MAKING SENSE OF PRESCHOOL RESEARCH: A MULTI-PAPER DISSERTATION ON THE IMPLEMENTATION AND EFFECTIVENESS OF PRESCHOOL CURRICULUM INTERVENTIONS

By

CATHERINE L. DARROW

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Learning, Teaching and Diversity

August, 2010

Nashville, Tennessee

Approved:

Professor David K. Dickinson

Professor Dale C. Farran

Professor Carin Neitzel

Professor David Cordray

To Noelle who put up with it all

ACKNOWLEDGEMENTS

This dissertation and the work leading up to it would not have been possible without the support of many individuals and organizations. I have been fortunate to benefit from the financial support of the Experimental Education Research Training Program (ExpERT), funded through the Institute of Education Sciences (IES), and Peabody College’s Department of Teaching, Learning, and Diversity.

I am particularly grateful to the people who made this possible. I owe a special thanks to Dr. David Dickinson, Dr. Dale Farran, Dr. David Cordray, Dr. Carin Neitzel, and Dr. Mark Lipsey for their encouragement, guidance, and patience. I have learned to be a more conscientious student, a more effective writer, and a more critical researcher by working with these professors.

Certainly, the overall experience would have been less enjoyable and meaningful without my family, classmates, and friends.   First, I would like to thank Tanya Flushman and Jill Freiberg for being there and laughing with me whenever possible. Also, I cannot communicate how helpful Dr. Kerry Hofer was throughout the program as she served the important role of friend, confidant, and mentor statistician. And lastly, I would never have even made it to Tennessee or stayed long enough to finish without Noelle. Noelle, it’s your turn.

TABLE OF CONTENTS

Page

Appendix

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Preface

The following multi-paper dissertation addresses issues of effectiveness and implementation inherent in preschool curriculum interventions. This document consists of an overview of challenges faced by educational researchers and issues addressed by the three separate papers that follow. These papers include: (1) a meta-analysis examining the effectiveness of preschool curriculum interventions; (2) an analysis of fidelity instruments used in preschool studies; and (3) a report illustrating the development and use of fidelity and quality measures.

Preschool as an Educational Intervention for Children from Poor Families

Improving the lives of children through education has been a long-standing goal of the United States. Researchers, policy makers, and practitioners have focused on the quality of education offered to children and young adults for decades. This focus on effectiveness has increased in recent years. With the legislation of No Child Left Behind in 2001, both federal and state governments have placed teaching quality and student learning in positions of high priority. Moreover, greater attention to the ways in which young children, in particular, have been prepared for school has come to the forefront of educational research. A focus on early school preparedness introduces issues related to how well programs prepare children and in which type of skills, how effectively teachers deliver newly developed interventions aimed at improving those skills, and whether any particular curriculum is more effectively delivered than what a teacher would do on his or her own.

Children from poor families have historically entered formal schooling with lower language skills than their more economically advantaged peers. The Family and Child Experiences Survey (FACES) of 2006 reported that children were entering Head Start far below the national levels. On average, children from poor families were one standard deviation below their peers in receptive vocabulary and anywhere from one-third to two-thirds of a standard deviation below in letter-word identification, early writing, and applied problems (Tarullo, West, Aikens & Hulsey, 2008). As a result of this persistent problem, many preschool programs similar to Head Start have attempted to erase this pattern of low achievement.

The need for a larger number of preschool programs has increased over the years as more women enter to the workforce as the sole wage-earning adult in the family. Likewise, a shift has also occurred in that workingwomen often share the responsibility of providing care and education experiences of their children with preschool teachers and other childcare staff. This trend started as early as the 1960s when an increasing number of women sought part- and full-time employment. Thus, more families were in need of services that provided secure childcare and early educational experiences. By 1993, the mothers of 10 million children were employed, and 47% of their children were enrolled in either center- or family-based childcare programs (Hofferth, 1996). The rate of attendance in preschool has been steadily increasing over the past few years, reaching as high as 80% in 2008 (National Institute for Early Education Research [NIEER], 2008).

Given the frequency with which families rely on others to provide care, preschool programs and other childcare services in conjunction with parents are now responsible for enriching the lives of young children.

Competing Goals of Early Preschool Programs

Early childcare services for families have been prevalent in Europe even before the turn of the last century. Issues of childcare, however, have been complicated in the United States by the government's hesitancy to participate in caring for young children - historically deemed the right and responsibility of parents, and mothers in particular (Condry, 1983; Scarr, 1986). However, as more women entered the workforce, alternative views arose over who should be involved in providing care, safety, nurturance, and rich early experiences to children while mothers are working.

In addition to the economic need for preschool programs, issues around equity and access emerged in the early 1960's. The education research community increased its focus on issues of achievement and school readiness for children from poor families. Condry (1983) attributes the increased focus to two events: 1) the recently accepted view that intelligence and achievement were not fixed traits, but modifiable, often influenced by environmental factors and 2) the recent research depicting differential levels of achievement among ethnic groups. President Johnson endeavored to address this discrepancy by funding Project Head Start through his War on Poverty in 1965. Head Start was originally established to provide poor children with health and nutritional services (Beatty, 1995) and subsequently to prepare poor children for school and thereby attempting to change the downward trend of school failure (U.S. Department of Health and Human Services, Administration for Children and Families, 2010).

The goals of childcare and preschool programs have always varied, and the priority of one set of goals over another has historically been a hotly debated topic. Well before the establishment of Head Start and subsequent state-funded preschool services, early childhood education fell into three types of programs: kindergartens, day nurseries (i.e., the modern day version of childcare services), and nursery schools. Despite some differences in the original goals of these three programs, they all stressed social and emotional development in children (Goffin, 1994). The methods by which programs attempted to nurture children were at times divergent, however. Some programs viewed early childhood programs as a source of enrichment while others looked at these programs as a means to prepare children for school (Condry, 1983; Vinovskis, 1993).

Programs that stressed cognitive development increased in number in the 1960s as issues of school preparedness in poor children came to the forefront. Studies like the report by Coleman (1968) provided evidence that clear variations in achievement among ethnic and racial groups existed in the K-12 systems. The Coleman report coupled with an increasing acceptance of Piagetian claims that intelligence is not fixed but rather modifiable, gave rise to the idea that early educational experiences could lead to academic improvement in young children from poor families (The Consortium for Longitudinal Studies, 1983). The political and social climate of the 1960s provided an ideal setting for the rise of early intervention educational programs ranging in size from programs as large as Head Start and to smaller community-based services.

The perspective that early childhood programs could and should address the academic needs of young children regained momentum in the 1980s. At that time, many scholars and educators working in the field of early childhood education championed Vygotsky's views that cognitive function in children is influenced by social and cultural experiences (Davydov, V. V., & Kerr, S. T., 1995; Ramey, C. et al., 2000). Many scholars looked to preschool programs to provide such experiences. In addition, preschool programs were more often assessed on how well they provided rich, positive educational environments using newly developed observation tools like the Early Childhood Environment Rating Scale (ECERS) (Hyson, 1991). Moreover, the ongoing rise of working middle-class mothers and educational reform in the early childhood arena amplified interest in the academic contribution of preschool programs (Goffin, 1994; Vinovskis, 1993). By the end of the 1980s, however, research emerged reporting mixed effects of early childhood programs. Studies found that the type of preschool program and the curriculum used influenced children's social-emotional development both while enrolled in a program (DeVries & Goncu, 1987) as well as behavior and engagement in later years (Schweinhart, Weikart, & Larner, 1986). Although the effect of preschool programs on children's academic gains was still in question by the end of the 1980s, concern continued over the additional contributions of particular curricula and philosophies of learning.

The Changing Role of Curriculum

In the 1990s and 2000s two major emphases materialized further. Preschool programs like Head Start began focusing even more on academic benefits obtained by children when enrolled in preschool. The focus on academic outcomes was particularly evident when the Advisory Panel for the Head Start Evaluation Design Project clearly recommended that research studies measure short- and long-term effects on child outcomes (Collins, 1990). Additionally, many programs viewed the adoption of curricula as a necessary step in ensuring the delivery of high quality educational experiences to young children. In 2002, the National Association for the Education of Young Children (NAEYC) issued a position paper stating that the implementation of an evidence-based, well-developed curriculum would promote academic growth in children (NAEYC, 2002).

Today, many states require preschool programs to adopt a specific curriculum as a way to ensure higher levels of quality in instruction and as a means to meet or exceed the state standards. In some states, like North Carolina, preschool programs must select from a list of curricula deemed acceptable by the state (Office of School Readiness, 2008). The use of a curriculum is not only mandatory, but the number of approved curricula is also small. Alternately, many states are similar to Illinois, which mandates that preschool programs employ a curriculum geared toward young children without endorsing a specific one (Illinois State Board of Education, 2009). As evident through examples of statewide endorsement of published curricula, there are expectations and assumptions that commercially available curricular packages are key factors in delivering high-quality educational experiences to young children.

A consistent focus throughout many state standards has been language and literacy development in preschool and the early grades. As an example, the Connecticut Preschool Framework contains 15 indicators of performance standards directly related to language and literacy development (Connecticut State Department of Education, 2006).

The indicators range in topic and scope from "demonstrate understanding of basic conversational vocabulary" to "use symbols or drawings to express thoughts, feelings and ideas" (Dowaliby, 2006). With language and literacy development becoming an increasing priority, published curricula have been adopted and implemented by many preschool programs to increase children's academic skills, with a major focus on the areas of language and literacy.

Growing numbers of curricula. Over the past few decades, commercially published preschool curricula have increased in number and accessibility. The range and number are evident in the collection of 77 early childhood interventions inventoried by the What Works Clearinghouse (WWC; U.S. Department of Education, 2009). Among these intervention reports, 19 include eligible studies that measure the effects of a particular curriculum on children's academic and social development; 13 commercially available curricula are represented in these reports. This means that when preschool administrators and practitioners utilize WWC intervention reports as a way to inform their decisions around which curriculum will best support the children in their programs, they have at least 13 published curricular packages from which to choose.

The expanding quantity of preschool curricula is also evident in the material included in the curriculum packages made available by individual publishers. Many preschool programs continue to use instructional packages that are developed by the teachers themselves who work in preschool classrooms. However, preschool administrators and practitioners also have the convenience of purchasing a curriculum online through established publishing companies. Specifically, Pearson, one of the largest international publishing companies, offers three preschool curricula: *Opening the World of Learning*^{TM} (OWL), *Sing, Spell, Read & Write© 2004*, and *Waterford Early*

Learning^{TM}. In addition to promoting their own curriculum entitled *Houghton Mifflin PRE-K©: Where Bright Futures Begin!,* the publishers of Houghton Mifflin Harcourt also offer two additional preschool curricular packages. The number and varying foci of available preschool curricula are large and steadily increasing.

Trends in the adoption and implementation of specific curricula. As a larger number of curricula surface, the variety of curricular packages adopted by program administrators and implemented by preschool teachers has also increased. The diversity of curricula used in preschool programs is evident in the curriculum adopted by Head Start programs over time. In the late 1990s and early 2000s, the majority of Head Start programs used either *High/Scope Curriculum* or *The Creative Curriculum ® for Preschool* (Shaul et al., 2003). By 2003, *High/Scope* and *The Creative Curriculum* continued to be widely used, yet over 40 percent of teachers in the Head Start Family and Child Experiences Survey (FACES) reported using an alternative curriculum. Other preschool programs, like those funded by states, are often required to adopt a curriculum. In these cases, program directors are faced with an assortment of curricula from which to choose.

Evaluating Effectiveness of Curricula

The emphasis on academic gains in children by means of easily defined, easily replicable curricula continues to grow. Furthermore, interest in measuring the additive contributions of particular preschool curricula on developmental outcomes of preschool children has also increased in recent years (Schweinhart & Weikart, 1997, 1998). In the past decade, several reviews of interventions that implement commercially available curricula have been published and are easily accessible by researchers and practitioners working with preschools.

The WWC provides an assessment of selected curricula, many of which focus on language and literacy development in preschool children. The results of these assessments vary from showing significantly positive results in the *Literacy Express* report (2007) to inconclusive effects in the *Curiosity Corner* report (2009). *The Creative Curriculum* and *High/Scope* are widely used throughout Head Start and other preschool programs. The WWC has produced an intervention report for *The Creative Curriculum* that includes results from three primary studies. Overall, the report concluded that children receiving *The Creative Curriculum* experienced no substantive gains in oral language and print knowledge or any other measured outcome. The focus and the accessibility of the intervention reports provided by WWC provides evidence that concern over the effectiveness of curriculum in facilitating school readiness in preschool children is widespread.

Language and literacy development have increasingly become the dominant focus of preschool intervention. This emphasis can be seen in the number of commercially available curricula that primarily contain methods and materials used to teach children language and literacy skills as well as being the focus of federally funded preschool interventions. Of the 14 curricula implemented in the 15 Preschool Curriculum Effectiveness Research (PCER) studies funded by the Department of Education’s Institute of Education Sciences (IES), 10 focused primarily on language and literacy development (National Center for Education Research [NCER], 2008). Questions continue to exist regarding both the effectiveness of curriculum interventions for different groups of children and which components of these preschool programs and interventions produce the most substantial gains in language- and literacy-related outcomes in particular. As evident from the WWC reviews and the recent NCER report, the effectiveness of individual preschool curricula targeting language and literacy development, among other competencies, varies dramatically.

Are Preschool Curriculum Interventions Effective in Improving Children's Language and Literacy Development Overall?

Given that so much attention and funding has been devoted to identifying curricular packages that bring about the greatest positive gains in preschool children, it is necessary to step back and examine the overall effect of published curricula. Research has shown that when examined individually specific interventions have produced mixed results, yet more attention is warranted to determine how well these efforts might be when examined collectively.

Examination of the aggregate contribution of preschool curricula in increasing children's readiness for school, especially as it relates to the language and literacy skills, is missing. Many programs that receive state and federal funding look at well-developed curricular packages as a way to deliver high-quality instruction to young children (Dodge, 2004). Preschool curricula are portrayed as the necessary element in programs geared toward making children ready for school through better-developed language and literacy skills.

Given the considerable number of studies reporting curricular effects on children's development, it is now possible to determine the overall effectiveness of curriculum interventions. By analyzing this vast collection of studies, it is also now possible to identify the specific types of curricula that have more effects on language and literacy outcomes, as well as which types of programs benefit most from curriculum implementation, and which gains in language and literacy are made by children from diverse backgrounds. The first of three papers to follow addresses questions related to the effectiveness of preschool curricula though a meta-analytic analysis specifically examining the language and literacy development of preschool children.

Are Issues of Effectiveness Associated with Fidelity of Implementation?

One important issue in intervention research is fidelity of implementation. Researchers should not assume that teachers are delivering critical elements of a curriculum with requisite frequency and quality. One difficulty with determining fidelity is that developers have not always identified the critical elements of the implemented curriculum that teachers must deliver faithfully, nor have they evaluated teachers' levels of implementation through the use of well-developed, precise measures of these critical elements.

Over the years, many scholars working in the field of early childhood education have determined that poor children who are enrolled in preschool programs have a higher likelihood of school success than children who receive no formal care. In a study measuring the effects of Oklahoma's Universal Pre-K program, Gormley, Gayer, Phillips, & Dawson, B. (2005) found that children from diverse backgrounds benefited from attending preschool and were better prepared to enter school as evident from their higher achievement scores. Preschool programs like those included in the Oklahoma study and also those that receive state and federal funding frequently adopt curricula as a means to increase academic skills in the children they serve. Unlike the encouraging findings that preschool attendance has positive effects, recent reports of preschool curriculum interventions targeting children already enrolled in preschools have revealed that few studies have found additive, positive effects in children coming from specific curricular packages (Barnett, 2008; NCER, 2008). This lack of success could be due to the ineffectiveness of the curricula tested, but it is also possible that teachers did not fully implement the curriculum under scrutiny. These two very opposed interpretations of the effectiveness of a curriculum make it essential to determine first if the curriculum was implemented before determining its success or failure.

There are a number of reasons for measuring the fidelity of implementation within any intervention. First, fidelity is one of several essential measures used to confirm that outcomes of interest could indeed be related to the delivery of a treatment. In other words, if researchers cannot confirm that essential elements of an intervention were indeed delivered as intended, then positive outcomes cannot be attributed to the treatment. Gresham, Gansle, and Noell (1993) claimed that evidence of high levels of fidelity is a necessary but not sufficient ingredient in establishing the relationship between modifiable independent variables and the resulting dependent variable. Second, by using fidelity instruments to calculate levels of implementation, and to compare treatment classrooms with those serving as comparisons on the same measures, researchers are able to determine the achieved relative strength of an intervention (Hulleman & Cordray, 2009). Knowing the strength of an intervention as it was actually delivered (e.g., how different the treatment was from the comparison), researchers can better understand how varying levels of implementation can help interpret intervention effectiveness. Thus, accurate measures of implementation may help to explain why an intervention did or did not produce the expected results.

In particular, many preschool curriculum interventions have shown no effects on children's language and literacy development. For example, one of the PCER studies implemented the curriculum *Doors to Discovery* in treatment classrooms. Children in the treatment classrooms receiving *Doors to Discovery*, however, did not experience significant positive gains in targeted outcomes as compared to children assigned to the control classrooms (NCER, 2008). One possible reason for this may be that the interventions are not being implemented to the degree that the curriculum developer originally intended. Alternatively, lack of effects may have occurred because teachers adequately implemented a curriculum that was ineffective at increasing children's language and literacy skills. On the other hand, a lack of effects could have resulted from less than ideal levels of implementation in that teachers failed to deliver the critical elements of the curriculum necessary to increase children's skills.

How Different are Interventions than Business as Usual?

Precise measure of fidelity used in conjunction with well-defined curriculum interventions can be employed to highlight differences between the treatment condition that received the curriculum and the counterfactual where teachers engaged in their usual instructional practices. Effective measures of fidelity should include items that assess teachers' delivery of critical elements of the curriculum hypothesized to positively influence children's development. All teachers, across conditions, therefore should be evaluated on the degree to which and the quality with which they deliver these items. If this were done, researchers and curriculum developers would be able to identify patterns of implementation and to associate specific elements of the curriculum with child outcomes. Without confirming that particular practices were implemented, researchers are not able to attribute child gains to curricular and instructional elements. On the other hand, fidelity measures may also indicate that experimental teachers and comparison teachers are providing similar instruction. In this case, any gains experienced by the children in these classrooms cannot be directly linked to unique practices of the experimental teachers since differentiation in the instructional practices of teachers is not evident.

Reviews of intervention studies suggest that it is rare to use measures of fidelity to differentiate the instruction being delivered in classrooms assigned to different conditions. Dane and Schneider's review (1998) highlights this infrequency by reporting that only 6 percent of the total sample of studies included aspects program differentiation. In addition to Dane and Schneider's review of school-based prevention programs, a review of drug abuse prevention research in school settings by Dusenbury, Brannigan, Falco, and Hansen (2003) found no representation of program differentiation among the studies they included.

Historically, researchers and practitioners working in early childhood education have focused on instructional practices that they believe will provide children with the best opportunity to develop appropriate skills and behaviors. Research has suggested that meaningful and effective instruction delivered by teachers for children preparing to enter school must be rooted in developmentally appropriate practice (Epstein, Schweinhart, &

McAdoo, 1996). Consequently, many preschool teachers share a common goal in that they want to provide early educational experiences that facilitate children's learning while being age-appropriate.

Given these commonalities in many preschool programs, it becomes important for intervention researchers who ask teachers to implement new curricula to be able to identify unique elements of those curricula that differ from typical developmentally appropriate practices. Therefore, when a new curriculum is adopted and the unique elements are identified, researchers must be able to assess the degree to which those elements are present in both the treatment and comparison classrooms.

The second and third papers to follow address issues related to the effectiveness of fidelity measures in identifying and assessing the degree to which teachers implement unique, causal elements of a curriculum. Specifically, the second paper looks critically at how effectively fidelity instruments used in preschool research represent unique elements of the associated curriculum. The third and final paper in this collection illustrates how reliable measures of implementation fidelity were developed and the ways in which these measures were used to understand the complex, multi-dimensional nature of fidelity.

Understanding and Making Progress in Preschool Research

Researchers, policy makers, and practitioners continue to view children's attendance in preschool programs as a way to increase school readiness. Moreover, many preschool programs are adopting published curricula to bring about high quality instruction. Therefore, it is paramount that the educational community understand which programs and curricula are most effective in increasing child outcomes, how well teachers implement new curriculum deemed to be valuable, and the degree to which newly implemented published curriculum produce better effects on children's learning than the instructional practices typically used by preschool teachers. The collection of papers that follows examines issues of effectiveness, implementation, and program differentiation in preschool research. Specifically, this collection of three manuscripts addresses three primary research questions:

1. Are preschool curriculum interventions effective in improving children's language and literacy development?
2. How effectively do curriculum developers and intervention researchers use measures of fidelity to assess the implementation of critical curricular components?
3. How can measures of fidelity be developed and applied to most effectively represent critical components of a curriculum intervention?

CHAPTER II

LANGUAGE AND LITERACY EFFECTS OF CURRICULUM INTERVENTIONS FOR PRESCHOOL SERVING ECONOMICALLY DISADVANTAGED CHILDREN: A META-ANALYSIS

Background

At a time when teacher quality and student achievement are central issues in education, more effort and energy is being directed to identifying aspects of schools and instruction that are most effective in increasing children's academic development. This concern has made its way to the youngest members of our society. In order to improve early childhood education with the aim of better preparing young children to enter school, many states have developed standards of learning in early childhood education. By 2002, more than 25 states had created standards for early childhood education (National Association for the Education of Young Children [NAEYC], 2002) and all but a few states had developed standards by 2007 (Scott-Little, Lesko, Martella & Milburn, 2007).

Many preschool programs, with guidance from the state, have adopted a curriculum as a way to ensure higher levels of quality in instruction and as a means to consistently meet or exceed the state standards. In some states, like North Carolina, preschool programs must select from a list of curricula deemed acceptable by the state. The use of a curriculum is not only mandatory, but the number of approved curricula is also small. Alternately, many states are similar to Illinois, which mandates that preschool programs employ a curriculum geared toward young children, without endorsing a specific one. As evident through examples of statewide endorsement of published curriculum, there are expectations and assumptions that commercially available curricular packages are key factors in delivering high-quality educational experiences to young children.

A consistent focus throughout many state standards has been language and literacy development in preschool and the early grades. As an example, the Connecticut Preschool Framework contains 15 indicators of performance standards directly related to language and literacy development. The indicators range in topic and scope from "demonstrate understanding of basic conversational vocabulary" to "use symbols or drawings to express thoughts, feelings and ideas" (Dowaliby, 2006). With language and literacy development becoming an increasing priority, published curricula have been adopted and implemented by many preschool programs to increase children's academic skills, especially in the areas of language and literacy.

The prevalence of language- and literacy-focused curricula is evident in the recent publication by NCER reporting the effects of several preschool curriculum interventions (NCER, 2008). Of the 15 separate curriculum evaluations included in the report, nine examined the effects of language- and literacy-based curricula. Only one evaluated curriculum targeted mathematical development. Evidence of this focus goes beyond the field of research and is also apparent in statewide policies. Of the seven commercially available curricular packages endorsed by North Carolina's Office of School Readiness (Office of School Readiness, 2008), six curricula were developmental, comprehensive, and integrated across domains. The only domain-specific curriculum included in that list targeted language and literacy development. No specific math- or science-focused curricula are currently endorsed by the state of North Carolina.

Emphasis on Language and Literacy Development

Research has provided evidence that levels of language and literacy development in young children are important factors in academic success at later points in life. Delays in language development often predict literacy problems for children in later grades (Harris & Herrington, 2006; Lee & Burkam, 2002). In particular, low performance in early vocabulary acquisition has a detrimental impact on language and literacy competencies in later grades (Dickinson & Tabors, 2001; National Institute of Child Health and Human Development [NICHD] Early Child Care Research Network, 2005; Spira, Bracken, & Fischel, 2005). Although many preschool children across the nation experience difficulties in language acquisition and use, issues around vocabulary plague poor children more often than any other group (Dickinson & Tabors, 2001; Gormley, Gayer, Phillips, & Dawson, 2005; U.S. Department of Health and Human Services, 2005).

Children's early understanding of letters also influences language and literacy development and has profound implications for school success. Research suggests that young children who have difficulty recognizing and naming letters are likely to have reading difficulties later (O'Connor & Jenkins, 1999). A study by Catts, Fey, Zhang and Tomblin (2001) linked alphabet knowledge in kindergartners to reading achievement in second grade. Moreover, letter knowledge is connected to other elements of language development, which also influence later success in reading. Research has shown that instruction that combines training in letter identification and phonological awareness has greater effects on reading development in young children (Bus & van Ijzendoorn, 1999).

