CHAPTER 7

# An Eye to Efficient and Effective Fidelity Measurement for Both Research and Practice

*Chrishana M. Lloyd, Lauren H. Supplee, and Shira Kolnik Mattera*

Fidelity is one of many components (including the contrast from usual practices, the context in which the program is being delivered, and who the intervention serves) that may be necessary for achieving program impacts. It is a multidimensional construct that includes adherence to, dose of, quality of, and participant engagement in an evidence-based practice (Dusenbury, Brannigan, Falco, & Hansen, 2003). The evaluation of fidelity captures the gap between the intended evidence-based practice and the actual implementation of the practice. It is specifically the gap between intended and actual practice that may prevent the promise of evidence-based programs from coming to fruition.

To date, measurement of fidelity has been limited, focused primarily on intervention adherence and occasionally on dose. Increasingly, however, research suggests that understanding and measuring across the *spectrum* of fidelity components is critical to program implementation (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005). Capturing implementation processes via quality fidelity measurement and understanding how to use what has been learned to improve practice has the potential to strengthen both research and practice and can be useful for policy makers and funders who are increasingly making decisions about where to invest limited fiscal resources based on program effectiveness. One of the many challenges that evidence-based policy brings, though, is how to ensure that programs that have shown efficacy on a small scale within tightly controlled conditions will continue to produce effects when implemented more widely. Implementation science, located at the nexus between the science of evidence-based programs and the applied settings and populations intended to use and benefit from them, becomes a useful mechanism for understanding this and other intervention implementation issues related to fidelity.

As interventions go to scale, both researchers and practitioners need tools to help assess fidelity within research contexts and in applied settings. In fact, Schoenwald and colleagues (Schoenwald, 2011; Schoenwald et al., 2010) have made strong arguments that researchers and practitioners need to develop useful and practical tools that support intervention monitoring and quality implementation as a potential means to obtaining positive impacts.

Measuring fidelity, therefore, becomes important to researchers for ensuring an intervention is tested fairly and that the impacts are real—internal validity—as well as allowing one to understand the potential for similar impacts should the intervention be replicated—external validity (Moncher & Prinz, 1991). Moreover, Type III errors, in which studies fail to find impacts for a program because the intervention was not implemented as planned, may be reduced, increasing the likelihood that conclusions about the actual efficacy of the intervention theory are accurate (Dusenbury et al., 2003). In sum, measuring fidelity in a research context is important for ensuring the evidence-based program is occurring as intended, documenting how implementation unfolds, and understanding differentiation between the program of interest and usual practice (Gearing et al., 2011; Moncher & Prinz, 1991; Schoenwald et al., 2010).

Understanding how to achieve and monitor fidelity is equally important for practitioners as interventions move from being implemented in research settings to real-world contexts. In a meta-analysis of impacts across a range of social services interventions, Wilson and Lipsey (2001) found that intervention implementation quality was strongly associated with achieved effect size of the programs. This raises the question: How do practitioners support high-quality implementation? Gearing and colleagues (2011) conclude that monitoring of implementation is a core component of fidelity that allows for adjustments to training and intervention delivery to improve performance. Others have found that monitoring and assessment of fidelity provides critical information to program staff, such as trainers, coaches, and supervisors, about implementation efforts on the ground and can inform training and coaching efforts (Elliott & Mihalic, 2004; Fagan & Mihalic, 2003; Fixsen et al., 2005). These findings suggest that practitioners need not only to actively monitor fidelity but also to understand the levels of fidelity that need to be achieved to affect change.

This chapter directly addresses the aforementioned research and practice considerations, using an implementation science framework to explore fidelity in the Head Start CARES (Classroom-Based Approaches and Resources for Emotion and Social skill promotion) demonstration project. The chapter provides an overview of the rationale and creation of standardized fidelity measures and tools in the Head Start CARES demonstration, offering key lessons learned about the process of developing fidelity measurement tools and reviewing how they were created to accomplish the goals and needs of researchers and practitioners within the context of scaling up and replicating previously existing interventions. It concludes by offering lessons for those interested in crafting fidelity measurement tools for both research and practice, making the final recommendation that fidelity measurement, though challenging, should move beyond intervention-specific, one-shot research trials, given the potential value and usefulness in both scientific and applied contexts.

## Overview of the Head Start CARES Demonstration Project

Head Start CARES is a large-scale national research demonstration designed to test the effects of three social-emotional program enhancements in Head Start settings over the course of one academic year. MDRC, a nonprofit, nonpartisan, education and social policy research organization, coordinated the demonstration and provided technical assistance for implementation. The demonstration tested the enhancements at the field test stage, meaning it examined the scaling up of programs across multiple and diverse settings that had previously shown promise at a smaller scale.

