Early Childhood Research Quarterly 25 (2010) 284–298

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

# Implementation quality: Lessons learned in the context of the Head Start REDI trial

Celene E. Domitrovich*, Scott D. Gest, Damon Jones, Sukhdeep Gill, Rebecca M. Sanford DeRousie

*The Pennsylvania State University, Prevention Research Center, 109 South Henderson Building, University Park, PA, 16801, United States*

ARTICLE INFO

*Keywords:*
Preschool
Intervention
School readiness
Implementation

ABSTRACT

This study uses data collected in the intervention classrooms ($N = 22$) of Head Start REDI (Research-based, Developmentally Informed), a randomized clinical trial testing the efficacy of a comprehensive preschool curriculum targeting children's social-emotional competence, language, and emergent literacy skills delivered by teachers who received weekly coaching support. Multiple dimensions of implementation (Dosage, Fidelity, Generalization, and Child Engagement) were assessed across curriculum components. Results indicated that REDI Trainers perceived significant growth in teacher implementation quality over time but that patterns differed by implementation dimension. Dosage and Fidelity of all intervention components were high at the beginning of the year and remained so over time while Generalization was low at baseline but increased significantly across the year. Variation in implementation was associated with variation on several child outcome measures in the social-emotional domain but not in the language and literacy domains.



## 1. Implementation quality: lessons learned in the context of the head start REDI trial

Research on the prevention of problem behaviors and promotion of positive adjustment in children and youth has risen dramatically during the past several decades. A growing number of sophisticated models have emerged from longitudinal research to describe how risk and protective factors contribute to the development of these problems over time. Preventive interventions based on these models have been tested empirically and shown to be effective in improving outcomes in targeted domains (Catalano et al., 2002; Greenberg, Domitrovich, & Bumbarger, 2001; Hahn et al., 2007). The dissemination of these findings has caused a surge in the number of communities adopting evidence-based preventive interventions in order to maximize the likelihood of achieving similar outcomes.

Research linking implementation quality with outcomes suggests that positive results are more likely to be replicated when the quality of implementation is high (Durlak & Dupree, 2008). Unfortunately, in community settings, programs often are not implemented in the same way or with the same quality as when they are first evaluated (Dusenbury, Brannigan, Falco, & Hansen, 2003; Gottfredson & Gottfredson, 2002). There are a variety of factors that affect implementation including the characteristics of the intervention and the nature of the support system (Domitrovich et al., 2008). Ideally, communities using evidence-based interventions will monitor implementation and use this information to adopt strategies that maximize quality. While this is essential for successful diffusion, little research exists to guide this process. Additional research is needed to help communities know "when" information on implementation should be collected to be

* Corresponding author.
*E-mail address:* cxd130@psu.edu (C.E. Domitrovich).

0885-2006/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ecresq.2010.04.001

most useful, "what" dimensions of implementation are most important to monitor, and "who" can provide this information.

Although the evidence base for prevention programs is growing rapidly, the science regarding program implementation under real-world conditions is still under-developed (Greenberg, Domitrovich, Graczyk, & Zins, 2001). A significant portion of implementation research comes from studies of interventions in the mental health, education, and substance abuse prevention domains (Durlak & Dupree, 2008; Dusenbury et al., 2003; Fixen et al., 2005; O'Donnell, 2008). Even though there are a number of evidence-based interventions that target young children prior to school entry, implementation research on these is especially limited. Given the heightened emphasis on accountability and moving science to practice at younger developmental levels, the number of intervention trials with this segment of the population has grown. This provides an opportunity to also extend implementation research. There are enough similarities between schools and preschool settings that the educational literature should inform these early childhood studies. Indeed, the primary purpose of the present study is to explore the implementation quality achieved in the context of a randomized trial of the REDI (REsearch-based Developmentally Informed) intervention, a comprehensive preschool curriculum conducted in Head Start classrooms and delivered by teachers who received on-going implementation support from experienced trainers.

### *1.1. Defining implementation*

The level of implementation is the degree to which an intervention is followed as prescribed by the developer or as conducted under rigorous conditions such as in an efficacy trial (Yeaton & Sechrest, 1981). Given this definition, the measurement of implementation requires the specification of an intervention model and measurement of actual effort against the ideal, in order to determine the degree to which the intervention is successfully replicated in the field. Traditionally, the measurement of implementation focuses on two dimensions: fidelity and dosage (Dusenbury, Brannigan, Hansen, Walsh, & Falco, 2005). Fidelity is the degree to which the core elements of an intervention are conducted as planned and dosage is the amount of exposure participants have to an intervention. The latter is often presented in terms of specific units of an intervention (e.g., number of lessons delivered) or amount of time that a participant is exposed to an intervention (e.g., hours of contact).

In addition to fidelity and dosage, research suggests that process-oriented dimensions such as the quality of intervention delivery are an important aspect of implementation that has the potential to facilitate or undermine outcomes (Dane & Schneider, 1998; Dusenbury et al., 2005). Unfortunately, this construct is examined less frequently because it is difficult to assess without conducting observations which are time-consuming and expensive. Quality of delivery has been conceptualized in a variety of ways. Most frequently, it is assessed in terms of how the intervention content is delivered and responded to, particularly for universal interventions that utilize explicit lessons. Given the interactive nature of these interventions, the interpersonal style of the implementers and their sensitivity to participants' needs have the potential to impact intervention delivery which, in turn, may impact participant responsiveness. Several studies have combined multiple indicators of implementers' delivery behavior (e.g., enthusiasm) and participant reaction (e.g., engagement) to assess this construct (Dusenbury et al., 2005; Resnicow et al., 1998). However, high-quality delivery also requires generalization of the content outside of the intervention's scripted lessons. Such application requires "deep structure" knowledge of the intervention model and skill to contribute to student improvement (Han & Weiss, 2005). Further, generalization is more difficult to capture through self-report measures and requires more frequent observation as the behavior is spontaneous and dependent on specific conditions.

### *1.2. Implementation measurement challenges*

Collecting information on the implementation of preventive interventions in community settings is challenging for a number of reasons. Some measures, such as observations, are more difficult to collect and may not be permitted in some settings. Implementation data are also costly and time-consuming. Most research studies that plan to include implementation variables in analyses gather assessments on more than one occasion (e.g., Spoth, Guyll, Trudeau, & Goldberg-Lillehoj, 2002). This strategy is based on principles of observational research that suggest that reliability of measurement increases with the addition of at least one time point (Hamre et al., 2009) but also reflects the assumption that levels of implementation have the potential to change over time. Improvement in implementation is expected in trials such as the one evaluated in the current study that provide on-going professional development support to teachers during the implementation period in order to address potential implementation challenges.

There is limited research regarding how many implementation assessments are needed to capture this process of change adequately. One study that compared single and dual assessments of fidelity and quality of process found stronger associations between the mean ratings of the two time points and outcomes than a single rating (Resnicow et al., 1998). However, there are few studies comparing levels of specificity and sensitivity for implementation measures that could be used to guide community-based practice. Another challenge to implementation research is that measures have the potential to vary in their validity. Two studies that included assessments of dosage gathered from both self-report and observational methods, reported positive associations between measures of dosage and student outcomes when observational measures were used, which disappeared when implementer self-report ratings were examined (Lillehoj, Griffin, & Spoth, 2004; Resnicow et al., 1998).

### 1.3. Linking implementation and child outcomes

Studies that have empirically assessed quality of delivery confirm that it is related to variation in outcomes (Abbott et al., 1998; Botvin, Dusenbury, Baker, James-Ortiz, & Kerner, 1989; Conduct Problems Prevention Research Group, 1999; Resnicow et al., 1998; Rohrbach, Graham, & Hansen, 1993). Assessments of fidelity and dosage are also related to participant outcomes but the findings are not always consistent. For example, several researchers have noted that observational ratings of implementer fidelity are associated with student outcomes (Battistich, Schaps, & Wilson, 2004; Botvin, Baker, Dusenbury, Tortu, & Botvin, 1990; Kam, Greenberg, & Walls, 2003; Resnicow et al., 1998; Rohrbach et al., 1993). Other studies, in which fidelity was assessed using observational methods, found limited or no associations with student outcomes (Cho, Hallfors, & Sanchez, 2005; Elias et al., 1986; Spoth et al., 2002). One interpretation of the lack of association is that observations were not conducted frequently enough to detect covariance of fidelity ratings and outcomes (Spoth et al., 2002). Further, in studies where the overall level of implementation quality is high, a smaller range of scores reduces the likelihood of being able to detect associations with outcomes (Cho et al., 2005; Spoth et al., 2002).