These works support the idea that children's understanding of letters impacts other language development and later literacy achievement.

Recent published analyses suggest that instructional interventions can produce positive effects on children's development in language and literacy. Mol, Bus and Jong (2009) found that effective interactive book reading increased children's expressive and receptive vocabulary. Piasta and Wagner (2010) found that multicomponential and pure alphabet instruction increased children's letter knowledge. Meta-analyses published by National Early Learning Panel (2008) and Mol et al. (2009) mirrored Piasta and Wanger's results when they found that code-focused instruction and interactive reading interventions positively affected children's alphabet knowledge.

Evaluating Effectiveness of Curricula

The over-arching emphasis on academic gains in children via well-developed, easily defined, easily replicable comprehensive curricula continues to grow. Furthermore, interest in the additive contributions of preschool curriculum has increased in recent years (Schweinhart & Weikart, 1997, 1998). Consequently, researchers, educators, and policy makers have targeted preschool programs use of comprehensive curricula and supplementary curriculum-based interventions as a means to increase school readiness and lessen the achievement gap. These programs and interventions often are driven by an attempt to increase the language skills of young at-risk children. In 2002, the Institute of Educational Sciences (IES) set a goal to provide information on the effectiveness of educational interventions that would help professionals in the field make decisions about classroom and program-wide practices. The WWC conducts reviews of curriculum-based interventions and has reported an assessment of several programs addressing language and literacy interventions in preschools. The results of these individual assessments vary from showing significantly positive effects on preschool child outcomes for an intervention that used Literacy Express (U.S. Department of Education, 2007) to inconclusive effects for an intervention that implemented Curiosity Corner (U.S. Department of Education, 2009).

Several other initiatives, such as NIEER and PCER, were formed with the goal of investigating and evaluating solutions to the problems faced by children from poor families upon school entry. Consequently, the United States has seen a surge in the development of curricular interventions targeting language and literacy development for preschool children over the past few decades. Such interventions often focus on the implementation of curriculum as a means to direct teachers' attention to particulars of language and literacy essential for school success stability and growth in adulthood. These interventions have had mixed results, ranging from substantial positive gains to limited and, at times, no influence on child outcomes (NIEER, 2008). Questions continue to exist regarding both the effectiveness of curriculum interventions for different groups of children and, moreover, which components of these preschool programs and interventions produce the most substantial gains in language- and literacy-related outcomes. As evident from the WWC reviews the effectiveness of individual preschool curricula targeting language and literacy development, among other competencies, varies dramatically.

A recently published meta-analysis provides some insight into the effectiveness of preschool curricula implemented in the United States. Camilli, Vargas, Ryan, and Barnett

(2010) analyzed 123 studies of early childhood interventions and found that children who received early intervention in preschool experienced significantly higher cognitive gains than did children who received no intervention or inconsistent, unsystematic interventions. Additionally, Camilli et al. found that curricula that required more direct instruction from teachers and that afforded more opportunities for individualized instruction produced higher positive outcomes. Their synthesis exposes the benefit that preschool has for the cognitive development of young children. However, more analysis is needed to tease out preschool curriculum effects on literacy and language development in particular. For example, Camilli et al. combined intelligence and cognitive/reading achievement into a single outcome labeled “cognitive” thereby confounding the impact of preschool on specific literacy and language gains. Although their synthesis is quite broad in its scope, as it contains studies published between 1965 and 2003, they place a stronger emphasis on early studies. Of the 123 included studies, 38 were published in the 1960s and only four were published in 2002 or later.

Implications of this Synthesis

The meta-analysis reported in this paper aims to broaden understanding of various interventions that implement entire curricula or add elements, like materials, methods, and instructional foci, to pre-existing curricula. Results of this synthesis reveal the overall effectiveness of preschool curriculum targeting children from high-poverty families. This synthesis adds to the contributions of the Camilli et al. (2010) analysis by further examining the moderating effects of program characteristics on literacy and language development in preschool children, as well as representing interventions administered after 2002, a pivotal year in early childhood education research as federal funding targeting preschool research increased. Additionally, this synthesis more closely examines the effect of curriculum on language and literacy outcomes in particular. Unlike the Camilli et al. synthesis, this analysis does not aggregate outcomes into a single global cognitive outcome, but rather targets preschool children's vocabulary acquisition and alphabet knowledge, along with the moderating effects of child background and program and treatment characteristics. Lastly, this meta-analysis goes beyond examinations of individual curricula like those of the WWC (2007, 2009) and NCER reports (2008) to look at the overall impact of curriculum.

The objective of this report is to assess the effectiveness of curricular interventions on the language and literacy development of preschool children to improve vocabulary and print awareness. The following research questions guide this review:

- Do preschool curriculum interventions have significant effects on preschool children's vocabulary development and alphabet knowledge? By the end of preschool? By the end of kindergarten?
- Are the effects of preschool curriculum interventions on vocabulary and alphabet knowledge affected by intervention or program characteristics? By the end of preschool? By the end of kindergarten?

Method

Criteria for Inclusion

In order to be considered for inclusion, research studies had to report vocabulary and alphabet knowledge outcomes from a preschool curriculum intervention. Although assessments of vocabulary and alphabet knowledge had to be norm-referenced, different assessments could have been used to measure the outcomes of interest. Studies eligible for inclusion had to supply information on assessments used to measure particular language and literacy competencies. Focusing on vocabulary and alphabet knowledge considerably restricted eligibility for inclusion in this meta-analysis; the choice to limit the focus was intentional. Vocabulary and alphabet knowledge are the two most robust language measures predicting later school success.

Studies also had to employ either random assignment with at least one treatment and one control group or quasi-experimental studies with comparison groups. Due to the nature of preschool programs and the manner in which children are often assigned to teachers and classrooms by the program itself, it was acceptable for researchers to assign on the classroom level. For studies that used a quasi-experimental design, it was essential that they reported children's pre-test scores on the language and literacy measures.

Studies included had to target children between the ages of 3:0 and 5:11 years of age who attended a preschool program serving children from low-income or high-poverty communities with high risk of academic failure (e.g., Title I, Head Start). Children in the studies included could represent a variety of cultural, linguistic, and racial/ethnic backgrounds. No exclusions were made based on children's backgrounds or ability status; thus studies that included but did not exclusively target English Language Learners (ELLs) or children with Individualized Education Plans (IEPs) were eligible for inclusion.

Lastly, studies in this review could report the use of curricula that were either (a) entire, self-contained curricula or (b) additional curricular elements that were used as supplementary components to a pre-existing program of instruction. No restrictions were placed on the deliverer of the intervention; therefore, teachers, support staff, specialists, researchers, and parents in school settings delivered the interventions. Lastly, no restriction as to the duration of the intervention was considered for inclusion of any study.

Search Procedure

Several methods were employed to compile the studies included in this synthesis. Electronic databases were searched using keywords and descriptors. In particular, the following databases were targeted: ERIC (Education Resources Information Center), PsycINFO, Dissertation Abstracts (ProQuest Digital Dissertations), and Education Abstracts. Specific keywords and descriptors were entered to narrow the search. Searches were grouped into three concepts: Literacy, curricular intervention, and preschool. The primary keyword strings used are shown in Table 1.

Table 1

Primary Keyword Strings Used in Searches

Concepts (and)	Search Terms / Descriptors (or)
Language/Literacy	Linguistic awareness Reading Letter sound correspondence Phonemic awareness Phonemes Word recognition Phonological awareness.
Curriculum or Intervention	Outcomes of education
Preschool	Preschool education Early childhood education

Studies with undesirable samples (e.g., kindergarten and Grade 1) or mismatched outcomes (e.g., aggression, anti-bias surveys) were eliminated. Additionally, only studies that occurred after 1990 and before December 2008 (when the search process for this review was completed) were included in this review. The 1990s saw an increase in the number of published preschool curriculum available and a resurgence of interest in language and literacy development and school readiness in preschool children, especially those labeled as high risk for school failure.

Supplemental searches produced a list of additional related studies. The What Works Clearinghouse publishes intervention reports with varying curricular foci. The studies mentioned in the reports targeting language and literacy curriculum-based interventions were examined. Additional studies were found when examining reference lists of studies initially included in this synthesis. These subsequent studies were included when selection criteria were met. Reports such as the recently published NCER report

(2008) and the results of the PCER studies (2008) were examined. Related literature reviews and meta-analyses were also searched for eligible studies.

Despite the breadth of coverage capable of online databases, hand searches are an essential process in acquiring studies that have potential for inclusion. Journals that focus on early childhood education and instruction, and on curriculum in particular, were searched. These journals included, but were not limited to:

- Early Education and Development
- Early Childhood Research Quarterly
- Journal of Early Intervention
- Reading and Writing: An Interdisciplinary Journal
- Early Childhood Education Journal
- Journal of Educational Psychology

Inclusion Coding

Title, abstract, and full text screening. A primary reviewer initially read through titles and abstracts of studies located through the search process to further determine the eligibility of each study. An independent reader then read through the titles and abstracts of a random selection of 10% of those studies. Any disagreements between the primary reviewer and verifier around study inclusion were resolved to consensus. The full text of the collected studies was read and examined for inclusion. Again, an independent reader then read through the full text of a random selection of 10% of those studies. There was 100% agreement between the researcher and verifier in indicating which studies met the inclusion criteria and which did not. Finalized codes collected for each study included were entered into a spreadsheet for documentation and later imported into statistical software for analysis.

Assessment of methodological quality. It is most desirable to include studies that employ high quality methods and valid measures in this review. In order to help maintain high quality, selection criteria mandated that only studies using random assignment or quasi-experimental designs be included in this synthesis. Qualitative studies were not considered for the synthesis. However, qualitative data regarding sample and intervention characteristics were included for studies that meet inclusion criteria and were analyzed and coded for essential information whenever possible.

Examples of exclusion. Several studies were initially included but failed to meet all required criteria. The 2006 study by Landry, Swank, Smith, Assel, and Gunnewig, for example, examined teacher influences on preschool children’s language and literacy development. However, the focus of this study was on the professional development model and not curricular elements of instruction. Likewise, Starkey, Klein, and Wakeley (2004) presented findings from their report of a preschool mathematics intervention. This study qualified for inclusion in every way except that it did not include assessments of vocabulary or alphabet knowledge. The Landry et al. (2006) and Starkey et al. (2004) studies contribute to the growing body of work in preschool research; however, neither study fully met the inclusion criteria for this synthesis.

Statistical Analysis

A coding scheme was produced (see Appendix A) and used to calculate effect sizes in determining each intervention effect as well as to identify specific characteristics of curriculum interventions. In the instance that a study reported the results of various assessments representing a single construct, the most widely used standardized assessment across studies was used to calculate effect sizes. For example, one study reported scores on both the Peabody Picture Vocabulary Test (PPVT) and the Expressive Vocabulary Test (EVT) among other measures. In this case, pre- and post-test scores from the PPVT were used to calculate the effect size associated with vocabulary.

Calculating effect sizes. Using the Statistical Package for the Social Sciences (SPSS 17.0), effect sizes were initially calculated as the standardized mean difference by finding the difference between the two group means and dividing by the pooled standard deviation of the subjects—children, in this case (Borenstein, Hedges, Higgins, & Rothstein, 2009; Lipsey & Wilson, 2001):

$$ES_{sm} = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{s_p}$$

All studies included in this review provided the mean scores of two or more contrasting groups on various outcomes relating to language and literacy achievement. The Hedges' g correction for small sample size bias was used for each primary study on each outcome of interest to account for the upward bias for studies with studies that had smaller sample sizes:

$$ES'_{sm} = [1 - \frac{3}{4N - 9}]ES_{sm}$$

ES'sm = unbiased standardized mean effect size adjusted for sample size
ESsm = biased standardized mean effect size
N = total sample size

This adjustment lessens the bias of smaller studies on the final calculation of the weighted grand mean.

Lastly, adjustments of the standard errors were also employed. Because random assignment was made at the classroom level and this synthesis reports effect sizes at the child-level, a standard error adjustment using Hedges' correction procedure (Hedges, 2007) was made to account for the influence of sample clustering (McHugh, 2004). An interclass correlation coefficient (ICC) of 0.10 was used, as is typical for analysis involving preschool children. The ICC value represents the degree to which members of the same group may be related. In this case, the analysis must take in to account that children enrolled in one preschool program are more similar with each other than they are with children enrolled in another program. The ICC value of 0.10 used in this statistical synthesis is conservative and likely to be an over-representation of inter-grouping similarities at the preschool level. Because the primary studies included in this synthesis frequently assign programs and classrooms to experimental conditions yet report child outcomes this adjustment must be included when calculating individual effect sizes and weighted grand mean effect sizes across studies. Without this cluster adjustment, standard errors would likely be inflated and the likelihood of statistical significance would be disproportionately greater.

Ultimately, a weighted grand mean was calculated across studies by using Comprehensive Meta-Analysis software (Version 2.2; Borenstein, Hedges, Higgins, & Rothstein, 2005). Random effects analysis was used, as the aim of this synthesis was to formulate conclusions about the effectiveness of preschool curriculum overall. An argument could be made that this study should use a fixed-effect model since it is a less conservative approach. Preliminary analyses were run to compare results of fixed- and random-effects models and no differences in effect sizes and homogeneity statistics were found.

Moderating Factors

In addition to identifying the oveall effect size, analyses were also conducted to detect any influence on the overall effect generated by moderating factors. Several sample, intervention, and program-related characteristics were examined for their additional influence on child outcomes. In particular, the following moderators were initially investigated:

- Type of preschool / funding source (e.g. Head Start, Title I, etc.)
- Predominant race/ethnicity in preschool program
- Parental involvement
- Use of teaching mentoring/coaching in intervention
- Measure used to assess outcome
- Length of intervention

Accounting for Bias with Preliminary Data

To avoid unrepresentative influence of extreme effect sizes, studies were examined for the presence of outliers with regard to sample size and individual effect sizes. In order to detect possible outliers, distributions of both sample sizes and effect sizes were created for the collection of studies. Using Tukey's formula (Hoaglin, Mosteller, & Tukey, 1983), no sample size (N) outliers were identified. However, two outliers were identified when effect sizes were examined. Studies #5 and #6 (Assel, Landry, Swank, & Gunnewig, 2007) produced extreme negative effect sizes, -1.27 and -.24 respectively. Both effect size values were winsorized to -0.538. The act of winsorizing outliers reduces the disproportionate impact of any one study without entirely eliminating the impact of the study from the final analysis.

Publication bias, a form of sampling bias, can distort the overall findings of meta-analyses. Studies reporting large, positive effects have higher likelihood of publication, have higher visibility, and subsequently have a higher likelihood of being included in a statistical synthesis. In order to ensure accurate representation of all interventions, including both published and unpublished reports reporting a range of sample sizes and effect sizes, analysis of publication bias was conducted. Examination of Funnel Plots and Trim and Fill analysis indicated that the potential for such bias was weak.

Results

Description of Included Studies

After a thorough search, 28 studies originating from eight separate reports were included in the synthesis (see Appendix B for a summary table of included studies). The total number of studies included in this synthesis was smaller than that of other meta-analyses (e.g., Camilli, 2010) in part due to a more limited publication date range and more stringent inclusion criteria. All studies included vocabulary outcomes, the majority of which reported pre- and post-test scores or adjusted scores of receptive vocabulary by administering the PPVT. These PPVT scores were selected as the sole representative of vocabulary outcomes whenever possible. In the cases when PPVT scores were not reported, data from alternative vocabulary assessments were used (e.g. EVT, Expressive One-Word Picture Vocabulary Test [EOWPVT], or Mullen Scales of Early Learning – Receptive Language [MSEL - RL]). All included studies dealt directly with preschools serving low-income families. Eighteen studies reported pre- and post-test scores of assessments measuring alphabet knowledge. The vast majority of these studies (i.e., 16 of 18) used the Woodcock Johnson Achievement Test, 3rd Edition (WJ III), Letter Word Identification subtest to measure children's alphabet knowledge.

Of the 28 studies included, almost half were implemented in Head Start programs. In addition, the sample of children across studies was representative of diverse racial and ethnic backgrounds. Most of the interventions assigned groups to the treatment condition at the classroom level, as it is difficult in school settings to randomly assign children to treatment and control groups. Table 2 displays several characteristics of the 28 studies.

Table 2

Characteristics of Included Studies

Characteristic	N	%
Publication Type		
Journal Article	11	39
Report	16	57
Dissertation	1	4
Program Type		
Head Start (+ Combo)	14	50
Non-Head Start	14	50
Curricular Focus		
Lang & Literacy	22	78
Other	6	22
Measure of Vocabulary		
PPVT	18	64
EVT	6	22
EOWPVT	2	7
MSEL - RL	2	7
Measure of Alphabet Knowledge		
WJ Letter Word	16	89
Dev Skills Checklist	1	5.5
Other	1	5.5
Total Sample Size		
<100	6	22
100-199	14	50
200-299	6	22
300+	2	7

Each included study compared the effects of the targeted curriculum implemented in the treatment classrooms to another curriculum used in the control classrooms. The comparison groups across studies varied considerably. Thirty-two percent of the included studies compared the treatment curriculum to control classrooms using High Scope. A large percentage of studies either omitted descriptions of the comparison curriculum (18%) or described them as nonspecific, teacher-developed curriculum (23%). Additionally, 13.5% of the interventions in this synthesis provided an ambiguous depiction of the comparison curricula by labeling them as the typical curriculum used in Head Start. The remaining 13.5% of studies compared the treatment curriculum to classrooms using Creative Curriculum.

Mean Effects of Interventions on Vocabulary Development

The objective of this synthesis was to measure the effect of preschool curriculum interventions on the vocabulary development of children at the end of preschool and kindergarten. Random-effects analysis produced a grand mean effect size for vocabulary at the end of preschool of 0.038 ($p>.05$). The 95% confidence interval indicates that this result is not statistically significant since the range -0.036 to 0.112 contains 0 (see Table 3 for complete results). These calculations signify that, overall, there were no significant effects of these interventions on the vocabulary outcomes of preschool children. This finding is illustrated graphically in the Stem and Leaf Plot (see Table 4) showing that the majority of interventions had little to no effect on vocabulary as their individual effect sizes hovered around zero.

Table 3

Weighted Mean Effect Sizes for Vocabulary

Outcome (Time)	*k*	*g*	95% *CI*	$Q_{between}$	*p*	I^2
Vocabulary (end PK)	28	0.038	-0.036,0.112	17.11	0.92	0.00
Vocabulary (end K)	13	0.048	-0.067,0.158	15.25	0.22	21.35

Note. k = number of studies; *CI* = confidence interval; *g* = Hedges' *g* with cluster adjustments

This statistical synthesis produced similar results when examining the grand mean effect size for vocabulary at the end of kindergarten. The grand mean effect size across the 13 studies that included kindergarten results was 0.048. Similar to the results at the end of preschool, the impact of curriculum interventions at the end of kindergarten was positive, but weak. Despite being positive, however, the effect size was neither significant nor that far from zero in effect.

Table 4

Stem and Leaf Plot of Effect Sizes

Stem	Vocabulary	
	End of Preschool	End of Kindergarten
0.40		
0.35		
0.30		20
0.25	3,16	28
0.20	28,4,17,25,7	18
0.15		26
0.10	12,20,1,22,13,	16,17,25
0.05	10,15,14,23,24	23,22
0.00	9,26,18,2	24,21
-0.05	11,8,21	
-0.10	27	
-0.15		
-0.20		
-0.25	19	
-0.30		27
-0.35		
-0.40		
-0.45		19
-0.50		
-0.55	6,5	

Mean Effects of Interventions on Alphabet Knowledge

An additional aim of this review was to calculate the overall effects of curriculum interventions on the development of children's alphabet knowledge. Results show that the impact of these interventions was nearly undetectable. The weighted grand mean representing the overall effect of the 18 interventions with alphabet knowledge outcome scores at the end of preschool was 0.029. This value was not significant with a confidence interval ranging from -0.056 to 0.114. The impact on alphabet knowledge at the end of kindergarten was comparably small, but negative. The weighted grand mean for 13 studies at the end of kindergarten was -0.034 with a confidence interval of -0.139 and 0.070 (See Table 5 for complete results).

Table 5

Weighted Mean Effect Sizes for Alphabet Knowledge

Outcome (Time)	*k*	*g*	95% *CI*	$Q_{between}$	*p*	I^2
Alphabet Knowledge (end PK)	18	0.029	-0.056,0.114	13.39	0.70	0.00
Alphabet Knowledge (end K)	13	-0.034	-0.139,0.070	13.20	0.35	9.09

Note. k = number of studies; *CI* = confidence interval; *g* = Hedges' *g* with cluster adjustments

These results point to a very limited effect of preschool curriculum interventions on children's immediate growth in vocabulary and alphabet knowledge. Moreover, the interventions lacked effects over two time points spanning two years of development. Although several individual studies showed moderate positive effects and a few others actually demonstrated negative effects, the bulk of studies revealed null effects.

Of the 28 studies reporting effect sizes on vocabulary, seven studies involved treatment groups that shared a common control group, which can lead to effect sizes that are not independent of each other. This lack of independence can lead to underestimated standard errors and overestimated statistical significance. For all seven cases, a redundant effect size was removed and analysis was regenerated. The random effects mean of this subset of studies was comparable to the results generated with the entire sample; therefore, the issue of statistical independence was not of concern and all 28 studies were included in the final analyses. As a result, the weighted grand mean effect sizes generated for the two outcomes over two time periods are justifiable representations of curriculum effects.

Heterogeneity Analysis

As was noted previously, individual effect sizes associated with the primary studies varied. Heterogeneity analysis was performed to determine if the difference between and among studies was merely a result of sampling error or brought on by alternate sources, beyond the expected variation of random sampling. Significant heterogeneity among studies supports the need for moderator analysis.

A test of homogeneity was performed using the calculated effect sizes from primary studies reporting gains by the end of preschool. Main effects analysis produced non-significant Q-statistics for studies reporting effect sizes relating to vocabulary gains (Q = 17.11, DF = 27, p = .92) and alphabet knowledge (Q = 13.39, DF = 17, p = .70) at the end of preschool. These resulting Q-statistics prohibit researchers from rejecting the null hypothesis that assumes homogeneity. Since the differences between studies could be caused by mere sampling error, explaining variation through moderator analysis is not justified. As an additional measure of heterogeneity, an I^2 value equal to 0.00 was produced by the CMA software for studies reporting effect sizes on vocabulary and alphabet knowledge outcomes at the end of preschool. An *I*-squared statistic *(I^2)* of this value signifies that 0% of the total variance is represented by the between study variation. This additional result further precludes the use of moderator analysis.

An equivalent analysis was run for studies reporting effect sizes at the end of kindergarten. The Q-statistic for vocabulary gains (Q = 15.25, DF = 12, p = 0.22) is also non-significant and prohibits us from rejecting the null hypothesis that there is homogeneity among these studies. In much of the same way, a Q-statistic for alphabet knowledge was also calculated (Q = 13.20, DF = 12, p = 0.35) and found to be non-significant. The I-squared values for vocabulary (I^2 = 21.35) and alphabet knowledge (I^2 = 9.09) differ from those representing heterogeneity of studies reporting effect sizes at the end of preschool in that they reveal that some degree of variation between studies exists. In the case of studies reporting effect sizes of vocabulary and alphabet knowledge outcomes by the end of kindergarten, roughly 21% and 9% of the total variance is represented by the between study variation, respectively. Neither of these values is particularly large; however, evidence of a sufficient degree of variation between studies reporting vocabulary outcomes at the end of kindergarten supports the need for further exploration. Because of this result, moderator analysis was used to explore that variation. The ensuing moderator analysis helped to clarify which studies and which particular elements of those studies produced larger effects on vocabulary gains at the end of kindergarten.

Regression Analysis of Impacts for Subgroups of Studies on Vocabulary Outcomes

Several moderators of initial interest were highly correlated. In particular, the variable representing receptive (e.g., PPVT) and expressive (e.g., EVT, EOWPVT) vocabulary was highly correlated with both the type of publication reporting the study (r = -0.62, p = 0.00) and the presence of a mentoring program (r = 0.69, p = 0.00). The decision to drop the differential contribution of measurement type from the model, while maintaining both type of publication and the presence of mentoring, was necessary to pursue the initial research questions and to perform sensitivity and bias tests. Nevertheless, all other variables were included in the final analysis, primarily to aid in answering questions of interest.

The author used meta-regression to analyze the influence of program and intervention characteristics on the intervention's effectiveness (see Table 6 for results). The full model included four covariates: the type of preschool program (i.e., Head Start or an other program), the predominant ethnicity of the children enrolled in the program (i.e., children of color as opposed to children from European backgrounds), the focus of the treatment curriculum (i.e., language and literacy versus other), and the presence or lack of mentoring or coaching systems for teachers.