Lead and assistant Head Start teachers were trained and coached together throughout the year in one of the three social-emotional enhancements:

1. The Incredible Years Teacher Training
2. Preschool PATHS (Promoting Alternative Thinking Strategies)
3. Tools of the Mind

The teachers' centers were randomly assigned to one of the three enhancements, with each center implementing one of the enhancements, or to a control condition, which continued as usual. The three enhancements that were part of the demonstration were chosen because 1) they had shown evidence of effectiveness in improving children's social-emotional development in previous efficacy trials conducted with low-income preschool children, 2) they conceptually fit within the structure and operation of Head Start grantees and classrooms, and 3) they had implementation support available, including training and technical assistance, written materials, and fidelity monitoring.

The Center on the Social Emotional Foundations for Early Learning defines social-emotional development as the developing capacity of a child, from birth through 5 years of age, to form close and secure adult and peer relationships; to experience, regulate, and express emotions in socially and culturally appropriate ways; and to explore the environment and learn—all in the context of family, community, and culture. The enhancements' theories of change emphasized distinct strategies for supporting children's social-emotional development. The Incredible Years Teacher Training enhancement focused on professional development of teachers to promote positive social development and discourage problem behaviors; Preschool PATHS focused on providing children the language and skills to positively engage with emotions and peer interactions; and Tools of the Mind used dramatic play and specific teaching practices to scaffold children to develop self-regulation. The three enhancements overlapped conceptually because of their focus on social-emotional outcomes, but there were clear distinctions in their theories of change, primarily around the focus and activities utilized to build and strengthen children's social-emotional skills.

In addition, as a component of each enhancement, teachers and administrators received instruction and technical assistance from enhancement developers and trainers on the enhancements, and Head Start programs hired coaches (who attended enhancement training with their assigned teachers) from the local community to support the teachers in their implementation of them. Technical assistance was also provided to sites, enhancement developers, trainers, and coaches (see Table 7.1).

## Developing and Creating the Head Start CARES Fidelity Measures

In general, researchers who test evidence-based programs in trials create tools that are useful for the specific study, but they may be burdensome to implement

**Table 7.1.** Description of Head Start CARES key players

| Function | Role in Head Start CARES | Employment and supervision |
|---|---|---|
| Coach | · Attended training sessions with teachers<br>· Received ongoing content-related support from trainers and developers<br>· Observed and met with teaching teams weekly to discuss implementation | · Employed by the grantee liaison<br>· Supervised by the developer/ trainer *and* grantee liaison/ center director |
| Trainer | · Delivered training sessions to coaches and teachers on enhancement content<br>· Visited classrooms to support coaches and teachers with implementation<br>· Provided supervision and regular feedback on coach performance | · Employed and supervised by the enhancement developer |
| Grantee supervisor | · Recruited, hired, and supervised coaches<br>· Monitored implementation throughout the year | · Employed and supervised by the grantee |
| Teacher[a] | · Attended training sessions alongside the coach<br>· Received ongoing support from coaches and trainers throughout the year<br>· Had responsibility for classroom implementation | · Employed and supervised by the grantee |
| MDRC research and technical assistance team | · MDRC launched and researched the Head Start CARES demonstration. MDRC project staff also provided ongoing technical assistance to grantees, enhancement developers, trainers, and coaches throughout the year. | |

From Lloyd, C.M., & Modlin, E.L. (2012). *Coaching as a key component in teachers' professional development: Improving classroom practices in Head Start settings*. New York, NY: MDRC; adapted by permission.

[a]"Teacher" refers to both lead and assistant teachers.

outside of a research trial (Schoenwald et al., 2010). Given the increased interest in evidence-based practices and outcomes, the reality is that early childhood settings and other community-based agencies that receive external funding will likely have an increasing number of evidence-based programs in place over time. The challenge, then, is to create measurement tools that capture fidelity of implementation, that are multifaceted, that can be integrated into community-based programs, and that are ecologically valid (Schoenwald et al., 2010).

The creation of coherent and crosscutting fidelity measures in the Head Start CARES demonstration was influenced by the need to have data that were useful for the Head Start CARES research and technical assistance team and that could be integrated into the Head Start program and used by intervention staff. This meant that the project needed to document and receive data from *multiple reporters* across *different settings*.

Two primary groups of people reported on the fidelity of enhancement implementation: coaches and trainers. Coaches worked in classrooms weekly, observing teachers for an hour and meeting with them to debrief on their observations for 30 minutes. Trainers, who provided enhancement intervention support to coaches, visited the classrooms periodically as experts in the program, and they also trained teachers and coaches throughout the year on the enhancements. (See Office of Planning, Research & Evaluation [n.d.] for a more thorough review of the implementation measures used in the Head Start CARES demonstration.) In terms of settings, coaches and trainers worked in a range of sites and classrooms, including schools and community-based facilities.

The different enhancements, reporters, and settings necessitated development of measures that were common to the three enhancements but could still be used to identify unique features of each enhancement and setting that was thought to affect change. To accomplish this task, structured templates called "fidelity logs" were created by the research team in close partnership with the enhancement developers. These logs provided a space for coaches and trainers to report on enhancement implementation practices by highlighting two components of fidelity.