The findings regarding the relationship between assessments of dosage and participant outcomes are similarly mixed. In some cases, ratings of dosage gathered from implementer report are associated with student outcomes (Aber, Jones, Brown, Chaudry, & Samples, 1998; August, Bloomquist, Lee, Realmuto, & Hektner, 2006; Botvin et al., 1990; Spoth et al., 2002), whereas no association was reported by others (Basch, Sliepcevich, Gold, Duncan, & Kolbe, 1985; Lillehoj et al., 2004; Resnicow et al., 1992, 1998). Explanations similar to those provided for a lack of association between fidelity ratings and outcomes are also cited for studies of dosage. However, when self-report ratings of dosage are obtained, there is additional potential for social desirability to cause implementers to inflate ratings, resulting in lower associations between ratings and outcomes and, thus, masking the actual magnitude of association (Connell, Turner, & Mason, 1985).

This research review illustrates the complexity of measuring implementation and highlights a number of issues that warrant further examination. A limited knowledge base currently exists regarding different methods for documenting implementation. Very few studies include assessments of multiple dimensions of implementation, collect information from more than one source, or gather information at multiple time points. This type of data, when related to child outcomes, allows for the exploration of the relative usefulness of different dimensions of implementation quality.

### 1.4. The present study

The present study has two goals. The first is to examine patterns of implementation of the REDI intervention across multiple dimensions (Dosage, Fidelity, Generalization, and Child Engagement) that were measured over time for each of the four REDI curriculum components. Given that the REDI project utilized several techniques for promoting high-quality implementation including manualized materials, training, and ongoing support (Domitrovich et al., 2008), the quality of implementation on multiple dimensions was expected to be relatively high overall but still show growth over time.

The efficacy of the REDI intervention was established in a randomized clinical trial (Bierman et al., 2008). At the end of the one-year intervention, children who received REDI had significantly higher scores on a variety of language, literacy, and social-emotional measures compared to children who received the standard Head Start program (Bierman et al., 2008). While the basic integrity of the intervention and its implementation has been confirmed, there is still variation between teachers that could be examined in more detail (Bierman et al., 2008). The second goal of the present study is to determine, within each curriculum component, how different indicators of implementation relate to child social–emotional, language, and literacy outcomes. We expected child outcomes to vary by levels of implementation quality.

## 2. Method

### 2.1. Participants

Data included in this study are drawn from assessments collected as part of the Head Start REDI program, a randomized trial of an enriched Head Start intervention that included 44 classrooms across three Head Start programs in Central Pennsylvania. Participants in this study included two cohorts of 4-year-old children ($N = 192$; 17% Hispanic, 25% African American; 54% girls) from the 22 intervention classrooms.

### 2.2. Data collection procedures

Child assessments were conducted at schools by trained interviewers, during two individual 30–45 min "pull-out" sessions. To give children time to acclimate to the classroom setting, assessments began three weeks after school began and continued through the end of October. End-of-year child assessments were conducted in March and April.

One lead and one assistant teacher in each classroom provided independent ratings of child behavior. Baseline child ratings were collected from teachers at the end of September and end-of-year ratings were collected in April. Teachers were compensated US$ 20 to provide general information about themselves and their classrooms, and they received an additional US$ 7 per child for completing behavioral ratings.

### 2.3. Intervention design

The REDI intervention was divided into four components, each of which included curriculum-based lessons, center-based extension activities, and training in "teaching strategies" delivered by classroom teachers and integrated throughout the day into their ongoing classroom practices. Half of the participating classrooms used High/Scope (Hohmann, Weikart, & Epstein, 2008) as their base curriculum; the others used Creative Curriculum (Dodge, Colker, & Heroman, 2002). Project staff developed "crosswalk" tables to illustrate how the REDI target skills and methods mapped onto each of these base curricula. Three of the REDI components targeted language and emergent literacy skills while the fourth component targeted social-emotional skills. The underlying theory linking the REDI intervention to specific child outcomes is described elsewhere (Bierman et al., 2008).

#### 2.3.1. Language/emergent literacy skill enrichment

Several language and emergent literacy skills were targeted in REDI including vocabulary, syntax, phonological awareness, and print awareness. Three intervention components were developed to target these skills including an interactive reading program, a set of activities to promote phonemic awareness, and a center designed to foster increased letter recognition.

The Dialogic Reading (DR) component was based on the shared reading program developed by Wasik and Bond (2001) and Wasik, Bond, and Hindman (2006) which was, in turn, an adaptation of the dialogic reading program developed by Whitehurst and colleagues (Whitehurst et al., 1994). The curriculum included two books per week, which were scripted with interactive questions. Each book had a list of targeted vocabulary words, presented with the aid of physical props and illustrations. In addition to presenting these materials in a systematic way during the week, teachers received mentoring in the use of "language coaching" strategies, such as expansions and grammatical recasts, to provide a general scaffold for language development in the classroom (Dickinson & Smith, 1994). The overall goal was to improve teachers' strategic use of language in ways that would increase child oral language skills including vocabulary, narrative, and syntax.

The second curriculum component was a set of 61 Sound Games (SG), modeled primarily upon the work of Adams and colleagues (Adams, Foorman, Lundberg, & Beeler, 1998), to promote children's phonological awareness and print knowledge. The games were organized developmentally, moving from easier to more difficult skills during the course of the year (e.g., listening, rhyming, alliteration, words and sentences, syllables, and phonemes). Teachers were asked to use a 10–15 min Sound Game activity at least three times per week.

For the third component, teachers were provided with a developmentally sequenced set of activities and materials to be used in Alphabet Centers (AC). These included letter stickers, a letter bucket, materials to create a letter wall, and craft materials for various letter-learning activities. Teachers were asked to make sure that each child visited the alphabet center several times per week, and were given materials to track the children's acquisition of letter names.

#### 2.3.2. Social-emotional skill enrichment

The Preschool PATHS Curriculum (Domitrovich, Greenberg, Cortes, & Kusche, 2005) was the fourth curriculum component and targeted the promotion of children's social-emotional skills. The domains included: (1) prosocial friendship skills, (2) emotional understanding and emotional expression skills, (3) self-control (e.g., the capacity to inhibit impulsive behavior and organize goal-directed activity), and (4) problem solving skills, including interpersonal negotiation and conflict resolution skills. The curriculum is divided into 33 lessons that are delivered by teachers during circle time. These lessons include modeling stories and discussions, and utilize puppet characters, photographs, and teacher role-play demonstrations. Each lesson includes extension activities (e.g., cooperative projects and games) that provide children with opportunities to practice the target skills with teacher support. Teachers taught one PATHS lesson and conducted one extension activity each week. Generalized teaching strategies were encouraged with mentoring, including positive classroom management, use of specific teacher praise and support, emotion coaching, and induction strategies to promote appropriate self-control.

### 2.4. Implementation support

Teachers received detailed manuals and kits containing all materials needed to implement the intervention. A three-day professional training was conducted in August, prior to initiating the intervention, and a one-day "booster" training session was conducted in January. Teachers also received weekly mentoring support provided by local educational consultants (REDI Trainers), experienced master teachers who were supervised by two project-based senior educational trainers. The weekly consultations were intended to enhance the quality of implementation through modeling, coaching, and providing ongoing feedback regarding program delivery. REDI Trainers spent an average of three hours per week ($SD = .18$) in each classroom observing, demonstrating, or team teaching lessons. They also met with the lead and assistant teacher for one hour each week outside of class.

## 3. Measures

### 3.1. Implementation quality

#### 3.1.1. REDI trainer implementation ratings

The REDI Trainer Implementation Ratings (TIR) were completed by REDI Trainers each month during the academic year (eight times) to assess the extent to which teachers implemented the four curriculum components included in the REDI intervention model (Dialogic Reading, Sounds Games, Alphabet Center, and PATHS). REDI Trainers based their monthly rating on weekly visits to the classroom. Measures of Fidelity, Generalization, and Child Engagement were ratings made on 4-point Likert scales with wording that varied slightly across curriculum components depending on the nature of the materials and activities. For each curriculum component, each facet of implementation was rated with a single item, with the exception of Dialogic Reading, as noted below. Fidelity described the degree to which the lessons/activities were delivered as intended (e.g., for Sound Games, "Were the activities delivered as written?"; for PATHS, "Did the teacher cover the core elements of the written curriculum?"; 1: a little, 2: somewhat, 3: mostly, 4: completely and consistently). Generalization described the degree to which the teacher encouraged and reinforced the application of relevant concepts outside of the formal lesson time (e.g., PATHS "Did you observe generalization of PATHS concepts throughout the day?"; two items for Dialogic Reading, "Did the teacher reinforce vocabulary introduced in the lesson?" and "Did the teacher provide recasts to reinforce new grammatical structures?"; 1: very little or rarely, 2: some of the time, 3: most of the time, 4: nearly all of the time). Generalization was not rated for Sound Games because the intervention was restricted to delivery in scheduled small groups. Child Engagement described the degree to which children were interested and engaged in the lessons/activities (e.g., "How many of the children were positively engaged and interested?"; 1: very few, 2: some, 3: most, 4: nearly all).

