Published in final edited form as:
Early Child Res Q. 2010 ; 25(3): 314–328. doi:10.1016/j.ecresq.2010.03.001.

Examining Different Forms of Implementation and in Early Childhood Curriculum Research

Samuel L. Odom[1], **Kandace Fleming**[2], **Karen Diamond**[3], **Joan Lieber**[4], **Marci Hanson**[5], **Gretchen Butera**[6], **Eva Horn**[2], **Susan Palmer**[2], **Janet Marquis**[2], and **Children's School Success Project**

[1]FPG Child Development Institute, University of North Carolina [2]University of Kansas [3]Purdue University [4]University of Maryland [5]San Francisco State University [6]Indiana University

Risk conditions associated with poverty create developmental and educational disparities between young children experiencing these conditions and those from more economically advantaged circumstances (Halle et al., 2009). A primary prevention approach to reducing these disparities has been to provide early childhood education programs focusing on school readiness to children living in poverty and/or from ethnically diverse backgrounds (Farran, 2000; Winsler et al., 2008). In early childhood education, there is an increased emphasis on the development and evaluation of curriculum models and increasing contemporary evidence that use of these curricula will positively affect the development and learning of young children (Bierman et al., 2008; Clements & Sarama, 2008;). For the promise of such findings to be realized, especially for children with the most pressing developmental and learning needs, efficacious curricula must be implemented well by practitioners in the community (Odom, 2009). Assessment of implementation has been conceptualized in a variety of ways (Durlak & DuPre, 2008), but to date, there have been few examinations of different approaches to implementation and their association with child outcomes.

Although different conceptualizations or features of implementation appear in the early childhood education literature (see Odom et al., in press for a review), they all revolve around the types of instructional practices teachers use with children and/or the curriculum content that teachers deliver in the classroom. In the current study, early childhood education curriculum is conceptualized as having an established content as specified by a scope and sequence of activities and instructional procedures that describe expected teacher behavior and practices. Clements and Samara (2008) provide one example of that approach in their study of the *Building Blocks* early math curriculum. They identified the math topics that teachers covered along with the percentage of weeks that teachers spent on the topics, described large group, small group, computer, and home activities through which teachers delivered content, and delineated the specific instructional behaviors in which teachers engage when delivering the curriculum.

In contrast, there are some early childhood instructional or intervention models that are "curriculum-independent" in that they are designed to be used in classes implementing different curricula. For example, strategies for increasing use of math-mediated language in

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

the classroom (Rudd, Lambert, Satterwhite, & Smith, 2009), dialogic reading by caregivers (Briesch, Chafouleas, & Lebel, 2008), or positive behavior support (Benedict, Horner, & Squires, 2008) are designed to be used in settings that may have different curricular emphases. This distinction is important because different approaches to assessment of implementation will be differentially appropriate. That is, when a specific curriculum is implemented, assessment of that curriculum's implementation should emphasize delivery of the quantity of specified content and teachers' delivery of the critical instructional components of the curriculum, whereas implementation of curriculum-independent approaches may well focus on teachers' instruction of critical intervention components.

O'Donnell (2008) highlighted that distinction in her meta-analysis of instructional implementation for school-age children and outcomes. She found that implementation could be classified as structural (i.e., focusing on an amount of instruction or number of lessons provided) or process (i.e., focusing on delivery of key procedural features of the curriculum and/or quality of delivery) variables.

In the early childhood curriculum evaluation literature, researchers have applied structural and process-oriented approaches to the measurement of implementation as well. In their analysis of teachers' use of curricula in their Reading First research, Al Otaiba et al. (2008) assessed implementation by the amount (e.g., in time) of reading instruction that was delivered during the day. To promote physical activity of preschool-aged children, Williams, Carter, Dibbe, and Dennison (2009) assessed the amount of time and number of lessons a preschool teacher used *Animal Tracker* curriculum activities. Further, in the evaluation study of the Infant Health and Development Project (IHDP) designed to promote the development of children born with low birth weights, Ramey et al. (1992) created a family participation index to assess implementation. The index consisted of the number of home visits, parent meetings attended, and child attendance in centers. The content of the *Early Partners* (i.e., for the youngest participants) and *Partners for Learning* curricula were delivered in those contexts.

In some early childhood intervention studies, intervention or curricula are delivered across nationally or regionally dispersed sites. For example, in the studies of the efficacy of IHDP, eight nationally dispersed sites were centers for implementation. The *Animal Trackers* curriculum in the Williams et al. (2009) study took place in 32 Head Start classes located in four different counties in New Mexico. In such curriculum or model evaluation studies, it is important to assess whether implementation is similar across sites, for both practical and methodological reasons. Practically, the implementation data should allow researchers to determine sites in which implementation is not occurring as well as expected or as well as at other sites. For example, Ramey et al. (1992) displayed the variation in Family Participation Index scores (i.e., a measure that included number of home visits, number of attendances at parent group meetings, days child attended child development center) across sites. Methodologically, it would inform researchers' decisions about including site as a factor in their data analysis models. In the Ramey et al. (1992) study, site was entered as a variable into the regression model they used to assess program impact.

Implementation in early childhood curriculum research has been assessed through process measures as well. Often researchers rely on observation of teachers in classrooms and use checklists or observer rating scales containing key procedural components of the curriculum. In their analysis of implementation of the Language Focused Curriculum approach with preschoolers, Pence, Justice, and Wiggins (2008) constructed and used a procedural checklist during observations in the class. To assess implementation of a literacy curriculum in programs for preschoolers from low-income families, Zvoch, Letourneau, and Parker (2007) constructed a checklist that documented the teachers' delivery of procedural features of the curriculum during classroom observations. Using a multilevel analysis, they noted that the relationship between degree of implementation and growth on literacy measures was generally positive, but also that it was substantially affected by a small number of low implementation sites. As part of an efficacy study of the Head Start REDI curriculum, Bierman et al. (2008) had research staff observe monthly in classes and rate the quality of implementation of four program components. In several studies of the use of the Incredible Years program to promote social problem solving and reduce conduct problems for children attending Head Start classes, Webster-Stratton and colleagues have used an implementation rating scale that documents quality of the delivery of the curriculum activities and procedures (e.g., Webster-Stratton, Reid, & Stoolmiller, 2008).

An added dimension to the examination of implementation is the time variable. Early childhood curricula are usually designed to be delivered across a specified time period, which is often a school year. Although measures of implementation are often collected at several time points during the year, the consistency of implementation across a year is rarely reported. An exception is Zvoch’s (2009) evaluation of two early childhood literacy curricula implemented in 52 Head Start classrooms. Research staff collected process implementation measures (i.e., a checklist of key instructional features in each curriculum) at three points in time across the year. Using a multi-level growth curve analysis model to examine the trends in implementation across time and at different sites, he found that sites differed at the onset of interventions (i.e., at observation 1 or the intercept), and trends across time also differed. Zvoch proposed that collecting measures at different points in time was an important dimension of implementation assessment.

When judging the impact of a curriculum on development and learning, it is logical to assume that children are affected by both the amount of the curriculum content they experience and the quality of instruction as intended by the curriculum developer. That is, in addition to a scenario in which a high proportion of the curriculum is delivered at a high quality, it is also possible that a small amount of a curriculum could be delivered at a high quality or a large amount of a curriculum could be delivered at a low quality. In the latter two scenarios, if researchers were measuring only quality of implementation in the former case (i.e., a process measure) or quantity of curriculum in the latter case (i.e., a structural measure), it would appear that implementation was occurring at a high level. The children’s experience of the curriculum may be overstated, and the curriculum’s potential impact on child outcomes may be understated. In previous early childhood curriculum evaluation studies, researchers have tended to assess either the structural or the process features of implementation. It may well be that: a) structural and process assessments of implementation may be differentially associated with specific impact of the curriculum; and/or b) a multiplicative composite of structural and process implementation approaches may provide a measure of implementation that is more strongly associated with curriculum impact than either approach when used alone.