The logs focused on the practices of teachers and measured both adherence—how *much* or to what extent teachers implemented the program—and quality—how *well* teachers implemented it. For instance, logs reported on whether the teachers were using many of the identified enhancement's strategies and if "modifications or additions are consistent with" the enhancement's goals and objectives (Lloyd, 2009). Reports were then created out of the fidelity logs and used for collecting implementation data throughout the year and guiding the work of the MDRC research and technical assistance team (who worked with developers, trainers, and coaches) in improving the quality of implementation. In this way, fidelity was inherently tied to both research and ongoing practice.

The following section provides a detailed overview of the process of creating fidelity measures in the context of the Head Start CARES demonstration and reviews four key strategies and considerations:

1. The development of fidelity definitions, anchors, and measures that were internally valid to the individual program and externally valid across similar evidence-based programs
2. The creation and determination of meaningful thresholds for those instruments so that both researchers and practitioners knew whether implementation was going well or not
3. The collection of data from multiple sources to accurately document program implementation
4. The process of training on the measures to ensure that practitioners were able to use the data generated to monitor and respond to implementation challenges on the ground

## General Measurement, Specific Measurement, and Anchors

Defining fidelity for both research and practice involved creating *universal* and *specific* measurement tools from which to monitor fidelity within and across enhancements. The logs were devised to ensure that information could be shared across enhancements (universal fidelity), with one section clearly differentiating the core enhancement components (specific fidelity). The measurement and monitoring of fidelity work in the Head Start CARES demonstration allowed for nuanced documentation of implementation at scale that took into account both *general* aspects of fidelity—those that were important to the delivery of all evidence-based programs—and *specific* components of fidelity that were important to the individual model used. To ensure that both tools provided clear, defined measurement, clearly articulated *anchors* were created for each item in the tools. These anchors allowed developers and the research/technical assistance team to work together to define each item, with clear demarcations between weak, average, and strong implementation for each item.

***Universal Measurement*** Components of fidelity that were *universal*—the same across all of the models—were identified to facilitate comparisons of the quality of implementation across enhancements and to provide a standardized metric for evaluating fidelity. Fidelity in this case was operationalized as implementing the program to such a degree that it was clear that teachers, children, and classrooms were steeped in their respective social-emotional enhancements. For example, one item asked whether observers could easily tell when they "enter this classroom and look around" that the identified enhancement was being used (Lloyd, 2009). Another asked whether "the children are actively engaged" in the enhancement all day long, rather than enhancement implementation being "just seen as a special event" (Lloyd, 2009).

These measures assumed a basic level of skill from the teacher that allowed for a new initiative to be implemented. For example, teachers needed to be able to understand the theory underpinning whichever enhancement they were implementing in order to be able to use and generalize it throughout the day. For the research/technical assistance team, understanding the prerequisite knowledge and the skills necessary for successful intervention implementation by teachers was critical for determining how to craft fidelity measures that captured intervention implementation practices accurately.

***A Head Start CARES Demonstration Example*** In the Head Start CARES log, the universal fidelity sections, "Modeling and Generalization" and "Fidelity of Teaching and Supporting Children," were each made up of five scaled items. Along with the items presented earlier about teacher practice, the items also addressed issues of child receipt of the enhancement, asking whether the children were responsive to the particular enhancement's strategies and if the strategies were effective in the teacher's classroom. The 10 questions asked in the universal fidelity section for a Tools of the Mind enhancement are listed next (Lloyd, 2009).

***Enhancement-Specific Measurement*** In order to support and improve the delivery of the individual enhancements, components of fidelity that were *specific* to each were also identified and measured. For example, the Preschool PATHS enhancement used lessons that had to be delivered weekly, while the Tools of the Mind enhancement asked teachers to aid children in completing plans for their playtime daily. These items were unique to the identified enhancement: Preschool PATHS was the only enhancement that used weekly large-group lessons to deliver enhancement content, and the Tools of the Mind enhancement was the only one to have children create play plans.

**Modeling and Generalization of Tools of the Mind**

1. It is clear when you enter this classroom and look around it is a Tools of the Mind classroom.

2. The teachers have taken extra steps to extend the Tools of the Mind concepts into other parts of the Head Start program by designing special activities or adapting standard activities to be consistent with Tools of the Mind themes.
3. The children are actively engaged in Tools of the Mind throughout the day. It is not just seen as a special event.
4. The teachers use Tools of the Mind as part of their strategies for managing conflicts, as part of classroom procedures, and to help build positive relationships between the children.
5. The teachers model and actively promote Tools of the Mind and praise the children when they use Tools of the Mind techniques.