Table 6

Regression Model for Effect Size Moderators on Vocabulary (End of Kindergarten)

Moderator	B	z	*p*	Beta
Type of Program *	.258	1.760	.078	.518
Predominant Ethnicity	.191	1.333	.182	.480
Curricular Focus *	-.244	-1.959	.050	-.607
Mentoring *	-.282	-1.903	.057	-.694

Three of the four moderators included in the meta-regression were significant at the p <0.10 levels. The type of preschool program, curricular focus, and use of mentoring/coaching within curriculum-based interventions all have independent significant relationships with vocabulary gains at the end of kindergarten above and beyond other predictors. Moderator analysis through the use of meta-regression provides a best-case scenario. These results illustrate that curriculum interventions in Head Start programs that implement curricula emphasizing something other than language and literacy, and that do not utilize mentoring or coaching, on average, have larger positive effects on children's vocabulary gains.

Discussion

The primary goal of this meta-analysis was to indicate the extent to which preschool curricular interventions increase vocabulary and alphabet knowledge in children. Results of this synthesis showed that interventions with this focus viewed in the aggregate have no significant effects on children's vocabulary and alphabet knowledge by the end of preschool and by the end of kindergarten. Although vocabulary and alphabet knowledge are only two aspects of language and literacy development in children, research suggests that children's ability to identify and understand words and letters is a proximate measure of other important language skills. In addition, the extent to which young children acquire vocabulary and understanding of the alphabet predicts achievement in later years. Certainly, more research on the effects of preschool curriculum interventions on other language and literacy-related competencies is warranted, given the emphasis often put on curriculum as a means to meet early childhood educational standards.

Previous research has shown that improving levels of vocabulary in children is a difficult task; vocabulary levels are remarkably stable from early ages to later grades (U.S. Department of Health and Human Services, 2005; Gormley, Gayer, Phillips, & Dawson, 2005). The most recent Head Start Impact Study (2010) found that children who attended Head Start experienced slightly higher vocabulary by the end of preschool than did children who attended other childcare programs or none at all. This equated to an effect size of 0.09. This increase, however, virtually disappeared by the following year and the effect size dropped to 0.04, similar to the one obtained in this larger analysis.

Conversely, some studies have reported strong effects on vocabulary. Mol, Bus and Jong (2009) found that interactive book reading instruction in approximately 20 studies produced an overall effect size of 0.62 on children's expressive vocabulary and 0.45 on receptive vocabulary. Targeted instructional methods, like those included in Mol et al.'s 2009 meta-analysis, provide a promising picture for work involving vocabulary in young children. Effects of targeted interventions on children's alphabet knowledge have met with additional success. Piasta and Wagner (2010) found that instruction for children in preschool, kindergarten, and early elementary grades that included multicomponential or pure alphabet instruction produced effects on letter knowledge ranging from 0.14 to 0.65. Similarly, the meta-analysis published by the National Early Literacy Panel (NELP) in 2008 included 24 code-focused early childhood interventions and found an overall effect size of 0.38 on alphabet knowledge. Mol et al. (2009) also found an overall effect size of 0.39 from 13 interactive reading interventions on alphabet knowledge. These studies reveal that some interventions do positively affect children's alphabet knowledge. However, the implementation of a curriculum intervention, in and of itself, does not have an equally positive impact.

It is important to note possible explanations for the differential patterns of effects between previously published meta-analyses and the synthesis presented in this paper. For example, the Mol et. al. (2009) analysis included studies that implemented activity-specific interventions as compared to more comprehensive curricula. They only examined book reading interventions, half of the studies were implemented by experimenters, and of the 31 studies Mol and others reviewed only nine included the full group, with five studies involving interventions that included 1-1 interventions. In contrast, the studies included in this review involved the full class of children, and were delivered by classroom teachers. The interventions reviewed by Mol and colleagues (2009) almost certainly were delivered with much greater fidelity to the intended model than those reviewed in this paper because they targeted specific language skills in a particular activity (i.e. book reading), and were more often delivered by experimenters rather than practicing teachers.

Limitations

There are two types of limitations associated with this synthesis. One comes from the analysis itself and the other issue is inherent in the study reports and in curriculum interventions overall. Although this meta-analysis includes a number of studies, no internationally based studies were considered. This body of research is important and may provide a clearer picture of curriculum-based intervention effectiveness. Moreover, in this synthesis, the constructs of language and literacy were solely represented by assessments of vocabulary and alphabet knowledge. There are numerous measures associated with language and literacy in young children and the inclusion of a larger number of these measures in this synthesis might produce a fuller description of intervention effectiveness and children's development.

The other limitation in this review is intimately related to the ways in which curriculum-based intervention research is reported. The state of the counterfactual in these interventions (i.e. the instruction and environment established in classrooms assigned to the control condition) creates a complicated situation. For one, although several studies provide detailed pictures of the treatment curriculum, few describe the comparison curriculum with an equivalent amount of specificity. Secondly, the comparison curriculum in four studies (e.g., *The Creative Curriculum*) was the same used as the treatment curriculum in three other studies. Because of the ambiguity inherent in descriptions of comparison groups, the degree to which the curricula in the treatment and control classrooms differ is unclear. Because of that, the relative effectiveness of one curriculum over another is confounded (Whitehurst, 2009). Clear descriptions of specific instructional components must be provided and consequently accounted for in the analysis in order to meaningfully calculate the effect of one curriculum over another.

Practical Implications

The resulting weighted grand means for both outcomes and both time points calculated in this synthesis were quite small and non-significant. These statistics can be interpreted in practical terms. The PPVT is a nationally normed assessment of receptive vocabulary with the standardized mean score equal to 100 and a standard deviation of 15. The results from this meta-analysis revealed that the overall effect of curriculum-based interventions targeting preschool children from low-income families compares to a 0.6-point increase in the PPVT, with an effect size of 0.038. The effect of preschool curriculum interventions amounts to a 0.75-point increase in PPVT scores by the end of kindergarten. The practical interpretation for gains in alphabet knowledge is equally minimal. On average, preschool children attending a program in which a curriculum intervention has been implemented should experience a gain of 0.45 points on the Woodcock-Johnson (WJ III) Letter–Word Identification subtest at the end of preschool. However, this slight advantage is reversed by the end of kindergarten and children in the control condition experience a gain of 0.45 points.

As is the case with all research involving vulnerable children and attempts to facilitate their development, it is desirable to use statistical analyses like the ones used in this synthesis to inform practice and policy. It must be acknowledged that the findings of this report are not enough to justify widespread policy adoption regarding preschool interventions. The overall impact of preschool curriculum on children's language and literacy development is virtually undetectable. The increased focus on implementing published curricula in preschool programs across the country and the expectation that the use of published curricula will be instrumental in helping teachers meet state-endorsed early childhood education standards is questionable. Closer attention must be directed to the instructional elements of these curricula and how they differ in a positive manner from the typical and varied instruction delivered in preschool programs. It is apparent that the mere implementation of a preschool curriculum is not enough to improve language and literacy achievement in young children.

CHAPTER III

MEASURING FIDELITY IN PRESCHOOL INTERVENTIONS: A MICROANALYSIS OF FIDELITY INSTRUMENTS USED IN CURRICULUM INTERVENTIONS

Background

A major challenge in interpreting the results of an intervention stems from the difficulties inherent in confirming that specific and essential components of that intervention were delivered as intended. Well-designed and defined interventions coupled with measures that accurately assess the level of fidelity to implementation are necessary to achieve accurate delivery. Research on the effects of preschool has seen an increase in studies committed to measuring intervention effectiveness, yet the procedures and instruments for measuring fidelity to implementation remain underdeveloped.

This paper has three primary functions. First, it will review trends in the behavioral psychology literature related to implementation fidelity including a description of the manner and frequency with which studies in this field have defined treatment, measured fidelity to the implementation of that treatment, and analyzed the relationship between treatment integrity and behavioral outcomes. Evidence of a historical trail of inconsistency and ambiguity within behavioral psychology will emerge. The second section of this paper will review ways in which the issue of treatment fidelity has been addressed in reports published within the last 25 years that involve school-aged children and educational interventions. The same problems faced by researchers in behavioral sciences are mirrored in the field of education. Finally, this paper will detail measurements of implementation fidelity in recent preschool curriculum interventions.

Reviews of K-12 interventions have been quite informative, yet a review of preschool studies remains outstanding. Analysis of 16 measures of fidelity used by 12 recently funded preschool curriculum interventions highlights how research in preschools is and is not effectively measuring fidelity of implementation.

Historical Issues in Defining and Measuring Fidelity

For decades, issues of defining and measuring fidelity of implementation have existed and researchers have had difficulty presenting implementation fidelity in a clear, comprehensive, and consistent manner. Earlier intervention work in psychotherapy and other behavioral sciences often ignored such issues of implementation, creating confusion when reporting details about how treatments were administered. Due to the omission of detailed measures of implementation, it was often unclear how different the treatment was to the control condition in an intervention (Moncher & Prinz, 1991). Complete agreement regarding the measurement of fidelity to the intended delivery has been slow in forming and adoption of standard practices associated with program adherence has been sluggish and inconsistent across fields and over time. Despite this lack of consensus in how to measure and lack of consistency in reporting fidelity, several researchers upheld the need to more closely consider fidelity of implementation. Several decades ago, Yeaton and Sechrest (1981) wrote that it is essential to consider treatment strength and program integrity when evaluating interventions (as cited in Cordray & Pion, 2006). Despite the work of early proponents like Yeaton and Sechrest and the developing conceptualization and operationalization in fields such as behavioral psychology, disagreement and inconsistencies continue. In recent years, educational researchers have used developments in other fields as a guide in conceptualizing implementation fidelity.

Labels used to represent fidelity. Terms and definitions representing the degree to which a program has been delivered as originally intended have varied considerably. Labels such as “treatment integrity”(Cordray & Pion, 2006; McIntyre, Gresham, DiGennaro, & Reed, 2007) , “program integrity” (Dane & Schneider, 1998), “treatment fidelity” (Moncher & Prinz, 1991), “adherence” (Waltz, Addis, Koerner, & Jacobson, 1993) and “fidelity of implementation” (Dusenbury, Brannigan, Falco, & Hansen, 2003; O'Donnell, 2008; Songer & Gotwals, 2005) have been used. Similarly, the ways in which treatment adherence has been measured have varied both within fields and across fields. Assessment of treatment implementation has taken many forms over the years. It is no surprise that such variation exists as treatment implementation is a multi-dimensional construct (Moncher & Prinz, 1991) that forces researchers to consider using multiple measures in a variety of methods.

The importance of operationalizing and measuring fidelity. There are several primary reasons for measuring the fidelity of implementation demonstrated by individuals who deliver the treatment within any intervention. First, fidelity is one of several essential measures used to confirm that outcomes of interest could indeed be related to the delivery of a treatment. Gresham, Gansle, and Noell (1993) claimed that evidence of high levels of fidelity is a necessary but not sufficient ingredient in establishing the relationship between modifiable independent variables and the resulting dependent variable. Second, defining the treatment and indicating the level of fidelity necessary for successful outcomes allow researchers, practitioners, and other interested individuals a higher probability of successfully replicating the treatment in another setting at another point in time.

With a detailed analysis of the relationship between an independent variable and an outcome, researchers may be able to identify particular components of the treatment that have a greater influence on the dependent variable (McIntyre et al., 2007). Researchers may hypothesize that certain elements of an intervention are more important than others. For example, changing the way an adolescent views drugs and friends who use drugs may be more important in preventing drug use than altering an adolescent's access to the drugs themselves. In this case, accurate measures of how well elements related to the first component of the intervention will aid researchers in justifying such a claim. Thus, it may become apparent that particular components of a treatment have a stronger correlational relationship with desired outcomes than others within the same treatment. Lastly, using measures of fidelity may identify components that appear to be important to success, but require additional support in order to be delivered in a high caliber fashion (e.g., additional professional development for treatment deliverers).

Related issues in defining the treatment. Defining a treatment and assessing the degree to which a treatment was implemented as intended are critical. Without confirmation that a treatment was indeed delivered accurately, interpretation of the results of an intervention may be obscured. Furthermore, understanding treatment characteristics, also referred to as independent variables of an experiment, enables researchers to deliberately and accurately manipulate elements of the treatment (Kazdin, 1980). Thus, effective measures of implementation fidelity can only be created if the essential elements of the independent variable(s) are clearly identified.

With this clarification, deliverers may be better prepared to implement the intervention. Likewise, the use of well-designed fidelity instruments that correspond to elements of the independent variable can be used to evaluate to what extent the intervention was administered.

Moreover, interpreting the effectiveness of treatments and interventions requires a report of treatment fidelity to know to what extent the treatment was actually implemented (Wheeler, Baggett, Fox, & Blevins, 2006). Despite frequent appeals for researchers to consider and include data on treatment integrity, reports of treatment integrity are inconsistent at best, and often missing entirely from intervention reports. Studies that do include analysis of treatment fidelity do so on a superficial level and do not link fidelity to treatment outcomes in a systematic manner (Dane & Schneider, 1998; Dusenbury et al., 2003). Many procedures and measures for analyzing fidelity exist in the behavioral sciences, yet no standards have been set and instruments are not uniformly used throughout the field (Waltz, et al., 1993).

Research focused on behavioral psychology is related in many ways to work in the field of education. In both fields, researchers or program developers rarely deliver interventions. Counselors and teachers are the primary deliverers of interventions. Major variability exists among counselors as well as teachers (Dusenbury et al., 2003). They have different training and education, work in an array of different communities and interact with children and young adults from a wide range of backgrounds. Consequently, individual counselors and teachers are likely to deliver programs with different levels of fidelity. The need to account for such variation in how these programs are delivered is especially meaningful, as there can be a disconnect between the developers and the deliverers of the treatment.

Connections Between Behavioral Psychology and Educational Research

Reviews of studies in the fields of behavioral psychology and counseling have indicated that defining treatment characteristics and measuring fidelity to the treatment by those individuals delivering the intervention are difficult and complex matters (Moncher & Prinz, 1991; Peterson, Homer, & Wonderlich, 1982). Even with highly trained professionals, manipulation and delivery of treatment are challenging when interventions are carried out in natural settings, outside the controlled environment of a laboratory. In the case of educational interventions, these difficulties abound. The challenge is two-fold as practitioners and researchers are both faced with challenging tasks. Many educational interventions have multiple components and a variety of desired outcomes. Well-prepared and organized training may be required in order for practitioners to be prepared to deliver the program adequately. Practitioners are often asked to implement a multitude of activities and strategies with which they have little familiarity. On the other hand, researchers may face obstacles when measuring fidelity of implementation to the ideal delivery demonstrated by the practitioners. First, researchers must thoroughly understand the characteristics of the treatment itself. It is then essential that they use efficient and effective methods in observing and assessing how well practitioners deliver components of the intended intervention. Because of these issues and the challenges brought about by the natural and ever-changing environments of schools, problems in treatment implementation are prevalent in educational research (McIntyre et al., 2007).

With the passing of No Child Left Behind Act of 2001, policy makers have emphasized the effectiveness of educational programs while holding educators accountable for student achievement. This emphasis has influenced educational researchers and program developers in that both groups now must justify that their programs and interventions have the strength to increase student learning, often measured through standardized tests. Because educational researchers must have a comprehensive plan for demonstrating the positive effects of the implemented program, it is even more important to ensure effective delivery of essential program components by practitioners. Thus, the need for educational researchers to utilize accurate and comprehensive measures of fidelity has become even more apparent.

Fidelity of Implementation in Behavioral Psychology

There are two particularly important aspects associated with measuring intervention effectiveness: 1) defining the causal components of the treatment 2) and measuring the degree to which these components are delivered with fidelity. It is helpful to examine the ways in which studies grounded in the behavioral sciences have dealt with these issues as a means to better understand and improve upon educational research. Several reviews in the behavioral sciences reveal interesting trends in the frequency with which studies define the treatment (typically identified as the independent variable in experimental studies) and the quality and depth of the measures. These reviews also point to a paucity of studies that measure treatment fidelity.

Defining treatment. In behavior analysis, Peterson et al. (1982) evaluated the number of studies that operationally defined the treatment of interest in experimental studies. Peterson and colleagues reviewed articles published in the Journal of Applied Behavioral Analysis (JABA) from 1968 to 1980 and coded them in one of three categories. Of the 539 articles that met inclusion criteria, they calculated that fewer than half of all studies that reported experiments define the independent variable (i.e. treatment). Peterson et al. concluded that researchers were quite able to identify causal components of their experiments and thereby had opportunities to clearly define the independent variable, but only a small number of researchers actually did so. Likewise, Wiese (1992) targeted overlapping years when she examined articles published from 1975 to 1990. Wiese included published articles from a variety of psychology related journals that reported on parent training interventions related to behavioral problems between parents and their children. Of the 148 studies taken from 18 journals that met inclusion criteria, 58% defined the treatment by describing in detail the training programs delivered to parents. Thus, almost half of these studies did not define the independent variable.

As a continuation of the original review by Peterson and colleagues in 1982, Gresham et al. (1993) and then McIntyre et al. (2007) also reviewed articles in the Journal of Applied Behavioral Analysis. Using similar inclusion criteria, Gresham and colleagues found that only 34% of articles published between 1980 and 1990 comprehensively defined the treatment. They concluded that the mandate put forth by Peterson et al. that reports of experimental studies should include such a definition was largely ignored, as the majority of articles published in subsequent years did not include these definitions. A review of articles by McIntyre et al. (2007) published in JABA between 1991 and 2005 produced far more hopeful results. Of the 152 studies included in their review, an overwhelming 95% satisfactorily defined the treatment – a meaningful increase from the results generated by the Peterson et al (1982) and Gresham et al (1993) reviews.

Measuring treatment fidelity. Identifying causal components and defining the independent variable of an intervention are essential precursors in the process of measuring the deliverer's fidelity to the intended intervention. Several reviews have demonstrated that historically few studies clearly describe the specific characteristics of the intervention, but the numbers of studies that do include those details are on the rise. Unfortunately, these reviews do not reveal a comparable positive trend over the years with regard to the frequency with which these same studies measure treatment fidelity. Peterson et al. (1982) found that no more than 30% of reports published in JABA in any year between 1968 and 1980 assessed the level of treatment fidelity. On average, across the 13 years represented in their review, only 20% examined treatment integrity. Gresham et al. (1993) found that less than 16% of the reports published in JABA from 1980 to 1990 measured treatment fidelity, and the review by McIntyre' and colleagues (2007) estimated that 30% of the studies in JABA (1991-2005) assessed treatment integrity. Although their review showed an increase in studies reporting essential details about the intervention, it is clear that the majority of articles published in JABA from 1968 to 2005 did not include analysis of treatment integrity in their reports. This trend is fairly consistent across four decades, with only a slight increase in the frequency of assessment of treatment integrity over the past 14 years.

Supporting evidence of this trend can be found in Moncher and Prinz's (1991) review of psychosocial interventions published between 1980 and 1988. Of the 359 treatment studies, only 45% of studies considered issues of treatment fidelity. Although the Moncher and Prinz review found a higher frequency of treatment fidelity than the reports by Peterson et al. (1982) and Gresham et al (1993), the majority of studies in their review essentially ignore treatment integrity. Unlike trends across the years in defining independent variables, the rate at which intervention studies assess treatment fidelity has been relatively stable and quite low.

Recent Work Related to Education and School Settings

These reviews of treatments in the behavioral psychology press have brought attention to key omissions in the analysis of intervention effectiveness. Dane and Schneider (1998) provided a framework that has helped researchers in educational settings more clearly understand the different yet related multiple dimensions of fidelity. Dane and Schneider reviewed published reports of prevention-based interventions that were delivered to schoolchildren in primary and early secondary grades. The focus of the interventions included in this review related to behavioral and academic problems demonstrated by children in school settings. Dane and Schneider analyzed the reports, published between 1980 and 1994, and synthesized the ways in which fidelity had been understood and measured. They categorized the manner in which studies conceptualized fidelity and then used instruments to the measure fidelity. Their comprehensive analysis resulted in identifying five major components of fidelity: exposure, adherence, quality of delivery, program differentiation, and participant responsiveness. This categorization is often cited in literature focused on fidelity of implementation, although individual interpretation of these categories varies considerably.

Further coding and analysis of studies in the Dane and Schneider (1998) review yielded additional conclusions. They found that there was no real consensus among researchers across the field of prevention related to the types of measures needed to accurately assess fidelity of implementation. Because of these inconsistencies, no solid understanding how to define and then measure program integrity emerged. The Dane and Schneider review was able to clarify the ways in which fidelity had been defined, but also clearly illustrated the gaps in understanding among researchers working with school-aged children. Even though Dane and Schneider were able to identify five major components of implementation fidelity, few researchers included all five in their conceptualization of fidelity. Moreover, researchers had divergent definitions of these components. Researchers' individual definitions of exposure, adherence, quality of delivery, program differentiation, and participant responsiveness often varied the definitions put forth by Dane and Schneider.

Unfortunately the conceptual clarity that Dane and Schneider brought to the idea of implementation fidelity has not yet been translated into widespread action. Very few educational studies report measures of fidelity, and even fewer account for the relationship between fidelity and student outcomes. Dusenbury and colleagues (2003) further endorsed this conclusion with their own review of studies that were published between 1980 and 1994, delivered in school settings, and related to mental health, prevention, personal and social competence promotion, education, or drug abuse treatment and prevention. Dusenbury et al. categorized 25 years of studies using Dane and Schneider's framework and found inconsistencies across studies with regard to how treatment was defined and how fidelity was measured. Only 24% of 162 studies assessed fidelity of implementation. Of those, only one-third looked at the effect of fidelity on outcomes. The authors were unable to identify a single published study that included representations of all five components of fidelity defined by Dane and Schneider. Moreover, they found that researchers more typically described the methods they used to ensure higher levels of fidelity than details about the measures used to assess fidelity elements themselves.

Inconsistencies in the way fidelity is defined and ambiguity around how to measure fidelity also exist in K-12 curriculum intervention studies. In her review, O'Donnell (2008) provided evidence that few reports of curriculum interventions included details on measures of fidelity of implementation. However, she emphasized that more attention has been paid to issues around fidelity in the recent years. O'Donnell had several objectives: two of her primary goals relate directly to this paper. First, her review identified K-12 studies that define and quantitatively measured fidelity of implementation. Secondly, O'Donnell revealed studies that linked quantified measures of fidelity to child outcomes.

Overall, O'Donnell found 23 studies that defined fidelity as including both concepts of adherence and integrity. She found that studies targeting K-12 curriculum interventions had a greater likelihood of referencing instructional quality as an element of fidelity. Despite the paucity of available studies, O'Donnell described a two-fold trend in the manner in which studies measured fidelity. These 23 studies identified critical components that were considered the mainstay of the intervention. Moreover, they also indicated an acceptable range of variation with regard to these components. On the other hand, very few studies included in the review linked levels of fidelity to outcomes. In fact, O'Donnell found that only 5 of the 23 selected studies that met most of the selection criteria actually measured the effect of implementation fidelity on outcomes. Thus, O'Donnell's review provided evidence that K-12 research has yet to suitably account for implementation fidelity and its inherent relationship with treatment outcomes.

Researchers have recently turned their attention toward using measures of fidelity to calculate the actual strength of an intervention. For example, Hulleman and Cordray (2009) established methods to measure the relative strength achieved by the intervention as it was actually delivered as compared to its ideal delivery. In their study that examined differences in the effectiveness of the same treatment when implemented in a laboratory environment and in a university classroom, Hulleman and Cordray emphasized both quality of instruction and responsiveness as the primary components of fidelity. More specifically, they combined Dane and Schneider's version of adherence and quality of delivery into a single construct referred to as "Specific Fidelity". This perspective on the definition of fidelity differed from the traditional view proposed by Dane and Schneider in 1998 (Cordray, 2009; Lipsey, 2009). Assessment of this more developed definition of quality enables researchers to quantify how well the deliverer (e.g. teacher) met expectations fidelity represents not only the essential components of the intervention but also characteristics of the ideal delivery.

In past years, Dane and Schneider (1998) presented the concepts of "adherence" and "quality of delivery" as two distinct components of fidelity. According to Dane and Schneider, adherence represents the degree to which particular components associated with the intervention are carried out. Quality of delivery, on the other hand, is an assessment of implementation that focuses on enthusiasm and attitudes that are not tied to a specific program, but are more general in nature. Dane and Schneider's definitions have largely been accepted, and have been used to structure reviews such as those published by Dusenbury et al. (2003) and O'Donnell (2008). However, the definition presented by Hulleman and Cordray (2009) is the particular definition of specific fidelity (i.e., adherence) used in the analysis that follows.

