluency (January and May)
- Comprehension and vocabulary assessment – Expedition days 5 and 10
- Semester Exams – middle and end of the year (end of *Expedition* 7 and 14)
- Online vocabulary technology self-assessment – online, days 5 and 10.

Outcomes

Short-term

At least 50% of the treatment students will improve at least one reading level, as measured by standardized assessments

Long-term

Gains in 9th grade persist as students move to upper grades as measured by standardized assessments

RESOURCESS

Personnel

6 FTE Reading Intervention Teachers ← principal or assistant principal ← 4 LEA Project Coordinators ← Project Director

Classroom *School* *LEA* *State*

Technology and supplies

- *Schools*: computers, DVD projector, school library
- *Journeys III* : books (high-quality, high interest, leveled by reader ability) in print, audio and e-books, teacher guides, student workbooks, other supporting materials

Measuring Fidelity

At this point in the evolution of the field of implementation research, there is no consensus on the "best" way to measure fidelity of implementation. Increasingly, a distinction is being made between studies of "fidelity of implementation" and studies of "fidelity of intervention." In the former, researchers examine the fidelity of the *structure* of an intervention, such as the difference between the "delivered" program inputs and activities compared to the "planned" inputs and activities. In the latter, researchers examine the fidelity of the *process* of implementing the intervention, such as differences among the "observed" mediators of changes in student behavior and knowledge compared to the "expected" mediators of student behavior and knowledge (Goodson & Darrow, 2013). Within this framework, what needs to be measured clearly depends on whether a researcher is examining fidelity of implementation or fidelity of intervention. The specific methods used to measure fidelity, in contrast, can be used in both types of systems.

Whichever type of fidelity study one plans to conduct, some common issues need to be addressed, including:

a. deciding to what extent the developer of the intervention will be involved with the creation of the fidelity system and with data collection;
b. determining which components of the intervention are key to implementation with fidelity;
c. identifying indicators of implementation of the key components and selecting which indicators feasibly can be measured;
d. pinpointing which sources of data will provide the "best" information on each indicator;
e. choosing whether to adopt existing data collection instruments or develop new ones;
f. deciding whether it will be necessary to measure treatment-control contrast;
g. developing a method to calculate a fidelity score for each component, including assigning ratings and weights and choosing fidelity thresholds; and
h. choosing whether and how to roll-up the fidelity scores from each component to the program level in order to generate one overall fidelity score.

In the sections that follow we describe the process we went through to address these issues within the context of the broader points of deciding what to measure and determining how to measure it.

Deciding What to Measure for Fidelity

The targeted nature of the intervention provided a relatively straightforward framework from which we could begin developing our system to measure fidelity of implementation.[2] We chose to focus our implementation study on collecting data on the teacher PD and PRJ classroom implementation components in treatment group classrooms, as discussed below.

The first step in developing our system to measure fidelity was to decide what we could reasonably expect to measure within the budgetary, time and contextual constraints of the study; that is, within the context of a student-level RCT being conducted within schools. Major considerations in deciding what to measure for fidelity included: (a) data collection needs and capabilities, (b) the number of schools and teachers participating in the intervention, (c) district research requirements,

[2] At the time we were developing our system we were not aware of the distinction between the terms "fidelity of implementation" and "fidelity of intervention." For this reason, although our system contains elements that Goodson and Darrow (2013) might classify as "fidelity of intervention" we consider our study to have focused primarily on fidelity of implementation and will thus use this term throughout our paper.

including requirements for obtaining parental and student consent, and (d) whether to measure the treatment-control contrast.

We knew that we would need to collect information about teacher professional development and classroom implementation of PRJ, so we decided to focus our fidelity measures on these two components. Indeed, the logic model depicted in Figure 1 focused primarily on these two aspects of the intervention; the different school-level structures and supports and district and school administrative support elements were listed as "resources" that would support implementation of the intervention yet not be included in our measurements of fidelity.

The limited number of schools and teachers participating in the intervention (n = 15) made it possible for us to include in our evaluation plan in-person informational meetings and data collection activities at each participating school. Since this was a state-level project that was being implemented in selected districts, we did not have much difficulty creating buy-in for the study. Nevertheless, we visited all schools during the planning period in order to meet with the school principals and data collection liaisons, explain the study requirements, and answer any questions they may have. In addition, we budgeted three classroom observations per teacher per year, as well as monthly phone calls with teachers for the duration of the project. We also included interviews with the PRJ coaches and the state project directors as sources of data triangulation.