Examinations of curriculum and intervention effects have sometimes revealed that children who are low performing at the beginning of a study may benefit more from curriculum effects that other children in the group (Boocock, 1995; Currie & Thomas, 1999; Ramey, Campbell, & Ramey, 1999). One example of this association was found by the Cost, Quality and Outcome Study Team (1995). They documented that the effects of child care quality on children’s development were most pronounced for children who were the most disadvantaged. By inference, it may be possible that the effects of different levels of implementation could have differential effects on children performing at different levels at pretest. To date, there have not been early childhood curriculum studies that have examined the interaction between children’s pretest performance and types or levels of implementation

In this study, we use data from a larger early childhood curriculum evaluation study to describe the collection of both structural and process implementation measures, examine the levels of process implementation across time and sites, calculate a multiplicative composite implementation metric that combines structural and process measures, and determine the association of the measures of implementation with child outcomes. The specific research questions addressed in this study are: 1) What are the levels of structural and process implementation for the CSS curriculum? 2) Are there site differences for the structural and process measures? 3) What are the levels and site differences of a multiplicative composite measure of implementation that includes both structural and process measures? 4) What are the trends for the process measure across time? 5) What are the associations between the structural, process, and combined measures of implementation and child outcomes? Are there interactions between pretest performance on outcome measures and associations with different measures of implementation?

Method

In this study, we have used data from a larger research study conducted by the Children's School Success Project (CSS) (2004). The CSS Project is a randomized evaluation of an integrated curriculum model designed to promote school readiness for children from low-income families or who are at-risk for poor school performance for other reasons (e.g., having English as a second language, having an identified disability). To examine different approaches to implementation, we draw only from data collected for the treatment group (i.e., children receiving the CSS curriculum) in the larger CSS study.

Curriculum

The CSS Curriculum (Children's School Success Research Group, 2009) was designed to promote academic and social competence of preschool children. During its development in a planning year for this project, the researchers surveyed curricula that had evidence of efficacy, consulted with experts in the field about the most efficacious curricula, drew content and process from the various curricula to establish an integrated set of curriculum activities, and developed lessons for 133 school days. The daily lessons included large group activities, small group activities, plans for extending discussion of themes across the day, and procedures for individualization. The content is briefly described in subsequent sections.

Language and literacy—In this curriculum, language and literacy activities focused on vocabulary development, phonemic awareness, letter recognition, listening and comprehension. These activities were informed by research on responsive book-reading (Lonigan, 2006; Whitehurst et al., 1994), approaches to large-group teaching in preschool classrooms (Powell & Diamond, 2004), and phonemic awareness instruction (Adams, Foorman, Lundberg, & Beeler, 1998). The literacy component of the curriculum occurred every day.

Early math—Content for math activities was drawn from the *Building Blocks* curriculum by Clements and Sarama (2003) and the professional literature (Copley, 2000). The curricular goals focused on teaching beginning numbers and operations, geometry and spatial sense, measurement, pattern/algebraic thinking, and displaying and analyzing data. Small group math activities occurred every day.

Science—Content for science was drawn from the French, Conezio, and Boynton (2003) *Science Start* curriculum materials. Specifically, activities were drawn from the units on Color and Light, Measurement and Graphing, Neighborhood Habitat, and Properties of Matter. The Science Start curriculum focuses on a problem solving process for discovering and enhancing scientific concepts, learning specific science concepts and relationships, and language development through vocabulary enrichment. Science lessons occurred every other day.

Social competence—The Dinosaur School curriculum from the larger Incredible Years Program, developed by Webster-Stratton (2000), was the basis for the social competence component. This curriculum component included large group and small group activities focused on social challenges children encounter in the classroom, and it made use of puppets to illustrate concepts. The unit themes were: emotional literacy, empathy and perspective taking, friendship skills, anger management, interpersonal problem-solving, and being successful in school. Emphasis was placed on the prevention of challenging behavior and prosocial problem-solving strategies. Dina School activities occurred every other day.

Curriculum integration—The contents and procedures from each of these curriculum sources described previously were integrated into scripted daily lessons. Although literacy and math activities occurred in each daily lesson, the primary themes addressed in large group and small group activities alternated daily between social competence and science. Each day, the teacher also reviewed the content that was presented the previous day during the large group activity.

Training for teachers and coaching—To introduce the curriculum, research staff provided three days of training to teachers and assistant teachers of early childhood classes at the beginning of the year and another day of training 6–8 weeks after the teachers had used the curriculum. Also during the initial training sessions, research staff planned with teachers how to use the curriculum in their classroom. Research staff who served as site supervisors visited classrooms at least once per week. These research staff members had Masters or doctoral degrees in early childhood education or special education and experience working with young children in classroom settings. During visits to classrooms, they observed, modeled curriculum lessons, provided feedback to teachers, discussed with teachers the activities they had conducted in the previous week, and planned the activities they would conduct in the next week. Teachers agreed to conduct curriculum lessons every day classes were held with the exception of occasional special days (e.g., Halloween, Valentines, field trips), began the lessons within the first week of school, and used the curriculum across the entire school year. Classes met either 4 or 5 days per week, depending on the program. Most (71%) of the classes were “half day,” meeting from 2 ½ to 3 ½ hours per day, with the remainder being full day classes.

Setting

A total of 51 early childhood classes for preschool-age children were involved in the study. Information about sites and participants appears in Table 1. These classes were located in California (CA), Kansas (KS), Indiana (IN), West Virginia (WV), and Maryland (MD) (i.e., 10–11 classes in each state). The rationale for this site selection was to provide a test of the CSS curriculum in sites that displayed some of the regional and ethnic diversity that exists in the United States. The CA site was in an urban setting that had a large percentage of Latino families. The KS site had classes in suburban or small-town settings and primarily nonminority families. The IN classes were located primarily in rural communities and included primarily nonminority families. The WV site was also rural but located in the Appalachian mountains, with families being primarily nonminority. The MD site was urban and consisted of a higher proportion of African American families than occurred in other sites. It should be noted that while the CA site had the highest proportion of children who were English Language Learners (ELL), all sites had some children who were ELL (i.e., primarily from Spanish speaking families). The programs were designed for children “at risk” for school failure, and primarily fell into five classifications: Head Start (68%), Head Start and Public School combination (10%), state pre-kindergarten (12%), private (2%), and other (8%). They were located in urban, suburban, and rural locations. Teachers varied in education level, training, and certification (i.e., range = Associate’s to Master’s degree). The average number of children per classroom was 17.1 (*SD* = 3.10), with typically 2–3 teachers and/or assistant teachers in each classroom. Also, it should be noted that classes entered the study across a four-year time period, with three entering in the first three years and one entering the last year (i.e., the IN site was an exception in that it had two classes in this study in the fourth year of the project).

Child Participants

To participate in the study, children had to meet the age-criteria for entering kindergarten in their respective communities the year following their participation in the study. The mean age of children at the beginning of the school year was 53.19 months (*SD*=4.30). Three-year old children were sometimes in the classrooms and participated in curriculum activities, but they were not included as study participants. Three hundred fifteen boys and 261 girls participated in the study. The sample included a mix of children who qualified solely because of the low-income level of their families (*n* = 369), children who were English Language Learners (*n* = 112), and children with identified disabilities (*n* = 89). The latter two classifications of children usually came primarily from low-income families.

Implementation Measures

Implementation measures were designed to capture both structural and process features of implementation (O’Donnell, 2008).

Proportion of curriculum completed (Structural variable)—The structural measure of implementation in this study was the number of lessons the teacher completed during the school year in respective components of the curriculum. Site supervisors, during their weekly visits to the classrooms, took notes on the teachers’ use of the curriculum and other events in the class. From these notes and their observations, the site supervisors reported the number of lessons teachers completed. Using this information, they calculated the proportion of the literacy, math, science, and social lessons the teachers taught (e.g., number of literacy lessons taught/total number of literacy lessons).