**Fidelity of Teaching and Supporting Children in Tools of the Mind**

1. The teachers are prepared for Tools of the Mind activities and seem familiar with what to do.
2. The teachers use many of the Tools of the Mind strategies, and modifications or additions are consistent with Tools of the Mind goals and objectives.
3. Material is presented in an engaging manner. The teachers are positive, energetic, and enthusiastic about Tools of the Mind. There is flexibility in the presentation and the teachers appear comfortable with Tools of the Mind.
4. The teachers are patient and sensitive to the skill level of the children and adapt their style of presentation and pacing to match the children.
5. The children have fun during, and enjoy doing, Tools of the Mind activities. They are attentive and engaged during Tools of the Mind activities.

This more specific definition of fidelity provided a means to identify the presence of elements unique to and between the enhancements. That is, teachers not only had to generally implement the enhancements well, they also needed to adhere to a core set of program-unique skills and/or lessons that developers identified as being critical to their enhancements' effectiveness.

***A Head Start CARES Demonstration Example*** In the Head Start CARES logs, a third section, "Fidelity of Programmatic Activities," asked coaches and trainers to rate the lead and assistant teachers on *how well* each was implementing the activities, strategies, or other programmatic activities for each enhancement. These differed by number and type for each enhancement, with Incredible Years having 13 items, Preschool PATHS 10, and Tools of the Mind 6. In this section, as well, coaches and trainers were asked to rate teachers on a five-point scale from 1 (*strongly disagree*) to 5 (*strongly agree*).

The specific fidelity sections were much more individualized, so items between enhancements did not match. For instance, the first item of the specific fidelity section asked about building relationships with students for Incredible Years, PATHS lessons for Preschool PATHS, and play planning and scaffolded writing for Tools of the Mind. However, the ratings for this section still followed a graduated pattern across enhancements. A rating of 1 meant that there was no evidence the teacher was using the strategy, or that its use was flawed. A rating of 5 meant that the teacher was using the strategy frequently, consistently, or in an exemplary fashion.

***Creation of Anchors*** Along with the scaled items in each section, concrete examples of each rating, called "anchors," were created to help understand if the enhancements were being implemented as intended. The research and technical assistance team worked with developers to create detailed anchors for each item so that coaches and trainers could be trained to observe similar practices and rate them in a reliable and valid way.

Two primary tenets guided the scaling and anchoring work. First, the team wanted anchors that would accurately and meaningfully reflect the core components of the enhancements. This would allow for distinguishing between enhancements, but would also help the team to understand implementation practices, including faithful and nonfaithful implementation of the intervention. The anchors explicitly defined each rating so that the ratings could have the same meaning, regardless of who the reporter was or for which enhancement.

Second, the team wanted scales that would be sensitive to changes throughout the year, including variation in implementation and settings. Each item had a set of five anchors that accompanied its five-point scale. These anchors were created to match the conceptualization of fidelity outlined earlier. They were a mix of adherence (how much) and quality (how well) of implementation of the enhancement. The anchors were graduated, with a rating of 1 meaning that something happened never or with poor quality, while a rating of 5 meant that something happened very often or with exceptional quality. As teachers became more familiar with their assigned enhancement, the expectation was that change would occur in a positive direction on the rating of their implementation of the enhancement.

The anchors were a key component for allowing generalization across enhancements in the general fidelity of the logs. In some general fidelity items, the anchors looked the same across all three enhancements. In others, they differed in order to allow necessary enhancement-specific details to emerge. For example, in the general fidelity item "The teachers are prepared for [their identified enhancement] activities and seem familiar with what to do," anchors were the same for all enhancements (Lloyd, 2009):

- If the teachers are never prepared or familiar with [their enhancement] activities, *select "1 = strongly disagree."*
- If the teachers only rarely are prepared or familiar with [their enhancement] activities, *select "2."*
- If the teachers occasionally are prepared or familiar with [their enhancement] activities, but not consistently, *select "3."*
- If the teachers are usually prepared or familiar with [their enhancement] activities, *select "4."*
- If the teachers are exceptional in their preparation or familiarity with [their enhancement] activities, *select "5 = strongly agree."*

Therefore, although the details were different, the process of creating items and anchoring them to the definition of adherence and quality were the same.

In another general fidelity item, asking whether the teachers use their identified enhancement "as part of their strategies for managing conflicts, as part of classroom procedures, and to help build positive relationships between the children," the details of the anchors varied by enhancement. The anchors still followed a five-point scale, with "1" meaning that something almost never happened, "3" that it occasionally happened, and "5" that it happened an "exceptional" amount. But even though the anchors were the same, the verbal descriptions of the anchors differed by enhancement. The meaning of the item stayed the same, but the anchors

allowed for individual definitions of what quality implementation of that enhancement component looked like. For instance, the anchors for a rating of 1 for each enhancement were as follows (Lloyd, 2009):

- [For Incredible Years] If teachers never use IY strategies for managing conflicts, as part of classroom procedures, or to help build positive relationships between the children, *select "1 = strongly disagree."*
- [For Preschool PATHS] If teachers never use PATHS routines…, materials (feeling faces or posters), or strategies (reflecting feelings, cuing turtle, supporting problem solving) for managing conflicts, as part of classroom procedures, or to help build positive relationships between the children, *select "1 = strongly disagree."*
- [For Tools of the Mind] If teachers never use Play Planning to work out disputes before they escalate, use mediators to help children take turns or focus on specific attributes, or use attention focusing activities to promote self-regulation during large groups or transitions, *select "1 = strongly disagree."*