We chose to gather all student outcome data from district or school administrative records or from VPORT, Cambium's online data system. This allowed our study to be considered exempt from human subjects protections, since it entailed collecting student data as part of students' regular education program. We did obtain passive consent from parents and students in selected districts, as necessary.

Although our study was an RCT with students randomly assigned to treatment and control groups, when developing our fidelity system we chose not include measurement of treatment-control contrast. We felt this was a reasonable decision because schools signed a Memorandum of Agreement with the states stipulating that treatment group students would be assigned to special classes that incorporated the PRJ intervention, while control group students would be assigned to non-reading related elective classes. School staff was required to agree with the terms of the randomized study and not provide any supplemental reading instruction beyond the typical instruction provided to struggling readers in each school. We used class rosters to monitor the integrity of the random assignment throughout the study period and did not detect any serious issues with crossover.

Determining How to Measure Fidelity

The next step in the process was determining how we were going to measure implementation of the different components. At this point we needed to consider (a) the data sources for each indicator of implementation of the key components of the PRJ intervention; (b) whether we could use existing data collection instruments; (c) if we could not use existing instruments, how we would develop our own instruments; (d) how we would conduct the data collection activities; (e) how we would calculate fidelity for the different key components; and (f) whether and how we would create an overall school-level or program-level fidelity score.

Answering these questions required gaining an in-depth understanding of the intervention and the mechanisms by which it could be expected to achieve its desired results. We accomplished this by conducting multiple meetings with the developers, conducting in-person meetings with staff at all of the participating schools, and obtaining copies of all of the PRJ instructional materials. We enlisted two literacy specialists to attend the initial PRJ teacher training, conduct a thorough review of the instructional materials, and guide our thinking about fidelity based on the instructional materials and their deep knowledge of literacy instruction. The information included in the logic model was the direct result of our efforts and pointed to specific indicators that we would need to collect as part of our fidelity system.

We identified four primary data sources for our study of fidelity: classroom observations, monthly check-ins with teachers, interviews with PRJ coaches and project directors, and VPORT. Although we would have liked to administer student surveys to obtain information regarding their experience with and perceptions of the intervention—thereby allowing us to measure student-level mediators of the intervention—we chose not to in the end because of the data collection and consent requirements.

Due to the lack of existing data collection tools that we could use for the classroom observations, check-ins, and interviews, we had to develop our own instruments, as described below. VPORT provided information that we used to inform the calculation of fidelity for the teacher PD component.

Developing the data collection instruments

Classroom observation tool. Our PRJ classroom observation rubric was created in collaboration with the developer, based on the developer's own implementation rubric, and designed to be completed by members of an evaluation team who may or may not have extensive knowledge of the PRJ intervention. During our meetings with the Cambium research team we learned that PRJ coaches used Cambium's *Classroom Status Rubric* (CSR) to measure "fidelity of implementation" whenever they conducted a classroom visit. The CSR covered five areas: quality of instruction, amount of instruction, differentiation, classroom management, and use of assessments.

After a careful review of the rubric, we decided to create our own rubric for the following reasons. First, the CSR was not intended as a true fidelity measure, but rather it was a formative assessment of teacher practice. Second, it required significant background knowledge of PRJ and literacy instruction to score, making it difficult for a member of the evaluation team to complete. Third, it was a subjective measure with unclear weights for each item, adding additional complexity to the scoring. Fourth, it included a mixture of classroom instruction items and overall PRJ implementation items that did not fit into a coherent fidelity index. Finally, it excluded elements of the classroom environment (such as having sufficient space to carry out the different activities) that were part of the intervention as outlined in the instructional materials.

Rather than abandoning Cambium's rubric altogether, however, we used it as a starting point to develop an instrument that incorporated the key elements of the CSR and closely followed the PRJ lesson structure. We had numerous meetings with the developer to discuss the intervention, the CSR, and the intervention components that they considered key to fidelity. The resulting rubric included four sections: classroom environment, general and lesson-specific information about lesson planning and delivery, and classroom management and behavior.

An important benefit of our observation rubric beyond the CSR is that it is lesson-specific and can be completed easily by members of an evaluation team, rather than by the PRJ coaches. To achieve this, our literacy specialist created an individualized outline for each lesson based on a thorough review of the instructional materials. The methodologists on our team then worked to translate the outlines into specific indicators and ratings for each aspect of fidelity. We made rubrics for all 10 lessons, although some of the lessons followed a similar structure, so it was not necessary to make 10 different instruments. We used our first round of observations in one state to test the rubric and made adjustments to the rating scales based on feedback from the different observers. Appendix A presents the observation rubric for Lesson 1.