Quality rating (Process variable)—Implementation quality rating scales were developed to assess the degree and quality of the teachers’ implementation of the literacy, math/science1, and social features of the curriculum. The rating scales had items assessing teachers’ preparation of large group and small group activities, the skill with which the teachers delivered the lessons, the teachers’ integration of concepts into activities across the day, teachers’ responses to children, and other teaching related to the specific curriculum. The literacy rating scale had 19 items; the math/science rating scale had 25 items, and the social rating scale had 61 items. The social rating scale had substantially more items because we used, with little adaptation, the scale that Webster-Stratton and colleagues developed for the Dinosaur School Curriculum (Webster-Stratton et al., 2008), whereas we developed the other implementation rating scales based on elements of the CSS curriculum that we had developed. The scales were based on a five-point, Likert rating system with ratings ranging from “not well” (a rating of 1) to “extremely well” (a rating of 5). Also, there was an “N/A”

[1] The assessments of math and science implementation were included on the same rating form because of logistic issues. There were different sections for math and science on the form. Because we did not have a logical child outcome measure for the science component of our curriculum, we have limited our report in this study to the math implementation measure.

designation if a specific item was not intended to be implemented during a specific phase of the curriculum or if the site supervisor was unable to observe the item.

At the beginning of the project, the rating scales were introduced to the site supervisors and scale items were discussed. Site supervisors practiced the scale and brought questions back to the group for discussion on cross-site conference calls. Site supervisors had periodic (i.e., usually weekly) conference calls after the project began and discussed their interpretation of items when questions arose. Because the site supervisors were located at regionally distant sites, we were not able to collect cross-site, inter-rater agreement for the measures.

Site supervisors completed the rating measures seven times during the year in each classroom, at approximately the same points (across sites) during the curriculum implementation. Their ratings were based on their observation of an entire day. The site supervisors also used the observations to discuss with teachers their implementation of the curriculum. Cronbach alphas were calculated to assess internal consistency of the items for each of the three scales at each observation. The alpha coefficients were .98 for social competence (range from .93–.99 at individual sites), .94 for literacy (range from .82–.94 at individual sites), and .98 for science/math (range from .90 to .99 at individual sites).

Multiplicative composite measure—It was possible that the quantity of curriculum content that a teacher delivers and the quality with which the teacher delivers the curriculum could in combination affect children’s learning and development differently from either variable in isolation. To address this possibility, a multiplicative composite score was created for each curriculum area by multiplying the proportion of curriculum completed by the average quality rating for the specific curriculum area. For example, if a teacher delivered .75 of the literacy component of the curriculum with an average quality item rating of 4.0, the multiplicative composite score would be 3.0.

Child Outcome Measures

To assess child outcomes, research staff conducted child assessments in the fall of the year and again in the spring. Staff members were trained to conduct assessments by following a common protocol across sites. Assessment information was sent to and aggregated at a common data center at the Scheifelbusch Life Span Institute at the University of Kansas.

A variety of dependent measures were collected to monitor child outcomes. These assessments were collected in the fall and the spring. For literacy and language, we collected the Peabody Picture Vocabulary Test-III (Dunn & Dunn, 1997), an expressive letter naming task employed in Head Start’s Family and Child Experiences Survey (FACES; Zill & Resnick, 2006), the Woodcock-Johnson (WJ) Word Attack (McGrew, Woodcock, & Mather, 2000), the Picture Naming Expressive Language/Individual Growth and Development Indicator (McConnell, 2004), and the Purdue Emergent Writing assessment developed by Diamond, Gerde, and Powell (2008). For early math, three subtests from the Woodcock Johnson were selected: Applied Problems, Quantitative Concepts A, and Quantitative Concepts B. For the science area, there were no specific outcome measures that had been developed to assess early science knowledge or problem solving. The developer of *ScienceStart* used the PPVT as a primary outcome measures (French, 2004), although it is primarily a measure of receptive vocabulary. So in this study, we do not include an implementation-child outcome analysis for science because an appropriate measure had not yet been developed.

To assess the social outcomes, teachers completed the Social Skills Rating System (Gresham & Elliott, 1991). As part of a larger study, the research team conducted a factor analysis of the social skills and problem behavior scales of the SSRS and established the following factors: Social Participation, Self Regulation, Classroom Survival Skills, Externalizing, Internalizing. All scales met minimal criteria for internal consistency (alphas ranged from .82 to .90). Also, to assess children's problem solving, the research staff administered an adaptation of the WALLY (Webster-Stratton, 1990). The WALLY is an analogue assessment in which research staff present 13 scenarios of social problem situations (e.g., What would you do if someone took your toy on the playground?) and show picture cards that illustrate the situation. Children describe their solution to these problems. All WALLY data were sent to the IN site and coded by a single set of research staff members (i.e., rather than by staff members at all five sites). Staff classified the solutions as positive or negative, using a coding system developed by Webster-Stratton (1990). For this analysis, inter-rater agreement on solutions identified was collected and calculated for 25 % of the data. The formula for calculating inter-rater agreement was agreement by the two raters on a child's response as positive or negative divided by the total number of responses issued by a child. Inter-rater agreement was 84.67%. For this study, the mean number of positive solutions to problems and mean number of negative solutions were the measures employed.

Results

Proportion of Curriculum Completed (Structural Measure)

The mean proportion (and standard deviation) of curriculum components that the teachers delivered at each of the regional sites and the total mean appear in Table 2. For the literacy and social components of the curriculum, teachers delivered about three-fourths of the curriculum during the year, and for math, teachers delivered about two-thirds of the curriculum content.

To determine if there were site differences in implementation ratings, we conducted an analysis of variance using three implementation variables for each of the three curriculum components (nine variables total) as dependent variables, The single factor was site with five levels. Planned contrasts compared site means to the grand mean.

Quantity of Implementation (Structural Variable)

For literacy and math proportion variables, a Levene's Test revealed heterogeneity of variance across sites, so a Welch's correction was used for the ANOVA. For the quantity variables there were no significant differences across sites (Literacy, $F(4,22) = 1.2$, $p > .05$; Math, $F(4,46) = .98$, $p > .05$; Social, $F(4,22.7) = 1,27$, $p > .05$).

Quality of Implementation (Process Variable)

For literacy quality, there were significant site differences in the quality ratings [$F(1,46)= 3.57$, $p < .05$], and planned comparison indicated that the IN and KS sites had significantly higher scores ($p < .05$) than the grand mean. There were also significant site effects for math quality (F (4, 22.22)= 6.20, $p < .01$), with planned comparisons indicating the KS ratings significantly above the mean ($p < .05$) and the MD site marginally below the mean (p, < .05). The social component quality ratings also yielded significant effects for site (F (4, 22.57) = 4.24, $p < .05$), with planned comparisons indicating the IN site ratings significantly above the mean ($p< .01$), and the MD site ratings significantly below the mean ($p < .01$). Given that the MD site had lower quality ratings for Social and the IN and KS sites had consistently higher ratings, the quality ratings for individual classrooms were inspected. At the MD site, three classes had extremely low levels of both quality and quantity especially for social implementation. These teachers (one in each of three different years) were extremely low implementers of the curriculum and appear to account for the difference in mean ratings for MD. When the quality rating data for these three were removed from the whole MD sample, the average ratings for that site approximated the total group mean (i.e., Literacy = 3.91, Math = 3.92, Social = 3.63). This would suggest that the ratings given by the MD raters were similar to those given by the other sites except in the instance of teachers with extremely low levels of implementation. An examination of both the IN and KS sites did not reveal similar outlier classes that substantially increased the mean.

Multiplicative composite Measure

A similar analysis of site differences was conducted for the multiplicative composite measure. No significant site effects were found for literacy [*F* (4,46)= 1.61, *p*> .05], math [*F*(4,46)= 1.10, *p* > .05], or social [*F* (4,22.80) = 2.22, *p* > .05].