## Thresholds

As part of the process of creating the logs, thresholds for the ratings were developed, agreed upon, and clearly defined for each section of the logs in an effort to support the standards for acceptable levels of practice. The thresholds were integral to the conceptualization of fidelity in the Head Start CARES demonstration and were designed to take into account the complexity of real-world implementation and adaptation. For each of the log sections, the research and technical assistance team provided benchmarks for the ratings. The ratings became a point of reference against which the research and technical assistance team, developers, trainers, and coaches could monitor and assess the implementation of the enhancement and respond as necessary. This practice is different from the usual process of documentation of implementation because it set a standard level of implementation that all intervention implementers had to meet.

Coaches were the primary and most frequent raters of implementation quality within the Head Start CARES demonstration, so trainers, enhancement developers, and the research and technical assistance team worked closely with them to ensure they understood the principles of the intervention and the importance of providing reliable and valid data in advance of program implementation. Although the coaches would be working closely with teachers, it was clearly explained that documentation of fidelity was not about judging teachers or making them look good or bad, but was instead a necessary means of recording clearly and accurately what occurred during the year for both research and programmatic purposes.

In the Head Start CARES demonstration, items were considered implemented at an acceptable threshold and with fidelity if a score of 3 or above was given on a scale of 1 to 5. Throughout the year, the data were monitored on an ongoing basis, and scores below 3 were flagged. Definitions of what a "3" meant differed by question; in general, it signaled that an act had occurred, but that it had occurred only inconsistently or with moderate quality. The MDRC research and technical assistance team worked with enhancement developers, trainers, and sites not only to identify teachers or sites with below-threshold scores but also to brainstorm and implement steps to remedy the implementation challenges. The process of comparing fidelity of implementation practices based on what was happening in the classrooms occurred throughout the intervention year by researchers and practitioners alike.

### Multiple Reporters and Intervention Documentation

Despite the various sources of data and the already complex conceptualization of fidelity, it was deemed important that multiple viewpoints of teacher fidelity be collected. As mentioned previously, the use of both coaches and trainers as reporters of teachers' performance in delivering the enhancements resulted in the measurement of fidelity from multiple perspectives across all three enhancements. Trainers were considered expert in their enhancement, but had fewer day-to-day interactions with teachers and classrooms, seeing them only during training and technical assistance visits. In comparison, coaches were not experts in the enhancements, but were in the classrooms weekly, working with teachers.

Coaches rated teachers' implementation of the enhancement through weekly logs documenting their classroom observations and contact with teachers (not discussed in this chapter) and monthly logs documenting fidelity of enhancement implementation in the classroom. Trainers' documentation of teachers' implementation of the enhancement occurred via a classroom visit log that assessed teachers' fidelity in the classrooms visited. Importantly, the coach monthly logs and trainer visit logs mirrored each other so that coaches and trainers reported on the same items. Each log consisted of approximately 15–25 scaled items that used a 1 (*strongly disagree*) to 5 (*strongly agree*) Likert scale to assess the critical facets of the implementation process and the core components of the enhancements. The perspectives of the two different reporters were thought to provide a more accurate assessment of teachers' implementation processes and how they compared to the enhancement models.

While not systematically assessed, the team also hypothesized that coaches and trainers would reach general consensus about ratings throughout the year, particularly as coaches became more knowledgeable about the enhancements. A preliminary review of the ratings indicated that trainers and coaches generally did agree on the extent to which a teacher was implementing an enhancement with fidelity. Trainers typically rated teachers a point lower on the five-point Likert scale than coaches did, but ratings between them followed the same pattern and were consistently in the same direction. For example, in one month, a coach might have rated a teacher a 3 on a particular question, while the trainer rated her a 2; in the following month, their ratings of that teacher might have been a 4 and a 3, respectively.

In short, the fidelity logs were designed to provide observational data—the gold standard for measurement of fidelity (Hamre et al., 2010)—and served as a comparative framework that afforded an objective, systematic, and structured way to provide feedback about program implementation, while facilitating understanding of implementation processes, trends, and outliers. Throughout the year, the research and technical assistance team and enhancement developers graphed and reviewed the ratings over time using the thresholds discussed above to evaluate and determine acceptable levels of fidelity of the enhancements.

### Training

The experiences in Head Start CARES suggests that ratings derived from multiple raters—in this case, trainers and coaches—are useful for both researchers and practitioners in understanding fidelity. However, to ensure the reliability and usefulness of the data, it was important to train the various reporters on how to document and rate their observations.