As can be seen, the instrument is broken down into six sections: an introductory section that provides instructions to the observer, an overview of the observation context, and the four sections related to classroom environment, general and lesson-specific information about lesson planning and delivery, and classroom behavior and management described below.

Section A covers aspects of the classroom environment and is to be completed at the beginning of the lesson. Observers indicate whether the environment has sufficient space, designation of instructional areas, and materials on display (0 = no, 1 = partially, 2 = yes). The purpose of Section B is to gather a general idea of whether the teacher follows the instructional guidelines as planned (rating options are 0 = no, 1 = partially, 2 = yes, and n/a = not observed). This section includes items that originally were included in the developer's observation rubric. For measurement reasons we would have chosen to eliminate or alter some of the poorly worded or unclear items (e.g., Item 6: Pace is brisk and business-like, yet personal), but in the end we decided to leave them unchanged since the developer felt they were important elements of the PRJ intervention. To address possible inconsistencies that may have arisen in the ratings, we worked with the Cambium research team to develop consensus about how to rate these items. Section B should be completed at the end of each lesson, to allow the observers to make a rating based on a holistic view of the lesson.

Section C changes depending on the lesson. Using the outline of each lesson as a guide, we created specific indicators that allow raters to objectively identify if the teacher is following the instructional guidelines as expected. For example, the instructional materials provide guidelines for the amount of time that each part of the lesson should take, so our instrument has places for the observers to record the start and end times of each part. During the lesson, observers record whether the primary components of each lesson (in the case of Lesson 1, whole group instruction and independent work) are delivered in order, whether the steps that make up each component are delivered in order, and whether the components are delivered within the allotted time. For the first two columns, the raters assign a rating from 0-3, based on whether the teacher follows the order or makes modifications that are/are not allowed (0 = not in order, 1 = modifications that are not allowed, 2 = modifications that are allowed, 3 = in order). Observers are expected to keep detailed notes of the observation and any modifications to instruction to help make determinations of whether a modification is allowable in a given lesson context.

Finally, Section D gathers information about classroom behavior and management. Since classroom behavior and management are variable, observers are asked to record data in 10-minute intervals during the lesson, adding more intervals as needed. For each interval, the observer rates the item on a 3-point scale or indicates that the item is not applicable for that specific time interval. The observers then calculate the total score for each item as the proportion of time that the behavior was observed relative to the total possible time intervals. For example, if a behavior was observed "frequently" (a score of 2) in 3 out of 5 10-minute intervals and observed "not at all" (a score of 0) in the remaining intervals, the score for that item would be 0.6. In the case of block scheduling, there is an expectation that a teacher will complete two lessons during the period, so we ask observers to complete two separate observation protocols for each lesson.

Teacher check-in protocol. We designed the teacher check-in protocol to collect information related to the teacher PD and PRJ classroom instruction components, as well as school-level factors that might influence implementation. The protocol asked teachers questions related to:

- The quantity and focus of any PD and coaching they had received that month;
- The reasons for not obtaining any PD or coaching (from the PRJ coach or the state project director) that month;
- The quality of the coaching support and the relationship with the coach or project director;
- The number of students enrolled in their PRJ classes;
- Student attendance in their classes;
- School closures and class cancellations;
- Pacing of the PRJ intervention;
- Types of optional activities chosen for selected PRJ lessons; and
- Teachers' additional work assignments beyond teaching the PRJ intervention.

Interview Protocol. The interviews with the PRJ coaches and the state project directors were designed as tools to triangulate data on fidelity obtained through the monthly check-ins and the classroom observations. As such, these open-ended interviews asked questions related to the type and focus of coaching support, PRJ classroom implementation problems, and coaches' perceptions of the school-level structures and supports.

Collecting Data

Data collection for fidelity started in the fall of the first year of classroom implementation of the PRJ intervention. Since we did not measure treatment-control contrast, the data collection activities described below were limited to treatment group teachers and classrooms.

Classroom observations. We budgeted three observation visits per teacher, per year, and during each visit the evaluation team observed more than one lesson for each teacher. There were two observers for each visit: a methodologist and a literacy specialist. During the lesson each observer was expected to complete the observation rubric independently. Each observer had a copy of the teacher handbook and was expected to follow along with the instruction. This allowed the observers to know if the teacher was following the different components and steps of the instruction as planned (in Section C of the observation rubric).