Process Measure (Quality) Across Time

To examine differences among sites across time, we plotted the mean quality rating at each site for each of the seven quality ratings. Figure 1 contains the mean quality ratings at each of the sites for each of the measures. We conducted a growth curve analysis to model the repeated quality scores over time within a classroom. Site, time, and the interaction between site and time were modeled to determine if there were site differences in intercept (initial quality level), if there were differences in quality ratings across time, and if the slopes (trajectories) of quality ratings differed between sites. For math quality, there were significant differences in intercept across sites, [*F*(4,51.5)=3.60, *p* <.01], with KS and IN continuing to have higher intercepts but there were not significant effects for time or the time by site interaction. For literacy quality, there were significant differences in intercepts [*F*(4,73.9)=9.01, *p*<.001] and slopes, [*F*(4,53.9)=6.54, *p*<.002]. IN (4.27) and KS (4.30) had higher intercept ratings than MD (3.80), WV (3.70), and CA (2.82), Once again, the overall trend across time was not significantly different from zero [*F*(1,55.3)=.30, *p* > .05]. For the social quality measure, there were significant differences for intercept across sites, [*F*(4, 62.8)=9.68, *p* < .001] and slopes [*F*(4, 56.2)=7.23, *p*< .001]. Examination of the parameter estimates from the model indicated that IN (4.57) and KS (4.23) sites had higher intercept ratings than CA (2.97), WV (3.81) and MD (3.36). CA had a positive slope, but the trends in the other sites were zero or negative such that the overall trend over time was not significantly different from zero [*F*(1,56.4)=3.14, *p* >.05].

Association among measures—Correlations between the proportion of curriculum component completed, mean rating of quality of implementation and the multiplicative composite measure were quite high and significant, as can be seen in Table 3. The constituent measures (i.e., proportion and quality) are always more highly correlated with the multiplicative composite measure of their curriculum component than with the other constituent measure of the same curriculum component or any measures of implementation of the other components. For example, the social quality rating variable is more highly correlated with the social multiplicative composite measure than it is with the social proportion measure or any of the math/science and literacy variables.

Model Building Procedures

To examine the utility of the various models of implementation, we followed a model building process described by Snijders and Bosker (1999). With this process, we compared three different implementation models to a base model in order to determine which implementation variables were significantly related to scores at the end of the preschool year. Because child outcomes (a Level 1 variable) were nested within classrooms and classroom measures of implementation (Level 2 variables) were predictors, we used a Multilevel Model approach for the modeling. The Base model for student outcomes at posttest included terms for the categorical variables of site, gender, prekindergarten (preK) IEP, English Language Learner (ELL), race/ethnicity2, and the grand-mean centered pretest score on the variable of interest. Random intercepts for each class were also included in the model. SAS Proc Mixed was used to run all analyses.

With this model building approach, we utilized the Base model as the basis for comparison of the other implementation models. The first model, called the Quantity Model, added proportion of curriculum completed to the terms in the Base model (i.e., the predictor terms in the Quantity Model were site, gender, prek IEP, ELL, race/ethnicity, pretest score, and percentage of curriculum completed). A second model, called the Quality Model, added quality of implementation to the terms in the base model. A third model, called the Multiplicative Composite Model, added the multiplicative composite to the terms in the Base model. Each of these implementation models were compared to the Base model to determine if the addition of the implementation variable(s) resulted in a significantly improved model fit. When predictors were added to the models, a main effect for the implementation variable and an interaction between pretest and implementation were entered as a first step. For some of the outcomes, the interaction was significant and needed to remain in the model, for other outcomes the interaction was not significant and the implementation variable was evaluated as a main effect. If the interaction between pretest and the fidelity variable added significantly to the model, the interaction was "probed" using the tools provided by Preacher, Curran, and Bauer (2006) to identify the range of values on the pretest where the fidelity variable was significant (regions of significance). Computing regions of significance is advantageous because rather than just picking arbitrary values of the variable at which to examine the significance of simple slopes, knowing the region of significance tells the user the results of all possible simple slopes tests.

The match between the child outcome variable to include in the analysis of the implementation variables was determined by the component of the curriculum being examined. For example, when examining the math component of the curriculum, the WJ Math Concept A was a dependent variable assessing math child outcomes, and the proportion of math curriculum completed, math quality, and math multiplicative composite were the explanatory variables used to create the model comparison for that component. The explanatory variables being compared in the models were all grand-mean centered (the mean for the sample on each variable was subtracted from each person's score) to allow for easier interpretation of the results.

Fit indices—To compare the implementation models with the Base model, two types of fit statistics that are standard SAS PROC MIXED output were used: the deviance statistic and Akaike's Information Criterion (AIC). When models are nested within one another and use identical data, model comparisons can be made through the use of the deviance statistic (Snijders & Bosker, 1999). Two models are nested if both contain the same terms and one has at least one additional term. The deviance statistic is calculated by computing the difference between the −2 log likelihood values obtained for the models being compared. The deviance statistic is distributed asymptotically as a χ^2 statistic with the degrees of freedom equal to the difference in the number of model parameters estimated. When the deviance statistic exceeds the critical value based on degrees of freedom, the predictor model has significantly better fit than the comparison model. The critical value for one degree of freedom is 3.84. AIC values are also fit indices that can be used to compare models when the models are not nested models, and a lower value indicates better fit of the model. Hox (2002) indicates that parsimony or simplicity is an important criterion in the comparison of nonnested models and that the AIC was developed to compare nonnested models, adjusting for the number of parameters estimated. Comparisons between the base

[2] The race/ethnicity variable was designed to control for racial/ethnic diversity in our sample. The classification collapsed children into white and other groups for this analysis.

model and the predictor models are nested. Comparisons between predictor models are not nested.

For some variables, an interaction occurred between student scores on the pretest measure and implementation. That is, the effect of implementation was significant only for students scoring in certain regions of significance on the pretest. As noted previously, when this occurred, comparison of the implementation models to the Base model required inclusion of two terms (i.e., the grand-mean centered implementation variable and the interaction between implementation and grand-mean centered pretest). Thus, the critical value for the deviance test is that of a Chi Square with two degrees of freedom rather than one, and the deviance value necessary for a significant improvement in the model is higher (i.e. 5.99).

Math Implementation and Math Outcome

In Table 4, the base model for the WJ Test 18A Quantitative Concepts variable appears with the −2 log likelihood value listed next to it (i.e., 1729.1). For the comparison of each implementation model with the base model, the deviance statistic is calculated and appears in the second column of the table. All three implementation models were significant improvements over the base model. The parameter estimates for all three implementation variables were positive, indicating that as the implementation variables increased, student outcomes at posttest increased. The implementation variables were not associated with the other math child outcome variables.

Literacy Implementation and Vocabulary Outcomes

For the PPVT variable, the model including quality and the interaction between quality and pretest PPVT score yielded a significantly improved model over the Base model (see Table 4). The negative parameter estimate for the interaction indicated that as pretest scores increased, the effect of quality of implementation was reduced, that is for children scoring below the mean on the pretest, the effect of quality was stronger. The effect of quality was significant for those scoring 78 or lower on the pretest (the lowest scoring 22% of the sample. The Quantity model was not a significant improvement over the Base model. Although the Multiplicative Composite Model had a relatively high deviance statistic, it did not have significantly better fit than the Base Model ($p > .05$). The implementation measures were not associated with the other child literacy and vocabulary outcome variables.

Social Implementation and Outcomes

For two of the SSRS variables, self-regulation and classroom survival skills, each of the implementation variables significantly improved the model over the base model. Comparison of the AIC indices indicated that the Multiplicative Composite Model had the lowest AIC value and would be the model of choice. For the SSRS Problem Behavior Internalizing variable, the Quality model had the lowest AIC, indicating that it would be the preferred model. Also, for this variable there was an interaction with the pretest scores. Follow-up tests on the significant interaction between pretest internalizing behavior and social curriculum quality for regions of significance revealed that quality was positively associated with outcomes for children who scored greater than 1.89 (the 68% of the sample with the least internalizing behavior). Quality was inversely associated with posttest internalizing behavior scores for children scoring less than 1.28 at pretest, which was the 10% of the sample with the most internalizing behavior. For this analysis the ratings were reverse scored, so that higher ratings on the 0–2 scale, indicated less internalizing behavior.

For the Wally, tests on the interaction for the regions of significance indicated that quality of implementation was significant for children who scored 0 or 1 at the pretest for total positive solutions (12% of the students)or greater than 12 (only 1% of the students) on the pretest.

Children who gave the fewest positive responses on the WALLY at pretest were most affected by quality of implementation with higher quality resulting in increased positive responses at posttest. The implementation variables were not associated with posttest scores for any of the other social outcome variables.