The validity of the ratings depended in large part on the individuals completing them, their knowledge of the enhancement, and their reliability in reporting what they saw. To facilitate reliable ratings, coaches were trained to rate the core specific intervention components by developers and trainers in advance of program implementation. They also received training on the universal fidelity ratings by the research and technical assistance team.

The enhancement developers and trainers were encouraged to use the anchors as a tool for dialoguing with coaches throughout the year. Grounding the conversations with the anchors helped to ensure that coaches fully understood the principles of the interventions and provided a common framework and language for trainers and enhancement developers to discuss the ratings, ensuring that enhancement rater "drift" was less likely to occur.

## Fidelity Documentation and Measurement Challenges

While there is a clear need in the field for a method to assess fidelity across different programs by different reporters, the rating scales used in the Head Start CARES demonstration raised some challenging but not insurmountable measurement issues that may inform future measurement efforts. In general, it was a long and complex process to create measures that tapped into universal enhancement implementation components while allowing for adaptations. Developers and researchers needed to clearly articulate their definitions of fidelity. This meant spending a significant amount of time negotiating each partner's perspective and incorporating it into a definition of fidelity that was agreed upon by all key stakeholders in advance of implementation.

Once fidelity was defined, individual items were needed both to encompass this general operationalization of fidelity and to create individualized anchors for each enhancement. For instance, although the item "It is clear when you enter this classroom and look around it is a [specific enhancement] classroom" (Lloyd, 2009) can be applied to any program, the anchors that specified what a PATHS classroom looked like were different from those that identified a Tools of the Mind classroom. Additionally, the five-point Likert scale anchors needed to denote the same incremental change between a rating of 1 and a rating of 2 for the scales to be comparable across enhancements. The ability to compare across enhancements is incredibly powerful, but requires considerable time and thoughtfulness about the measurement tool during development. In addition, creating the measures was, by necessity, an iterative process that used information during initial data collection to improve the measure over time.

As the tools were used throughout the intervention year, challenges emerged. The conceptualization of fidelity for the Head Start CARES demonstration included a mix of both adherence and quality. This occurred in numerous ways: Observers may have had different numbers of opportunities to observe an item happening (regardless of quality), and teachers may have actually used a strategy varying number of times. For instance, in some logs, the ratings were created in such a way that a score of 1 could be interpreted as indicating low or poor fidelity or that the reporter did not have a chance to observe that aspect of fidelity. In other logs, the ratings were defined such that a "1" could mean that a teacher used the strategy poorly or did not use the strategy at all.

The ability to separate adherence (i.e., did this component of implementation happen at all and to what extent?) from quality (i.e., when this component happened, was it done well?) taps into a larger question within the field of implementation research. Currently, the implementation measurement field is beginning to call for measurement that distinguishes between the multiple components of fidelity. The measurement process in the Head Start CARES demonstration took a first step toward teasing apart these components; however, further work is needed to clearly differentiate among these aspects of fidelity in a reliable and valid manner.

## Considerations for the Early Childhood Field

Intervention research in the early childhood field is typically aimed at improving outcomes for children and families. By selecting and implementing programs that demonstrate evidence of effectiveness, the research and practice communities may better achieve their goal. However, implementation of evidence-based programs is complex. Both researchers and practitioners need tools to help support the monitoring of fidelity to support and strengthen program quality. The following section outlines key lessons learned from our efforts to document fidelity work in the Head Start CARES demonstration, providing suggestions for future research and practice efforts.

### Lessons Learned for Research

The large-scale measurement and monitoring of fidelity in the Head Start CARES demonstration suggests some lessons that researchers can apply in future evidence-based intervention research scale-up efforts.

- *Development of fidelity measures requires collaboration and is a continuous process.* The Head Start CARES research and technical assistance team and enhancement developers agreed on appropriate measures of fidelity across interventions. Moreover, they planned for continued dialogue about the measures to ensure that they were capturing fidelity accurately. The creation of the measures and logs was an iterative process between the team and the developers, guided primarily by the experiences of the raters. The process included working with enhancement developers and practitioners to assess the face validity of both the universal and specific components of fidelity for the individual enhancement models. Over the course of the demonstration, the research and technical assistance team refined the instruments by changing the language to make the logs more reflective of actual implementation experiences, and enhancement developers continued to update and streamline their anchors for various items to make them more explicit and clear. By the end of the demonstration, the team felt confident the data collected through the logs would allow for analysis on fidelity across enhancements.