Status of Fidelity in Preschool Studies

No review of preschool studies and the ways in which they define treatment and assess fidelity of implementation has yet been published. Therefore, it is difficult to detect if the problems inherent in K-12 studies are also evident in preschool interventions. The reviews of behavioral and educationally based interventions that targeted school-aged children discussed earlier in this paper provide evidence that few studies define implementation fidelity. Moreover, it was rare that these studies linked levels of fidelity to child outcomes. Analysis of how researchers who are engaged in interventions at the preschool level deal with issues of implementation fidelity is missing, yet it remains a necessary part of our complete understanding of effective educational research.

A new focus on educational effectiveness has become increasingly apparent in preschool research, as the movement to establish effective preschool programs has gained momentum. The fact that some children are not academically prepared to enter kindergarten is a source of great concern for educational researchers, practitioners, and policy makers (NCER, 2009). This concern prompted the U.S. Department of Education to invest $36 million in grants for evaluations of preschool curricula (Barnett, 2008). Increased funding for effectiveness studies and heightened attention to the importance of implementation drives the need to fully understand the multiple components associated with fidelity of implementation.

In an effort to evaluate how well preschool curricula are improving the social and academic development of children, The Preschool Curriculum Evaluation Research Initiative (PCER) supported by the Institute of Education Sciences funded a group of projects. Specifically, this initiative was established to assess the efficacy of preschool curricula. To accomplish this task, PCER put forward three primary research questions (NCER Report, 2008, p. xxxi). First, funded studies were to determine the effect of each curriculum on child outcomes related to language, literacy, mathematical knowledge, and behavior. Secondly, studies were required to follow children through Kindergarten to calculate the curricular affect on children at the end of Kindergarten. Lastly, studies were to account for the effect of curriculum on “classroom quality, teacher-child interaction, and instructional practices.”

To address these questions, The PCER study funded 12 research projects to evaluate a total of 14 different curricula. An additional requirement was that the preschools were to serve children from economically disadvantaged families. Preschool classrooms were randomly assigned to treatment and control conditions. Across all sites, control conditions varied from using teacher-developed curriculum (in 7 sites), a non-supplemented version of *Creative Curriculum* (1 site), a combination of commercial and teacher-developed curriculum (2 sites), High Scope curriculum (4 sites), and components gathered from multiple commercial curriculum (1 site).

As the funder of PCER, the Institute of Education Sciences provided specific guidelines to projects for measuring and reporting rates of implementation. In a meeting of PCER grantees IES officials stated the importance of measuring implementation fidelity (U.S. Department of Education Office of Educational Research and Improvement, 2002). Accurate depictions of the degree to which the curricula were implemented as intended were to be included in the analysis models of curricular impacts. All projects were required to collect fidelity data on both the treatment and control classrooms. At the end of the implementation year, all projects reported global implementation ratings for all classrooms that either implemented the treatment curriculum or were assigned as the control condition. Classrooms were rated on a four-point scale ranging from "0" to "3" with "0" representing "No Implementation" and "3" representing "High Implementation".

The PCER grants provide an interesting collection of pre-kindergarten experimental projects mandated to measure fidelity of implementation. Thus, this national study appears to provide an excellent opportunity to determine the contribution of implementation fidelity to curricular effects. However, beyond the stipulation to use some fidelity measure, little guidance around specific measures and the frequency of measurements was provided by IES. Consequently, all projects used site-specific measures and reported average scores for treatment and condition classrooms, yet the measures themselves varied dramatically in breadth, focus, and precision. The analysis that follows examines these measures at both the instrument- and item-level. The objective of this study is to determine the degree to which the fidelity measures used throughout the PCER study truly measure fidelity.

Method

Projects

The 12 research projects were located across the country in the following states: Tennessee, North Carolina/Georgia, New Hampshire, Florida/Kansas/New Jersey, Texas, Virginia, California/New York, Wisconsin, Missouri, New Jersey, and two located in Florida. Across these projects, 14 separate curricula were implemented. The analysis presented in this paper includes 10 of the 12 projects and 12 of the 14 curricula because the measures used by the projects implementing *Early Literacy and Learning Model* (ELLM) located in Florida and *Curiosity Corner* based in Florida, Kansas, and New Jersey were not available.

Although PCER guidelines did not require studies to include children exclusively from economically disadvantaged backgrounds, they gave preference to applicants who targeted poor communities. Therefore, the majority of interventions described in this paper had samples consisting mostly of low-income children. The 12 projects implemented curricula in various types of preschool programs. Five worked with public preschools, four with both public pre-kindergarten and Head Start programs, two collaborated with Head Start programs, and two worked with child-care centers. Projects used several techniques to randomly assigned groups to conditions. Three projects randomly assigned preschool centers as a whole to conditions, whereas nine projects randomly assigned classrooms.

Curricula

This paper focuses on information related to 12 curricula (See Appendix C for curriculum characteristics). For the purpose of this report, the combination of *DLM* and *Open Court* was treated as a single curriculum because the researchers in this project used *DLM* in conjunction with *Open Court* and measured fidelity to the combination of curricula by means of a single measure. The 12 curricula varied in scope and focus. Eight curricula addressed only language and literacy, one focused on math, two were inquiry-based and stressed cognitive development; the remaining curriculum had a more general developmental approach. All curricula typically spanned the course of an entire preschool year, but their structures varied considerably. Fifty percent were organized around sequential or thematic units, while 4 of the 12 curricula were structured around particular learning areas or activities. The framework of one curriculum was centered on children's developmental goals, whereas another was organized around instructional techniques employed by teachers.

Instruments

Each project submitted the specific fidelity measures used to assess degree of implementation to Institute of Education Sciences (IES) along with descriptive data and results. Overall, each project used between one and three measures. All but one of the 16 measures submitted were in a checklist format. For the analysis presented in this paper, the author coded all 16 measures at the instrument-level and coded the15 measures that were in a checklist format at the item level. One of the instruments (i.e., *Ladders to Literacy* – Scaffolding) used in the project evaluating the *Ladders to Literacy* curriculum was not in a checklist format and could not be coded at the item level. Researchers with each project required checklists to be completed in vivo by selected observers with the aim of evaluating environmental and instructional components of classrooms and teachers. Moreover, each instrument contained anywhere from 8 to 314 items and were organized into as many as 10 sections (e.g. classroom organization, teacher-child interaction, etc.).

Coding Procedures

The coding scheme used in this study was created to capture the general characteristics of each measure overall, as well as the specific nature of the items within each measure. Thus, there were two types of coding schema: 1) coding by instrument and 2) coding by item (see Appendix D for detailed code sheet). At the instrument level, each measure was summarized as to the total number of items, the number of sections divided thematically (e.g. teacher-child interactions), and the inclusion of a teacher interview. The author also determined if the directions included with the measure indicated the optimal length of an observation, and if observers were requested to provide the length of observed individual activities.

In order to determine the degree to which these measures represented fidelity to the curriculum, each item was analyzed independently (see Figure 1 for the item-by-item coding schema). More specifically, every item within the 15 measures received a code from four primary categories. Starting from the left of the tree in Figure 1, the first code for each item related to curricular reference in that each item received a score for whether it directly referenced the targeted curriculum (Y) or represented more general qualities of instructional or environmental aspects of classroom practices (N). In addition, if the item asked coders to indicate the quality of delivery or depth of a particular feature of the environment or instruction it was coded as adequate (AD). If the item only asked coders to make an objective observation based on the presence or absence of an object or event, it was coded as inadequate (IN). Effective quality of delivery items require the classroom observers to make judgments about the quality of instruction and make conjectures about nuances in a teacher's instruction or intended purpose for classroom materials. As an example, the item "Adapt methods/materials, as needed, for children with disabilities" required the observers who were completing this checklist to draw conclusions on the quality and effectiveness of a teacher's attempt to use different strategies and materials with children with special needs. This was determined to be an "Adequate" item as it was written in a way to give observers opportunities to indicate the depth, breadth and quality of the teacher's instructional methods. A preceding item that asked if teachers "allow each child in the class to have a turn as the cashier or store owner" was considered "Inadequate" in that observers would be able to determine if the teacher did or did not perform this task completely on the basis of observable behaviors with little regard for quality of delivery.

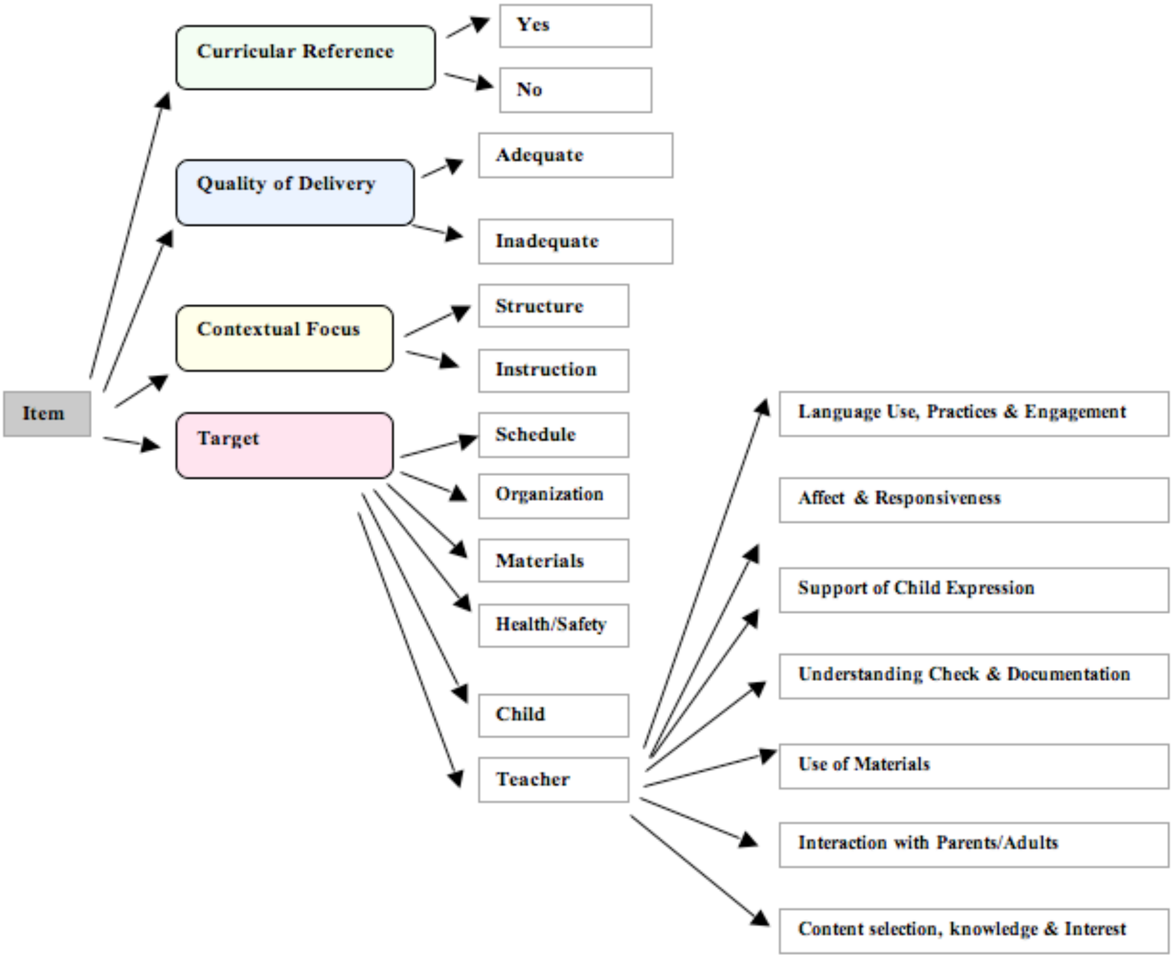

Figure 1. Graphical coding schema for each item.

Beyond these initial two categories, each item of a fidelity scale then received a code from both the contextual focus and target categories. Within the contextual focus category, items were designated as relating either to structural elements of the classroom setting (S) or to instructional features associated with teacher or child behaviors (I). For example, two items in the *Doors to Discovery* Curriculum Fidelity Checklist provide a good comparison for the kinds of codes within this category. The item "Language – Scrapbook: The scrapbook is displayed in the classroom and implementation of activities is evident" was coded as "S" in that it represented structural elements of the classroom. The next item in the checklist "Language – Story Character Props: Teacher uses story character props: in large group, to model for story retelling, etc." was coded as "I" in that it represented instructional features related to teacher behavior.

Items also received an additional code indicating the target of interest (i.e., target). Items were subsequently marked as being related to "scheduling", "organization", "materials", "health/safety", "child", or "teacher". Scheduling items were associated with the length of activities and integration of activities into the daily or weekly classroom calendar. Items labeled as organizational related to the presentation of materials, the presence of teachers in particular classroom settings, and the ways in which children were grouped. The code for materials, as expected, was associated with the presence of and access to items such as puzzles, books, paper and writing tools. Items associated with health and safety were used exclusively to evaluate how the environment helped to promote the physical well being of children. For example, both the presence of toothbrushes at the sink area as well as the inclusion of hand washing at points in the daily schedule would be coded in that manner.

Additionally, the items such as "Children are asked to predict what will happen in the story" were coded as relating to supporting children's expression (C) because such requests made by teachers encourage children to express themselves. Lastly, items designated as "T" specifically referred to teachers' behavior and varied dramatically in focus and specificity. Therefore, items targeting the teacher were subsequently assigned one of the following seven codes indicating a specific facet of instruction: (a) teacher language use, practices, and engagement, (b) affect and responsiveness, (c) support of child expression, (d) assessing and/or documenting children's understanding, (e) use of materials for learning objectives, (f) interaction with parents, specialists, other adults, and (g) selection of content, knowledge of child development and/or content, and interest. For example, the item "Teacher delivers story in an engaging manner" in the *Language-Focused Curriculum* (LFC) Fidelity Checklist was related to the affect and responsiveness of a teacher. Another item "Story is related to the daily theme" was associated with a teacher's selection of content, as observers were asked to assess teachers' selection of a book that related to content emphasized throughout the week.

Reliability

Each item in every measure was coded across four major categories. All items received a single code for the curriculum reference, quality of delivery, contextual focus, and target categories. Items that were coded as targeting teachers then received a code from the instructional facet category. In total, 1,113 items were coded. Fifteen percent of the total number of items were randomly selected and coded by a second independent researcher. Exact percent agreement and Cohen's *Kappa* were calculated for each of the four primary coding categories (see Figure 2 for reliability results). Reliability was also calculated for the subsequent code for items targeting teachers. Exact percent agreement and Cohen's *Kappa* was calculated as 78.4% and 0.71 respectively for this subordinate category. Both types of reliability calculations are reported to provide a more comprehensive picture of the reliability. Percent exact agreement is commonly reported in studies of a similar nature. However, Cohen's *Kappa* calculation, a more conservative representation of reliability, adjusts for the possibility of chance agreement between two observers (Banerjee, Capozzoli, McSweeney, & Sinha, 1999). Of the four primary categories coded, reliability between the coder and verifier was lowest for quality of delivery. Cohen's *Kappa* for this category was calculated as 0.50. Researchers have not yet reached wholesale agreement on acceptable levels of *Kappa* calculations. However, some research suggests that levels ranging from 0.40 to 0.75 are considered adequate (Fleiss, 1981). Despite the relative low reliability, this category was included in the analysis because it represents elements of implementation fidelity that are important but also hard to capture.

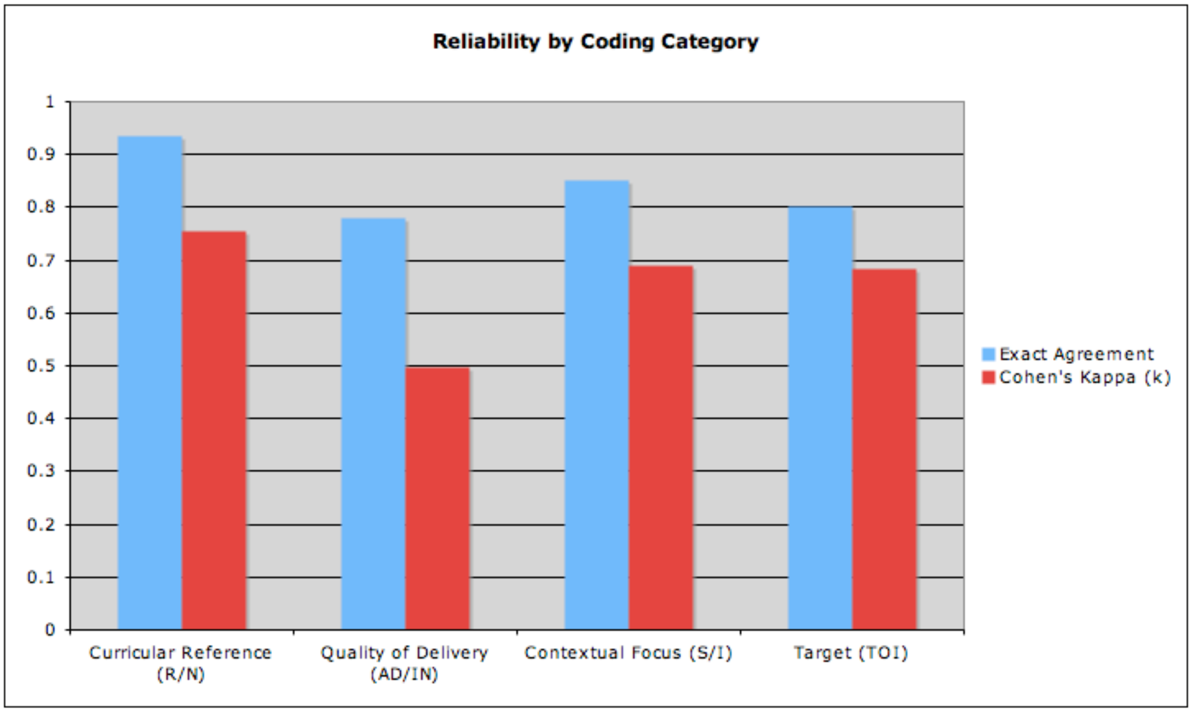

Figure 2. Reliability as percent of exact agreement and Cohen's *Kappa* (*k*).

Interpretation of Coding Schema

Interpreting adherence versus general instructional quality. Fidelity of implementation has been defined in a variety of ways across fields and even within education research itself. For the purpose of this paper, the coding schema was used to analyze how measures represented the two primary elements of fidelity: adherence and exposure. The schema also coded for participant responsiveness, a possible moderator of fidelity. The definition of adherence used in this analysis has been borrowed from Cordray's (2009) work and refers to structural and instructional elements that are both tied specifically to a particular curriculum and assess quality of delivery. Therefore a distinction was made between curriculum specific adherence and general instructional quality.

Curriculum specific adherence included items that represented the quality of the instruction and environment as it related to essential components of the curriculum. Therefore, items coded as both "R" (i.e., have a direct reference to the curriculum) and "AD" (i.e., assess levels of quality, breadth, and depth) are identified as representative of adherence. Items that refer to the curriculum but do not assess the quality of delivery are considered inadequate measures of adherence, despite their connections to an individual curriculum. General instructional quality items, however, assessed the breadth and depth of observable instructional and environmental factors but did not directly refer to the curriculum. Rather, these items represented a commonly accepted set of instructional and environmental characteristics expected in any preschool classroom. Items coded as "N" under Direct Reference (meaning they had no direct reference to the curriculum) represented general instructional quality and did not qualify as measures of adherence.

Interpreting exposure. Analysis of treatment exposure is also an integral part in the comprehensive understanding of what is being delivered and what children are potentially receiving. Within these PCER instruments, the author of this paper focused on items that captured the degree to which children were exposed to structural and instructional elements observed in classrooms. Specifically, she developed a system by which items were coded to indicate the degree to which children were exposed to the treatment condition. In particular, items labeled as "scheduling" included references to the sequence or schedule of behaviors or opportunities, duration of activities, and integration of activities into the daily schedule. These items, therefore, represent children's exposure to the curriculum.

Interpreting participant responsiveness. The third and final construct of interest in this paper is associated with participant responsiveness. The responsiveness of teachers in delivering and of children in receiving curricular elements may influence levels of implementation fidelity. Agreement on the definition of participant responsive is contested. The reviews of both O'Donnell (2008) and Dane and Schneider (1998) specifically examine student responsiveness. However, other more traditional views include teacher buy-in and attitude as the primary component of responsiveness (O'Donnell, 2008). In the present study, instruments and items were analyzed for ways in which both teacher and child responsiveness were represented. Teacher responsiveness could be represented through one of several more qualitative measures that some instruments included. Specifically, instruments were examined for the inclusion of teacher interviews. Additionally, the author read the directions and guidance contained in the instrument documentation and identified instances in which classrooms observers were requested to ask teachers for clarification or advised to look at lesson plans or classroom records. Due to the age of the sample, however, observers were unable to use the same methods of assessing levels of responsiveness by the preschool children as they did for teachers. Therefore, in order to have some measure of child responsiveness, all items that were coded as targeting children (C) were considered representative of child responsiveness.

Results

Adherence and General Instructional Quality

Of the 1,113 items across the 15 checklist measures, 32.8% made direct references to the respective curriculum and 28.4% measured quality of delivery. Overall, items that were coded as having a direct reference to the curriculum and assessing quality of delivery were rare. Only the items that were coded with this combination of attributes were considered to represent adherence to the specific curriculum. In total, only 69 items were coded as directly referring to the associated curriculum as well as requiring coders to assess teachers on the level of quality with which they delivered aspects of the curriculum. Therefore, by the strictest definition, only 6.2% of the items across the 15 instruments truly measured curriculum specific adherence. Across all 15 measures, 26.5% of the items made direct references to the curriculum but did not assess the quality of the delivery. In those cases, items coded as such represented an incomplete definition of adherence. Items that did not have any direct reference to the curriculum were designated as representing general elements of the environment or instruction that may exist in a preschool classroom. These items, however, did not represent any conception of adherence. As a result, approximately 67% of the items contained in fidelity measures only evaluated general instructional quality. For instance, an item in the *Ready, Set, Leap!* Fidelity Observation Checklist asked observers to indicate if "the teacher provides feedback in a positive manner". This item asked observers to indicate if teachers were providing the type of instruction one might expect in any quality preschool classroom. However, the act of providing feedback and doing so in a positive manner is not a unique requirement of a teacher implementing the *Ready, Set, Leap!* curriculum.

The 69 items that truly represent adherence were contained within only 5 of the curriculum fidelity measures. The *LFC* Fidelity Checklist had by far the highest proportion of items measuring adherence; 21 of the 45 items (46%) were coded as such. Other measures also contained items representing adherence, but many fewer. The percent of items measuring adherence in instruments used to assess fidelity to *Ladders to Literacy* (Classroom Activities), *Doors to Discovery*, *Let's Begin with Letter People*, and *Pre-K Mathematics* were 15.6%, 4.2%, 32%, and 10.2% respectively. Despite containing items relating to adherence, most of the items in these five measures did not evaluate how well teachers had implemented specific components of the curriculum. Moreover, the remaining 10 measures did not contain any items that specifically evaluated how critical aspects of the respective curriculum were implemented.

Of the 69 items that represented adherence, 92.8% were coded as instructional, and all of those focused on teachers, not children. Sub-analyses indicated that nearly half of these items specifically evaluated teachers' support and scaffolding of the manner in which children expressed themselves. For example, one item included in the *Pre-K*

Mathematics Fidelity of Implementation Record Sheet specifically asked observers to indicate the quality of scaffolding provided by the teacher during small group activities. Observers could select one of three options: “too specific or detailed”, “about right”, or “too general or vague”. The next highest proportion of items in this subset examined teachers’ use of language and engagement in practices specific to the activities observed. For example, of the twenty-one items in the *LFC* Fidelity Checklist that related to adherence, eight referred specifically to teachers’ use of language. For instance, the sixth item in the *LFC* Checklist focuses on teacher-child interaction and asked observers to confirm that the teacher used open questions. Overall, less than 6% (i.e., only 4 of 69) of the adherence items actually assessed the degree to which teachers had command over and interest in the content or how well they were able to select appropriate topics within a curriculum activity. Three of the four items of this type were found in the *LFC* Checklist. One *LFC* item asked if the teacher “organizes daily lessons around a particular theme”. When items focus on things like the teacher’s selection of content, they have more potential to evaluate how well teachers understand and implement important topics tied to the curriculum.

Exposure

Each measure was also analyzed to verify the inclusion of elements related to exposure, such as the frequency in which curricular activities and instructional techniques were implemented. On average and including items with and without direct curricular references, 10.1% of all items across instruments were related to the time, duration, or integration of activities into the daily or weekly schedule. Nine of the 15 fidelity measures included these kinds of scheduling items. Across measures, the percentages of items devoted to scheduling ranged from 1.7% in the *Ladders to Literacy* Classroom Activity checklist to 54.1% in the *Pre-K Mathematics* Fidelity Record Sheet (see Appendix E for additional percentages). Items related to scheduling were not found in the fidelity instruments related to the following curricula: *DLM* / *Open Court*, *Literacy Express*, *Doors to Discovery*, *Let's Begin with Letter People, Early Mathematics Classroom Observation* (EMCO), and *Project Construct*. Items that represented exposure and made direct references to the associated curriculum were even more rare. Only 63 items (i.e., 5.7%) evaluated the degree to which children were exposed to instructional and structural elements particularly associated with a curriculum.