Correlations Between Implementation Variables and Child Outcomes

Another way to examine relationships among the implementation variables and child outcomes is to look directly at the correlations between the two. This is somewhat complicated by the fact the child outcomes are Level 1 variables and implementation variables are Level 2. In order to obtain appropriate child outcomes to correlate with the implementation variables, the base models were run for each outcome with all of the covariates (site, gender, IEP, ELL, race/ethnicity, and pretest score). The model predicted score at posttest for each child was saved into a data file. These predicted outcomes were then aggregated to obtain the average predicted outcome for students in each class. These average outcomes for each class were then correlated with the implementation variables for each class. The correlations appear in Table 5. These correlations display somewhat similar relationships for the model analysis. For Math, the quality and multiplicative composite measures had the strongest associations, although these were not significant for the correlational analysis. For the literacy measure, the model including quality was the best fit and the correlation between quality and the PPVT was the strongest, although again this correlation failed to reach significance. For the SSRS-Self Regulation variable, all three correlations with implementation variables reached significance and all three models including the implementation variables were significant improvements over the base model. A similar tend occurred for the SSRS-Classroom Survival skills variable, except that the correlation coefficient for quality did not reach significance. For the SSRS-Internalizing Behavior variable, the correlations did not reach significance, although in the model analysis, the quality model in interaction with the pretest was a significantly better fit than the base model. For the Wally Positive Responses, the model including the quality implementation variable in interaction with the pretest was a significantly better fit than the base model and the quality variable was most highly and significantly correlated with the Wally variable.

Discussion

The purpose of this study was to examine the different forms of implementation assessment in preschool curriculum evaluation research. Building from O’Donnell’s (2008) conceptualization of structural and process implementation approaches and the work Zvoch (2009) completed on implementation across sites and time, we investigated five questions related to the assessment of implementation and its association with child outcomes. These questions were addressed through a larger evaluation study of the CSS curriculum, a multi-component preschool readiness curriculum.

The first question addressed the degree of curriculum implementation, as measured by structural (i.e., proportion of curriculum delivered) and process (i.e., quality ratings of teacher implementation) variables. Across curriculum components, the proportion of curriculum delivered ranged from .67 to .77 and the mean quality ratings ranged from 3.6 to 3.7 out of a possible rating of 5. In their review of implementation research in a broad range of community mental health and substance abuse prevention treatment studies, Durlak and DuPre (2008) noted that full implementation rarely occurred, with the typical level of implementation varying between 60% and 80%. The current study extends Durlak and DuPre’s findings by noting similar levels of implementation in a national early childhood curriculum research study.

The second research question addressed site differences in implementation. Site variation did occur in this study, although significant site differences were only found for the mean quality ratings. The IN and KS sites were significantly higher than the group mean on two of the three components of the curriculum. The MD site had significantly lower quality ratings than the group mean on one of the variables and marginally significantly lower ratings on another variable. A closer examination of the classroom variation for these three sites found a group of extremely low implementation classes at the MD site that negatively affected the site mean, but in the KS and IN sites there were not similar outlier classes that positively affected the site means. In an evaluation of an early reading curriculum noted previously, Zvoch et al. (2007) also examined individual classroom levels of implementation when attempting to interpret their group findings. In findings similar to ours, they found a small number of “low performing” classrooms that substantially affected group results. The implication of the findings of Zvoch et al. (2007) and the current study are that implementation researchers conducting cross site studies may well wish to examine closely individual classroom levels of implementation that may exert undue influence on site means.

The third research question addressed the calculation of a multiplicative composite measure that would incorporate information from both the structural and process measures. This multiplicative composite measure, as expected, was generally correlated more highly with its constituent measures than with other implementation measures and was not affected by site differences. To our knowledge, this is the first attempt to incorporate both structural and process measures within one multiplicative composite measure. The appeal of such a metric is that it may more directly represent the concept of dosage for curricula that have a specified set of content and clearly articulated procedures.

The fourth research question addressed the pattern of the process measure data across time. In a study comparing the effects of two early literacy curricula, Zvoch (2009) collected process implementation data at three time points, and his growth curve analysis generated intercept and trend differences across sites, with one set of classes showing negative trends (or decreasing implementation) across time. In the current study, our growth curve analysis documented intercept differences that had also been reflected in our analyses of mean differences in quality ratings across sites (i.e., KS and IN having significantly higher intercepts than other sites). Also, it documented that for most sites, the quality of implementation was relatively consistent across time, as indicated by a zero or flat trend. The exception was the CA site, in which there was a general pattern of increasing quality of implementation across time. The contributions of this analysis, as well as Zvoch’s (2009) study, are in the conclusions that implementation can be a dynamic variable, it should be assessed (and reported) across multiple time points, and the examination of variation across time may contribute to a better understanding of the implementation process.

The fifth research question addressed the association of structural, process, and the multiplicative composite implementation variables to child outcomes. Following a model comparison process, associations between different forms of implementation and child outcomes occurred. For the math component, both the quantity and multiplicative composite models respectively, generated the best fit for one math outcome variable. For the literacy component, the quality variable generated a significantly better fit for the PPVT outcome than the base model for children who performed low on the pretest. For the social component, the multiplicative composite and/or quality implementation variables generated a significantly better fit than the base model for four of the variables (i.e., two in interaction with pretest scores). From these findings, we propose that both the structural and process measures of implementation are important to include in preschool curriculum evaluation research because they appear to be individually associated with different types of child outcomes and they allow for the calculation of a multiplicative composite measure. The conceptually important features of the multiplicative composite measure is that it incorporates information about how much of the curriculum children receive with how well the curriculum is delivered in an intuitive and relatively easily calculated form.

Given the interaction that sometimes occurs between pretest performance and outcomes for children, with low performing children benefiting more from an intervention that middle or high performing children, we examined how pretest performance interacted with different forms of implementation in predicting child outcomes. For the literacy variable (i.e., PPVT) and the Wally, children who were lower performing at pretest benefited more from higher levels of implementation than children from the rest of the group. This occurred even after controlling for race/ethnicity, disability, and status as an English Language Learner. The pattern of low-performing children benefiting most from readiness curriculum is not unique (Boocock, 1995; Currie & Thomas, 1999; Ramey, Campbell, & Ramey, 1999), although rarely have there been studies that examined the interaction of implementation and relative performance of children.

The SSRS Internalizing variable further illustrates the complex associations of implementation and outcomes for children that sometimes occur. For this variable an interaction with pretest scores was detected, but the direction of the association was different for children scoring at the higher and lower regions of the pretest score distributions. For children with few internalizing behaviors at pretest, which was the majority of the participants, there was a positive association between quality and adjusted posttest scores, perhaps suggesting that the curriculum served as a protective factor for internalizing behavior for this group. For children with the most pronounced internalizing behavior (i.e., the lowest 10%), quality of implementation was inversely related to adjusted posttest scores, suggesting that a more intense or different intervention approach may have been needed for these children. In studies of curriculum implementation and child outcomes, these findings suggest the important of employing an analysis that may be sensitive to interaction effects that may occur for children performing a both ends of the performance distribution at pretest.

In this study, we also used bivariate correlations to portray the relationship between implementation measures and child outcomes, which confirmed some of the relationships found in the model comparison analysis. This correlational analysis also illustrates the problem of using traditional correlation matrix comparisons in complex multisite studies of implementation. In this study, implementation was a level 2 variable and child outcomes were level 1 variables. To calculate the correlations, we had to convert child outcomes to a Level 2 variable. We did this by first creating predicted posttest scores that incorporated the necessary controls for variables that might otherwise affect outcomes and then calculating mean-adjusted posttest scores per classroom, which served as the metric included in the correlation. We propose that following this process may result in associations sometimes being “lost in translation” (e.g., the interaction effects are not adequately represented) and that the model comparison process is a more sensitive approach to detecting differences in the association of implementation and child outcomes.