We used these monthly REDI Trainer ratings in two ways. First, we used the eight monthly ratings to test for changes over time in implementation. Second, to test the associations between implementation and child outcomes, for each facet of implementation in each curriculum area, we created a single score by averaging scores across the eight months of REDI Trainer ratings. For example, the eight monthly REDI Trainer ratings of PATHS implementation Fidelity were averaged to create a single score characterizing PATHS implementation Fidelity across the year. This procedure yielded internally consistent composite scores with distributions that were close to normally distributed ($|\text{skewness}| < .50$). Psychometrics for measures of Fidelity were: PATHS $\alpha = .92$, $M = 3.16$, $SD = .52$; Dialogic Reading $\alpha = .82$, $M = 3.17$, $SD = .35$; Sound Games $\alpha = .90$, $M = 3.13$, $SD = .51$. For measures of Generalization: PATHS $\alpha = .92$, $M = 2.44$, $SD = .44$; Dialogic Reading $\alpha = .85$, $M = 2.43$, $SD = .36$; Alphabet Center $\alpha = .92$, $M = 3.15$, $SD = .57$. For measures of Child Engagement: PATHS $\alpha = .93$, $M = 3.34$, $SD = .53$; Dialogic Reading $\alpha = .90$, $M = 3.27$, $SD = .51$; Sound Games $\alpha = .81$, $M = 3.10$, $SD = .45$; Alphabet Center $\alpha = .76$, $M = 3.30$, $SD = .40$.

#### 3.1.2. Teacher reports of implementation quality

Teachers completed weekly logs in which they recorded information about their curriculum delivery. For each curriculum area, teachers recorded the number of intervention units delivered during a given week. The weekly reports of implementation Dosage were summed within each month for analyses of month-to-month changes in dosage, and the monthly dosage scores were summed across the entire year for analyses linking dosage to child outcomes (these data were summed rather than averaged to preserve the natural meaning of the count-based dosage scores.) Teachers used 3-point scales to rate implementation Fidelity and Child Engagement for some curriculum areas (e.g., "Did you deliver the PATHS lesson as written?" 1: major changes, 2: minor changes, 3: as written; "How engaged were the children in the PATHS lesson? 1: not at all, 2: somewhat, 3: very). A single score for each dimension was created by averaging scores across the eight months of teacher ratings. Preliminary analyses revealed strong ceiling effects in these ratings: across the year, mean ratings of Fidelity ranged from 2.84 to 2.92 and mean ratings of Child Engagement ranged from 2.75 to 2.93. Given that all teachers reported very high levels of Fidelity and Child Engagement throughout the year, we did not use these ratings to analyze within-year changes in implementation quality or to analyze the links between variations in implementation quality and child outcomes.

#### 3.1.3. Intercorrelations among measures of implementation

We examined intercorrelations among the REDI Trainer-based measures of Fidelity, Generalization, and Child Engagement; and then examined correlations between these dimensions and the teacher-reported dimensions (Dosage, Fidelity, and Child Engagement). For this purpose, we used the scores that were aggregated across the eight monthly reports. For PATHS, Generalization was not significantly correlated with Child Engagement, $r = .26$, *ns*, but Fidelity was strongly correlated with both Generalization, $r = .70$, $p < .001$ and Child Engagement, $r = .70$, $p < .001$. For Dialogic Reading, Fidelity, Generalization and Child Engagement were all moderately to strongly intercorrelated: $r_{\text{engagement·generalization}} = .50$, $p < .05$; $r_{\text{fidelity·generalization}} = .67$, $p < .001$; $r_{\text{fidelity·engagement}} = .66$, $p < .001$. For Alphabet Center, Generalization and Child Engagement were moderately correlated, $r = .48$, $p < .05$. Teachers' ratings of Fidelity were not significantly correlated with REDI Trainer ratings of Fidelity for PATHS, $r = .11$, *ns*, or for Alphabet Center, $r = .34$, *ns*. Similarly, teacher reports of Child Engagement were not correlated with REDI Trainer ratings of Child Engagement for PATHS, $r = .18$, *ns*; but teacher- and REDI Trainer-based reports of Child Engagement were significantly correlated for both Dialogic Reading, $r = .52$, $p < .05$, and Sound Games, $r = .67$, $p < .001$. None of the REDI Trainer-based measures of Fidelity, Generalization, or Child Engagement were significantly correlated with teacher reports of Dosage.

### 3.2. Child outcomes

Outcome measures included in this paper represent the core domains targeted by the different curriculum components: language skills targeted by Dialogic Reading; emergent literacy skills targeted by Sound Games and Alphabet Center; and emotional understanding, social problem solving, social behaviors, and learning engagement targeted by PATHS. Measures were collected through direct child assessment and teacher ratings. Details regarding scoring procedures and scale reliability are described in more detail in the original efficacy study (Bierman et al., 2008).

#### 3.2.1. Language skills

Two tests were administered directly to children to assess their language skills. In the *Expressive One-Word Picture Vocabulary Test* (EOWPVT; Brownell, 2000), children gave the word that best described pictures they were shown ($\alpha = .94$). The *Grammatical Understanding* subtest of the *Test of Language Development (TOLD*; Newcomer & Hammill, 1997) assessed syntax comprehension. Children listened to a sentence and chose one of four pictures that "best matched" the meaning of the sentence ($\alpha = .80$). Prior research has documented strong test-retest reliability for TOLD subtests ($r = .90$) and concurrent validity with comprehensive language assessments (Newcomer & Hammill, 1997).

#### 3.2.2. Emergent literacy skills

Three subscales assessing emergent literacy skills were drawn from the Test of Preschool Early Literacy (TOPEL; previously labeled the Pre-CTOPP; Lonigan, Wagner, Torgesen, & Rashotte, 2007). The *Blending* subtest assessed phonological processing. Children were asked to combine different parts of a word, such as "hot" and "dog" or "b" and "air" and point to the correct picture or say the full word ($\alpha = .86$). On the *Elision* subtest of the TOPEL, children deconstructed compound words and pointed to the correct picture (e.g., Point to 'snowshoe' without 'snow'; Say 'airport' without 'air'; $\alpha = .83$). In the *Print Knowledge* subtest of the TOPEL, children identified pictures of letters or words and named letters ($\alpha = .97$). Prior research has reported correlations in the range of .43 to .88 between these three subscales and the acquisition of initial reading skills (Lonigan, 2006).

#### 3.2.3. Emotional understanding and social-cognitive skills

Two measures were used to assess emotional understanding. On the *Assessment of Children's Emotion Skills* (ACES; Schultz, Izard, & Bear, 2004), children determined whether the facial expressions in 12 photographs reflected happy, mad, sad, scared, or no feelings. The score was the total number correctly identified expressions ($\alpha = .57$). On the *Emotion Recognition Questionnaire* (Ribordy, Camras, Stafani, & Spacarelli, 1988), children listened to 16 stories describing characters in emotionally evocative situations, and identified their feeling by pointing to pictures of happy, mad, sad, or scared faces. Children received a score of two for correctly identifying the feeling and a score of one for correctly identifying the valence ($\alpha = .63$).

Social problem-solving skills were assessed using a variation of the Challenging Situations Task (CST; Denham, Bouril, & Belouad, 1994). Children were presented with pictures of four peer scenarios (e.g., a peer knocking down blocks, being hit, entering a group, and a peer taking a ball). After each scenario, children were asked what they would do in the situation. Their open-ended responses were coded as *competent* (i.e., appropriately asserting oneself or calmly negotiating a solution, $\alpha = .68$), *aggressive* (i.e., verbal or physical antagonism, intimidation, or force, $\alpha = .77$), or *inept* (i.e., passive avoidance, $\alpha = .68$). Agreement between research assistants and data checkers on 1616 responses was high (kappa = .94).

#### 3.2.4. Social-emotional behaviors

*Social competence* was measured with 13 teacher-rated items describing prosocial behaviors such as sharing, helping, and understanding others' feelings, as well as self-regulatory behaviors such as resolving peer problems independently (Social Competence Scale, Conduct Problems Prevention Research Group [CPPRG], 1995). Internal consistency across the 6-point Likert scales ("never" to "almost always") was high ($\alpha = .94$); ratings by lead and assistant teachers were averaged ($r = .56$).

Seven items from the Teacher Observation of Child Adaptation-Revised (TOCA-R; Werthamer-Larsson, Kellam, & Wheeler, 1991) assessed *overt aggression* (e.g., stubborn, yells, fights). Six items from the Preschool Social Behavior Scale—Teacher Form (PSBS; Crick, Casas, & Mosher, 1997) assessed *relational aggression* (e.g., "Tells other kids he/she won't be their friend unless they do what he/she wants"). Items were rated on a 6-point Likert scale ("almost never" to "almost always"; $\alpha = .88$ and .93, respectively). Ratings from lead and assistant teachers were averaged ($r = .68$ for overt aggression; $r = .51$ for relational aggression). Teachers also completed the ADHD Rating Scale (DuPaul, 1991). This *activity level* scale includes 14 items reflecting difficulties with impulse control, distractibility, and sustained attention (e.g., "Is easily distracted," "Has trouble following directions") each rated on a 4-point scale. Lead and assistant teacher ratings were averaged ($\alpha = .94$; $r = .76$).