Alternatively, all 15 measures asked observers to record instances when teachers grouped children, set up activity areas, were present in particular settings, or managed children's behavior. These instances were coded as organizational items. Across the instruments, 13.8 % of the items depicted teachers' involvement in classroom organization. The percentages of items devoted to organization across measures ranged from 6.1% in *Ladders to Literacy* Classroom Activity checklist to 44.4% in the *DLM* / *Open Court* checklist.

Of the items coded as "structural", the majority asked observers to identify the ways in which teachers facilitated children's use of materials, as well as the presence of materials in certain settings across the school day. Over 20% of the structural items were related to materials. Comparing these results to those obtained for adherence items, it appears that greater attention was paid to the participation of teachers in classroom organization and management of materials than to the frequency with which children were exposed to particular elements of the curriculum.

Participant Responsiveness

Few items within all the included fidelity instruments captured teacher responsiveness to the implemented curriculum. There are, however, several cases in which instruments were augmented with teacher interviews. Three of the measures contained teacher interviews. In addition, two instruments required observers to examine teacher reports or records, and one suggested that observers ask teachers for clarification on items that were difficult to evaluate. Both the *Creative Curriculum* and *Bright Beginnings* measures included a teacher interview with 15 and 5 questions respectively. These questions varied in focus and depth. Questions asked teachers to identify components of curriculum they used and if they would use these in the future. They also asked teachers to describe the strategies they used to scaffold children's learning and the ways in which they incorporate instructional activities into the class day. Such questions allowed observers to become more aware of the extent to which teachers valued and endorsed key components of the curriculum. Unfortunately, only three measures contained teacher interviews as a means to identify participant responsiveness.

In much the same way, most of the measures virtually ignored children's response to the instructional and environmental elements associated with the respective curriculum. Only three measures included items that specifically targeted child behaviors. The *Project Approach* fidelity instrument was anomalistic in that 5 of 26 items (19.2%) asked observers to assess children's behaviors (e.g. language, involvement, etc.). None of these five items, however, made direct references to the curriculum. Therefore, the *Project Approach* fidelity checklist required observers to evaluate children's response to general instruction, but did not ask observers to gauge children's responses to particular elements of the *Project Approach* curriculum. Likewise, *EMCO* and the PCER checklist used by the project that evaluated both the *Doors to Discovery* and *Let's Begin with Letter People* curricula contained a few items that assessed child responsiveness. Again, none of these items directly referred to the respective curricula. Despite the inclusion of child related items in these three measures, the vast majority of measures included in this analysis essentially omitted this element of fidelity, as both teacher and child responsiveness were underrepresented.

Discussion

The analysis of preschool curriculum fidelity measures presented in this paper demonstrates reveals limitations in how these measures actually evaluate fidelity of implementation. Analysis reveals that the majority of measures and the items within each of the fidelity measures over-represented general instructional quality. Few items and measures referred directly to unique elements of the curriculum; consequently they provided an inadequate gauge of implementation fidelity.

The results of this study illustrate trends in measuring fidelity in a micro-analytical manner in that the specific fidelity measures were analyzed rather than reviewing the ways in which researchers reported fidelity of implementation. These results parallel the conclusions reached in the literature reviews in both the fields of behavioral psychology and K-12 education. Both the review by Dusenbury and colleagues (2003) of drug prevention studies and the synthesis of K-12 curriculum studies by O'Donnell (2008) indicated that few studies measured implementation fidelity. Authors of both studies asserted that no consistent and useful conceptualization of fidelity has been established. Problems concerning the lack of prevalence and quality of fidelity measures have plagued the fields of education and behavioral interventions. Item-by-item analysis of this smaller sample of instruments representing 10 preschool projects reveals a comparable pattern in preschool curriculum research. Only 5 of the 15 checklists in this study contained items that assessed curriculum specific adherence. Additionally, no single measure contained a combination of items that would enable observers to provide data on adherence and exposure, as well as participant responsiveness.

An ideal measure of implementation fidelity used in preschool projects like those funded in the PCER study would assess the frequency and quality with which teachers' deliver structural and instructional elements that are unique to the curriculum. The use of such effective measures would enable researchers to confirm differences between the treatment and control conditions. Hulleman and Cordray (2009) demonstrated the importance of confirming differences between conditions by quantifying the achieved relative strength (ARS) of a motivation intervention in both a controlled lab setting and a less proscribed classroom environment. The authors cautioned that ARS, the actual strength of an intervention calculated as the difference between the treatment and control conditions, might be substantially different than that of the theoretical intervention delivered in ideal settings. It is only through the deliberate use of well-verified fidelity measures that the achieved relative strength can be calculated. When those measures are employed properly in both treatment and control conditions the difference indicates how closely the actual treatment matches expectations and signifies the difference between the true implementation of the treatment and the true implementation of the control.

The results of this analysis revealed that the measures used by the 10 projects did not adequately represent elements specific to the curriculum. Two primary challenges are involved in the attempt by researchers and evaluators to create an effective curriculum fidelity instrument. For one, researchers must be thoroughly aware of the unique characteristics of the curriculum used in the intervention. A well-developed conceptualization of the change model that fully represents the intervention, primary constructs, causal elements of the implemented curriculum, related components, and desired outcomes is a necessary precursor to effectively measuring implementation fidelity. Moreover, it is mandatory for the instruments used to assess levels of implementation fidelity to specifically evaluate teachers' delivery of those components in order to confirm that the essential elements of the change model are indeed present throughout the intervention.

Second, in addition to identifying key components of the curriculum and the intervention overall, classroom observers assigned to complete the fidelity instruments must be extremely familiar with the curriculum. When classroom observers evaluate teachers, they are required to make conjectures about complex matters such as the quality and breadth of instruction, the intent of the teacher, and the nuances of complex curricula. Thus, it is challenging for observers to accurately capture these elements inherent in the physical environment of the classroom and in the instruction delivered by teachers. Not only do observers need to interpret the behaviors of the teachers but they must also possess a comprehensive understanding of the critical elements inherent in the curriculum.

For the analysis presented in this paper, when coding measures at the item level, both the initial coder and verifier had difficulty attaining reliability within the “quality of delivery” category. The coder and verifier were required to draw one of two conclusions: items either objectively assessed the presence of materials or instructional strategies, for instance, or they represented more pedagogical, critical aspects of learning such as asking if the teacher were reading a book written at an appropriate level for the children. Ambiguity in the purpose of the items themselves led to ambiguity in the ways they were coded. In many cases, items did not fall clearly into one of these two scenarios and the coder and verifier were often faced with assigning a code that they felt was the better, but not the best option. Issues in reliability with items related to causal components and quality of delivery highlight how difficult it is for measures to represent such a complex phenomena.

Limitations

This analysis examined the ways in which fidelity is measured in preschool research through the use of a small sample of specific classroom observation measures. There are other ways for researchers in any given intervention to measure the elements of fidelity highlighted in this analysis. This paper reports on instruments submitted as fidelity measures for representation of adherence, exposure, and participant responsiveness. Studies included in this analysis may have employed additional ways of evaluating fidelity beyond these particular classroom observation measures.

For one of the 10 projects included in this analysis, Justice, Mashburn, Pence, and Wiggins (2008) provided more detail on implementation fidelity. Justice and colleagues measured children's exposure to targeted instructional techniques. In addition to using the 45-item fidelity checklist coded for this analysis and teacher-submitted weekly plans, researchers also gathered 50-minutes of video from each classroom. These videos were then coded for frequency of use of each of the seven-targeted techniques. Researchers coded each technique as not being used, used once, used two or three times, or used four or more times by the teacher. Additionally, data were collected on child attendance. Results indicated that greater exposure to techniques observed through videotapes and higher attendance had the strongest effects on child language outcomes. This level of evaluation provides a more accurate picture of exposure than does the sole use of the fidelity instrument reported in the NCER report.

In similar ways, Assel, Landry, Swank and Gunnewig (2007), researchers of the Texas studies indicated that teachers reported more ease in implementing *Let's Begin with Letter People* than *Doors to Discovery* due in large part to a single, more convenient and comprehensive teacher guide. Although few additional details about the teacher report were given in the article, there was an implication that researchers gathered information about levels of teacher responsiveness through personal communication with the teachers. Teacher perceptions were also gathered by Pence, Justice and Wiggins (2008) who used a teacher questionnaire to capture teachers' perceptions of the quality of the delivery of and their level of comfort with the curriculum. Although this questionnaire was submitted anonymously and did not provide data at the individual teacher level, a picture of the overall level of teacher responsiveness was produced. In these two cases, researchers went beyond the use of classroom observation and coding sheets to gather data on the level of teacher enthusiasm and endorsement. Other opportunities to measure implementation fidelity exist. The analysis presented in this paper looks specifically at the ways classroom observation checklists do or do not represent fidelity.

Next Steps

As noted earlier, any measure of fidelity is intimately tied to the change model representing the intervention. Further analysis of how successfully researchers and developers conceptualize change models related to their intervention and how successfully they create those models are both related to the development and use of fidelity measures. In this study I have highlighted problems in the accuracy of measuring fidelity. However, the core basis of the issue may be related to ambiguity in change models themselves and the lack of identification of causal components in the intervention.

CHAPTER IV

THE DEVELOPMENT AND APPLICATION OF FIDELITY AND QUALITY MEASURES IN A PRESCHOOL CURRICULUM INTERVENTION

Background

The goal of many preschool interventions is to better prepare children for the academic demands of school. Curriculum-based interventions endeavor to improve the ways in which teachers deliver instruction and developers of curriculum interventions often base curriculum design on assumptions about teaching and learning, build these into the content and structure of the activities, and provide teachers guidance that is designed to help foster learning in ways consistent with the curriculum. Program directors and practitioners often select curricular packages based on researchers' claims that they are effective in increasing children's academic development. Yet, in order for preschools programs to experience success, users of the promising curriculum must be able to identify the essential instructional and structural components, recognize the specific goals of the curriculum, and deliver the level of quality with which the curriculum must be implemented to accomplish these goals.

Researchers and curriculum developers must explicate the causal elements of a curriculum in order to identify intervention components responsible for the desired change. When elements are identified, deliverers of the intervention gain a better understanding of their role. A clear, accurate conceptual model acts as a guide to practitioners as they venture to deliver curricula in a manner intended by the developer. Likewise, researchers are able to use this model as a guide when creating measures that assess teachers on the degree to which they implement the causal components. Lastly, conceptual models allow researchers to determine how much the intervention in the treatment condition differs from business as usual in the control condition.

Researchers often grapple with the multidimensional nature of teaching and learning and the inherent difficulties in implementing effective educational interventions. Models of interventions, as well as the measures developed to ensure teachers have adequately implemented the causal components, can be equally complex. Despite efforts to create effective measures of fidelity, the instruments produced by many curriculum developers and educational researchers and the methods in which they have been used have frequently fallen short.

Three major problems have arisen in recent attempts by educational researchers to effectively employ fidelity measures in experimental interventions. Few educational interventions illustrate a conceptual model that identifies critical components and includes the relationship between these and targeted outcomes. When measures are used to assess levels of implementation among program deliverers, they rarely represent the critical components of the intervention illustrated in the conceptual model and are seldom subjected to tests of reliability and validity. Secondly, researchers who do account for issues of fidelity when interpreting the effectiveness of an intervention often do so on a superficial level by defining fidelity as a simple, one-dimensional concept. Measures of fidelity must represent both process-related and structural elements inherent in the intervention and account for the potential variation in fidelity across a comprehensive intervention as well as within specific aspects of it. Lastly, researchers often intermingle naturally distinct constructs of teaching quality and fidelity of implementation measures by mistaking global measures of instructional and environmental quality, such as Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008) and The Early Childhood Environment Rating Scale (ECERS; Harms & Clifford, 1980) as measures of fidelity. It is important for researchers to evaluate general teaching quality as it may relate to the ways in which teachers faithfully implement a new curriculum; however, general teaching quality and fidelity to a specific curriculum are necessary, but separate, constructs.

Creating a Conceptual Model

There is a paucity of conceptual models presented in educational research and a dearth of fidelity measures used to assess the ways in which teachers deliver structural and instructional constructs depicted in models. In 2003, Mowbray, Holter, Teague, and Bybee asserted "interventions are expected to specify the model—a scientifically sound program theory or theory of action, explicating mechanisms through which the program will achieve its desired outcomes" (p. 315). Despite their edict nearly a decade ago, few researchers heeded their call. Hulleman, Cordray, Nelson, Darrow, and Sommer (2009) reviewed reports presenting effects of elementary math interventions. Of the 46 studies that employed fidelity measures, only half included a complete representation of the intervention change model.

Representing Critical Components

Fidelity measures must include indicators that directly represent the mechanisms of change illustrated in any well-developed conceptual model. The indicators within a single or set of measures need to specifically determine if teachers are delivering the critical components of the curriculum within all curricular activities, with adequate levels of quality (Cordray, 2009; Hulleman & Cordray, 2009), and at the ideal frequency. Both Mowbray et al. (2003) and O'Donnell (2008) claimed that fidelity measures should include items that represent both structural and process-related elements of the intervention. They caution against using limited measures that represent a single component of fidelity.

Odom et al. (2010) heeded O'Donnell (2008) and further encouraged researchers to incorporate fidelity measures that represent both the structure (e.g., quantity of lessons delivered) and process (e.g., quality of delivery) of implementation. Odom et al. created a multiplicative composite fidelity score across elements of structure and process. In doing so, they were able to examine the patterns of implementation within these individual constructs for each of the disciplinary foci of the curriculum (i.e., literacy, math, and social skills) as well as overall. Analysis of structure and process measures individually revealed across-site differences in the frequency and quality with which teachers were likely to implement activities related to literacy, math, and social behavior. Analysis using the composite measures that combined structural and process-related items enabled the authors to confirm that overall implementation was higher in literacy activities.

Even though Odom's study (2010) broke down fidelity of implementation by disciplinary focus, it did not identify which individual instructional aspects were less likely to be implemented by classroom teachers. The analysis presented in the Odom study revealed differences in the way literacy and math portions of the curriculum were implemented, but it did not uncover the particular instructional elements within those disciplinary-focused activities that teachers were more successful in delivering. Pence, Justice, and Wiggins (2008) provided a more fine-grained analysis of patterns of implementation in their report of a *Language-Focused Curriculum* (*LFC*) intervention involving 14 preschool teachers. In this study, fidelity measures were used to assess the degree to which teachers implemented structural elements (i.e., activity context) as well as instructional processes. Instructional processes were divided into seven categories (e.g., modeling, open question) that corresponded with the instructional techniques unique to the treatment curriculum.

Establishing Reliability and Validity of Fidelity Measures

As issues of implementation fidelity come to the forefront of educational research, there is an increasing need to establish reliable and valid measures. The final step in confirming adequate levels of implementation in an educational intervention is to establish fidelity criteria by confirming the reliability and validity of the fidelity indicators (Mowbray et al., 2003). Internal consistency must be established to confirm that scaled fidelity measures are truly representing the intended construct (Durlak, 2010). The need for reliable measures has been substantiated in recently published reports involving research in early childhood education.

The most recent edition of Early Childhood Research Quarterly (ECRQ) edited by James A. Griffin (2010) included five individual reports related to issues of implementation in early childhood education. This collection of research provided a snapshot of current perspectives and approaches in defining, measuring, and analyzing fidelity of implementation. Of the five published studies, four reported calculations of internal consistency for measures of fidelity (see Baker, 2010; Domitrovich, 2010;

Hamre, 2010; Odom, 2010). Typically, researchers have reported only inter-rater reliability calculations for observational measures. However, the collection of studies in ECRQ reveals that the need to validate measures of fidelity is gaining greater acceptance.

Using Measures to Differentiate Fidelity across Components and Conditions

Many researchers fail to identify program differentiation as they frequently miss the opportunity to use measures to assess the multi-dimensional aspects of fidelity and to employ measures in both treatment and control classrooms. Griffin (2010) highlighted that the papers included in the special edition of ECRQ assessed multiple dimensions of implementation. As a collection, these studies did indeed represent multiple aspects of fidelity, however issues remain with individual studies. Specifically, in their evaluation of a readiness program geared toward Head Start children entering kindergarten, Baker, Kupersmidt, Voegler-Lee, Arnold, and Willoughby (2010) measured the relationship between teachers' characteristics and backgrounds and their participation in the intervention. Unfortunately, Baker et al. operationalized fidelity only as the frequency with which teachers participated in the designated lessons, a structural component of implementation. The work of Baker and colleagues contributes to our understanding of how teacher backgrounds and program characteristics relate to the frequency with which teachers deliver the curriculum. These authors did not, however, operationalized fidelity in a manner that allows researchers to relate these factors to multiple dimensions of fidelity such as the quality with which teachers delivered aspects of the targeted curriculum.

In many cases, researchers evaluate teachers assigned to the treatment condition, but do not identify and assess instruction that occurs in classrooms assigned to the control condition. This limited approach can provide some detail on the instructional practices of teachers assigned to the experimental group and can accurately assess the level of fidelity demonstrated by these teachers. However, when researchers fail to use equivalent fidelity measures in control classrooms, they are unable to confirm differences between treatment and control classrooms.

The importance of confirming program differentiation between treatment and comparison groups has been stressed for over a decade. Dane and Schneider (1998) made a strong appeal for educational researchers to guard against program diffusion. Mowbray et al. (2003) encouraged the use of fidelity instruments to ensure children in treatment and control classrooms have received critical elements of the respective curricula. O’Donnell (2008), as well as Hulleman and Cordray (2009), most recently reiterated the importance of determining degrees of program differentiation.

Within the last three years, two contrasting examples have emerged in preschool curriculum research (Pence et al., 2008; Odom et al., 2010). Pence and colleagues used their set of fidelity measures to assess the degree to which treatment and control teachers demonstrate methods specific to the treatment curriculum. Odom et al. (2010) effectively employed fidelity measures to represent critical elements of the treatment curriculum, but did not collect fidelity data on the control classrooms. In a comparable manner to the work by Pence et al. (2008), the study presented in this paper took a similar approach to fidelity in that teachers in both treatment and control classrooms were assessed on the degree to which, and quality with which, they deliver multiple components of the treatment curriculum, as well as the curriculum designated to be used in the control classrooms.

Distinguishing Between Quality and Fidelity

Instructional and environmental quality is an important consideration in any effectiveness study situated in an educational setting. Researchers can gain deeper understanding of intervention effectiveness when they examine how instructional quality and a teacher's ability to manage children and the environment relates to a teacher's level of implementation fidelity and subsequent outcomes of interest. The need for measures of instructional quality, classroom management, and fidelity of implementation is essential in that they are distinct, yet potentially related, constructs.

The development of effective measures of instructional quality is time consuming and expensive. Because of this and other challenges, many researchers use previously developed measures of classroom quality like the Early Language and Literacy Classroom Observation (ELLCO; Smith, Dickinson, & Sangeorge, 2002) and CLASS to differentiate the quality of instruction occurring in classrooms. Unfortunately, some researchers confuse the two constructs and mistake global classroom quality measures for measures of implementation fidelity specifically tied to a unique program or curriculum. Cautions against this confusion have been made in recent years:

> Developers of fidelity criteria also need to be aware of the fact that fidelity to program standards can be confounded with the competence of the program implementers (Clarke, 1998); skillful practitioners may implement intervention

> models better and achieve superior results (Luborsky, McLellan, Diguer, Woody, & Seligman, 1997). (Mowbray et al., pp. 327)

There are benefits to measuring the teaching ability and environmental quality of teachers and classrooms involved in any educational intervention. Knowing how well teachers are teaching can help researchers identify possible interactions between quality and fidelity which could lead to better understanding of how teaching quality and environmental resources relate to issues in implementation. For example, data collected on teaching quality can be used to identify the set of characteristics most likely shared by high implementers and efficient adopters of a new curriculum. On the other hand, it may be more important for a highly scripted program to be delivered by highly compliant teachers rather than teachers with exceptional overall teaching abilities. Measures that distinctly assess quality and fidelity are needed to identify the ways in which these two constructs are related.

A recent report of a preschool curriculum intervention revealed the ways in which quality of teaching and aspects of implementation were confounded. Hamre et al. (2010) reported data collected from measures representing dosage, adherence, and quality. In an effort to measure dosage and adherence, these researchers used teacher self-reports of the frequency and length of curriculum activities and observer reports on the extent to which teachers adhered to scripts and used materials as required. Hamre and colleagues (2010) noted the importance of evaluating teachers' quality of delivery. They used the CLASS to represent quality of delivery and fidelity to the intervention. As a result, they confounded quality of teaching, in general, with the quality of implementation.

Research Questions

In order to address several issues that arise when assessing levels of implementation and teaching quality while confirming differentiation between treatment and control conditions through the use of well-developed measures, fidelity data collected from a preschool curriculum intervention have been analyzed to address the following research questions:

- Is it possible to create a conceptual model of the intervention that reflects the instructional model inherent in the curriculum in a way that captures process and structural elements across setting in a psychometrically strong manner?
- Can a comprehensive measure allow researchers to view fidelity as something beyond a unitary construct by differentiating implementation of instructional constructs and activities across conditions?
- Can a set of tools be used to distinguish between different constructs – fidelity of implementation, teaching quality, and classroom management?

Methods

Research Site and Participants

Data in this study were collected as part of Teacher Enhanced Language and Literacy (TELL)[1] that involved a large Head Start program in a medium sized southern city. In this four-year intervention, teachers implemented the *Opening the World of*

[1] The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324E060088A to Vanderbilt University.

Learning (*OWL*) curriculum (Schickedanz & Dickinson, 2005) in conjunction with *The Creative Curriculum* (Dodge, Colker, & Heroman, 2002) in 36 treatment classrooms. Teachers in sixteen additional classrooms used *The Creative Curriculum* exclusively as the primary curriculum (i.e. business as usual). The Head Start program targeted for this intervention consisted of 13 individual centers organized into six clusters. Program administrators organized centers into clusters so that centers and classrooms could be more conveniently and effectively managed. In several cases clusters contained both larger centers with more classrooms and smaller centers located nearby. Five of the six clusters contained three Head Start centers and the remaining cluster contained only two centers.

The 52 total classrooms were assigned to one of two conditions in the experiment. The majority of data (e.g., classroom observations) were collected from a single lead teacher in each of the 52 classrooms. However, there were five classrooms in which the lead teacher changed during the year of implementation. Consequently, there were some classrooms in which video data were collected on assistant teachers. Therefore, data were collected on more teachers than the number of classrooms would typically indicate. In total, videotaping targeted 65 teachers representing these 52 classrooms.

Three activities were videotaped at two time points during the school year for each of the 52 classrooms. Because the unit of analysis for this study is the classroom, teacher scores were aggregated as a representation of classroom implementation. Specifically, observers evaluated two sessions of each videotaped activities. Scores from these two sessions were then averaged to create a classroom score.

Implementation of the Treatment Curriculum

OWL is a comprehensive curriculum with emphasis on developing language and literacy skills in preschool children. The curriculum spans a full year and is separated into six units focused on themes ranging from *Family* to *Things that Grow*. The preschool day is organized into six activities: Morning Meeting, Centers, Small Group (SG), Book Reading (READ), Group Literacy Instruction (GLI), and Let's Find Out About It/Let's Talk About It.

Of all the activities included in the *OWL* curriculum and implemented by teachers assigned to the treatment condition, this study includes analysis of data obtained from (a) Small Group, (b) Book Reading, and (c) Group Literacy Instruction. These three activities in particular are designed to increase children's exposure to concepts of language and literacy considered to be integral to their development. Because of this, it is important to understand how well and to what extent teachers are implementing lessons within these three activities.

Small Group (SG). Teachers were asked to divide children into three small groups of approximately five to six children. *OWL* requests three small group activities for each day over the course of a three-day period. All children are expected to participate in each of the three groups over the 3-day schedule. The focus of activities within Small Group varies considerably as some lessons relate to language and literacy and others are associated with mathematical and scientific concepts. The *OWL* manual suggests that small groups last 20 minutes.