Although we had two observers for each lesson, we chose not to calculate inter-rater reliability for each lesson. Rather, we decided that the two observers should talk through their ratings at the end of each lesson and see if they could come to agreement. We decided to come to interobserver agreement in this way because each observer brought different strengths to the observation context and we felt that simply calculating indices of inter-rater reliability might allow these strengths to be overlooked. The methodologists had extensive experience conducting observations for evaluation purposes. The literacy specialists had extensive knowledge of the intervention and the instructional guidelines, and thus were able to inform whether modifications to the delivery of the lesson were appropriate at a given time. Therefore, in Section C, in most cases the literacy specialist could "override" the methodologist in deciding whether to assign a specific rating based on modifications.

In Section D, we also asked the observers to try to reach agreement, but in some instances one observer saw something that the other observer missed, so agreement could not be reached and the ratings remained different.

Initially, we tried to schedule the visits without the knowledge of the teacher so that she could not anticipate an observation and change her instruction accordingly. Scheduling difficulties made that extremely problematic, however, so in the end we had to coordinate with teachers to schedule the visits. We hoped that by observing multiple lessons during a visit and by conducting more than one visit we could get a better idea of the way that teachers actually taught on a daily basis, but we cannot be sure that teachers did not tailor their instruction specifically for our visits. In addition, although we had budgeted for three visits per teacher, scheduling difficulties related to snow closures, assemblies, testing, and other school activities made scheduling visits extremely difficult. In the end, we only were able to complete an average of two visits per teacher.

Teacher check-ins. The monthly check-ins originally started as phone calls, but difficulties with scheduling led us to change the format to an online survey that teachers were supposed to complete at the end of each month. The questions were the same in both formats, but the online format made it much easier to get teachers to comply with the monthly check-ins, and required less follow-up than phone calls. Due to resource and time constraints, we did not examine whether the response patterns were different in the phone check-ins compared to the online check-ins.

Interviews. The interviews were conducted mid-way through the school year, with follow-ups planned at the end of each school year.

VPORT. At the end of the first implementation year we collected data on teacher PD from VPORT.

Calculating the Fidelity Scores

The process of developing a system to calculate fidelity scores involves (a) selecting the range for the fidelity scores (e.g., between 0-1, 10-20, or 0-100); (b) determining the index score for each indicator of an intervention component (e.g., 0-1, 0-3); (c) deciding whether to weight all items related to a certain component equally (e.g., correcting for uneven number of measures for each component or reflecting relative importance of each component); (d) combining sub-component measures to create an index score for each component; (e) deciding whether to weight the different components that make up the intervention equally; and (f) determining the thresholds for different levels of fidelity at the component level (including considerations of the possible values that might be obtained based on sample size or number of observations).

Fidelity of the PRJ classroom instruction component. Table 1 presents the system for calculating the PRJ classroom instruction fidelity score for each section of the observation rubric and overall. To fully understand the scoring system, it may be helpful to refer back to the observation rubric presented in Appendix A. As can be seen in Table 1, we chose to weight the different sections of the observation rubric to reflect their relative importance to the overall intervention. These weights were determined through discussions with the developer. Based on the weighting, the total possible score for the PRJ classroom instruction component was 1.00. In addition, in collaboration with the developer we established the following fidelity thresholds: < 0.70 = inadequate or low fidelity; 0.70-0.89 = medium fidelity; ≥ 0.90 = high fidelity. We determined that medium fidelity would be considered "adequate" in the context of the PRJ intervention.

Table 1. Calculating the PRJ classroom instruction fidelity score

Section	Weight		Section Score	Total Possible Weighted Score
A	.20	x	X_A/8	.20
B	.30		X_B/(total possible score for section)	.30
C	.30		X_C/8	.30
D	.20		X_D/8	.20
			Total possible score	1.00
Fidelity Thresholds:	0.0 – 0.69 = low		0.70 – 0.89 medium	0.90 – 1.0 = high

The score for Section A was calculated as the teacher's score divided by 8, which is the sum of all of the items in the section (since the maximum score for each item is 2). In Section B, the score was calculated as the total teacher score divided by the total possible score for the section. To calculate the total possible score, we took the total number of items (6) minus the number of items that were rated not applicable and then calculated the total possible score accordingly (the total possible points for each item was 2). So, if one item was not applicable to a given lesson, the total possible score for that section was 10. If a teacher received a score of 8 during that lesson, then her score for Section B would be 0.8.