*School readiness* was measured with an 8-item scale developed for this study. Items were rated on a 6-point Likert scale ("strongly disagree" to "strongly agree"), and reflected self-regulation (e.g., "Has the self-control to do well in school" and "Can follow the rules and routines that are part of the school day"), learning motivation and involvement (e.g., "Seems enthusiastic about learning new things"), and compliance (e.g., "Is able and willing to follow teacher directions"). Lead and assistant teacher ratings were averaged ($\alpha = .95$, $r = .70$).

**Table 1**
Linear growth (slope) in implementation quality across eight monthly ratings.

| | PATHS | Dialogic Reading | Sound Games | Alphabet Center |
|---|---|---|---|---|
| Fidelity | .04 (.024) | .08*** (.017) | .01 (.022) | NA |
| Generalization | .08*** (.019) | .09*** (.019) | NA | .05* (.019) |
| Child Engagement | .01 (.016) | .04* (.015) | −.04 (.031) | .06*** (.016) |

*Note*. Entries indicate linear growth (standard deviation) expressed as the change in rating scale units per month.
* $p < .05$.
*** $p < .001$.

### 3.3. Observation of teaching quality

The Emotional Support scale from the Classroom Assessment Scoring System (CLASS; La Paro & Pianta, 2003) was used as a control variable in some analyses to clarify whether curriculum implementation quality was associated with child outcomes after taking into account more general patterns of supportive teaching. The CLASS is an observational measure that assesses 10 dimensions of teaching quality. Each dimension is rated on a 7-point Likert scale after each of four 20-min observation periods on a single day. Ratings for each dimension are averaged across the four observation periods. Research assistants were trained and certified as reliable by the CLASS developers. Intraclass correlations for the five dimensions describing emotional support indicated very good inter-rater agreement: positive climate ($r = .95$), negative climate (reversed; $r = .95$), teacher sensitivity ($r = .83$), overcontrol (reversed; $r = .79$), and behavior management ($r = .81$). Scores on these five dimensions were averaged to yield a single score describing *Emotional Support* ($\alpha = .86$).

## 4. Results

### 4.1. Changes in implementation across eight months of implementation

Changes over time in implementation quality were examined using the monthly ratings completed by REDI Trainers with the exception of Dosage which was only rated by teachers. We used multi-level regression models (time period nested within teacher) to analyze growth patterns separately for Dosage, Fidelity, Generalization, and Child Engagement for each of the curriculum components. Preliminary 3-level multi-level models indicated no statistically significant variations in growth parameters across program sites, so site was not included as a structural level in further statistical analysis. Non-linear growth terms were tested when the data suggested them, but were excluded from the final models as none of these were statistically significant. Final models were 2-level multi-level models with random intercepts and slopes. Observed data points are depicted in Fig. 1, with linear growth terms summarized in Table 1.

#### 4.1.1. Dosage

Growth in Dosage was not expected, and plots of means across time suggested no increase or decrease in levels across the year. Growth models confirmed that there were no statistically significant changes in mean Dosage levels for any of the curriculum components.

#### 4.1.2. Fidelity

Implementation Fidelity (Fig. 1a) for each curriculum component was moderately strong at the start of the intervention, averaging approximately 3 on the 1 to 4 rating scale. Ratings of Fidelity increased significantly over the year only for Dialogic Reading ($p < .001$), with the slope of .08 indicating a total increase of approximately .56 scale points across the intervention year. Linear changes in intervention Fidelity were not statistically significant for PATHS or Sound Games.

#### 4.1.3. Generalization

Generalization (Fig. 1b) started at relatively high levels for Alphabet Center (nearly 3 on the 4-point scale) and increased further across the intervention year ($p < .05$). In contrast, levels of Generalization were below the scale mid-point for both PATHS and Dialogic Reading at the start of the school year, but demonstrated statistically significant linear growth across the intervention year (.56 scale points for PATHS, .63 scale points for Dialogic Reading; both $p < .001$).

#### 4.1.4. Child engagement

Ratings of Child Engagement (Fig. 1c) were relatively high for each curriculum component at the start of the intervention, with average ratings falling between 3 and 3.5 on the 4-point scale. Child Engagement increased over time for both Dialogic Reading ($p < .05$) and Alphabet Center ($p < .001$). Child Engagement did not change over time for PATHS, with ratings remaining quite high across the full year (i.e., Mean ratings ranged from 3.3 to 3.55). Ratings of Child Engagement during Sound

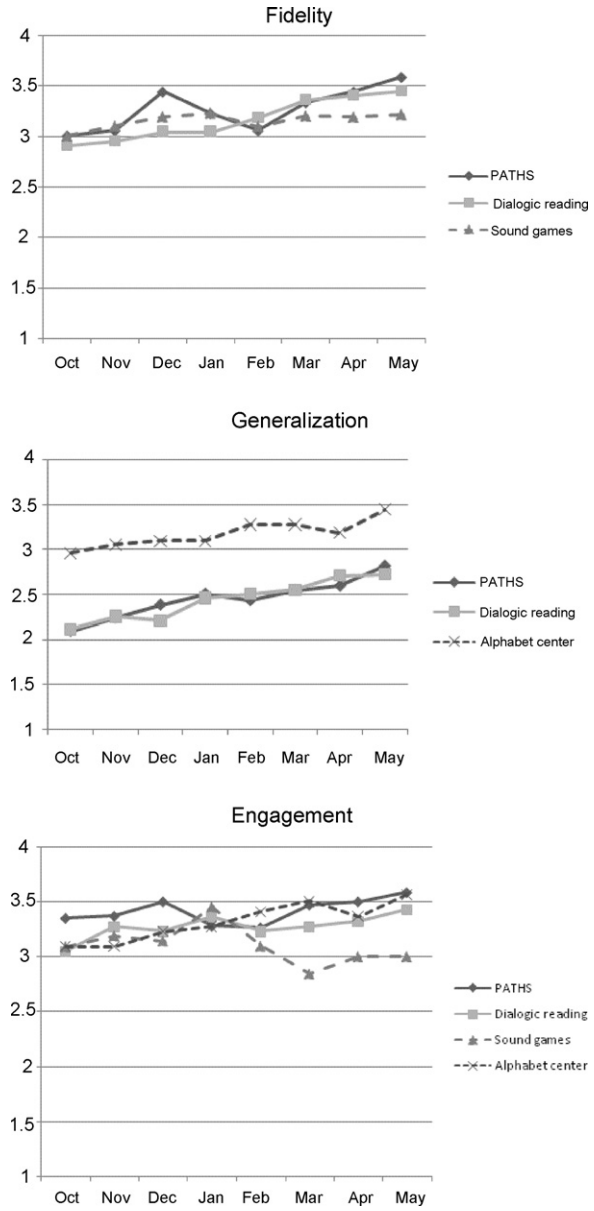


**Fig. 1.** Plots of REDI Trainer monthly ratings of implementation quality.

Games suggested a negative but statistically nonsignificant slope, so that engagement in Sound Games was lower (although still near 3 on the 4-point scale) than for other curriculum components by the end of the intervention year.

### *4.2. Association between implementation and child outcomes*

In each model testing the associations between implementation and a child outcome, a specific child outcome served as the dependent variable and separate regressions were conducted with different implementation dimensions as the independent (predictor) variable of interest. As described previously, each curriculum component was designed to impact a specific set of child outcomes, therefore, for each set of child outcomes, only the implementation dimensions for the corresponding curriculum component were tested (e.g., social-emotional outcomes were regressed on ratings of PATHS Fidelity, Generalization, and Child Engagement). Covariates in each model were pre-intervention levels of the child outcome variable (when available), child sex, child race, study site (urban vs. rural), and study cohort. For dependent variables with approximately normal distributions, we used multi-level regression models with a random intercept specified to account for within class clustering (SAS Proc Mixed). Outcomes with highly non-normal distributions were rescaled to create ordinal values, and then modeled using multi-level categorical regression models (using GLLAMM in Stata). For models revealing

**Table 2**
Child outcomes as a function of REDI-trainer ratings of fidelity, generalization, and child engagement: PATHS implementation and social-emotional outcomes.