The purpose of this study was to examine different forms of implementation, and it revealed variation across classes and teachers. In a qualitative study that paralleled the current study, Lieber et al. (2009) examined the factors that influenced implementation of the CSS curriculum. From observation, interviews, and field notes collected during visits to the classrooms across the year, they identified teachers who were “high” and “low” implementers and then themes associated with implementation. From the qualitative analysis, the authors identified three primary sets of themes: 1) curriculum and instruction (i.e., the similarities of the previous curriculum approach used and CSS, teachers’ ability to manage the classroom, and integration/expansion of CSS concepts), 2) teacher (i.e., enthusiasm, partnership with CSS staff), and 3) beyond teacher (i.e., relationships among adults working in the classroom, administrators' support, events beyond the classroom, and receptivity to coaching). High implementers were identified as having a philosophy compatible with the CSS curriculum, being willing to engage in a partnership with the research staff, and being receptive to coaching. In contrast, low implementers had difficulty managing the classroom, were often not receptive to working with staff, and experienced difficulty among adults in the classroom. These factors contributed to the variations in implementation seen across classes.

Several limitations exist for this study. Because of the multi-site nature of this study, with sites being regionally located, it was not possible to collect inter-rater agreement data for the quality rating measures. Attempts to ensure consistent ratings across sites were supported by conference calls among raters, and the analyses of internal consistency of each rating measure were quite high and similar across site. Nevertheless, the absence of inter-rater agreement is a severe limitation of the study that has also been encountered by other investigators who have conducted cross-site curriculum evaluation research (Preschool Curriculum Evaluation Research Consortium, 2008; Zvoch, 2009; Zvoch et al., 2007).

As second limitation is that the associations between implementation and child outcome variables were identified for only a subset of the entire battery of child outcome measures collected in this study. The absence of associations with some child outcome variables may be interpreted in several ways. It may well have been some variables were more sensitive to the effects of the curriculum on children than were others. Alternatively, it is possible that some curriculum components, regardless of well they were implemented, did not affect child outcomes. The authors are now examining the overall efficacy of the CSS curriculum, which should provide information for the latter alternative explanation.

A third limitation is that there was not a logical child outcome variable that could be used to analyze the models of implementation of the science component of the CSS curriculum. The model developer had used the PPVT as an outcome measure (French, 2004), but it is a measure of receptive vocabulary and when the study began, no measures of preschool science knowledge or problem solving existed. The study would have been strengthened by the inclusion of a science child outcome measure. Last, the findings of this study are that the analyses of different forms of implementation assessment were confined to this one curriculum evaluation. Future research with other curriculum evaluations will be needed to determine the utility of the various approaches to implementation examined in this study.

In conclusion, this study contributes to the emerging literature on measurement of implementation in early childhood curriculum research by demonstrating two complementary approaches to assessing implementation, examining a novel approach to combining information from both measures, documenting the importance of the analysis of implementation data across sites and across time, and analyzing the association between different approaches to implementation assessment and child outcomes. Implementation science is emerging as a new and important dimension of applied early childhood research (Warren, Domitrovich, & Greenberg, 2009), and the questions that were once unacknowledged or viewed as subtle features of studies (e.g., the manner in which one assesses implementation) are emerging as prominent factors in curriculum evaluation research. Understanding the complexities of curriculum implementation in large, field-based studies may well inform the movement of research findings into practice in early childhood classrooms.

References

Adams, MJ.; Foorman, BR.; Lundberg, I.; Beeler, T. Phonemic awareness in young children: A classroom curriculum. Baltimore, MD: Brookes; 1998.

Al Otaiba S, Conner C, Lane H, Kosanovich ML, Schatschneider C, Dyrlund AK, Miller M, Wright TL. Reading First kindergarten classroom instruction and students' growth in phonological awareness and letter naming-decoding fluency. Journal of School Psychology. 2008; 46:281–314. [PubMed: 19083361]

Benedict EA, Horner RH, Squires JK. Assessment and implementation of positive behavior support in preschools. Topics in Early Childhood Special Education. 2007; 27:174–192.

Bierman KL, Domitrovich CE, Nix RL, Gest SD, Welsh JA, Greenberg MT, Blair C, Nelson KE, Gill, Sukhdeep G. Promoting academic and social-emotional school readiness: The Head Start REDI Program. Child Development. 2008; 79:1802–1817. [PubMed: 19037951]

Boocock S. Early childhood programs in other nations: Goals and outcomes. The Future of Children. 1995; 5(3):94–114. [PubMed: 8835516]

Briesch AM, Chafouleas SM, Lebel TJ. Impact of videotaped instruction in dialogic reading strategies: An investigation of caregiver implementation integrity. Journal of School Psychology. 2008; 45:978–988.

Chard DJ, Baker SK, Clarke B, Jungjohann K, Davis K, Smolkowski K. Preventing early mathematics difficulties: The feasibility of a rigorous kindergarten mathematics curriculum. Learning Disability Quarterly. 2008; 31:11–20.

Children's School Success. Children's School Success: An experimental study of an early childhood education curriculum model. Bloomington, IN: Indiana University; 2004 [Retrieved February 7, 2010]. from http://css.crlt.indiana.edu/index.html

Children's School Success Research Group. Children's School Success Curriculum (CSS). School of Education, Indiana University; 2009. Unpublished curriculum manuscript

Clements, DH.; Sarama, J. DLM Early Childhood Express math resource guide. Columbus, OH: SRA/McGraw-Hill; 2003.

Clements DH, Sarama J. Experimental evaluation of the effects of a research-based preschool mathematics curriculum. American Journal of Educational Research (45). 2008:443–494.

Copley, JV. The young child and mathematics. Washington, DC: National Association for the Education of Young Children; 2000.

Cost, Quality, and Child Outcomes Study Team. Cost, quality, and child outcomes in child care centers, Public Report. Denver: Economics Department, University of Colorado at Denver; 1995.

Currie J, Thomas D. Does Head Start help Hispanic children. Journal of Public Economics. 1999; 74:235–262.

Diamond KE, Gerde HK, Powell DR. Development in early literacy skills during the prekindergarten year in Head Start: Relations between growth in children's writing and understanding of letters. Early Childhood Research Quarterly. 2008; 23:467–478.

Dunn, LM.; Dunn, LM. Peabody Picture Vocabulary Test, Third Edition. Circle Pines, MN: American Guidance Service; 1997.

Durlak JA, DuPre EP. Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. American Journal of Community Psychology. 2008; 41:327–350. [PubMed: 18322790]

Farran, D. Another decade of intervention for children who are low-income or disabled; What do we know now?. In: Shonkoff, J.; Meisels, S., editors. Handbook of early childhood intervention (2nd Ed.). New York: Cambridge University Press; 2000. p. 510-548.

French L. Science as the center of a coherent, integrated early childhood curriculum. Early Childhood Research Quarterly. 2004; 19:138–149.

French, L.; Conezio, K.; Boynton, M. Using science as the hub of an integrated early childhood curriculum: The ScienceStart! Curriculum; Proceedings of the symposium in honor of Lilian G. Katz; Champaign, IL: ERIC; 2003. p. 303-312.ericeece.org/pubs/books/katzsympro.html

Gresham, FM.; Elliott, SN. Social Skills Rating System. Circle Pines, MN: American Guidance Service; 1990.

Halle, T.; Forry, N.; Hair, E.; Perper, K.; Wandner, L.; Wessel, J.; Vick, J. Disparities in early learning and development: Lessons from the Early Childhood Longitudinal Study – Birth Cohort (ECLS-B). Washington, DC: Child Trends; 2009.

Hill JL, Brooks-Gunn J, Waldfogel J. Sustained effects of high participation in an early intervention for low-birth-weight premature infants. Developmental Psychology. 2003; 39:730–744. [PubMed: 12859126]

Hox, JJ. Multilevel analysis: Techniques and applications. Mahwah, NJ: Lawrence Erlbaum Publishers; 2002.

Lonigan CJ. Development, assessment, and promotion of preliteracy skills. Early Education and Development. 2006; 17:91–114.

McConnell, SR. Picture naming/expressive language Individual Growth and Development Indicator. Minneapolis, MN: Early Childhood Research on Measuring Growth and Development, University of Minnesota; 2004.