The calculation of the score for Section C was based on whether the components and steps were delivered in order, and whether they were completed in the suggested timeframe. The teacher's score was thus divided by the total possible score for the three columns combined (8). Finally, in Section D we decided that items 4 and 5 should get double points because those were the elements that were completely under the teacher's control, thereby indicating effort to manage behavior appropriately. As such, the score was calculated by taking the teacher score divided by 8 (since the total possible score for items 1, 2, 3, and 6 was one and the total possible score for items 4 and 5 was two).

We started thinking through our system for calculating the fidelity score prior to beginning the classroom observations, but we made modifications to the system based on our experience conducting the observations. For example, after visiting a few classrooms we decided to make items 4 and 5 in Section D count more than the other items (as discussed above), since it was common to see students being disruptive or off-task during the lessons, but we noticed marked differences in how teachers were dealing with those behaviors.

Once we had the data from all of the observations, we followed three steps to calculate the fidelity score for each teacher: (1) all observation rubrics completed by every member of the observation team were entered into our observation database; (2) the different observations for each observation team were combined to get an average score across all observers for each lesson; and (3) the scores for both rounds of observations were then combined to get an average score for each teacher. Table B1 in Appendix B presents the results for all teachers involved in the two studies. As can be seen, all teachers but one achieved at least adequate fidelity and four teachers achieved high fidelity of the PRJ classroom instruction component.

Fidelity of the teacher PD component. Although teachers were offered both professional development and coaching as part of the intervention, our system for calculating the teacher PD fidelity score considered only the teachers' adherence to the required PD. The primary reason for this decision related to the inherent difficulty of determining what constitutes "fidelity" when services are based primarily on need. In the case of the PRJ intervention, states could purchase a variety of coaching packages from the developer, so the nature and frequency of coaching varied across states. Additionally, the PRJ coaches made determinations of the need for coaching based on their classroom visits, so some teachers were identified as needing more coaching while other teachers were identified as not needing much coaching at all. Resource constraints and the relative lack of information related to "best practices" regarding measuring fidelity of needs-based services led us to limit our calculation of fidelity of implementation to only the PD components.

Consequently, we calculated the fidelity score for the teacher PD component as the number of hours of PD the teacher completed relative to the number of hours required by each project. Table B2 in Appendix B presents the results for all teachers involved in the two studies. As can be seen, almost all of the teachers had fidelity scores greater than 1.0 (high fidelity), which reflected their participation in the optional PD activities that were offered as part of the intervention.

Determining school-level fidelity. For reasons described below, rather than calculating one overall school-level fidelity score, we kept the scores on the teacher PD and PRJ classroom instruction components separate and used both scores to serve as indicators of school-level fidelity at each school.

Our initial conceptualizations of fidelity related primarily to the teachers' adherence to the PD and PRJ classroom instruction components. As we considered how to develop a school-level fidelity score, our deliberations focused primarily on how to roll-up the teacher PD and PRJ classroom instruction components into one final score. We considered weighting the two components and simply combining them into one overall score, but it seemed that important information might be lost in the process. This led to numerous discussions with the technical assistance providers assigned to work with our projects about the challenge of calculating one overall score that could accurately represent fidelity of implementation across these different components.

In addition, in our observations and through the monthly check-ins with teachers we discovered a number of school-level factors not specifically related to the teacher PD or classroom instruction components that appeared to be affecting implementation of the intervention, including:

- Lack of support at the school level, such as not being given sufficient space to carry out the intervention as designed or school administrators being unwilling to address serious disruptive student behaviors;
- Technology issues, such as lack of access to computers for each student or problems accessing the internet;
- High numbers of school closures and class cancellations;
- Class periods cut short by interruptions, announcements, and other disruptions that hindered teachers' ability to finish one lesson per class period and to keep on pace with the lesson and Expedition schedule; and
- Lengthy student absences—including students who had been suspended or expelled from school—that affected teachers' ability to maintain the continuity of the lessons.

These elements had not been included in our original fidelity measurement system, but it seemed that they might be important parts of our implementation study. Following the first round of observations in one state we thus began collecting data related to school context factors that might influence implementation, including information on school closures, class cancellations, student attendance, technology support, and administrative support for the intervention. By the time the first implementation year was over it seemed that school context had an important role to play in the implementation of the intervention.

We were beginning to conceptualize a way to incorporate these school-level elements into our system to measure fidelity when the funding for all of the on-going Striving Readers projects was abruptly cancelled. With the untimely cancellation of the grant, we made a decision to stop all activities related to the implementation study and conserve the remaining evaluation funds for data analysis and reporting. Plans for further data collection on implementation were cancelled, and the remaining observation visits also were cancelled.