| | Pre-Int. Control | Fidelity | Generalization | Child Engagement | Dosage |
|---|---|---|---|---|---|
| PATHS | | | | | |
| Emotion Understanding (ACES) | Y | −.20<br>(.43) | .35<br>(.41) | −.48<br>(.32) | .37<br>(.48) |
| Emotion Understanding (ERQ) | Y | .15<br>(.14) | .05<br>(.14) | .08<br>(.11) | .21<br>(.16) |
| Problem-Solving: Competent | Y | .30<br>(.16) | .00<br>(.19) | .25*<br>(.11) | −.11<br>(.22) |
| with CLASS-ES | | | | .25*<br>(.10) | |
| Problem-Solving: Aggressive | Y | .05<br>(.24) | .20<br>(.20) | −.12<br>(.20) | .12<br>(.24) |
| Problem-Solving: Inept | Y | −.48***<br>(.15) | −.32*<br>(.14) | −.17<br>(.18) | .51*<br>(.20) |
| with CLASS-ES | | −.62***<br>(.17) | −.32*<br>(.14) | | .51*<br>(.20) |
| Teacher: Social Competence | | .54**<br>(.18) | .32<br>(.21) | .37*<br>(.16) | −.26<br>(.24) |
| with CLASS-ES | | .60**<br>(.22) | | .36*<br>(.18) | |
| Teacher: Overt Aggression | | −1.22*<br>(.51) | −.15<br>(.54) | −1.16**<br>(.36) | .55<br>(.62) |
| with CLASS-ES | | −.97+<br>(.58) | | −1.03**<br>(.40) | |
| Teacher: Relational Aggression | | −.59<br>(.64) | −.34<br>(.64) | −1.33**<br>(.44) | .06<br>(.72) |
| with CLASS-ES | | | | −1.58***<br>(.47) | |
| Teacher: Activity Level | | −.60<br>(.35) | −.08<br>(.42) | −.83***<br>(.20) | −.24<br>(.36) |
| with CLASS-ES | | | | −.85***<br>(.24) | |
| Teacher: School Readiness | | 1.82*<br>(.73) | 1.53<br>(1.06) | .51<br>(.82) | .51<br>(.82) |
| with CLASS-ES | | 2.38*<br>(1.08) | | | |

*Note*: Entries are regression coefficients (and SE) after conrolling for sex, race, cohort, setting (urban vs. rural) and, when available, pre-intervention levels of the dependent variable. Inclusion of the pre-intervention score is indicated by the letter "Y" in the first column of the table. For models in which a dimension of implementation was a statistically significant predictor of a child outcome, we also present coefficients from a more stringent alternative model in which the CLASS Emotional Support (CLASS-ES) rating is added as a control variable to account for the effects of a generally supportive teaching environment.

+ $p < .10$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

a statistically significant association between implementation quality and a child outcome, we added observer ratings of CLASS Emotional Support to the model to determine whether the effect of curriculum implementation remained statistically significant after controlling for a more general measure of supportive teaching. These results are summarized in Table 2 for PATHS implementation and social-emotional outcomes and in Table 3 for language and literacy curriculum components and child outcomes.

*4.2.1. Dosage*

There were no statistically significant associations between variations in teacher reports of curriculum Dosage for any curriculum area and the child outcome variables. Because the measures of Dosage were non-normally distributed, with most teachers reporting very high Dosage and a few teachers reporting somewhat lower Dosage, we explored whether a dichotomous measure of Dosage might provide different results. To do so, we distinguished children in "relatively moderate-high Dosage classrooms" ($N_{PATHS} = 145$, $N_{Dialogic\ Reading} = 136$, $N_{Sound\ Games} = 123$, $N_{Alphabet\ Center} = 155$) from those in "relatively low-dosage classrooms" classrooms ($N_{PATHS} = 47$, $N_{Dialogic\ Reading} = 56$, $N_{Sound\ Games} = 69$, $N_{Alphabet\ Center} = 37$) based on the distribution of Dosage scores for each curriculum area. Groups were created by dividing the sample at the 25th percentile for

**Table 3**
Child outcomes as a function of REDI-trainer ratings of fidelity, generalization, and child engagement: dialogic reading, sound games, and alphabet center implementation and language and literacy outcomes.

| | Pre-Int. Control | Fidelity | Generalization | Child Engagement | Dosage |
|---|---|---|---|---|---|
| Dialogic Reading | | | | | |
| Vocabulary | Y | −1.11 | .76 | −.44 | .59 |
| | | (1.73) | (1.67) | (1.27) | (1.84) |
| Grammatical | Y | −2.23* | −1.71 | −1.46 | −.65 |
| Understanding | | (.99) | (1.07) | (.74) | (1.24) |
| Sound Games | | | | | |
| Blending | Y | −.56 | NA | .33 | −.13 |
| | | (.80) | | (.82) | (.76) |
| Elision | Y | .15 | NA | .34 | −.61 |
| | | (.50) | | (.50) | (.47) |
| Alphabet Center | | | | | |
| Print Knowledge | Y | NA | −2.24* | −1.13 | .53 |
| | | | (.98) | (1.74) | (1.79) |

*Note*. Entries are regression coefficients (and SE) after conrolling for sex, race, cohort, setting (urban vs. rural) and, when available, pre-intervention levels of the dependent variable. Inclusion of the pre-intervention score is indicated by the letter "Y" in the first column of the table.
* $p < .05$.

each construct. The cut-offs (and ranges) in terms of total number of sessions delivered across the year were the following: for PATHS, cut-score 29 (range 26–47); for Dialogic Reading, cut-score 80 (range 57–92); for Sound Games, cut-score 80 (range 59–100); and for Alphabet Center, cut-score 27 (range 26–32). Regressions with the same set of covariates listed above revealed only one statistically significant effect for the dichotomous Dosage indicator: children in classrooms with a high Dosage of PATHS reported fewer inept problem-solving solutions. Model results are presented in the right columns of Tables 2 and 3.

*4.2.2. Fidelity, generalization, and child engagement*

For the PATHS curriculum, higher implementation Fidelity was associated with fewer inept problem-solving responses, higher teacher ratings of social competence, lower teacher ratings of overt aggression, and higher teacher ratings of school readiness. Higher levels of PATHS Generalization were associated with fewer inept problem-solving responses. Higher levels of Child Engagement in PATHS lessons were associated with more competent problem-solving solutions, higher teacher ratings of social competence, and lower teacher ratings of overt aggression, relational aggression, and activity level. When these models were re-run to control for CLASS observer ratings of general levels of Emotional Support, 9 of the 10 statistically significant associations between PATHS implementation and child outcomes remained so; only one association, between PATHS Fidelity and teacher ratings of overt aggression, dropped to non-significance ($p < .10$).

For the Dialogic Reading, Sound Games, and Alphabet Center curricula, none of the expected associations between implementation quality and child outcomes emerged as statistically significant. Two unexpected associations emerged: higher ratings of Fidelity in Dialogic Reading were associated with lower scores on Grammatical Understanding, and higher ratings of Generalization in Alphabet Center were associated with lower levels of print knowledge.

Most of these significant associations involved either PATHS Fidelity or PATHS Child Engagement, so to clarify the unique predictive power of these strongly correlated ($r = .70$) facets of implementation, we ran a final set of models that included both Fidelity and Child Engagement as predictors, retaining all other control variables including baseline CLASS Emotional Support. For the two outcomes that had been predicted significantly by Fidelity but not by Child Engagement, Fidelity remained a statistically significant unique predictor of lower inept problem solving ($b = -.67$, $SE = .19$, $p < .001$), but its unique association with school readiness was not statistically significant ($b = 2.81$, $SE = 1.85$, $p = .14$). For the three outcomes that had been predicted by engagement but not by Fidelity, Child Engagement retained statistically significant unique associations with lower levels of relational aggression ($b = -1.91$, $SE = .54$, $p < .001$) and activity level ($b = -.91$, $SE = .30$, $p < .01$), but not with competent problem solving. For the two outcomes that had been predicted by both Fidelity and Child Engagement, neither measure retained a unique predictive association at the $p < .05$ level. However, the unique association between Fidelity and teacher-rated social competence ($b = .49$, $SE = .27$, $p = .075$) was similar in magnitude to the effect obtained without Child Engagement in the model ($b = .60$, see Table 2); and the unique association between engagement and overt aggression ($b = -.91$, $SE = .48$, $p = .055$) was similar in magnitude to the effect from the model without Fidelity in the model ($b = -1.03$, see Table 2).

## 5. Discussion

As the current climate of educational accountability extends downward into preschool, early childhood programs such as Head Start are striving to improve their impact on children's social and cognitive development by incorporating evidence-based interventions into existing curricula. Replicating these interventions with quality is challenging because often resources are more restricted in community settings as opposed to universities and several contextual factors have the

potential to undermine the implementation process (Domitrovich et al., 2008). Implementation quality is the cornerstone of the dissemination process and communities are unlikely to replicate positive outcomes like those achieved in rigorous research trials without high levels of fidelity to the original intervention model (Durlak & Dupree, 2008; Greenberg et al., 2001). To facilitate the translation of research to practice, communities need to be aware of the importance of monitoring implementation and know the most effective ways to do this including *when* to conduct assessments, *who* to gather that information from, and *what* dimensions to measure.