- *When measuring fidelity across multiple models, clarity around similar and different components of the models is necessary.* Being clear about which intervention components are similar and which are different for multiple interventions is critical for fidelity data to be useful. It is nearly impossible to compare fidelity *across models* within or across research trials unless similar items and measures are used from the essential components of fidelity identified for each model. It is recommended that researchers identify a comparable set of general fidelity components across all of the enhancements they are using to create equivalence across them for comparison. This requires understanding of the general theory of change for the evidence-based practices being instituted. For instance, in the Head Start CARES demonstration, one program required teachers to present lessons, another asked teachers to do specific play-planning activities with children, and the third had teachers use a specific strategy to respond to a behavior. These components were unique to each enhancement, but all focused on supporting the social and emotional development of children—a common core component.
- *Fidelity measurement captures multiple components of implementation, and capturing these distinctions is important.* Implementation researchers should diligently attempt to distinguish between the multiple components of fidelity to better understand research outcomes. For example, distinguishing adherence from quality in separate rating scales may help to present a clearer picture of intervention implementation by providing insight into important distinctions between varying levels of adherence and quality implementation.
- *Developing comparable anchors across models to create equivalence across reporters, so that all can report on the same item with a high level of reliability, is suggested.* Ensuring rating anchors are meaningful to each enhancement allows the raters to report ratings and eventually thresholds with validity across enhancements. Viewpoints may differ, and the ability to combine data from multiple reporters across enhancements and know that their scores have roughly the same meaning is powerful. Moreover, using general fidelity tools with well defined anchors provides a way to compare the level of teacher performance for delivering enhancements across reporters.
- *Psychometrically testing fidelity measures may provide a common framework for researchers and practitioners to think about fidelity in evidence-based programs.* Some of this work has already been started. For example, in the Preschool Curriculum Evaluation Research Initiative (Preschool Curriculum Evaluation Research Consortium, 2008), 14 different evidence-based programs were tested. Implementation in these studies was assessed through both a program-specific measure and a global measure. However, clearly laying out the process through which these types of measures were created and used will help inform the field by integrating knowledge. For example, in the Head Start REDI (Research-based, Developmentally Informed) trial, trainers

assessed implementation across a number of different evidence-based *practices,* including Preschool PATHS, dialogic reading, and sound games (Bierman et al., 2008).

### Lessons Learned for Practice

The tools discussed in this chapter not only are beneficial for research but also offer some distinct contributions to the practice field broadly. First, the fidelity tools were developed to allow for easy training of coaches and trainers to complete and report their observations in a consistent manner. Given the limited time and fiscal resources in many community-based programs, this feature is very appealing. Second, the tools allow for observations of practice, minimizing disruptions in and burdens on the classroom. Third, the tools discussed are crosscutting. Within the early childhood field, programs are increasingly implementing a mix of curricula, enhancements, and practices within one setting. Having the framework for a tool that minimizes data collection challenges and that can cross multiple evidence-based programs is important.

- *Well-designed and flexible fidelity tools can help guide decisions about implementation in practice by providing a gauge of thresholds of quality and adherence, as well as a platform for discussion and feedback with the teacher or practitioner.* In many implementation studies, documentation about the intervention generally ends after the initial training and data collection are completed; however, practitioners may want to sustain interventions after the training is over. Alternatively, more community-based settings are being directed to choose evidence-based practices or programs, independent of a study. Flexible fidelity tools may help support the scale-up and sustainability of an enhancement within and across early childhood settings by providing practitioners with concrete data to guide training and technical assistance efforts toward attaining and maintaining fidelity to the model. Supervisors, coaches, or other professional development staff can use the fidelity tool ratings to benchmark the minimum levels of quality and adherence necessary to achieve the desired outcomes.
- *Training practitioners on how to assess and report fidelity is feasible even if they do not have specialized backgrounds in implementation research.* In addition to the instruments themselves, the project created training processes and tools that guided the use of the fidelity instruments. Neither coaches nor trainers in the Head Start CARES demonstration had specialized training in data collection, reporting, or implementation science. The demonstration showed that it is possible to quickly and, more importantly, effectively train a significant number of program staff to successfully use fidelity monitoring instruments.
- *Fidelity instruments can help practitioners make decisions about the feasibility of potential adaptations.* When evidence-based programs are disseminated, the practitioners are often interested in adapting programs for their own contexts to better serve the needs of their populations or communities. Though adaptation should be done in conjunction with the intervention developer, well designed fidelity tools make explicit the core components of a program. Having this information easily available may assist in navigating the chasm between *intentional adaptation,* which might improve the likelihood of intervention success, and *unintentional program drift,* which has the potential to detract from the anticipated outcomes.

- *To best monitor implementation, programs interested in implementing a number of evidence-based interventions or practices may find it important to isolate components unique to each intervention versus components necessary to attain fidelity across all evidence-based interventions.* Measures that focus on specific strategies are very useful to ensure unique components of the intervention are being implemented. However, if measures focus only on isolated strategies, it becomes difficult to tell if the implementer is generally doing a good job of using strategies across interventions.

## Future Directions

The Head Start CARES demonstration provided an opportunity to create a process for measuring social-emotional practices across multiple interventions, raters, and sites in the early childhood field. While the work in the demonstration was pioneering, there is still much that needs to be understood. The field must begin to pay attention to fidelity, including adherence, quality, dosage, and participant engagement. However, we should also not lose sight of the fact that other implementation processes are also integral to strong implementation, such as how implementation changes and unfolds throughout the year. The implementation science field needs to develop high-quality applied measurement tools to understand fidelity at the classroom level; the field should also examine 1) how factors such as fidelity are influenced by organizational culture and climate, 2) the relationship between the coach and teacher, and (3) the quality of professional development to fidelity in vivo.