Book Reading (READ). The *OWL* curriculum contains selected storybooks for each of the six thematic units. Each unit includes 5 storybooks. Teachers are expected to read each book four separate times. Each reading has a different purpose. The purpose of a first read is to introduce the story to the children and present the main events of the story as well as new vocabulary in an engaging way. In the second reading, teachers are asked to encourage children to recall events and characters while continuing to use and define suggested vocabulary. In the third and forth readings, teachers are asked to encourage children to chime in during the reading and to act out scenes from the story. Book reading should occur daily and last for approximately 20 minutes. This study examines the ways teachers in the treatment condition perform the first and second reading. Teachers should be very engaged in exposing children to word meanings as well as analytical elements of the story. Therefore, the first two readings of each story provide rich data from which to determine the ways in which teachers follow the guidelines of the curriculum. In one case, however, data were collected from a teacher's third reading, as no video of her first or second reading was available.

Group Literacy Instruction (GLI). This activity consists of several different elements such as songs, poems, and games. GLI is used to teach phonological and alphabet awareness as well as letter recognition. GLI lessons are scheduled to occur daily and should last up to 20 minutes per session.

Ongoing Delivery of *The Creative Curriculum*

For several years, the Head Start program involved in this study used *The Creative Curriculum* as the primary curriculum. *The Creative Curriculum* is a comprehensive preschool framework that stresses the socio-emotional, cognitive, and physical development of children. Program administrators communicated to the principal investigators that teachers in the program understood and valued multiple elements of *The Creative Curriculum*. Given that, teachers assigned to the treatment condition were asked to implement *OWL* while also maintaining core elements of *The Creative Curriculum*. Teachers assigned to the control condition were asked to continue teaching in a manner consistent with the objectives of *The Creative Curriculum*. Because of this agreement, it was expected that the instruction in treatment classrooms would represent components of both *OWL* and *The Creative Curriculum*, whereas the teaching in control classrooms would only reflect elements of *The Creative Curriculum*.

Process of Implementation

The implementation of *OWL* occurred over the course of the first two years of a four-year study. Teachers attended several days of professional development a few weeks before the start of school in the first year. Year 1 was viewed as a soft launch in which teachers became familiar with the structure and emphasis of the *OWL* curriculum and adopted elements of the curriculum at their own pace. Education specialists were trained to support teachers in the use of instructional techniques particular to *OWL*. Full implementation of *OWL* by the teachers occurred in the second year of the project. Professional development was ongoing and literacy coaches were hired to provide additional support for *OWL* teachers.

Instruments

Instruments were created to assess teachers on the frequency with which they implemented essential elements of the curriculum across the day. A series of checklists was developed to represent three daily activities: Book Reading, Small Groups, and GLI. Each checklist consisted of four categories: 1) *OWL*-specific fidelity items, 2) *The Creative Curriculum* fidelity items, 3) items representing general instruction, and 4) items representing general management.

***OWL* fidelity checklist.** With the guidance of the co-author of *OWL*, instructional items identified as uniquely related to the objectives of *OWL* were listed in the *OWL*-specific items section. The number of *OWL*-specific items in that section varied across activities. The Book Reading, Small Group, and GLI checklists contained 9, 8, and 7 items respectively. Items representing instructional subconstructs evaluated teachers on both the degree to which teachers delivered particular curricular elements as well as the quality with which these were delivered. Teachers received credit only when they met or exceeded ideal levels of quality. For example, a Book Reading item written as "teacher gives or elicits accurate definitions of *OWL* targeted vocabulary words, seven times or more" determines if teachers defined targeted words as the curriculum required and if teachers were doing so to the ideal extent. This item combines to evaluate teachers on their adherence to the curriculum and the quality and intensity of their delivery.

***The Creative Curriculum* fidelity checklist.** The checklists representing Creative Curriculum items compiled for the three corresponding *OWL* Activities were gathered using *The Creative Curriculum for Preschool* Implementation Checklist ® (Dodge, Colker, & Heroman, 2003). The author of this paper asked the Director of Education and two education specialists who were employed by the Head Start program and involved in the intervention, to identify items representing instructional and environmental elements included in *The Creative Curriculum for Preschool Implementation Checklist* ® that they would expect to see in control classrooms during the respective activities. When all three individuals agreed on an item, that item was included in *The Creative Curriculum* section of the checklist for the associated activity. For example, the Director of Education and both education specialists indicated that they would expect teachers using *The Creative Curriculum* to "Guide children in putting away materials where they belong (e.g., draw attention to the labels; play games to sort materials on shelves)" during Small Group activities. Therefore, that item was included in the section of the Small Group fidelity checklist representing *The Creative Curriculum* specific items. Similar to the *OWL* specific sections, the number of *The Creative Curriculum* specific items varied across activities. The Book Reading, Small Group, and GLI checklists contained seven, seven, and four items, respectively.

Instructional quality checklist. A separate but essential component of the instrument includes items that measure teachers' instructional quality. These items represent general instructional elements appropriate for preschool classrooms. This category was developed by identifying items that were common across the original *OWL* fidelity items (created by the developer) and *The Creative Curriculum for Preschool Implementation Checklist* ®. The Director of Education and educational specialists employed by the Head Start program involved in the study identified items of *The Creative Curriculum for Preschool Implementation Checklist* ® that they would expect to see teachers deliver in specific daily activities. The items identified by the Director of Education and educational specialists were then compared to items originally included in the activity-specific *OWL* fidelity checklists created by curriculum developers. Those items that were common across the two checklists were flagged and compiled as items representing general instructional quality. Items representing general instructional quality were assembled for each of the three activities highlighted in this study. For example, it is important for teachers to hold the book he or she is reading so the children can see during story time, but this instructional move is not unique to any one curriculum. The Book Reading, Small Group, and GLI checklists contained seven, six, and four items representing instructional quality, respectively.

General management checklist. The final component of the instrument contains items used to assess classroom management. Similar to the items representing instructional quality, these management-related items are also valued across preschool curricula. This measure of classroom management provides a means to analyze the ways in which teachers' instructional abilities are related to their management skills, as well as the ways in which the two constructs are related to levels of fidelity. Teachers were rated on the same nine items for each of the three activities. These items measure the degree to which teachers are able to maintain effective classroom management and child engagement. Measures of classroom management can be used to identify any differences in instructional climate that may occur in classrooms using varying curricula (see Appendices F, G, and H for checklist items by curricular activity).

Reliability of Implementation Measures

For each activity, teachers in the 52 classrooms were videotaped two times (i.e., fall and spring) over the course of the 2007-2008 academic year. Consequently, 104 videos per activity were coded by multiple research assistants for each teacher's fidelity of implementation to both *OWL* and *The Creative Curriculum*, as well the teacher's quality of general instruction and classroom management.

For each activity, a series of training workshops was conducted to introduce research assistants to the coding schema and to establish initial reliability between individual coder(s) and the verifier(s). The initial training sessions were considered complete when coders and verifiers arrived at the 85% or greater exact percent agreement. Twenty percent of the entire collection of videos was randomly selected for reliability, thereby providing a corpus of videos from which coders and verifiers could maintain reliability throughout the coding process. Once 85% reliability was initially established, coders began independently coding four video sessions. Both the coder and a verifier then coded a fifth video session. If the two individuals were at least 85% reliable on all items across the measure for the fifth video session, the coder would then be permitted to code an additional four video sessions independently. The process was then repeated until all coding was completed. In cases when the coder and verifier did not meet the 85% cutoff on the fifth video session, they would code another video from the corpus of reliability videos and compare scores. A coder was not permitted to continue until she reached the acceptable level of reliability (i.e. 85%) with the verifier. In cases of disagreement on individual items, the coder and verifier came to a consensus on the accurate score.

The number of coders and verifiers also varied across activity. Six individual research assistants (all doctoral students) and a single verifier (also a doctoral student) coded the Small Group video sessions. Three individual undergraduate students and two verifiers (both doctoral students) coded the GLI video sessions. Lastly, one graduate student and two verifiers (both doctoral students) coded the book reading video sessions (see Table 7 for reliability scores by overall instrument, category, and activity).

Table 7

Inter-rater Reliability Calculations by Category

Activity	Overall (%)	*OWL* (%)	*CC* (%)	GI (%)	GM (%)
Small Group	87	88	89	84	88
GLI	86	89	88	90	80
Book Reading	90	91	92	91	86

Results

Is It Possible to Create a Conceptual Model that Reflects Unique Curricular Elements in a Psychometrically Strong Manner?

The first of three research questions addresses the challenges in creating a comprehensive conceptual model that represents critical elements of the intervention and the corresponding indicators used to assess teachers' delivery of the intervention. The conceptual model for the intervention presented in this study is directly affiliated with *Opening the World of Learning* (*OWL*), the treatment curriculum. The model was created by collecting the original fidelity items, separated by activity, that were included in the curriculum package. The author of this paper, in collaboration with one of the authors of the curriculum, revised and enhanced the fidelity items to better represent essential instructional items endorsed by *OWL*. Subsequently, they completed two additional steps. First, they determined which items represented process- and structural-related components. Second, they grouped the sets of process-related items into instructional subconstructs (e.g., items used to assess teachers' support of language development in Small Group). The author then conducted internal consistency reliability testing at the activity-level (e.g., all items included in Book Reading) and at the subconstruct-level (e.g., all items included in the language development subconstruct within Book Reading).

This model accounts for three levels of the intervention: (1) curriculum-level, (2) activity-level, and (3) instructional- and structural-subconstruct level (see Figure 3). The curriculum-level of the model includes the three primary daily activities: Small Group, GLI, and Book Reading. The activity-level of the model includes measures of process- and compliance directly related to each activity. Lastly, the instructional- and structural-subconstruct level contains groups of fidelity indicators contained in the observational checklist for that activity (e.g., S1 refers to the first item in Small Group Checklist). There are some commonalities in the subconstructs within each of these activities, yet the delivery of specific curricular elements is also unique to the different activities and settings.

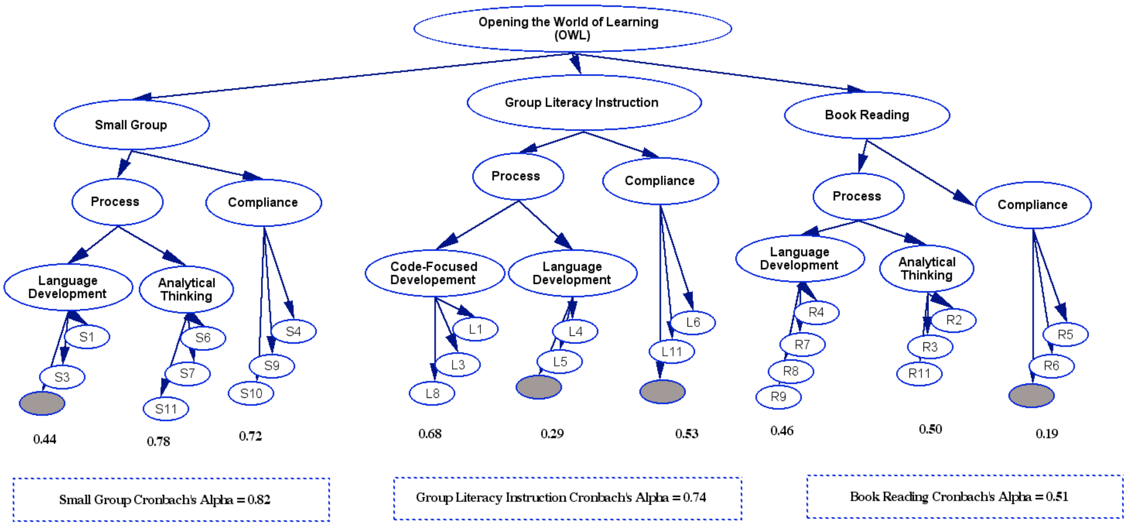

Figure 3. Change model for *OWL* curriculum intervention.

Reliability testing was conducted on the fidelity checklist used in this study to measure teachers' implementation of *OWL*-specific elements. As evident in the change model (refer to Figure 3), specific fidelity indicators were used to measure both process-related and structural (i.e., compliance) elements of implementation. The process-related group of items was then divided into instructional subconstructs linked to that specific curricular activity. Ideally, three or more indicators would comprise each subconstruct. In this case, only five of the nine subconstructs contained at least three fidelity indicators.

The author assessed internal consistency reliability by calculating a Cronbach's coefficient alpha for all items within each activity and for all items grouped by instructional and structural subconstruct within each activity. Figure 3 also shows that overall scales for Small Group, GLI, and Book Reading activities were 0.77, 0.69, and 0.57 respectively. The calculated Cronbach's alphas for instructional constructs within activities ranged from 0.19 to 0.78. Any construct with an internal consistency less than 0.40 was dropped from the final analysis. The results of internal consistency tests confirm that the majority of fidelity indicators reliably represent the critical elements contained within the conceptual model at both the activity- and subconstruct-level. Therefore, the fidelity measures used in this study provide a valid picture of implementation levels demonstrated by teachers.

Given the results of the internal consistency reliability testing, it is possible to conclude that the creation of accurate and valid models and associated fidelity indicators is complex and challenging. Although this model does represent key components of the curriculum, it has some limitations. More indicators are necessary to adequately represent constructs of interests. For example, only seven indicators were used to evaluate fidelity to GLI at the activity-level and only two indicators represent language development in Small Group at the instructional- and structural- subconstruct level. In addition, there were unacceptable internal consistency reliability values at the instructional- and structural- subconstruct level. A Cronbach's *alpha* of 0.19 was calculated for the group of compliance items in Book Reading. Together these indicators are not measuring "compliance" at an acceptable level. Because of low levels of internal consistency, several subconstructs were dropped from the analysis, thus weakening the overall accuracy of the model.

Can a Comprehensive Measure Allow Researchers to View Fidelity as Something Beyond a Unitary Construct?

The comprehensive set of fidelity measures used in this study provided a nuanced understanding of implementation, revealing differential levels of implementation between teachers assigned to the treatment and control groups. The author of this paper used measures to evaluate teachers' implementation through three different perspectives: overall fidelity across activities, fidelity by activity, and fidelity by structural and instructional subconstructs within activities.

A curriculum-level view: overall curriculum implementation by condition. As expected, teachers assigned to implement *OWL* had significantly higher scores than control teachers in overall fidelity to *OWL* across curricular activities. Treatment teachers implemented 50.2% (SD = 10.6) of the critical elements of *OWL*, averaged across two times and across three curricular activities. This level of implementation was less than ideal, yet it was significantly higher than that of the control teachers (M = 27.5%; SD = 8.1).

The overall rates of implementation of *CC*, the curriculum historically used by the Head Start program involved in the study, were consistent across conditions. Both groups of teachers implemented approximately 50% of the elements determined by program staff to be essential to *CC*. Through professional development and coaching, research and program staff supported the two groups of teachers in delivering a different instructional package, one with a combination of *OWL* and *CC* and the other with *CC* exclusively. Teachers in both conditions, nonetheless, delivered equivalent amounts of *CC*.

The view of overall fidelity provides a general picture of differentiation between the two experimental conditions. Despite the less than ideal levels of *OWL* implementation demonstrated by treatment teachers, the fidelity measures used in this analysis provide evidence that children in treatment and control classrooms had distinct educational experiences in that children in the treatment condition received instruction unique to *OWL* as well as elements of *CC*. Children in the control condition were exposed to equal degrees of *CC*, yet received limited exposure to *OWL*.

Zooming in: curriculum implementation at the activity-level. Unique patterns of implementation emerged, however, when examining levels of fidelity to both *OWL* and *Creative Curriculum* within specific curricular activities. Treatment teachers implemented *OWL*-specific elements of Small Group (M = 62.67) at higher rates than in any other activity. However, control teachers were also successful at implementing elements of Small Group instruction that were endorsed by *OWL* (M = 28.91) (see Table 8 for complete data). Despite delivering more than half of the curriculum-specific items in Small Group, treatment teachers did not differ much from the teachers assigned to the control condition, in which no *OWL*-specific instruction was expected. Thus, the apparently stronger levels of fidelity in the *OWL* classrooms were negated by the fact that the instruction delivered by treatment and control teachers in Small Groups was not as different as researchers predicted at the start of the study. In spite of this, the levels of fidelity in Small Group demonstrated by individual treatment teachers provide evidence that portions of the curriculum were implemented with a higher degree of fidelity. The weakest teacher participating in the intervention in this setting implemented 12.5% of OWL-specific items. However, 23 of the 36 treatment teachers implemented at least 60% of the *OWL* curriculum. Of those, two teachers reached perfect fidelity (100%).

Overall levels of fidelity by treatment teachers during GLI were similar to Small Group. On average, treatment teachers implemented roughly 53% of *OWL*-specific items, yet control teachers delivered 21.4% of the OWL curriculum. Although this difference between treatment and control is statistically significant (t = -5.90, p = .000, DF = 50), these levels of implementation do not differ to the extent that researchers expected. Thus, teachers in these experimental conditions differed statistically, but not to the extent that would be expected if the curriculum were being taught with a high degree of fidelity. In fact, one individual control teacher demonstrated a higher rate of *OWL* implementation (i.e., 50%) than did 14 treatment teachers (i.e., < 50%).

Data suggest that treatment teachers had the most difficulty in implementing *OWL* during Book Reading (M = 32.1). Differences between the *OWL*-specific instruction delivered in the treatment classrooms as compared to the control classrooms were also at their lowest during Book Reading. Additionally, *OWL* teachers had lower rates of implementation of *CC*-specific elements during Book Reading than did the control teachers. Low levels of fidelity to Book Reading by treatment teachers on the whole were underscored by less than ideal performances of individual teachers. Ten of the 36 treatment teachers implemented 27% or fewer of *OWL*-specific elements during Book Reading. Moreover, the two teachers with the highest level of fidelity in this setting delivered only 59.1% of the essential elements of the curriculum.

During Book Reading, control teachers delivered approximately 18% of instructional elements unique to *OWL* and roughly 51% of elements associated with *CC*. Explanations for why control teachers implemented elements of the treatment with higher rates than expected, and why their rates of *CC* implementation were lower, are varied and complicated. However, it is meaningful to note that the tools used in this analysis suggest the preschool children involved in the study were having similar experiences during Book Reading regardless of whether they were in a treatment or control classroom. Hence, methods of Book Reading in the treatment classrooms were not as different as researchers would have anticipated.

Table 8

Percentage Scores for Fidelity for each Curricular Activity by Condition

	Small Group		GLI		Book Reading	
	Treatment	Control	Treatment	Control	Treatment	Control
	Mean (SD)		Mean (SD)		Mean (SD)	
OWL FOI	62.7 (20.6)	28.9 (14.2)	52.6 (18.5)	21.4 (15.2)	32.1 (11.4)	18.5 (9.0)
CC FOI	52.8 (15.5)	54.5 (20.5)	94.8 (9.6)	80.5 (21.4)	42.3 (8.8)	51.3 (13.6)

The analysis of fidelity at the activity-level provides a more specific picture of differentiation between treatment and control classrooms in how teachers delivered instruction within each of the three curricular activities. Although rates of *OWL* implementation were significantly higher for treatment teachers in all three activities, data revealed more similarities between conditions than did the curriculum-level view of fidelity discussed earlier. In several cases, the degree to which treatment and control teachers differed in their implementation was less than ideal.

A nuanced view: analysis of structural and instructional subconstructs. The previous view of fidelity that examined specific curricular activities showed that patterns of implementation varied dramatically among teachers and across activities, yet a third and final analysis shows that patterns of implementation also varied with regard to specific instructional and structural elements highlighted in *OWL*. Researchers affiliated with this study hypothesized that some critical elements of the curriculum would be crucial in supporting language learning and would be more challenging for teachers to deliver. This was thought to be a likely scenario for the teachers involved in this study as they were asked to deliver multiple instructional methods across several activities in the course of a day.

Because the same set of validated tools was used to observe and assess both treatment and control teachers on the degree to which they implemented particular instructional methods and structural components, it was possible to focus on nuanced differences between the conditions. Researchers believed that treatment teachers would deliver all instructional methods and structural elements associated with *OWL* more often and with higher quality than control teachers. In several cases, analysis of fidelity data provided evidence for this assumption. Treatment teachers showed higher rates of fidelity in most of the instructional subconstructs related to Small Group and GLI activities. In particular, they were more supportive of children's language development and analytical thinking in Small Group activities. In addition, treatment teachers delivered code-focused instruction during Group Literacy Instruction with higher fidelity than did control teachers.

The set of fidelity measures produced unexpected results when data collected from the Book Reading activity were analyzed. There were no statistical differences in the ways that treatment and control teachers supported children's language development and analytical thinking while reading storybooks. Neither the selection of books nor the manner in which they were presented to the children produced differences in these instructional subconstructs. This counterintuitive result can be explained in one of two ways. For one, Book Reading activities may have been too difficult for treatment teachers to implement with adequate levels of fidelity. Therefore, the ways in which they presented instruction was comparable to the styles and methods used by control teachers. Problems with the accuracy of the Book Reading fidelity measure in representing unique elements of *OWL*, however, could be an alternative explanation for the lack of difference observed between these two groups of teachers. The tests of internal consistency resulted in lower alphas for Book Reading fidelity and may indicate weakness in this construct in particular.

Data collected on teacher fidelity also represented structural elements of implementation, referred to as compliance in this study. In a similar way to teachers' implementation patterns of instructional elements, teachers' levels of compliance to the treatment curriculum followed a mixed pattern. Treatment teachers demonstrated greater compliance than control teachers during GLI activities in that they presented activities that were specified by lesson plans for the recommended length of time. On the other hand, treatment teachers were statistically equivalent to control teachers when adhering to structural elements of Small Group activities. Treatment teachers did not consistently select the most challenging lessons scheduled for that day and engage children for the recommended length of time for Small Group activities. Again, lower rates of implementation may point to a lack of change in teacher behavior, but it may also indicate weakness in the measure itself.

The comprehensive and multidimensional set of fidelity measures used in this study provided researchers the opportunity to examine implementation through a variety of perspectives. When defining fidelity as a unitary construct by assigning each teacher a single value of fidelity across activities (i.e., at the curriculum-level), results suggest that treatment teachers implemented the intervention curriculum at significantly higher rates than did control teachers. When examining levels of fidelity within particular curricular activities, data revealed that the treatment teachers were the better implementers in all three of the activities but levels of differentiation between conditions varied by activity. Furthermore, this perspective provided evidence that treatment teachers had more difficulty implementing Book Reading than any other activity. The final and most nuanced analysis which reported fidelity at the instructional- and structural-subconstruct level revealed a different picture of fidelity by exposing instances in which treatment teachers did not implement instructional and structural elements unique to the intervention curriculum at higher rates than the control teachers.

Can a Set of Tools Be Used to Distinguish Between Fidelity of Implementation, Teaching Quality, and Classroom Management?

The third research question in this study examines the way in which a set of tools can be used to distinctly identify how well teachers deliver general instructional elements, manage the classroom and engage children in learning activities, and implement a new curriculum. These are three separate but related qualities, and observational measures were used in this study to identify differences among teachers, across activities, and across conditions.

Levels of instructional quality and classroom management. Varying patterns emerged upon examining levels of teaching quality when using the general instructional quality and general management checklists. Levels of instructional quality were lowest during GLI for treatment (M = 49.5, SD = 14.3) and control teachers (M = 33.2, SD = 10.6) as compared to other activities (see Table 9). In contrast, both groups of teachers delivered the highest quality of teaching during Book Reading. When comparing levels of instructional quality between conditions, treatment teachers demonstrated significantly higher levels than control teachers overall and within all three of the targeted curricular activities.

When evaluating teachers on their ability to manage the classroom and engage children through the general management checklist, levels remained relatively high and stable across conditions and across activities. This pattern diverged from patterns of teachers' general instructional quality. There were no significant differences between teachers by condition when levels of general management were calculated across activities and for each activity. Results show, however, that teachers in both groups had the most difficulty in managing children in Small Group (see Table 9). Both GLI and Book Reading are whole group activities that involve all the children. Small Group typically consisted of the teacher and 5-7 children, yet management was the greatest challenge in this smaller setting.

Table 9

Percentage Scores for General Instructional Quality (GI) and General Management (GM) Overall and for each Curricular Activity by Condition

	SG		GLI		READ		Overall	
	Tmt M (SD)	Ctrl M (SD)	Tmt M (SD)	Ctrl M (SD)	Tmt M (SD)	Ctrl M (SD)	Tmt M (SD)	Ctrl M (SD)
GI	77.3 (11.7)	70.6 (8.9)	49.5 (14.3)	33.2 (10.6)	91.4 (10.8)	90.6 (10.5)	73.3 (7.8)	64.1 (5.7)
GM	77.3 (16.6)	75.5 (15.1)	82.2 (13.2)	85.1 (11.8)	84.6 (12.5)	87.7 (9.4)	81.3 (11.0)	82.5 (6.8)

Relationship between instructional quality, classroom management, and fidelity. Beyond varying levels of instructional quality and management across conditions, these levels also related to each other and to curriculum-specific fidelity in different ways for treatment and control teachers. For control teachers, there was virtually no correlation between general instructional quality and general management (see Table 10), giving support to the claim that teachers who manage the classroom and engage children well are not always delivering high quality instruction. For example, one control teacher implemented roughly 34% of *CC*, earned mid-range scores for general instructional quality (61.2%), and demonstrated relatively high levels of classroom management (90.7%). Alternatively, another teacher implemented 58% of *CC*, one of the highest scores in the sample, yet earned comparably low scores in instructional quality (58.6 %) and management (77.8%).