The present study explored these issues by measuring four dimensions of implementation (Dosage, Fidelity, Generalization, and Child Engagement) in the context of an evidence-based comprehensive preschool enrichment program implemented by three Head Start programs. The REDI intervention included explicit lessons and center-based extension activities related to four curriculum components: PATHS, Dialogic Reading, Sounds Games, and Alphabet Center. These curriculum components, and associated teaching strategies supported through mentoring by REDI Trainers, were designed to foster discrete skill acquisition in social-emotional, language and literacy domains for children and facilitate high-quality language interactions, positive discipline, support for children's social-emotional development, and a responsive interaction style in teachers.

### *5.1. When to assess implementation quality: changes across the intervention year*

Consistent with our first hypothesis, monthly ratings of implementation quality made by REDI Trainers indicated that implementation quality generally increased across the school year. This was not surprising given that the REDI project drew from the literature regarding factors that affect implementation and used a variety of strategies to promote quality (Domitrovich et al., 2008). These included comprehensive professional development training that focused on a rationale for the intervention, the use of visually appealing, standardized materials, and on-going support from REDI Trainers. In the context of these positive overall trends, it is useful to consider some of the variation by specific dimension of implementation quality.

Teachers reported high levels of Dosage throughout the year. This may reflect the success of REDI Trainers' efforts to help teachers establish consistent classroom schedules that included all REDI activities, or strong feelings of accountability prompted by administrative support and the discussion of implementation efforts at weekly meetings with REDI Trainers. Alternatively, it is possible that teachers did not provide accurate reports of Dosage.

Implementation Fidelity began at similarly high levels for PATHS, Dialogic Reading and Sound Games (around 3 on the 4-point scale), but only increased significantly for Dialogic Reading. Dialogic Reading was the most procedurally complex curriculum component, so it is not surprising that teachers became more skilled over the year in delivering it as intended. The lack of significant growth in Fidelity for PATHS or for Sound Games may reflect the scripted materials available to guide PATHS implementation and the relatively simple nature of the Sound Games activities.

Ratings of curriculum Generalization indicated significant linear growth for Alphabet Center, Dialogic Reading and PATHS, but from markedly different starting levels (Generalization was not rated for Sound Games because professional development did not emphasize this practice). Teachers demonstrated high levels of Generalization in the Alphabet Center early in the school year and made additional modest increases throughout the year, perhaps due to the strong current emphasis on letter learning in Head Start and the simple nature of pointing out letters throughout the day.

In contrast, levels of Generalization for PATHS and Dialogic Reading were initially low (around 2 on the 1–4 scale) but rose substantially across the school year. Generalization is critical because both PATHS and Dialogic Reading lessons and activities introduce basic concepts that teachers should use as a scaffold to engage in generalized practices that change the broader social-emotional and linguistic ecology of the classroom. In the case of PATHS, teachers may only gradually come to recognize opportunities to reinforce and apply strategies for labeling and managing emotions and social conflicts on a regular basis, partly because this requires teachers to step back from emotionally charged situations to recognize them as "teachable PATHS moments." Similarly, it may take time for teachers to integrate Dialogic Reading principles of language recasting and vocabulary reinforcement into natural conversations with children throughout the day. Anticipating these challenges, REDI Trainers initially focused primarily on basic aspects of implementation dosage and fidelity (i.e., how to integrate the four curriculum components into teachers' weekly schedule and the mechanics of delivering core lessons). Over time, generalization became a primary focus of the mentoring provided by REDI Trainers (e.g., providing feedback about how to take advantage of teachable moments and apply the intervention concepts spontaneously as situations arose).

Ratings of Child Engagement were relatively high from the beginning of the year but showed variable patterns of growth across the year. Children responded especially positively to PATHS lessons from the outset, leaving little room for growth in engagement. This may reflect PATHS' long history of development and refinement and the fact that compelling puppet characters are introduced early in the year. In contrast, children became steadily more engaged in Dialogic Reading and Alphabet Center activities across the year, though perhaps for different reasons. Dialogic Reading was a complex intervention requiring the integration of multiple materials, activities and teaching strategies, so improvements in Child Engagement may reflect teachers' increasing success at achieving a natural and smooth integration of procedures. In contrast, the alphabet center was comprised of a simple activity sequence that left teachers with considerable freedom to choose specific materials, props, and generalization formats. Therefore, gains in engagement may reflect teachers' increasing success in finding the most engaging activities for their students. Child Engagement in Sound Games started at similar levels but did not increase across

the year. Unlike the other curriculum components, Sound Games activities increased steadily in cognitive complexity as the year progressed: it is possible that activities began to exceed the capacities of younger or less skilled children, dampening overall levels of engagement.

In sum, these findings highlight the importance of considering the nature of the intervention and the dimension of implementation being assessed when making decisions regarding the timing of data collection. An assessment made early in the year may be adequate if that dimension remains stable, but may lead to underestimates of other dimensions of implementation, particularly if intervention procedures are relatively complex or require the generalization of principles throughout the day. Assessments of child engagement may also be misleading if the cognitive demands of activities increase substantially across the year. Given that in this study, growth in implementation was reasonably summarized by linear trends, a compromise approach may be to assess implementation once relatively early in the year and once again later in the year. Additional research on these interventions with similar assessments of implementation quality is needed before these findings could be used to guide decisions in the field.

### *5.2. Who should assess implementation quality?*

In addition to making decisions about what dimension of implementation to assess, communities need to develop a strategy regarding who will provide the implementation data. Options for measuring some aspects of implementation are restricted. For example, it is difficult to envision how anyone but the classroom teacher can provide an estimate of intervention dosage: despite our REDI Trainers' presence in teachers' classrooms for 3 h per week, they were dependent on teachers' weekly implementation logs to learn how many lessons were actually implemented. With regard to Fidelity and Child Engagement, we attempted to gather information from both teachers and REDI Trainers so that we could compare how these two informants' reports were related to child outcomes. Consistent with prior research, however, teachers' ratings of and Child Engagement were so near the ceiling of the rating scales that they provided virtually no useful information about between-classroom variations. Teachers may be inclined to inflate ratings of their own implementation quality due to social desirability effects (e.g., wanting to please or appease the REDI Trainers) or fear of negative evaluation (e.g., from supervisors), but monitoring child receptivity while also delivering an intervention is a challenging task and teachers may not have a frame of reference for evaluating fidelity or child engagement relative to what may be taking place in other classrooms. It is also possible that an alternative rating format may have been more successful at generating more variance in teacher reports. In the absence of such evidence, however, it would appear risky to depend on teachers as the sole source of information about implementation quality.

### *5.3. What dimensions of implementation should be assessed?*

In other reports, we have described results of the randomized control trial documenting the positive impact of the REDI intervention on children's functioning in both the social-emotional and early literacy domains (Bierman et al., 2008). Here we focus on variations in child outcomes among children within the REDI intervention condition as a strategy to learn which aspects of curriculum implementation may have been especially relevant to these child outcomes.

Overall, we found little evidence of an association between the limited variation in Dosage we measured and child outcomes. The only reliable finding was an association between being in a high-PATHS-dosage classroom and providing fewer inept responses on a social-problem solving task. The predominance of null associations between Dosage and child outcomes likely reflects the very high levels of Dosage reported by teachers. The present results may not generalize to other studies in which a wider range of intervention dosage is observed.

For the PATHS curriculum, variations in implementation Fidelity and Child Engagement were associated with a range of child social-emotional outcomes, including the quality of children's problem-solving responses and teacher ratings of social competence, aggression, and school readiness. With the exception of child aggression, these associations held even after controlling for pre-intervention observations of teachers' general emotional support for students (i.e., positive climate, sensitivity, and behavior management). This finding is important because critics of explicit instruction in preschool suggest that these types of interventions are unnecessary in the context of high-quality teaching (Elkind, 2001). The present results indicate that high-quality implementation of a specific social-emotional learning curriculum is associated with variation in student outcomes that cannot be accounted for by teachers' general quality of teaching. The associations between implementation of PATHS and teacher ratings of attention and self control are consistent with other community replications of the intervention that demonstrated changes in student behaviors were mediated by improvements in these more immediate cognitive-behavioral skills (Domitrovich, Kam, & Small 2006).

Trainer ratings of the degree to which teachers generalized PATHS (i.e., supported children in the application of the concepts throughout the day) was significantly correlated with only one child outcome. REDI Trainers spent an average of 3 h per week in classrooms, which may not be enough time to adequately assess this dimension.

There were no instances in which Fidelity and Child Engagement each accounted for unique variance in a child outcome when they were tested in the same model, but it would be premature to consider these two dimensions to be synonymous or redundant. First, they tap fundamentally different phenomena—the teacher's adherence to lesson procedures versus children's affective and behavioral engagement. The strength of the correlation between these phenomena could be interpreted as measuring the program developers' success in refining curriculum procedures in a way that maximizes children's engage-

ment. Second, the small sample size in the current study ($n = 22$ classrooms) provides little statistical power to detect unique effects in highly correlated dimensions. Nonetheless, it may be that alternative measures of either fidelity or engagement could make them more distinct and less highly correlated.