McGrew, KS.; Woodcock, RW.; Mather, N. Woodcock-Johnson III. New York: Wiley; 2000.

Odom SL. The tie that binds: Evidence-based practice, implementation science, and early intervention. Topics in Early Childhood Special Education. 2009; 29:53–61.

Odom, SL.; Hanson, MJ.; Lieber, J.; Diamond, K.; Palmer, S.; Butera, G.; Horn, E. Prevention, early childhood intervention, and implementation science. In: Doll, B.; Pfhol, W.; Yoon, J., editors. Handbook of youth prevention science. New York: Routledge; in press

O'Donnell CL. Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. Review of Educational Research. 2008; 78:33–84.

Pence KL, Justice LM, Wiggins AK. Preschool teachers' fidelity in implementing a comprehensive language-rich curriculum. Language, Speech, and Hearing Services in Schools. 2008; 39:329–331.

Preacher KJ, Curran PJ, Bauer DJ. Computational tools for probing interaction effects in multiple linear regression, multilevel modeling, and latent curve analysis. Journal of Educational and Behavioral Statistics. 2006; 31:437–448.

Preschool Curriculum Evaluation Research Consortium. 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. Washington, DC: U.S.Government Printing Office; 2008.

Ramey CT, Campbell FA, Ramey SL. Early intervention: Successful pathways to improving intellectual development. Developmental Neuropsychology. 1999; 16:385–392.

Rudd LC, Lambert MC, Satterwhite M, Smith CH. Professional development = coaching = Enhanced teaching: Increasing usage of math mediated language in preschool classrooms. Early Childhood Research Journal. 2009; 37:63–69.

Snijders, TAB.; Bosker, RJ. Multilevel analysis: An introduction to basic and advanced multilevel modeling. London etc.: Sage Publications; 1999.

Warren, HK.; Domitrovich, CE.; Greenberg, MT. Implementation quality in school-based research: Roles for the prevention researcher. In: Dinella, L., editor. Conducting science-based psychology research in schools. Washington, D.C: 2009. p. 225-249.129–151

Webster-Stratton, C. Wally Game: A problem-solving skills test. University of Washington; 1990. Unpublished manuscript

Webster-Stratton C. The Incredible Years Training Series. Juvenile Justice Bulletin, June, 2000. 2000:1–23.

Webster-Stratton C, Reid MJ, Stoolmiller M. Preventing conduct problems and improving school readiness: Evaluation of the Incredible Years Teacher and Child Training Program in high-risk schools. Journal of Child Psychology and Psychiatry. 2008; 49:471–488. [PubMed: 18221346]

Whitehurst GJ, Epstein JN, Angell AL, Payne AC, Crone DA, Fischel JE. Outcomes of an emergent literacy intervention in Head Start. Journal of Educational Psychology. 1994; 86:542–555.

Williams CL, Carter BJ, Kibbe DL, Dennison D. Increasing physical activity in preschool: A pilot study to evaluate. Animal Trackers. Journal of Nutrition Education and Behavior. 2009; 41:47–56.

Winsler A, Tran H, Hartman SC, Madigan AL, Manfra L, Bleiker C. School readiness gains made by ethnically diverse children in poverty attending center-based child care and public school pre-kindergarten programs. Early Childhood Research Quarterly. 2008; 23:314–329.

Woodcock, RW.; McGrew, KS.; Mather, N. Woodcock-Johnson III-Tests of Achievement. Itasca, NY: Riverside Publishing; 2001.

Zill, N.; Resnick, G. Emergent literacy of low-income children in Head Start: Relationships with child and family characteristics, program factors, and classroom quality. In: Dickinson, DK.; Neuman, SB., editors. Handbook of early literacy research. Vol. Vol. 2. New York: Guilford; 2006. p. 347-371.

Zvoch K. Treatment fidelity in multisite evaluation: A multilevel longitudinal examination of provider adherence status and change. American Journal of Evaluation. 2009; 30:44–61.

Zvoch K, Letourneau LE, Parker RP. A multilevel multisite outcome-by-implementation evaluation of an early childhood literacy model. American Journal of Evaluation. 2007; 28:132–150.

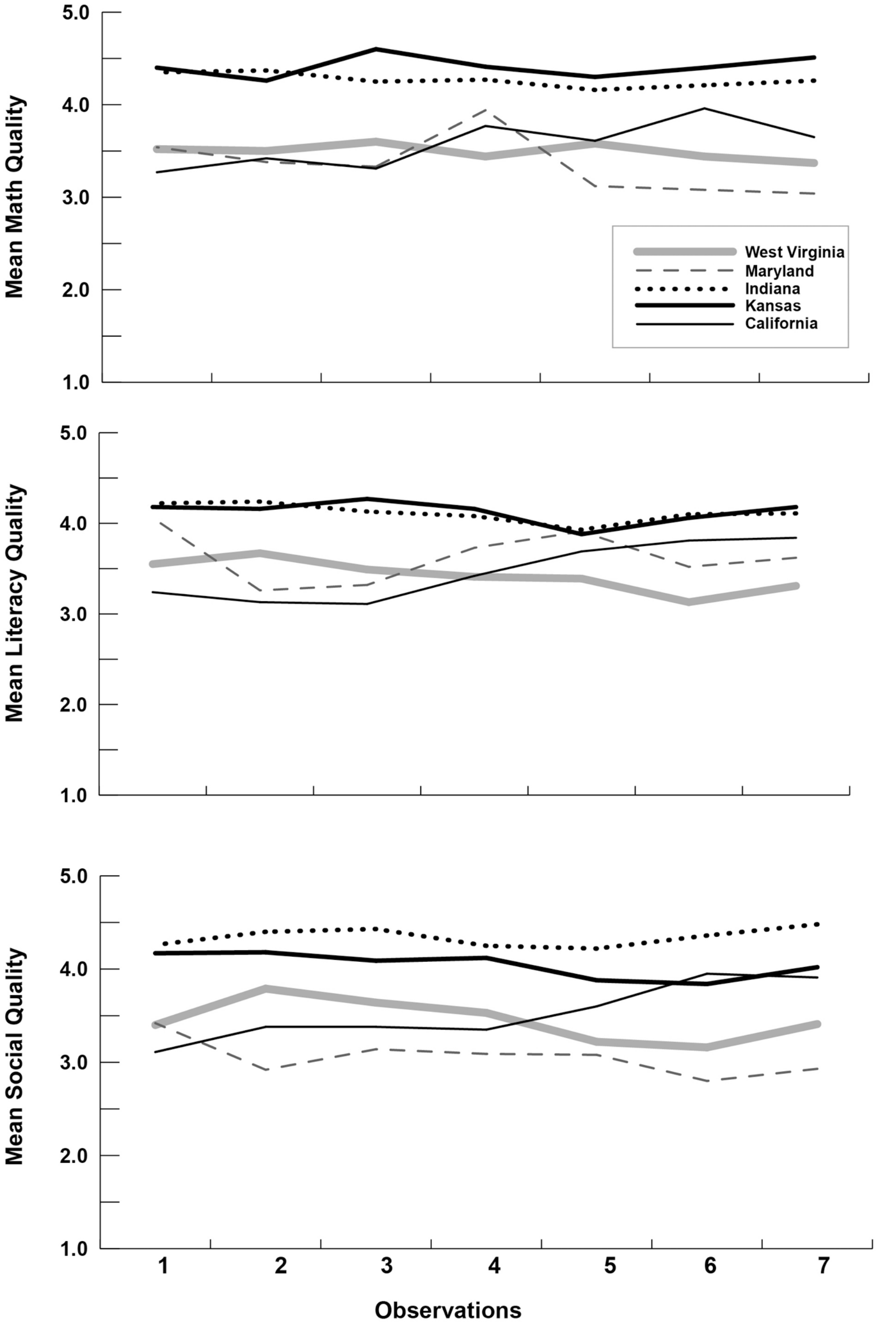

Figure 1.