For teachers in the control condition, there were no significant correlations between their instructional quality (i.e., good teaching) and *OWL* fidelity. This lack of relationship was not unexpected, as control teachers were not asked to implement *OWL*. Nonetheless, there was also no true relationship between “good teaching” and control teachers’ levels of fidelity in implementing *CC*, a curriculum that had been used in those classrooms for several years.

Treatment teachers’ levels of instructional quality and classroom management, on the other hand, did have significant, independent relationships with both *OWL* and *CC* fidelity (see Table 11). In addressing the question of whether a set of measures was able to capture differences in the ways teachers had demonstrated higher quality in instruction and management and the degree to which they implemented the curriculum, correlational analysis suggests that these qualities were distinct but linked for treatment teachers. In fact, the correlation between *OWL* fidelity and instructional quality for treatment teachers was moderate ($r = 0.54$, $p < .01$). Likewise, the correlation between *CC* implementation for these teachers and instructional quality was less, but still of moderate strength ($r = 0.45$, $p < .01$). The relationship between levels of general management, a measure used to represent classroom management, and curriculum-specific fidelity was also positive, significant, and moderate in strength ($r = 0.40$, $p < .05$).

In keeping with evidence brought to light by data collected on control teachers, correlational analysis associated with treatment teachers supports the claim that teaching quality and levels of implementation fidelity should be perceived as two constructs. The fact that there was a higher, significant correlation between treatment teachers' levels of quality and their fidelity to *OWL* suggests that the intervention curriculum may have provided greater alignment of effective practices. Correlational results suggest that teachers who delivered higher quality of instruction were higher implementers of OWL. Questions remain, however, as to the direction of that relationship. Are better teachers also better implementers or is it the case that teachers who implement *OWL* well become better teachers. Nonetheless, teaching quality and fidelity remain distinct components of the curriculum intervention, and data collected by the set of tools used in this study provided evidence that the relationship between these constructs differed by condition.

Table 10

Correlations of Instructional Quality and OWL Fidelity for Control Teachers (n = 16)

Variables		1	2	3	4
1	Overall *OWL* fidelity	-			
2	Overall *CC* fidelity	-0.02	-		
3	Overall General Instruction	0.32	0.38	-	
4	Overall General Management	0.34	-0.00	-0.19	-

Table 11

Correlations of Instructional Quality and OWL Fidelity for Treatment Teachers (n = 36)

Variables		1	2	3	4
1	Overall *OWL* fidelity	-			
2	Overall *CC* fidelity	0.31	-		
3	Overall General Instruction	0.54**	0.45**	-	
4	Overall General Management	0.40*	0.45**	0.57**	-

* p < .05.
** p < .01.

Discussion

The importance of confirming that interventions are delivered as originally intended has gained acceptance in the educational research community over the last several years. Our collective understanding of which measures to use and the manner in which to use them in accurately assessing fidelity remains in an immature state. Although most educational researchers familiar with issues of implementation accept the necessity to use measures to gauge levels of fidelity demonstrated by teachers involved in their study, few develop comprehensive conceptual models to represent unique causal components of the intervention. Even fewer researchers employ psychometrically tested measures that assess teachers' delivery of the casual components. Further, many researchers confound concepts of instructional quality with fidelity of implementation by erroneously using global, widely accepted measures of environmental and instructional quality as fidelity measures.

This paper addressed three research questions concerning issues around fidelity of implementation by presenting specific information on the development and application of a set of tools used to measure implementation fidelity and teaching quality. By laying out a conceptual model of the intervention and stepping through reliability tests of the measures, the author demonstrated that creating comprehensive conceptual models to represent unique causal components of interventions is necessary but difficult. Furthermore, it is challenging but possible to develop fidelity measures that reliably represent unique, causal elements of an educational intervention. In several cases, reliability tests confirmed that fidelity items truly represented intended constructs. However, the model and corresponding fidelity indicators were not perfect. Measures of *OWL* Book Reading were relatively weak and two instructional and structural subconstructs were dropped from the original model.

Three different analyses that examined fidelity at the curriculum-, activity-, and instructional- and structural-subconstruct levels were presented in this study. These analyses provide impetus for researchers to view fidelity of implementation as a complex, multidimensional construct. Approaching fidelity of implementation solely as a unitary construct can be beneficial when accounting for possible moderating effects of fidelity on intervention outcomes. Conclusions drawn from analysis of fidelity data at the activity-

and subconstruct-level provide a more detailed picture of the activities and elements of the curriculum with which teachers had the most difficulty. This view of fidelity is essential when developing resources such as materials and training used to support teachers in adopting the new curriculum.

When examining rates of implementation fidelity between experimental conditions at the curriculum-level, results suggested that there were significant differences. Additionally, the curriculum-level view of *CC* fidelity (averaged across activities and time) revealed that treatment and control delivered the curriculum to comparable degrees. Analysis of *OWL* fidelity in specific curricular activities added to the overall results in that treatment teachers were significantly better implementers of the *OWL* curriculum. On the other hand, analysis by activities showed that teachers in the two conditions were only equivalent in the degree to which they implemented *CC* in Small Group activities.

Results of analyses that adopted a more fine-grained perspective by targeting instructional and structural elements of the curriculum provided evidence that treatment and control teachers were not as different as initial analyses suggested. In fact, treatment teachers were not significantly different from control teachers in supporting children's language development and analytical thinking during Book Reading. The author hypothesizes that equivalence among conditions in educational interventions is rarely identified, not because it is absent, but because it is seldom measured effectively. In this case, by solely examining fidelity as a unitary construct, researchers could have been misled in thinking wholesale differences existed between the two conditions. The more nuanced perspective, which characterized fidelity as a multidimensional construct, actually exposed a lack of differentiation between conditions.

The results from this study mirror those found by Pence et al. in 2008. Fidelity measures used in the Pence study assessed teachers' implementation of contextual elements and instructional processes unique to *LFC*. Data collected three times over the course of the year provided evidence of program differentiation in some but not all aspects of the curriculum. When summing fidelity scores across seven targeted Language Stimulation Techniques (LSTs), overall rates of implementation between treatment and control teachers did not differ statistically. Moreover, of six specific activity contexts, teachers in the two conditions demonstrated varying rates of fidelity in only four of them. Despite the small sample size (i.e., seven teachers assigned to each of the two conditions), the analyses presented in the Pence study, as well this paper, prove that inclusion of fine-grained examinations of fidelity is necessary to produce precise and meaningful representations of program differentiation.

Despite some weaknesses, the measures of fidelity and quality used in this study were able to distinguish fidelity of implementation from instructional quality and classroom management. Divergent patterns of these constructs were evident across conditions and activities. In addition, correlational analysis of data collected on teachers' fidelity of implementation, instructional quality and classroom management suggest that the three attributes are related and that different patterns among teachers within those three constructs existed. In this study, teachers rated as having better overall teaching ability were more likely to implement the *OWL* curriculum with higher rates of fidelity. The reason for this relationship is not entirely clear. However, a curriculum like OWL requires teachers to support the language and literacy development of children by providing language-rich interactions. For teachers to implement *OWL* well they must have a firm grasp of the ways in which children learn and the ways in which teachers can provide the best opportunities for learning. Therefore, it comes as no surprise that teachers who provided high quality instruction, in general, were more successful at implementing a highly demanding curriculum.

Another possible explanation for higher levels of instructional quality among *OWL* teachers may be a consequence of professional development and coaching. Teachers may have become more adept at delivering high quality instruction as a result of the supports put in place during the intervention. Although professional development workshops and coaching sessions presented training on the specifics of implementing *OWL*, teachers may have become aware of more global elements of good instruction. They may have improved their teaching without developing into better implementers of *OWL*.

Limitations

There are several limitations to this study. For one, having such a small number of indicators per activity and per instructional and structural subconstruct calls in to question the accuracy of the measure. When components of a model are represented by such a small number of items, a single indicator can have a disproportionate effect on a teacher's score for that activity or subconstruct. For example, there are only two indicators in the compliance subconstruct in Small Group. One of those items assesses whether the length of the lesson is adequate. If a teacher fails to earn credit for that item, she is likely to be identified as non-compliant for that activity.

Additionally, low levels of internal consistency reliability highlight possible flaws in the conceptual model and associated fidelity indicators. Several subconstructs have less than ideal alphas. These sub par values indicate that those groups of items may not adequately represent the intended construct. Issues with reliability and validity must be addressed as the conceptual model and fidelity indicators used in this study undergo further development.

Lastly, moderate levels of *OWL* implementation demonstrated by control teachers raise a few questions. Evidence of instructional elements aligned with the *OWL* curriculum in the control classrooms can be explained in several ways. For one, these results may indicate that a certain level of contamination occurred. In some way, control teachers gained access to *OWL*, adopted particular practices endorsed by *OWL*, and delivered elements of the curriculum in their classrooms. However, the fact that data collected via fidelity measures indicate control teachers implemented elements of *OWL* also point to a lack of precision in the measures themselves. Some of the items designated as "unique to *OWL"* in the checklists may be more representative of general instruction. It may be more likely that control teachers were demonstrating higher levels of instructional quality rather than higher levels of fidelity to *OWL*.

Implications

This paper details the development and application of a conceptual model and a set of fidelity and quality measures used to assess rates of fidelity and differentiated instruction between experimental conditions. As discussed, views of fidelity as a unitary construct may obscure details about the specific ways in which treatment and control teachers vary. The addition of fine-grained analyses that examine the degree to which teachers from different experimental conditions deliver specific structural and instructional elements unique to the treatment condition is necessary to provide a robust and accurate representation of implementation. This analytical approach has implications for the ways in which educational researchers perceive the complexity of implementation fidelity and interpret intervention effectiveness.

Appendix A. Coding scheme

Study Identification

Study ID
- ##

Type of Publication
- Journal article
- Report
- Dissertation
- Other

Year of Publication
- ##

Sample Characteristics

Mean age of sample
- ##

Percentage of Children from low-income families
- Less than 30%
- Between 30% and 60%
- Over 60%
- Not Reported

Predominant Child Race in total sample
- More than 50% White
- More than 50% African American
- More than 50% Hispanic

Percentage of ELLs in total sample
- 0% ELLs
- 1-10% ELLs
- 11% + ELLs

Percentage of males in total sample
- Less than 40%
- Between 40% and 60%
- Over 60%

Percentage of children with learning disabilities in total sample
- 0% children with learning disabilities

1-10% children with learning disabilities
11% + children with learning disabilities

Average Teacher Experience (in years)
0-3
3-9
10-15
16+

Type of Preschool Program
Head Start
Title I
Universal
Public
Private
Mixed: HS & Private
Mixed: Public & Private
Other

Method of assignment
Random asgn + cntrl (w/out matching)
Random asgn + cntrl (matching)

Attrition Rate of Children in total sample
##%

Program characteristics

Average Classroom size (# students per classroom)
<12
13-20
21+

Total # of classrooms in sample
<15
16-25
26-35
36+

Treatment Characteristics

Type of Curriculum
Comprehensive
Supplementary

Primary Provider of Curriculum

- Classroom Teacher
- Specialist
- Researcher
- Other

Length of treatment

- ≤16 weeks
- 17 weeks to 1 year
- > 1 year

Level of Fidelity of Implementation reported

- Yes
- No

Total length of study (in years)

- < 1 year
- 1 year
- 2 years
- 3+ years

Most recent time of measures (following end of treatment)

- End of pre-K year n
- Fall of Grade 1
- Follow-up
- Other

Prof Development

- 1 x before intervention
- 2 x before and during
- 3 + before, during, other

Mentoring

- None available
- By curriculum trainers
- By researchers
- By master teachers / specialists

Parental Involvement

- No
- Yes

Effect Size Coding

Sample information:
# of classrooms (treatment)
# of classrooms (control)

Post-test Sample Size (children - treatment)
Post-test Sample Size (children - control)
Pre-test Sample Size (children - treatment)
Pre-test Sample Size (children - control)

Means and Standard Deviations:
Post-test Mean (treatment) - if reported
Post-test Adjusted Mean (treatment) - if reported
Post-test Std. Dev. (treatment) - if reported
Post-test Mean (Control) - if reported
Post-test Adjusted Mean (Control) - if reported
Post-test Std. Dev. (control) - if reported
Pre-test Mean (treatment) - if reported
Pre-test Std. Dev. (treatment) - if reported
Pre-test Mean (Control) - if reported
Pre-test Std. Dev. (control) - if reported

If only F-stat or T-star reported:
F-value reported
T-value reported
Page number ES-related data is reported
Calculated ES

Appendix B. Summary table of included studies

ID	Study & Authors	Description	Sample	Setting	Focus	Outcome Measure	Type of work
1	Assel, M. A., Landry, S. H., Swank, P. R., & Gunnewig, S. (2007)	Random assignment – classrooms to treatment or control. Treatment classrooms were then randomly assigned to mentor or no-mentor	245 PK children in HS; 10 Head Start classrooms, 5 w/ mentor	3 settings: Head Start, Title 1, and universal pre-kindergarten	Curriculum: Let's Begin with the Letter People Program: Head Start with mentoring	EVT	Journal Article *Reading and Writing*
2	Assel, M. A., Landry, S. H., Swank, P. R., & Gunnewig, S. (2007)	Random assignment – classrooms to treatment or control. Treatment classrooms were then randomly assigned to mentor or no-mentor	245 PK children in HS; 11 Head Start classrooms, 5 w/ mentor	3 settings: Head Start, Title 1, and universal pre-kindergarten	Curriculum: Doors to Discovery Program: Head Start with mentoring	EVT	Journal Article *Reading and Writing*
3	Assel, M. A., Landry, S. H., Swank, P. R., & Gunnewig, S. (2007)	Random assignment – classrooms to treatment or control. Treatment classrooms were then randomly assigned to mentor or no-mentor	213 PK children in TI; 8 Title I classrooms, 4 w/ mentor	3 settings: Head Start, Title 1, and universal pre-kindergarten	Curriculum: Let's Begin with the Letter People Program: Title I with mentoring	EVT	Journal Article *Reading and Writing*
4	Assel, M. A., Landry, S. H., Swank, P. R., & Gunnewig, S. (2007)	Random assignment – classrooms to treatment or control. Treatment classrooms were then randomly assigned to mentor or no-mentor	213 PK children in TI; 8 Title I classrooms, 4 w/ mentor	3 settings: Head Start, Title 1, and universal pre-kindergarten	Curriculum: Doors to Discovery Program: Title I with mentoring	EVT	Journal Article *Reading and Writing*
5	Assel, M. A., Landry, S. H., Swank, P. R., & Gunnewig, S. (2007)	Random assignment – classrooms to treatment or control. Treatment classrooms	145 PK children in UPK; 6 UPK classrooms, 3 w/ mentor	3 settings: Head Start, Title 1, and universal pre-kindergarten	Curriculum: Let's Begin with the Letter People Program: UPK with mentoring	EVT	Journal Article *Reading and Writing*

		were then randomly assigned to mentor or no-mentor					
6	Assel, M. A., Landry, S. H., Swank, P. R., & Gunnewig, S. (2007)	Random assignment – classrooms to treatment or control. Treatment classrooms were then randomly assigned to mentor or no-mentor	145 PK children in UPK; 6 UPK classrooms, 3 w/ mentor	3 settings: Head Start, Title 1, and universal pre-kindergarten	Curriculum: Doors to Discovery Program: UPK with mentoring	EVT	Journal Article *Reading and Writing*
7	Pietrangelo, D. J. (1999).	Random Assignment to treatment or control	129 PK children	10 Head Start classrooms in upstate NY	supplementary curriculum: direct and explicit instruction in prereading skills	PPVT	Dissertation
8	Whitehurst, G. J., Epstein, J. N., Angell, A. L., Payne, A. C., Crone, D. A., & Fischel, J. E. (1994)	Random assignment and pairing (equal number of each, and each center had treatment and control)	167 PK children (1992-1993)	4 Head Start centers in Suffolk County, NY	Add-on emergent literacy curriculum vs. HS curriculum only	PPVT	Journal Article *Journal of Ed. Psych.*
9	Fischel, J. E., Bracken, S. S., Fuchs-Eisenberg, A., Spira, E. G., Katz, S., & Shaller, G (2007)	Random assignment by classroom to treatment or control (3 cohorts)	185 PK children in treatment & 150 in control	12 Head Start classrooms in NY	Curriculum: Let's Begin with the Letter vs. High Scope	PPVT	Journal Article *Journal of Lit Research*
10	Fischel, J. E., Bracken, S. S., Fuchs-Eisenberg, A., Spira, E. G., Katz, S., & Shaller, G (2007)	Random assignment by classroom to treatment or control (3 cohorts)	172 PK children in treatment & 150 in control	12 Head Start classrooms in NY	Curriculum: Waterford merged with High Scope vs. High Scope alone	PPVT	Journal Article *Journal of Lit Research*
11	Davidson, M. R., Fields, M. K., & Yang, J. (2009)	Random assignment by classroom	254 PK children	27 classrooms in Public elementary schools in Newark, NJ	Evaluate effectiveness of Ready, Set, Leap in high-poverty schools	PPVT	Report
12	DeBaryshe, B. D., & Gorecki, D. M. (2007)	Blocked random assignment by site to treatment (1 of 2) or control	51 PK children (and 30 in control)	4 Head Start classrooms, within 9 sites, in Hawaii (3 control classrooms)	(LC) Experimental literacy curriculum – with focus on language outcomes	EOWPVT	Journal Article *Early Educ and Devel.*
13	DeBaryshe, B.	Blocked	44 PK	4 Head Start	Experimental	EOWPVT	Journal

	D., & Gorecki, D. M. (2007)	random assignment by site to treatment (1 of 2) or control	children (and 30 in control)	classrooms, within 9 sites, in Hawaii (3 control classrooms)	math curriculum – with focus on language outcomes		Article *Early Educ and Devel.*
14	Chambers, B., Chamberlain, A., Hurley, E. A., & Slavin, R. E. (2001)	Quasi, assigned by sites and matched on demographic characteristics	169 3-yr-old PK children	Private child-care centers, 106 children (with 63 control)	Implementation and evaluation of Curiosity Corner	Mullen Scales of Early Learning (MSEL): RL	Report / AERA paper
15	Chambers, B., Chamberlain, A., Hurley, E. A., & Slavin, R. E. (2001)	Quasi, assigned by sites and matched on demographic characteristics	147 4-yr-old PK children	Public PK, 100 children (with 47 control)	Implementation and evaluation of Curiosity Corner	Mullen Scales of Early Learning (MSEL): RL	Report / AERA paper
16	Farran, D. & Lipsey, M. (2008) (PCER)	Random assignment by classrooms	206 PK children total, 101 for CC	Public PK, 7 classrooms (with 7 control)	Creative Curriculum	PPVT	NCER Report
17	Farran, D. & Lipsey, M. (2008) (PCER)	Random assignment by classroom	208 PK children total, 103 for BB	Public PK, 7 classrooms (with 7 control)	Bright Beginnings	PPVT	NCER Report
18	Lambert, R.G. & Abbott-Shim, M. (2008) (PCER)	Random assignment of teachers (within centers)	194 PK children total, 97 for CC	Head Start, 9 classrooms (with 9 control)	Creative Curriculum	PPVT	NCER Report
19	Priest, J.S & Zoellick, L. (2008) (PCER)	Random assignment of classrooms	123 PK children total, 62 for LL	Head Start, 7 classrooms (with 7 control)	Creative Curriculum with Ladders to Literacy	PPVT	NCER Report
20	Chambers, B. & Slavin, R. (2008) (PCER)	Blocked random assignment by school across 3 sites	215 PK children total, 105 for CurCnr	SFA, Head Start, and Day Care Centers, 10 classrooms (with 8 control)	Curiosity Corner	PPVT	NCER Report
21	Fountain, C., Cosgrove, M. & Wood, J. (2008) (PCER)	Random assignment of classrooms	244 PK children total, 137 for ELLM	Head Start & mixed PK, 14 classrooms (with 14 control)	Early Literacy and Learning Model	PPVT	NCER Report
22	Justice, L., Pence, K. & Wiggins, A. (2008) (PCER)	Random assignment of classrooms	195 PK children total, 97 for LFC	Head Start & Public PK, 7 classrooms (with 7 control)	Language-Focused Curriculum	PPVT	NCER Report
23	Lonigan, C.J. & Schatschneider, C. (2008) (PCER)	Blocked random assignment by school	198 PK children total, 101 for DLM	Public PK, 5 classrooms (with 6 control)	DLM Early Childhood Express with Open Court Reading Pre-K	PPVT	NCER Report

24	Lonigan, C.J. & Schatschneider, C. (2008) (PCER)	Blocked random assignment by school	196 PK children total, 99 for LE	Public PK, 6 classrooms (with 6 control)	Literacy Express	PPVT	NCER Report
25	Powell, D. & File, N. (2008) (PCER)	Random assignment of teachers	204 PK children total, 114 for PA	Public PK, 7 classrooms (with 6 control)	Project Approach	PPVT	NCER Report
26	Thornburg, K.R., Mayfield, W. & Morrison, J. (2008) (PCER)	Blocked random assignment of centers (from convenience sample)	231 PK children total, 123 for PC	Full-day child-care centers, 10 classrooms (with 11 control)	Project Construct	PPVT	NCER Report
27	Cunningham, A. & Davidson, M. (2008) (PCER)	Blocked random assignment of classrooms (from convenience sample)	286 PK children total, 159 for PKM	Full-day preschool programs, 18 classrooms (with 21 control)	Ready, Set, Leap	PPVT	NCER Report
28	Starkey, P., Klein, A., Clements, D. & Sarama, J. (2008) (PCER)	Blocked random assignment by school	316 PK children total, 149 for RSL	Head Start & Public PK, 20 classrooms (with 20 control)	PK math + DLM	PPVT	NCER Report

Appendix C. Curriculum characteristics

Curriculum	Conceptual Emphasis	Skill(s) Focus	Structure
Creative Curriculum	Social/emotional, physical, cognitive, language development	N/A	10 “interest areas”: blocks, dramatic play, toys & games, art, library, discovery, sand & water, music & movement, cooking, computers (LOW)
Ladders to Literacy	Literacy and language development	Print awareness, Metalinguistic awareness, oral language	Approx. 20 activities within each of three skill areas
Pre-K Mathematics	Mathematical knowledge and skills	N/A	29 activities
Door to Discovery	Literacy	Oral language, phonological awareness, concepts of print, alphabet knowledge, writing, comprehension	8 Thematic Units
Let’s Begin with Letter People	Language & Literacy	Phonological awareness: rhyming, word play, alliteration, segmentation; oral language; letter knowledge	Thematic Units, plus “interest centers”, plus materials

(continued)

Curriculum	Conceptual Emphasis	Skill(s) Focus	Structure
Project Approach	Investigation of real-world topics	N/A	3 components of learning activities guided by children's interests: spontaneous play, systematic instruction, project work
Project Construct	Cognitive, Representational, sociomoral, physical	N/A	29 goals for children
Language Focused Curriculum	Language	N/A	Thematic organization with daily activities and 8 key instructional techniques
*DLM Early Childhood Express**	Social, emotional, intellectual, aesthetic, and physical development	N/A	36 weekly themes, 200 learning activities
*Open Court Pre-K**	Literacy	Phonological, phonemic, and print awareness, comprehension	8 thematic units, teacher read literature selections + activities
Literacy Express	Language and literacy	N/A	Thematic units with suggested activities, room arrangement, daily schedules, classroom management

(continued)

Curriculum	Conceptual Emphasis	Skill(s) Focus	Structure
Ready, Set, Leap!	Language and Literacy	phonological awareness, alphabetic knowledge, print awareness, reading, reading comprehension	9 thematic units, 120 lesson plans for each unit
Bright Beginnings	Language and literacy	(Lists 9 program components linked to units)	9 units

NOTE: Information taken from the NCER report (2008) and What Works Clearing House (WWC)
** DLM and Open Court curricula were used together in the treatment classrooms for 1 of the 13 studies*

Appendix D. Code sheet for items

Coding by Item

[CF] Contextual Focus

(S) Structure

Example: “Activity centers have been set up in the classroom”

“Children’s work is displayed on the walls”

(I) Instruction

Example: “Teacher introduces vocabulary words while reading”

[R] Reference

(R) Direct reference to curriculum materials, manual, etc)

Example: “Books designated by *Curriculum X* are located in library area”

“All materials are used during the *Letter are Friends Game* activity”

(N) If no reference exists

Example: “Teacher encourages children to participate in games”