For the language and literacy curriculum components (Dialogic Reading, Sound Games and Alphabet Center), we did not find any positive associations between implementation quality and child outcomes. In fact, there were two associations in the unexpected direction: Dialogic Reading Fidelity was negatively associated with grammatical understanding, and Alphabet Center generalization efforts were negatively associated with print knowledge. It is important to interpret these effects in the context of results from the randomized trial (Bierman et al., 2008), which indicated no overall effect of the REDI program on grammatical understanding scores and weak positive effects on print knowledge. The fact that the REDI intervention did not "harm" children's grammar skills and that it actually benefitted their print knowledge suggests, at the very least, that our measures of implementation quality for these curriculum areas did not adequately capture their "active ingredients."

It is noteworthy that some of the largest effects of the REDI intervention were on children's phonological awareness skills targeted by the Sound Games curriculum, yet we found no association between fidelity of Sound Games implementation and children's skills in these areas. It is possible that teachers were making adaptations that worked for their children but that were perceived by RT's as deviating from printed lesson directions. In the field of prevention science the fidelity-adaptation debate is one of the most important issues that needs to be addressed in future research (Bumbarger & Perkins, 2008). As intervention implementers gain knowledge and deep understanding of an intervention model, they are more likely to make adaptations that deviate from strict definitions of fidelity (e.g., not following printed directions) but that adhere to the underlying logic model of the intervention (Hord, 1987). To capture this phenomenon, it may be useful to supplement measures of "procedural" fidelity with measures that rate the degree to which deviations from printed protocols still adhere to the underlying logic of the intervention. For Alphabet Center, the lack of associations between implementation quality and child outcomes may reflect the fact that the current Head Start centers were already engaged in extensive efforts to promote letter knowledge prior to their participation in REDI, and these efforts were not rated by REDI Trainers.

Looking across curriculum areas, we found only one association between a measure of curriculum generalization and child outcomes: teachers who generalized the PATHS curriculum had children who provided fewer inept problem-solving solutions. This was unexpected because generalization throughout the school day is a particularly critical part of the logic models underlying the PATHS and Dialogic Reading curricula. We suspect that the predominantly null effects reflect the difficulty of obtaining valid measures of generalization from outside observers based on a very limited sampling of teaching practices across the full instructional day. Even though REDI Trainers' monthly ratings were based on 12 h of classroom observation each month, they may not provide a sufficient or representative window into teachers' typical use of generalization strategies. In contrast, REDI Trainer ratings of Fidelity and Child Engagement were based on a similarly limited sampling of classroom observation and yet were more consistently related to child outcomes. This suggests that valid measures of implementation fidelity and child engagement may be feasible even with relatively limited observations, whereas more extended observation samples may be necessary to obtain useful measures of generalization. Alternatively, observers could be given instructions to focus on generalization in particular classroom settings (e.g., use of Dialogic Reading strategies during lunchtime conversations with children).

### *5.4. Study limitations*

This study had several strengths including assessments of implementation quality by observers, a diverse set of ratings that reflect multiple dimensions of implementation quality supported in the literature, and a comprehensive set of child outcomes, but several limitations are noteworthy. First, while the findings regarding the usefulness of dosage ratings for distinguishing student outcomes is potentially important, variation was restricted on this dimension compared to others and may have prevented the ability to detect meaningful group differences. This raises an important caveat about the usefulness of implementation research conducted in the context of efficacy trials and the ability to generalize findings to effectiveness trials for the use of community programs. In these trials, the research team has significantly more control over implementers, and the level of support provided tends to be higher than what is typically available in community contexts. Validation of implementation measures should be conducted in the context of large scale replications where a full range of quality ratings is likely to be available to be examined in relation to student outcomes and a larger sample size would ensure adequate power for analyses.

A second limitation is the fact that we did not assess the reliability of the REDI Trainer implementation ratings. REDI Trainers were assigned to work with one Head Start program and only rated the levels of implementation for the teachers that they worked with. Meetings with REDI Trainers were held to discuss the ratings in an effort to ensure that they were interpreting scale items the same way. In future studies, a set of master video tapes of curriculum components will be coded using the implementation measures and trainers will have to demonstrate adequate reliability to these codes.

Third, this study only assessed implementation across one year. One might assume that outcomes assessed in the second year of implementation would be stronger if implementation quality improves from fall to spring, but very few studies assess implementation longitudinally. One exception is a study of a school-based, drug prevention curriculum (Ringwalt et al., 2009a). Sustainability studies are needed to determine for how long high levels of implementation quality would be maintained over time under normal community conditions. Our experience is that even in research trials, breaks in teaching caused by holidays or the seasons (i.e., summer) disrupt implementation quality.

Finally, the current trial is unique in that it utilized an intensive professional development support model that is not often present under typical community conditions and this may limit the extent to which the findings can be generalized. The use of "coaches" in educational settings is not a new concept (Joyce & Showers, 2002) but there is growing recognition of the potential benefit of this model for ensuring high levels of implementation of preventive interventions (Ringwalt et al., 2009b).

The data collected as part of the REDI trial allowed for initial exploration of the "*who*, *what*, and *when"* of implementation. The findings suggest that there are different patterns of implementation quality that appear to vary as a function of the intervention being assessed. Larger scale studies of interventions with greater implementation variability are needed to replicate these patterns in addition to randomized trials with conditions that vary specific dimensions of implementation (e.g., dosage) in order to establish the relevance of different dimensions of implementation for program outcomes. Research of this sort, that targets implementation as the outcome, is the next phase of prevention science that will facilitate a richer science and practice exchange.

## Acknowledgement

This project was supported by the National Institute for Child Health and Human Development grants HD046064 and HD43763 awarded to Karen L. Bierman of The Pennsylvania State University. The authors would like to thank the administration and teaching staff of the Head Start programs participating in the REDI project. We also extend our appreciation to the children and families for their participation.

## References

Abbott, R. D., O'Donnell, J., Hawkins, J. D., Hill, K. G., Kosterman, R., & Catalano, R. F. (1998). Changing teaching practices to promote achievement and bonding to school. *American Journal of Orthopsychiatry*, *68*, 542–552.

Aber, J. L., Jones, S. M., Brown, J. L., Chaudry, N., & Samples, F. (1998). Resolving conflict creatively: Evaluating the developmental effects of a school based violence prevention program in neighborhood and classroom context. *Development and Psychopathology*, *10*, 187–213.

Adams, M. J., Foorman, B. R., Lundberg, I., & Beeler, T. (1998). *Phonological sensitivity in young children: A classroom curriculum*. Baltimore, MD: Brookes.

August, G. J., Bloomquist, M. L., Lee, S. S., Realmuto, G. M., & Hektner, J. M. (2006). Can evidence-based prevention programs be sustained in community practice settings? The Early Risers advanced-stage effectiveness trial. *Prevention Science*, *7*, 151–165.

Basch, C. E., Sliepcevich, E. M., Gold, R. S., Duncan, D. F., & Kolbe, L. J. (1985). Avoiding type III errors in health education program evaluations: A case study. *Health Education Quarterly*, *12*, 315–331.

Battistich, V., Schaps, E., & Wilson, N. (2004). Effects of an elementary school intervention on students' "connectedness" to school and social adjustment during middle school. *Journal of Primary Prevention*, *24*, 243–262.

Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A., Greenberg, M. T., et al. (2008). Promoting academic and social-emotional school readiness: The Head Start REDI program. *Child Development*, *179*, 1802–1817.

Botvin, G. J., Baker, E., Dusenbury, L., Tortu, S., & Botvin, E. (1990). Preventing adolescent drug abuse through a multimodal cognitive-behavioral approach: Results of a three year study. *Journal of Consulting and Clinical Psychology*, *58*, 437–446.

Botvin, G. J., Dusenbury, L., Baker, E., James-Ortiz, S., & Kerner, J. (1989). A skills training approach to smoking prevention among Hispanic youth. *Journal of Behavioral Medicine*, *12*, 279–296.

Brownell, R. (2000). *Expressive One-Word Picture Vocabulary Test Manual*. Novato, CA: Academic Therapy Publications.

Bumbarger, B., & Perkins, D. F. (2008). After randomization trials: Issues related to dissemination of evidence-based interventions. *Journal of Children's Services*, *2*, 53–61.

Catalano, R. F., Berglund, M. L., Ryan, J. A., Lonczak, H. S., & Hawkins, J. D. (2002). Positive youth development in the United States: Research findings on evaluations of positive youth development programs. *Prevention & Treatment*, *5*, Article 15. Retrieved from http://journals.apa.org/prevention/volume5/pre0050015a.html.

Cho, H., Hallfors, D. D., & Sanchez, V. (2005). Evaluation of a high school peer group intervention for at-risk youth. *Journal of Abnormal Child Psychology*, *33*, 363–374.