In the end, rather than use resources trying to come up with a method to calculate overall school-level fidelity scores we decided to keep the fidelity scores for the teacher PD and the PRJ classroom instruction components separate. Each teacher's score on the two components thus became the school's score. The examination of how other contextual elements such as class cancellations, student attendance, and technology issues influenced fidelity of implementation was abandoned.

Discussion

In spite of our disappointment with how our studies came to an end, reflecting back on our experience gives us insight into the benefits of our approach. Our system was based on a thorough understanding of the intervention and of evaluation needs. In close collaboration with the developer, we operationalized key elements of the PRJ model that were not included in the developer's rubric, but that were important to measuring fidelity of classroom implementation. We leveraged content knowledge and methodological expertise to develop rubrics that were tailored to each PRJ lesson yet relatively simple to complete. We had on-going data collection for fidelity over the course of the school year, rather than just measuring fidelity at one point in time. We believe this gave us a better idea of what was actually happening in the classroom than if we had only made one visit to observe each teacher. Finally, our fidelity system incorporated elements of classroom context and teacher and student behavior and thereby allowed us to gather data on some of the mediators of student outcomes that we identified in our logic model.

Despite these benefits, our system nevertheless had its limitations. First, developing the rubric was a very labor intensive process that required substantial amounts of time and in-depth content knowledge not always available to evaluation teams working under time pressures or serious resource constraints. Second, although we piloted the observation rubric in the first round of observations and made changes accordingly, we did not have the time or the resources to establish the reliability of the instrument through a formal pilot test. In addition, we did not calculate inter-rater reliability following each observation, choosing instead to have the observers try to reach agreement regarding their ratings. We felt that the tradeoffs associated with not calculating inter-rater reliability were acceptable within the framework of our implementation study. Nevertheless, while we believe that our observation teams were able to consistently apply the ratings from one lesson to the next, it is unclear whether different observers might achieve similar results—especially if only one observer was conducting each observation and then the rubrics from multiple raters were being combined.

Third, because of resource constraints we chose to use self-report data for some of the components. For example, although we were able to access data on VPORT related to teachers' participation in some of the teacher PD activities, not all of the data on teacher PD was available through the online system. As such, we needed to rely on teacher reports of the type and amount of certain types of PD they had received each month. Fourth, our data collection process required travel to conduct the observations and quite a bit of follow-up with teachers in order to obtain good response rates to the monthly check-ins. This was possible because of the small number of schools and teachers involved in the intervention, but may not be feasible for studies of interventions being implemented across a larger number of schools or teachers.

Fifth, our systems for calculating fidelity scores of the teacher PD and PRJ classroom instruction components were developed based on what we thought was "reasonable," rather than on extensive experience conducting fidelity studies. At the time we were developing our system relatively little information was available about techniques or "best practices" for measuring fidelity. For reasons already discussed, we chose to include some elements of the intervention in the fidelity calculations (e.g., teacher participation in PD activities) and leave other elements (e.g., coaching) out. Despite the increasing interest in measuring fidelity of implementation in the research community, determining how to calculate fidelity of needs-based elements of an intervention remains an issue for evaluators attempting to develop systems to measure fidelity. Further, although our weighting system for the calculation of the classroom instruction fidelity score was carefully thought through and established in collaboration with the developer, the weights were "guesstimates" and not necessarily based on data.

Sixth, our system did not include measurement of treatment-control contrast, so we had no way to associate differences in student outcomes with levels of fidelity of implementation of the PRJ intervention. Lastly, our system did not incorporate measures of school-level structures and supports or district and school administrative support—elements which appeared to have affected implementation and student outcomes, even if not reflected in the fidelity scores for each school.

Notwithstanding these limitations, our experience can help to build knowledge of the tradeoffs involved in developing systems to measure fidelity of implementation and foster discussion among researchers and policymakers interested in the rapidly growing field of implementation research.

Implications for Future Studies

Our experience with the Striving Readers projects came at the beginning of the push toward greater emphasis on implementation science, and the lessons learned are relevant for researchers and policymakers planning for or currently conducting implementation studies. On a very basic level, the inclusion of a planning year in the program funding facilitated the process of preparing for implementation of the project by allowing the implementers to familiarize themselves with the intervention and the study, and opening the lines of communication among participants. Additionally, it gave our evaluation team time to gain an in-depth understanding of the intervention, plan our study, develop our system for measuring fidelity, and work with schools and districts to prepare for implementation.