Within the implementation process, it is clear that there are different stages of scientific inquiry into the evidence of effectiveness of an intervention, including piloting a new intervention, implementing it in a more controlled or clinical research setting, and moving to larger scale implementation. In each of these stages, it is essential that fidelity be measured thoughtfully, carefully, and clearly. However, the focus of that measurement might shift. In a pilot phase, fidelity measurement may be qualitative and more process oriented in order to inform future implementation and measurement efforts. In the research or efficacy phase, fidelity measurement may be more focused on deconstructing and understanding the various components of implementation (dosage, adherence, quality, general vs. specific) in a reliable and valid way. And when moving to a scale-up and replication phase, previously validated measures may need to be better contextualized with an eye toward appropriateness for the population using them, as well as how they mesh with other programs that are in place or might someday be added.

The fields of early childhood implementation research and practice would also greatly benefit from overarching measures that capture and operationalize fidelity to general early childhood recommended practices. For instance, what are the recommended practices regarding coaching or training that we can generalize across models? What models of coaching and training may be related to high fidelity? Can we create a definitive set of developmentally appropriate qualities or attributes that are necessary to achieve fidelity to any early childhood or professional development practice? To begin to develop a list of these attributes, measures are needed that clearly define and quantify dosage, adherence, and quality for each of the core components of effective implementation such as training and coaching. More research that begins to answer questions about how variability in these core components of implementation explains impacts of the intervention is also needed. By expanding research about *how* and *at what levels* various components of implementation influence impacts, the field can begin to create meaningful thresholds against which to determine fidelity.

In sum, early childhood implementation science is quickly growing, providing more clarification with each study about what best supports positive outcomes for children and families. This chapter has provided a glimpse into the detailed and nuanced processes initiated to move beyond a one-dimensional understanding of fidelity and fidelity measurement in Head Start CARES, a large-scale demonstration project. The process of documenting fidelity across multiple models, in multiple settings, and with multiple raters was challenging. The benefits, however, extend to and have relevance for both research and practice, making it clear that thoughtful, in-depth fidelity measurement is a worthy undertaking.

## References

Bierman, K.L., Domitrovich, C.E., Nix, R.L., Gest, S.D., Welsh, J.A., Greenberg, M.T., … Gill, S. (2008). Promoting academic and social-emotional school readiness: The Head Start REDI program. *Child Development, 79,* 1802–1817.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W.B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research, 18*(2), 237–256.

Elliott, D.S., & Mihalic, S. (2004). Issues in disseminating and replicating effective prevention programs. *Prevention Science, 5,* 47–53.

Fagan, A., & Mihalic, S. (2003). Strategies for enhancing the adoption of school-based prevention programs: Lessons learned from the Blueprints for Violence Prevention replications of the Life Skills Training Program. *Journal of Community Psychology, 31,* 235–253.

Fixsen, D.L., Naoom, S.F., Blase, K.A., Friedman, R.M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature* (FMHI Publication No. 231). Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network.

Gearing, R.E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gilles, J., & Ngeow E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review, 31,* 79–88.

Hamre, B.K., Justice, L.M., Pianta, R.C., Kilday, C., Sweeney, B., Downer, J.T., & Leach, A. (2010). Implementation fidelity of MyTeachingPartner literacy and language activities: Association with preschoolers' language and literacy growth. *Early Childhood Research Quarterly, 25,* 329–347.

Lloyd, C. (2009). *Head Start CARES Monthly Fidelity Form.* Washington, DC: Office of Planning, Research and Evaluation.

Lloyd, C.M., & Modlin, E.L. (2012). *Coaching as a key component in teachers' professional development: Improving classroom practices in Head Start settings.* New York, NY: MDRC.

Moncher, F.J., & Prinz, R.J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review, 11,* 247–266.

Office of Planning, Research & Evaluation. (n.d.). *Head Start CARES (Head Start Classroom-based Approaches and Resources for Emotion and Social skill promotion), 2007–2013.* Retrieved from http://www.acf.hhs.gov/programs/opre/hs/cares/index.html

Preschool Curriculum Evaluation Research Consortium. (2008). *Effects of preschool curriculum programs on school readiness (NCER 2008–2009).* National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Schoenwald, S.K. (2011). It's a bird, it's a plane…it's fidelity measurement in the real world. *Clinical Psychology, 18,* 142–147.

Schoenwald, S.K., Garland, A.F., Chapman, J.E., Frazier, S.L., Sheidow, A.J., & Southam-Gerow, M.A. (2010). Toward effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research, 38,* 32–43.

Wilson, D.B., & Lipsey, M.W. (2001). The role of method in treatment effectiveness research: Evidence from a meta-analysis. *Psychological Methods, 6*(4), 413–429.