Conduct Problems Prevention Research Group (CPPRG). (1995). *Teacher Social Competence Scale Technical Report*. Available from the Fast Track Project web site, http://www.fasttrackproject.org.

Conduct Problems Prevention Research Group (CPPRG). (1999). Initial impact of the Fast Track prevention trial for conduct problems: II. Classroom effects. *Journal of Consulting and Clinical Psychology, 67*, 648–657.

Connell, D. B., Turner, R. R., & Mason, E. F. (1985). Summary of findings on the school health evaluation: Health promotion, effectiveness, implementation, and costs. *Journal of School Health*, *55*, 316–321.

Crick, N. R., Casas, J. F., & Mosher, M. (1997). Relational and overt aggression in preschool. *Developmental Psychology*, *33*, 579–588.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control. *Clinical Psychology Review*, *18*, 23–45.

Denham, S., Bouril, B., & Belouad, F. (1994). Preschoolers' affect and cognition about challenging peer situations. *Child Study Journal*, *24*, 1–21.

Dickinson, D. K., & Smith, M. W. (1994). Long-term effects of preschool teachers' book readings on low-income children's vocabulary and story comprehension. *Reading Research Quarterly*, *29*, 102–105.

Dodge, D. T., Colker, L., & Heroman, C. (2002). *The Creative Curriculum for preschool* (4th ed.). Washington, DC: Teaching Strategies, Inc.

Domitrovich, C. E., Bradshaw, C. P., Poduska, J. M., Hoagwood, K., Buckley, J. A., Olin, S., et al. (2008). Maximizing the implementation quality of evidence-based preventive interventions in schools: A conceptual framework. *Advances in School Mental Health Promotion*, *1*, 6–28.

Domitrovich, C. E., Greenberg, M. T., Cortes, R., & Kusche, C. (2005). *The Preschool PATHS Curriculum*. Channing Bete Publishing Company.

Domitrovich, C. E., Kam, C. M., & Small, M. (2006, May). *Exploring the relationship between social-emotional learning and achievement.* Paper presentation at the Society for Prevention Research, Washington, DC.

DuPaul, G. (1991). Parent and teacher ratings of ADHD symptoms: Psychometric properties in a community-based sample. *Journal of Clinical Child Psychology*, *20*, 245–253.

Durlak, J. A., & Dupree, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*, 327–350.

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*, 237–256.

Dusenbury, L., Brannigan, R., Hansen, W. B., Walsh, J., & Falco, M. (2005). Quality of implementation: Developing measures crucial to understanding the diffusion of preventive interventions. *Health Education Research*, *20*, 308–313.

Elias, M. J., Gara, M., Ubriaco, M., Rothbaum, P. A., Clabby, J. F., & Schuyler, T. (1986). Impact of a preventive social problem solving intervention on children's coping with middle-school stressors. *American Journal of Community Psychology*, *14*, 259–275.
Elkind, D. (2001). *The hurried child*. Cambridge, MA: De Capo Press.
Fixen, D. L., Naoom, S. F., Blasé, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa, FL: University of South Florida. Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network (FMHI Publication #231).
Gottfredson, D. C., & Gottfredson, G. D. (2002). Quality of school-based prevention programs: Results from a national survey. *Journal of Research in Crime and Delinquency*, *39*, 3–35.
Greenberg, M. T., Domitrovich, C. E., & Bumbarger, B. (2001). The prevention of mental disorders in school-aged children: Current state of the field. *Prevention & Treatment, 4* Article 1. Retrieved from http://journals.apa.org/prevention/volume4/pre0040001a.html.
Greenberg, M. T., Domitrovich C. E., Graczyk, P., & Zins, J. (2001). *The study of implementation in school-based preventive interventions: Theory, research, & practice*. Report to Center for Mental Health Services (CMHS), Substance Abuse Mental Health Services Administration, U.S. Department of Health and Human Services.
Hahn, R., Fuqua-Whitley, D., Wethington, H., Lowy, J., Crosby, A., Fullilove, M., et al. (2007). Effectiveness of universal school-based programs to prevent violent and aggressive behavior: A systematic review. *American Journal of Preventive Medicine*, *33*(Suppl. 2), S114–129.
Hamre, B. K., Pinata, R. C., & Chomat-Mooney, L. (2009). Conducting classroom observations in school-based research. In L. M. Dinella (Ed.), *Conducting science-based psychology research in schools*. (pp. 79–105). Washington, DC: American Psychological Association.
Han, S. S., & Weiss, B. (2005). Sustainability of teacher implementation of school-based mental health programs. *Journal of Abnormal Child Psychology*, *33*, 665–679.
Hohmann, M., Weikart, D. P., & Epstein, A. S. (2008). *Educating young children* (3rd ed.). Ypsilanti, MI: High/Scope Press.
Hord, S. (1987). *Evaluating educational innovation*. New York, NY: Croom Helm.
Joyce, B., & Showers, B. (2002). *Student achievement through staff development*. Alexandria, VA: Association for Supervision and Curriculum Development.
Kam, C., Greenberg, M. T., & Walls, C. T. (2003). Examining the role of implementation quality in school-based prevention using the PATHS curriculum. *Prevention Science*, *4*, 55–63.
La Paro, K. M., & Pianta, R. C. (2003). *CLASS: Classroom Assessment Scoring System*. Charlottesville, VA: University of Virginia.
Lillehoj, C. J. G., Griffin, K. W., & Spoth, R. (2004). Program provider and observer ratings of school-based preventive intervention implementation: Agreement and relation to youth outcomes. *Health Education & Behavior*, *31*, 242–257.
Lonigan, C. J. (2006). Development, assessment, and promotion of preliteracy skills. *Early Education and Development*, *17*, 91–114.
Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2007). *TOPEL: Test of Preschool Early Literacy*. Austin, TX: PRO-ED, Inc.
Newcomer, P. L., & Hammill, D. D. (1997). *The Test of Language Development—Primary* (3rd ed.). Austin, TX: Pro-Ed.
O'Donnell, C. L. (2008). Defining, conceptualizing and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research*, *78*, 33–84.
Resnicow, K., Cohn, L., Reinhardt, J., Cross, D., Futterman, R., Kirschner, E., & Allegrante, J. P. (1992). A three-year evaluation of the Know Your Body program in inner-city schoolchildren. *Health Education Quarterly*, *19*, 463–480.
Resnicow, K., Davis, M., Smith, M., Lazarus-Yaroch, A., Baranowski, T., Baranowski, J., et al. (1998). How best to measure implementation of school health curricula: A comparison of three measures. *Health Education Research*, *13*, 239–250.
Ribordy, S., Camras, L., Stafani, R., & Spacarelli, S. (1988). Vignettes for emotion recognition research and affective therapy with children. *Journal of Clinical Child Psychology*, *17*, 322–325.
Ringwalt, C. L., Pankratz, M. M., Hansen, W. B., Dusenbury, L., Jackson-Newsom, J., Giles, S. M., et al. (2009). The potential of coaching as a strategy to improve the effectiveness of school-based substance use prevention curricula. *Health Education & Behavior*, *36*, 696–710.
Ringwalt, C. L., Pankratz, M. M., Jackson-Newsom, J., Gottfredson, N. C., Hansen, W. B., Giles, S. M., & Dusenbury, L. (2009). Three-year trajectory of teachers' fidelity to drug prevention curricula. *Prevention Science*. doi:10.1007/s11121-009r-r0150-0.
Rohrbach, L. A., Graham, J. W., & Hansen, W. B. (1993). Diffusion of a school-based substance abuse prevention program: Predictors of program implementation. *Preventive Medicine*, *22*, 237–260.
Schultz, D., Izard, C. E., & Bear, G. (2004). Children's emotion processing: Relations to emotionality and aggression. *Development and Psychopathology*, *16*, 371–387.
Spoth, R., Guyll, M., Trudeau, L., & Goldberg-Lillehoj, C. (2002). Two studies of proximal outcomes and implementation quality of universal preventive interventions in a community-university collaboration context. *Journal of Community Psychology*, *30*, 499–518.
Wasik, B. A., & Bond, M. A. (2001). Beyond the pages of a book: Interactive book reading and language development in preschool classrooms. *Journal of Educational Psychology*, *93*, 243–250.
Wasik, B. A., Bond, M. A., & Hindman, A. (2006). The effects of a language and literacy intervention on Head Start children and teachers. *Journal of Educational Psychology*, *98*, 63–74.
Werthamer-Larsson, L., Kellam, S., & Wheeler, L. (1991). Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology*, *19*, 585–602.
Whitehurst, G. J., Arnold, D., Epstein, J. N., An-gell, A. L., Smith, M., & Fischel, J. E. (1994). A picture book reading intervention in daycare and home for children from low-income families. *Developmental Psychology*, *30*, 679–689.
Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology*, *49*, 156–167.