On a more granular level, our experience points to several lessons that can be applied to future studies of implementation. Indeed, many of the lessons that arose from the Striving Readers program currently are being applied to on-going efforts to measure implementation fidelity in the i3 program evaluations. For example, the initial expectation of the Striving Readers program office was that we would combine fidelity scores for all intervention components to calculate school-level and program-level fidelity scores. This led to numerous discussions among evaluators, technical assistance providers, and Striving Readers program officers about the challenge of "finding a number" that could accurately represent fidelity of implementation across all of these components. Recognizing the challenge that developing one overarching program-level fidelity score poses to evaluators implementing a wide variety of different interventions, the i3 program office requires evaluators to identify the "core components" of each intervention and then calculate fidelity for each component separately. As evaluators conducting these types of studies know, this can be challenging enough in itself.

Another lesson that arose from our experience was the need to allow flexibility in the conduct of fidelity studies so that evaluators can make adjustments as they learn more about the way an intervention is being implemented on the ground. In our case, at the time we were developing our system to measure fidelity we did not realize just how much the different elements related to school-level structures and supports might influence the implementation of PRJ in the classroom. Despite being a quite scripted and "straightforward" intervention, the successful implementation of PRJ was adversely affected by school-level factors that were outside of the teachers' and state project directors' control. Ultimately, these factors affected the pacing of the intervention, resulting in fewer PRJ Expeditions being covered over the course of the school year and, therefore, less overall student exposure to the intervention. We originally had planned to incorporate these elements into our fidelity system, but for reasons previously discussed did not do so in the end. Surely the fidelity of implementation scores would have differed more if these school-level elements had been incorporated into the scoring system.

There is growing recognition that projects implementing interventions such as PRJ should include funding for a comprehensive study of fidelity of implementation. Nevertheless, these implementation studies should not focus only on whether the primary inputs and activities of an intervention are implemented with fidelity (what Goodson and Darrow, 2013, call "fidelity of implementation"), but they also should collect data on the extent to which the key short-term outcomes, or mediators, of an intervention are implemented as intended (what Goodson and Darrow call "fidelity of intervention").

In the context of RCTs, implementation studies should include an investigation of treatment-control contrast, so that evaluators can draw conclusions about the extent to which different levels of fidelity were associated with differences in student outcomes across groups. These studies are resource intensive, however, so funders must make a commitment to ensure that sufficient resources are available to conduct high-quality studies that provide useful information about "what works, for whom, and under what circumstances."

Finally, our experience points to the need for policymakers to ensure that funding for studies such as these does not get cut prematurely, before the intervention can be fully implemented and before any effects the intervention may have on outcomes realistically can be measured. Fixsen, Naoom, Blase, Friedman, & Wallace (2005)—recognized leaders in the field of implementation research—have concluded that "most evaluations of attempted program implementations occur during the initial implementation stage, not the full operation stage" (p. 18). Further, Fixsen et al. found that "evaluations of newly implemented programs may result in poor results, not because the program at an implementation site is ineffective, but because the results at the implementation site were assessed before the program was completely implemented and fully operational…. Only when effective practices and programs are fully implemented should we expect positive outcomes" (2005, p. 4).

References

Fixen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). Implementation research: A synthesis of the literature (FMHI Publication #231). Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network.

Global Implementation Conference-GIC. (2013). *Past and future GICs*. Retrieved from http://globalimplementation.org/gic/pastfuture

Global Implementation Initiative- GII. (2013). *About the GII*. Retrieved from http://globalimplementation.org/about

Goodson, B., & Darrow, C. (2013). *Fidelity of Implementation: Developing measures and linking to impacts.* Presented at the Spring meeting of the Society for Research on Educational Effectiveness, March 7, 2013, Washington, DC.

U.S. Government. (2013). *Federal Register*. Retrieved from: https://www.federalregister.gov

We would like to thank Patty Troppe for her valuable feedback on this paper.

Appendix A

Sample PRJ Observation Rubric – Lesson 1

Instructions to observer

1. Observers should stay for the whole class; if you need to leave the room before the end of the class, please check here and explain the reason in the back of this page _________________

2. Before starting the observation,

    a. Ask the teacher the number of the Expedition you will observe (from 1 to 10 in the Expeditions sequence).
    b. Within the Expedition, ask the teacher the number of the word study lesson she is teaching that day.

3. If you answer partially or no to any item in your observation forms, please use the back of the page to explain your answer.

4. In the classroom behavior/management table, if the behavior is not applicable during part of the observation time, write n/a and deduct that period from the score (e.g. instead of dividing by 5, divide by 4 or what is applicable).