Mean Quality of Implementation Across Time and Sites for Math Component of the CSS Curriculum

Table 1

Demographic Table

	CA	KS	IN	WV	MD	Total
Type of Program						
HS	10	0	10	11	9	40
PreK	0	5	0	0	1	6
Private	0	1	0	0	0	1
Other	0	4	0	0	0	4
Classroom Information						
Full day	3	1	0	10	1	15
Half day	7	9	10	1	9	36
Mean adults	2.4	2.3	1.9	2.4	2.1	2.2
per class	(1.3)	(1.4)	(.3)	(.5)	(.6)	(.9)
Mean children	20.0	18.0	18.3	12.2	17.9	18.25
per class	(0)	(7.15)	(1.4)	(.8)	(2.5)	(3.42)
Mean children	11.0	12.2	10.5	12.1	10.6	11.3
in study per class	(2.1)	(2.6)	(1.9)	(4.4)	(3.3)	(3.0)
Child Demographic Information						
Mean age in	50	54	55	54	53	53
months	(4.0)	(4.0)	(4.0)	(4.0)	(4.0)	(4.0)
Gender						
Boys	59	75	57	72	48	315
Girls	46	47	48	59	58	261
Child Study Status						
Low-income	20	90	87	108	82	387
Disabilities	8	28	10	24	16	86
ELL	87	6	9	1	8	111
Race/Ethnicity						
Caucasian	18	80	80	116	41	335
African American	7	25	4	5	48	89

	CA	KS	IN	WV	MD	Total
Latino/a	68	13	17	5	12	115
Other	17	4	4	7	5	37

Table 2

Mean Proportion of Curriculum Completed, Mean Rating of Quality of Implementation, and Mean Multiplicative Composite Scores by Site (*sd* in parenthesis)

	MD	WV	IN	KS	CA	Total
Proportion of	62.3	72.6	75.0	82.6	74.0	73.3
literacy completed	(29.6)	(23.5)	(14.8)	(13.1)	(12.0)	(20.1)
Proportion of	55.0	72.3	75.0	69.5	62.0	66.9
math completed	(36.2)	(25.3)	(18.6)	(22.8)	(25.8)	(26.2)
Proportion of	58.5	75.0	83.0	77.8	76.0	74.1
social completed	(35.2)	(20.6)	(11.1)	(19.8)	(13.1)	(22.4)
Quality rating	3.6	3.4	4.1[*]	4.1[*]	3.5	3.7
literacy[X]	(.74)	(.61)	(.53)	(.40)	(.55)	(.63)
Quality rating	3.3[*]	3.5	4.2	4.4[*]	3.6	3.8
math[X]	(1.43)	(.84)	(.71)	(.36)	(.49)	(.93)
Quality rating	3.1[*]	3.5	4.3[*]	4.0	3.54	3.67
social[X]	(1.18)	(.56)	(.63)	(.44)	(.60)	(.82)
Multi-Composite	2.4	2.6	3.1	3.4	2.6	2.8
literacy	(1.48)	(1.16)	(.91)	(.79)	(.80)	(1.09)
Multi-Composite	2.2	2.7	3.2	3.0	2.6	2.8
math	(1.80)	(1.41)	(1.19)	(1.04)	(.80)	(1.09)
Multi-Composite	2.1	2.7	3.6	3.2	2.7	2.9
social	(1.58)	(1.06)	(.94)	(1.03)	(.86)	(1.19)

[X] Main effect for sites $p < .0167$

[*] Planned comparisons revealed site M significantly different ($p < .05$) from Grand *M*

Multi-Composite = Multiplicative Composite

Table 3

Correlations Among Fidelity of Implementation Measures

	Social Quality	Social Quantity	Social Multi-Composite	Literacy Quality	Literacy Quantity	Literacy Multi-Composite	Math Quality	Math Quantity
Social Quantity	.761**							
Social Multi-Composite	.925**	.923**						
Literacy Quality	.843**	.683**	.829**					
Literacy Quantity	.788**	.828**	.847**	.755**				
Literacy Multi-Composite	.853**	.801**	.892**	.911**	.949**			
Math Quality	.903**	.715**	.840**	.866**	.792**	.863**		
Math Quantity	.697**	.802**	.799**	.727**	.844**	.839**	.726**	
Math Multi-Composite	.812**	.776**	.864**	.835**	.864**	.911**	.858**	.949**

** p < .01

Multi-Composite = Multiplicative Composite

Table 4

Implementation Variables and Child Outcomes

Math	− 2 Log Likelihood	Deviance$^{\#}$	AIC	Parameter Est.	
WJ 18a Quantitative Concepts				(SE)	
Base Model	1729.1		1753.1		
Quantity	1741.9	12.8^{***}	1742.3	$0.13\ (.004)^{***}$	
Quality	1736.8	7.7^{**}	1747.4	$.329\ (.12)^{**}$	
Multi-Composite	1740.8	11.7^{***}	1743.4	$.247\ (.07)^{**}$	
Literacy	−2 Log Likelihood	Deviance$^{\#}$	AIC	Parameter Est. (SE)	
PPVT (Interaction with Pretest)					
Base Model	2750.8		2774.8		
Quantity × Pretest	2753.4	2.6	2776.2	.019 (.031)	−.002 (.001)
Quality × Pretest	2758.4	7.6^{*}	2771.3	1.08 (1.10)	$-.115\ (.04)^{**}$
Multi-Composite × Pretest	2756.0	5.2	2773.6	.58 (.59)	$-.055\ (.03)^{*}$
Social	−2 Log Likelihood	Deviance$^{\#}$	AIC	Parameter Est.	
SSRS – Self Regulation				(SE)	
Base Model	235.1		259.1		
Quantity	241.8	6.7^{**}	254.4	$.003\ (.001)^{*}$	
Quality	242.5	7.4^{**}	253.7	$.096\ (.04)^{*}$	
Multi-Composite	244.7	9.6^{**}	251.5	$.071\ (.02)^{**}$	
Social	−2 Log Likelihood	Deviance$^{\#}$	AIC	Parameter Est.	
SSRS - Survival Skills				(SE)	
Base Model	177.6		201.6		
Quantity	186.3	8.7^{**}	197.4	$.003\ (.001)^{**}$	
Quality	183.8	6.2^{*}	194.9	$.080\ (.03)^{*}$	
Multi-Composite	188.1	10.5^{***}	193.1	$.067\ (.02)^{**}$	

Social	−2 Log Likelihood	Deviance#	AIC	Parameter Est. (SE)	
SSRS PB Internalizing (Interaction with Pretest)					
Base Model	358.3		382.3		
Quantity × Pretest	371.7	13.4**	372.9	.001 (.001)	.008 (.002)**
Quality × Pretest	331.0	27.3***	359.1	.041 (.04)	.266 (.052)**
Multi-Composite × Pretest	381.3	23.0***	363.3	.014 (.025)	.177 (.037)**

Social	−2 Log Likelihood	Deviance #	AIC	Parameter Est. (SE)	
Wally Total Positive (Interaction with Pretest)					
Base Model	2105.4		2127.4		
Quantity × Pretest	2106.5	1.1	2130.3	−.001 (.006)	−.002 (.002)
Quality × Pretest	2111.3	5.9*	2125.5	.045 (.16)	−.117 (.05) *
Multi-Composite × Pretest	2109.1	3.7	2127.7	−.013 (.11)	−.063 (.03) *

#Deviance Test is a 2 degree of freedom test with main effect and interaction added

* $p < .05$

** $p < .01$

*** $p < .001$

Multi-Composite = Multiplicative Composite

Table 5

Correlation between Predicted Posttest Scores Aggregated by Class and Implementation Measures

Measure	PPVT	WJ-18A	SSRS Self Reg	SSRS Classroom Survival Skills	SSRS PB-Internal	WALLY Positive Responses
Lit-Quantity	.10					
Lit-Quality	.25					
Lit-Comp	.20					
Math-Quantity		.20				
Math-Quality		.24				
Math Multi-Composite		.26				
Social Quantity			.35**	.38**	.02	.12
Social Quality			.28*	.22	.10	.35**
Social Multi-Composite			.34*	.32*	.09	.25

* $p< .05$

** $p < .01$

Multi-Composite = Multiplicative Composite