5. Use back of the observation page to enter comments. You don't need to comment in every aspect of your observation but make short comments about behaviors or events that catch your attention and can be relevant to a better understanding of why the lesson occurred the way it did.

Note

State A has two days for lesson 5 and two days for lesson 10. Each expedition takes a total of 12 days

OVERVIEW

School: ____________________ Grade ______ Teacher Name: ___________________

Observer: __________________ Observation Date: __________ Observation Time: _____________

Lesson Number: _______ Expedition Number: ______ No. of students: ________

Was the entire lesson completed in the class period? Yes_____ No _____

In addition to the teacher, is there another adult in the room? Yes _____ No _____

Who? (circle) Special education teacher Special education aide Voyager coach

School administrator School district staff State Project Director Other______________

A. Classroom Environment (complete at beginning of lesson)	**Yes (2)**	**Partially (1)**	**No (0)**
1. Teachers have sufficient space to conduct individual and/or group work			
2. Instructional areas are clearly identified (i.e. whole group, independent small group, word study)			
3. Teacher resources for the daily lesson are readily available			
4. All students have readily available materials, as needed			

B. Lesson Planning and Delivery – overview (complete at the end of lesson)	**Yes (2)**	**Partially (1)**	**No (0)**	**Not observed (n/a)**
5. Teacher closely follows the curriculum guide during instruction				
6. Pace is brisk and business-like, yet personal				
7. Skills are modeled correctly				
8. The steps of the correction procedures are followed as needed				
9. Teacher puts students into groups as indicated by the lesson				
10. Teacher uses built-in differentiated instruction strategies as needed : □ re-teach lesson □ word study lesson □ English Language Learner strategies □ challenge questions □ Paired reading				

Lesson 1 Form

C. Lesson Planning and Delivery-lesson specific **(**Check the box next to each activity you observe)	**Components delivered in order? (See scale)**	**Steps delivered in order? (See scale)**	**Components delivered within allotted time? (Y/N)**
WHOLE GROUP Introduce the Expedition (10-15 min.) □ Discuss probing questions Start time_____ End time______			
Before Reading (15 min.) □ Introduce vocabulary Start time_____ End time______ □ Introduce the target skill Start time_____ End time______ □ Introduce the passage Start time_____ End time______			
During Reading (10-15 min.) □ Students read text Start time_____ End time______			
After Reading (5-10 min.) □ Check comprehension Start time_____ End time______			
INDEPENDENT □ Students' practice vocabulary using the online technology component Start time_____ End time______ □ Students' select books for independent reading Start time_____ End time______			

C. Lesson Planning/Delivery Scale 0 = no, 1 = modifications not allowed, 2 = modifications allowed, 3 = yes

D. Classroom Behavior/Management	**Minutes**					**Total (Proportion of time – see scale)**
	10	**10**	**10**	**10**	**10**	
1. Half or more of the students are paying attention to teacher or following teacher instructions						
2. Half or more of the students are responding to teacher questions or prompts						
3. Half or more of the students are **actively** participating in the activities assigned by the teacher (group or individually)						
4. Teacher addresses student behavior promptly to minimize disruption in the classroom						
5. Teacher makes an effort to involve students who appear disengaged						
6. Students follow expectations for working in groups						

D. Classroom Behavior/Management Scale: 0 Not At All 1 Occasionally 2 Frequently

Appendix B

Table B1. Fidelity Scores for the PRJ Classroom Instruction Component

	Teacher														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Section A	0.20	0.17	0.17	0.17	0.20	0.17	0.20	0.20	0.15	0.2	0.17	0.2	0.2	0.2	0.13
Section B	0.29	0.27	0.27	0.28	0.27	0.28	0.24	0.28	0.28	0.29	0.19	0.26	0.3	0.28	0.11
Section C	0.25	0.22	0.27	0.24	0.29	0.22	0.19	0.25	0.23	0.12	0.26	0.15	0.3	0.23	0.12
Section D	0.15	0.16	0.08	0.15	0.20	0.20	0.18	0.17	0.18	0.18	0.16	0.19	0.18	0.19	0.17
PRJ Fidelity Score	**0.89**	**0.82**	**0.79**	**0.83**	**0.95**	**0.87**	**0.82**	**0.90**	**0.85**	**0.79**	**0.78**	**0.80**	**0.98**	**0.90**	**0.53**

Table B2. Fidelity Scores for the Teacher PD Component

	Teacher														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
PD Fidelity Score	1.10	1.10	1.00	0.50	1.10	1.00	1.00	1.20	1.00	1.24	1.47	1.42	1.16	1.00	1.58