[Q] Quality of Delivery

(IN) Inadequate (count, tally, presence of)

Example: “Give each child an opportunity to wear the conductor’s hat”

“Teacher uses finger puppets during book reading”

(AD) Adequate

Example: “Teacher gives background information to children before reading book”

“Selected content is interesting to children”

[TOI] Target (i.e., Targeted Individual or Object)

(G) General description

Example: For Coder: Provide general description of activity (Open response)

(S) Scheduling (time, duration, integration of activities into daily schedule, etc)

Example: For Coder: Provide dates on which activities were observed

“Each activity lasts between 20 and 30 minutes”

(O) Organization (Child grouping, activity areas, presence in setting, management)

Example: “Learning centers are defined and labeled”

“Furniture is arranged so children can easily navigate the classroom”

(M) Materials (use, presence, access)

Example: “Musical instruments are present and accessible to children”

“A variety of writing materials is accessible in the writing center”

(H) Promotion of healthy and safe physical environment & practices

Example: “All staff washes hands after meal time”

“Sharp knives are kept out of reach of children”

`

(T) Teacher

Example: "Teacher asks children to retell the story"

(C) Child

Example: "# of children in each center"
"Describe role of children during story reading"

[IF] Instructional Facet

(LUP) – Teacher language use, practice, and engagement (within any domain focus)

Example: "List target words used by teacher"
"Teacher asks open-ended questions"
"Teacher demonstrates how to hold a book, turn pages, and read from left to right"

(TAR) – Teacher affect (e.g. warm, caring, energetic, etc) & responsiveness

Example: "Teacher appears to enjoy the activity"
"Teacher praises children's responses"
"Teacher provides children with positive feedback"
"Teacher values child response regardless of accuracy and complexity"

(TS) – Teacher support or scaffolding of child expression; differentiate instruction (lang, writing, retelling, actions, etc)

Example: "Teacher encourages conversation between children"
"Encourages children to use language whenever possible"

(TUD) – Teacher performs Assessment / Check for understanding or Documents, collects, or records children's expression/knowledge

Example: "Teacher observes children within each activity"
"Teacher summarizes or re-teaches difficult lessons"
"Teacher uses child answers as evidence of understanding"
"Children's knowledge is documented"
"Teacher transcribes children's ideas"

(TM) - Teacher use of materials in targeted activities

Example: "Teacher prepares materials within activity centers in advance"
"Teachers uses counting bears to increase childrn's mathematical understanding"

(TP) – Teacher interaction with parents/other adults/specialists or related to teacher's assistant; materials sent home

Example: "Teacher communicates regularly with parents"
"Teacher shares weekly schedule with parents"

(TCK) – Teacher selection and use of content and concepts (e.g. brings children's attention to, provides opportunity for learning particular concepts) or knowledge and/or interest in content, curriculum, children's development

Example: "Teacher selects activities based on weekly theme"
"Teacher selects content that is developmentally appropriate for children"
"Teachers expresses curiosity for learning"
"Teacher understands the objective if each lesson

Appendix E. Summary of elements related to adherence, exposure, and participant responsiveness by measure

				Adherence		Exposure		Participant Responsiveness
Curriculum	Instrument	No. Items	Format	Curriculum Specific Fidelity (% of Items)	Duration of Observation	Duration of Activities	Scheduling (% items)	Teacher Interview / Artifact Examination
CC	Creative Curriculum Checklist	188	All items: yes/no	0.0	Yes, 3 hours with 20-minute interview (suggested)	No	13.4	Teacher interview, 15 Questions related to all 5 categories
BB	Bright Beginnings Checklist	314	2 forms: 1) weak, fair, good, excellent; 2) not present, not sufficient, sufficient	0.0	None indicated	No	6.0	Teacher interview, 3 primary questions

(continued)

				Adherence		Exposure		Participant Responsiveness
Curriculum	Instrument	No. Items	Format	Curriculum Specific Fidelity (% of Items)	Duration of Observation	Duration of Activities	Scheduling (% items)	Teacher Interview / Artifact Examination
LL	Ladders to Literacy Classroom Activities	179	All items: yes, no, not applicable	15.6	Yes, Obs enters start and end times	No	1.7	Not included
	Ladders to Literacy Scaffolding	N/A	Fill-in teacher utterances & focus	N/A	Yes, 30 minutes (suggested)	Yes, can be calculated	N/A	Not included
DLM + OC	DLM/Open Court Classroom Observation	9	Checklist: Specific activity or materials, No. of children, Facilitator's role, time	0.0	None indicated	Yes, entered by observer	0.0	Not included
LE	Literacy Express Classroom Observation	10	Checklist: Specific activity or materials, No. of children, Facilitator's role, and time	0.0	None indicated	Yes, entered by observer	0.0	Not included

(continued)

				Adherence	Exposure			Participant Responsiveness
Curriculum	Instrument	No. Items	Format	Curriculum Specific Fidelity (% of Items)	Duration of Observation	Duration of Activities	Scheduling (% items)	Teacher Interview / Artifact Examination
DD & Let's Begin (both with PCER Checklist)	Doors to Discovery Fidelity Checklist	24	Combination: 1) Rating 1-5 (5=all criteria met); 2) yes/no	4.2	None indicated	No	0.0	Not included
	PCER checklist	42	All items have variation of 1-5 scale (5=frequent, high quality, always, etc.)	0.0	Yes, Observer enters length of observation	No	2.4	Not included
	Let's Begin Fidelity Checklist	5	Combination: 1) Rating 1-5 (5=all criteria met); 2) yes/no; 3) yes/no/not applicable	32.0	None indicated	No	0.0	Not included
LFC	Language-Focused Fidelity Checklist	45	Checklist: mark (+) if present	46.6	Yes, 2 hours (suggested)	No	2.2	No interview; but teacher reports examined

(continued)

				Adherence		Exposure		Participant Responsiveness
Curriculum	Instrument	No. Items	Format	Curriculum Specific Fidelity (% of Items)	Duration of Observation	Duration of Activities	Scheduling (% items)	Teacher Interview / Artifact Examination
Pre-K Math + DLM	Early Mathematics Classroom Observation	15	Combination: 1) narrative; 2) duration; 3) checklist of type of activity; 4) Fill-in: # of children	0.0	None indicated	Yes, entered by observer	0.0	Not included
	Pre-K Mathematics Fidelity Record Sheet	109	Combination: 1) yes, no, not needed; 2) freq - no, some, usually; 3) date; 4) # of children	10.2	None indicated	Yes, observer enters start and end times	54.1	No interview; but teacher records/assessments examined
	Classroom Observation of Early Mathematics Environment and Teaching	26	All items: Scale SD, D, N, A, SA	0.0	None indicated	Yes, observer enters start and end times	3.8	Not included

(continued)

				Adherence		Exposure		Participant Responsiveness
Curriculum	Instrument	No. Items	Format	Curriculum Specific Fidelity (% of Items)	Duration of Observation	Duration of Activities	Scheduling (% items)	Teacher Interview / Artifact Examination
PA	Project Approach Fidelity Scale	26	Two formats: 1) no, some, moderate, or strong evidence; 2) source of info: observed, doc reviewed, journal	0.0	None indicated	Yes, project duration entered by observer	7.7	Study reported that interviews occurred; interview questions not submitted
PC	Project Construct Early Childhood Observation	59	All items: scale - No evidence, some evidence, extensive evidence	0.0	Yes, 2.5-3 hours (suggested). Observer enters start and end times	No	0.0	No formal interview, but observers are encouraged to ask teacher questions
RSL	Ready, Set, Leap! Fidelity Observation	44	All items: 6-point scale (1=never or not true at all; 6=all the time or very true)	0.0	Yes, 60-75 minutes (suggested)	No	2.3	Not included

Appendix F. Small Group *OWL* fidelity and quality measures

OWL Fidelity

Process [Yes/No]

Language Development

S1: Teacher uses precise vocabulary words when discussing materials and actions.
S2: Teacher helps children express themselves in words or actions by giving hints & telling when needed.
S3: Teacher prompts children to use vocabulary.

Analytical Thinking

S6: Instructional goals of the activity are conceptually based.
S7: Teacher makes a brief presentation that introduces key concepts, skills, and vocabulary.
S11: Materials are used to increase the conceptual understanding of the children

Compliance [Yes/No]

S4: Lesson is appropriate length (12 or more mins.).
S9: Teacher presents medium or high-support activities consistent with that week of the curriculum.
S10: Teacher presents the learning objective consistent with that week of the curriculum.

Creative Curriculum Fidelity

C1: Listen attentively to what each child has to say and respond respectfully to children at their eye level?

C2: Adapt instruction to include all children (e.g., offer challenging experiences, use clear visual cues for a child with a disability, use concrete objects and gestures with second language learners going through a nonverbal period)?

C3: Guide children in putting away materials where they belong (e.g., draw attention to the labels; play games to sort materials on shelves)?

C4: At the activity-level, teacher uses small-group times to address the needs and interests of the children (that fit the instructional context of small group time.)

C5: Teacher provides and engages children in using materials to either engage children in retelling, explore number concepts, or learn about spaces & geography.

C6: Encourage children to connect ideas to everyday experiences.

C7: Interact with children to support their understanding of (at least one): number concepts, patterns, geometry and spatial sense, or measurement

General Instructional Quality

GI1: Teacher has individual conversations with children during exploratory time.

GI2: Teacher gives individual children opportunities to respond verbally and/or nonverbally throughout activity.

GI3: > 50% Children have hands-on time with materials for > 50% of the time.

GI4: Information provided by the teachers is accurate.

GI5: Teacher manages materials effectively throughout lesson (e.g., introducing and labeling correctly, distributing efficiently, guiding children's use).

GI6: Teacher verbally summarizes/ reflects on the lesson before the transition to the next activity.

General Management

GM1: Teacher verbally praises students for appropriate behavior two or more times.

GM2: Teacher is positive and actively seeks to engage children throughout the activity

GM3: Teacher takes steps to address problems during the activity or no problems arise.

GM4: Behavioral challenges are addressed in a manner that minimizes disruption to the flow of the lesson or problems do not emerge.

GM5: Teacher has control of the group. Effective group-oriented strategies are used as needed.

GM6: Teacher maintains pace of the activity.

GM7: Visible and audible children attend and participate appropriately throughout the lesson.

Appendix G. Group Literacy Instruction OWL fidelity and quality measures

OWL Fidelity

Process [Yes/No]

Code-Focused Development

L1: If routines focus on letters, teacher points to & names letters or if counting activities, teacher says numbers clearly and actions make clear the number word meanings.

L3: For routines that focus on PA or other skills, teacher makes the skill explicit ("rhyme", "first sound") and emphasizes it in delivery.

L8: Whenever possible, teacher points to print, and tracks left to right.

Language Development

L2: Teacher encourages children to identify letters and/or numbers, say their names.

L4: Teaches word meanings: points to object/picture, says words, defines words, gives clear hints meanings.

L5: Teacher encourages children (as a group or individuals) to say key words.

Compliance [Yes/No]

L6: Teacher presents activities specified by lesson plans or changes retain skills focus in curriculum.

L11: Lesson is appropriate length (10 - 20 min.)

Creative Curriculum Fidelity

C1: Allow for flexibility within Group Literacy Instruction.

C2: *Transition* individual children from one activity within Group Literacy Instruction to another individually and in small groups as much as possible.

C3: Provide opportunities for children to explore music spontaneously.

C4: Provide opportunities for children to express themselves with movement.

General Instructional Quality

GI1: Teacher engages children with varied volume pace, expression, or gestures.

GI2: Teacher transitions smoothly and quickly from one GLI activity to the next.

GI3: Draw children’s attention to the sounds of language through playful songs, stories, rhymes, and chants to help develop *phonological awareness*?

GI4: Teacher has materials prepared and at hand.

General Management

GM1: Teacher verbally praises students for appropriate behavior two or more times.

GM2: Teacher is positive and actively seeks to engage children throughout the activity

GM3: Teacher takes steps to address problems during the activity or no problems arise.

GM4: Behavioral challenges are addressed in a manner that minimizes disruption to the flow of the lesson or problems do not emerge.

GM5: Teacher has control of the group. Effective group-oriented strategies are used as needed.

GM6: Teacher maintains pace of the activity.

GM7: Visible and audible children attend and participate appropriately throughout the lesson.

Appendix H. Book Reading OWL fidelity and quality measures

OWL Fidelity

Process [Yes/No]

Language Development

R4: Teacher elicits definitions from children.

R7: Teacher gives or elicits accurate definitions of OWL targeted vocabulary words.

R8: Teacher defines words using implicit strategies: pointing to pictures, voice, expressions, gestures, and expansions.

R9: Teacher prompts children to say words or is receptive to spontaneous use.

Analytical Thinking

R2: Teacher relates the story to other books or to the current theme (for read 1) or teacher encourages children to recall portions of the story before reading (for read 2).

R3: Teacher identifies the characters or elicits them from children.

R10: Teacher gives information about characters' feelings and emotions

R11: Teacher gives information about analytical issues (other than characters): event sequences, interpret pictures, cause-effect links.

R12: After reading, teacher asks thoughtful questions to assess and build student understanding of story.

Compliance [Yes/No]

R5: Teacher reads book assigned to that week in curriculum guide.

R6: Activity is appropriate length (10-20 minutes).

Creative Curriculum Fidelity

C1: Allow for flexibility within the activity.

C2: Allow for flexibility with individual children as much as possible.

C3: Draw children's attention to concepts of print, letters, and words and books book handling skills.

C4: Engage children in retelling a story using puppets, flannel board figures, or props; sing songs; employ fingerplay; engage in shared writing?

C5: Teacher has children take a picture walk through the story before reading.

C6: Teacher relates the story to children's personal prior experiences.

C7: Teacher prompts children to interact and respond to prompts to chime in on a predictable phrase.

General Instructional Quality

GI1: Teacher introduces book quickly, says title (may also add author & illustrator).

GI2: Teacher holds book so that the children can see.

GI3: Teacher reads in a manner designed to hold attention, and clarify meaning: varies volume, pace, may use facial expression or gesture.

GI4: Teacher defines educationally useful words explicitly (OWL or other words)

GI5: Teacher defines words implicitly (OWL or other words)

GI6: During the read, teacher elicits information from children about analytical issues

GI7: During the read, teacher elicits information from children about characters' feelings and emotions

General Management

GM1: Teacher verbally praises students for appropriate behavior two or more times.

GM2: Teacher is positive and actively seeks to engage children throughout the activity

GM3: Teacher takes steps to address problems during the activity or no problems arise.

GM4: Behavioral challenges are addressed in a manner that minimizes disruption to the flow of the lesson or problems do not emerge.

GM5: Teacher has control of the group. Effective group-oriented strategies are used as needed.

GM6: Teacher maintains pace of the activity.

GM7: Visible and audible children attend and participate appropriately throughout the lesson.

REFERENCES

References with an asterisk (*) are studies included in the meta-analysis.

*Assel, M. A., Landry, S. H., Swank, P. R., & Gunnewig, S. (2007). An evaluation of curriculum, setting, and mentoring on the performance of children enrolled in pre-kindergarten. *Reading and Writing: An Interdisciplinary Journal, 20*(5), 463-494.

Baker, C. N., Kupersmidt, J. B., Voegler-Lee, M. E., Arnold, D. H., & Willoughby, M. T. (2010). Predicting teacher participation in a classroom-based, integrated preventive intervention for preschoolers. *Early Childhood Research Quarterly, 25*, 270-283.

Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D. (1999). Beyond Kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, *27*(1), 3-23.

Barnett, W. S. (2008). Federal Pre-K curriculum study shows few stand-outs. Preschool Matters, 6, 4-5.

Beatty, B. (1995). *Preschool education in America: The culture of young children from the colonial era to the present*. New Haven: Yale University Press.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* West Sussex, UK: John Wiley & Sons.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive meta-analysis version 2*. Englewood, NJ: Biostat.

Bus, A. G., & van Ijzendoorn, M. H. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology, 91*(3), 403-414.

Camilli, G., Vargas, S., Ryan, S., & Barnett, W. S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record,* 112(3).

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools, 32*(1), 38-50.

*Chambers, B., Chamberlain, A., Hurley, E. A., & Slavin, R. E. (2001, April). *Curiosity Corner: Enhancing preschoolers' language abilities through comprehensive* *reform*. Paper presented at the meeting of the American Educational Research Association, Seattle, WA.

Clarke, G. (1998). Intervention fidelity in the psychosocial prevention and treatment of adolescent depression. *Journal of Prevention and Intervention in the Community, 17*, 19- 33.

Coleman, J. S. (1968). *Equality and Achievement in Education*. Boulder, CO: Westview Press.

Collins, R. (1990). *Head Start research and evaluation: A blueprint for the future*. Washington, DC: Administration for Children, Youth, and Families.

Condry, S. (1983). The history and background of preschool intervention programs and the Consortium for Longitudinal Studies. In The Consortium for Longitudinal Studies (Ed.), *As the twig is bent… Lasting effects of preschool programs*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Connecticut State Department of Education. (2006). *The Connecticut framework: Preschool curriculum framework.* Retrieved from http://www.sde.ct.gov/sde/lib/sde/PDF/DEPS/Early/Preschool_framework.pdf

Cordray, D. S. (2009). Identifying and assessing the cause in RCTs [PowerPoint lecture notes]. Nashville, TN: Vanderbilt University: IES/NCER Summer Research Training Institute.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 103-124). Washington DC: American Psychology Association.

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*(1), 23-45.

* Davidson, M. R., Fields, M. K., & Yang, J. (2009). A randomized trial study of a preschool literacy curriculum: The importance of implementation. *Journal of Research on Educational Effectiveness, 2*(3), 177-208.

Davydov, V. V., & Kerr, S. T. (1995). The influence of L. S. Vygotsky on education theory, research, and practice. *Educational Researcher, 24*(3), 12-21.

*DeBaryshe, B. D., & Gorecki, D. M. (2007). An experimental validation of a preschool emergent literacy curriculum. *Early Education and Development, 18*(1), 93-110.

DeVries, R., & Goncu, A. (1987). Interpersonal relations in four-year dyads from constructivist and Montessori programs. *Journal of Applied Developmental Psychology, 8*, 481-501.

Dickinson, D. K., & Tabors, P. O. (Eds.). (2001). *Beginning literacy with language: young children learning at home and school*. Baltimore: Brookes Publishing.

Dodge, D. T. (2004). Early childhood curriculum models: Why, what, and how programs use them. *Child Care Informational Exchange*, 72-75.

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

Dodge, D. T., Colker, L. J., & Heroman, C. (2003). *The Creative Curriculum for Preschool Implementation Checklist* ® Washington, DC: Teaching Strategies, Inc.

Domitrovich, C. E., Gest, S. D., Jones, D., Gill, S., & DeRousie, R. M. S. (2010). Implementation quality: Lessons learned in the context of the Head Start REDI trial. *Early Childhood Research Quarterly, 25*, 284–298.

Dowaliby, G. P. (2006). *The Connecticut framework: preschool curriculum framework*. Hartford: Division of Teaching and Learning Programs and Services.

Durlak, J. A. (2010). The importance of doing well in whatever you do: A commentary on the special section, “Implementation research in early childhood education”. *Early Childhood Research Quarterly, 25*, 348–357.

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*, *28*, 237-256.

Epstein, A. S., Schweinhart, L. J., & McAdoo, L. (1996). *Models of early childhood education*. Ypsilanti, MI: High/Scope Press.

*Fischel, J. E., Bracken, S. S., Fuchs-Eisenberg, A., Spira, E. G., Katz, S., & Shaller, G. (2007). Evaluation of curricular approaches to enhance preschool early literacy skills. *Journal of Literacy Research, 39*(4), 471–501.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley.

Goffin, S. G. (1994). *Curriculum models and early childhood education: Appraising the relationship*. New York: Macmillan College Publishing Company.

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-k on cognitive development. *Developmental Psychology, 41*(6), 872–884.

Gresham, F. M., Gansle, K. A., & Noell, G. H. (1993). Treatment integrity in applied behavior analysis with children. *Journal of Applied Behavior Analysis*, *26*(2), 257–263.

Griffin, J. A. (2010). Research on the implementation of preschool intervention: Learning by doing. *Early Childhood Research Quarterly, 25*, 267–269.

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

Harms, T., & Clifford, R. (1980). *Early Childhood Environment Rating Scale*. New York: Teachers College Press.

Harris, D. N., & Herrington, C. D. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education, 112*(2), 209-238.

Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics, 32*, 151-179.

Hoaglin, D., Mosteller, F., & Tukey, J. (Eds.). (1983). *Understanding robust and exploratory data analysis*. New York: John Wiley & Sons.

Hofferth, S. L. (1996). Child care in the United States today. *The Future of Children, 6*(2), 41-61.

Hulleman, C. S., & Cordray, D. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*(1), 88-110.

Hulleman, C. S., Cordray, D. S., Nelson, M. C., Darrow, C. L, & Sommer, E. C. (2009, April). *The state of treatment fidelity assessment in elementary math interventions*. Poster presented at the American Educational Research Association Annual Conference. San Diego, CA.

Hyson, M. (1991). The characteristics and origins of the academic preschool. *New Directions for Child Development, 53*, 21-29.

Illinois State Board of Education. (February, 2009). *Prekindergarten/preschool for all curriculum criteria.* Retrieved from http://www.isbe.state.il.us/earlychi/pdf/ECBG_PreK_Preschool_Curric

Criteria.pdf

Justice, L. M., Mashburn, A., Pence, K. L., Wiggins, A. (2008). Experimental evaluation of a preschool language curriculum: Influence on children's expressive language skills. *Journal of Speech, Language, and Hearing Research*, *51*, 983-1001.

Kazdin, A. E. (1980) *Research Design in Clinical Psychology*. New York: Harper & Row.

Landry, S. H., Swank, P. R., Smith, K. E., Assel, M. A., & Gunnewig, S. B. (2006). Enhancing early literacy skills for preschool children: Bringing a professional development model to scale. *Journal of Learning Disabilities*, *39*(4), 306-324.

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: social background differences in achievement as children begin school*. Washington, DC: Economic Policy Institute.

Lipsey, M. (2009). Specifying the conceptual and operational models and the research questions that follow [PowerPoint lecture notes]. Nashville, TN: Vanderbilt University: IES/NCER Summer Research Training Institute.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Luborsky, L., McLellan, A. T., Diguer, L., Woody, G., & Seligman, D. A. (1997). The psychotherapist matters: Comparison of outcomes across twenty-two therapists and seven patient samples *Clinical Psychology: Science & Practice, 4*, 53-65.

McHugh, C.M. (2004). *McHugh Cluster Calculator*. Alexandria, VA.

McIntyre, L. L., Gresham, F. M., DiGennaro, F. D., & Reed, D. D. (2007). Treatment integrity of school-based interventions with children in the Journal of Applied Behavior Analysis 1991−2005. *Journal of Applied Behavior Analysis*, *40*, 659-672.

Mol, S. E., Bus, A. G., & Jong, M. T. D. (2009). Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language. *Review of Educational Research, 79*(2), 979-1007.

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

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*(3), 315-340.

National Association for the Education of Young Children. (2002). Early learning standards: creating the conditions for success [Electronic Version]. *NAEYC Position Statement*. Retrieved April 3, 2010 from http://www.naeyc.org/positionstatements/learning_standards.

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

National Early Learning Panel. (2008). *Developing early literacy.* Washington, DC: National Institute for Literacy.

National Institute of Child Health and Human Development Early Child Care Research Network. (2005). Pathways to reading: The role of oral language in the transition to reading. *Developmental Psychology*, 41, 428-442.

National Institute of Early Education Research. (2008). Federal pre-k curriculum study shows few stand-outs. *Preschool Matters, 6*(3), 4-5.

No Child Left Behind Act of 2001, H.R. 1, 107th Congress.

Office of School Readiness, State Board of Education, North Carolina. (November, 2008). *North Carolina approved early childhood curricula*. Retrieved from http://www.osr.nc.gov/_pdf/NCApprovedEarly ChildhoodCurricula.pdf

O'Connor, R. E., & Jenkins, J. R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading, 3*(2), 159 - 197.

Odom, S. L., Fleming, K., Diamond, K., Lieber, J., Hanson, M., Butera, G., et al. (2010). Examining different forms of implementation and in early childhood curriculum research. *Early Childhood Research Quarterly, 25*, 314-328.

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*(1), 33-84.

Pence, K. L., Justice, L. M., & Wiggins, A. K. (2008). Preschool teachers' fidelity in implementing a comprehensive language-rich curriculum. *Language, Speech, and Hearing Services in Schools*, *39*, 329-341.

Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavioral Analysis*, *15*(4), 477-492.

Pianta, R.C., La Paro, K., & Hamre, B.K. (*2008*). *Classroom Assessment Scoring System*

(CLASS). Baltimore: Brookes.

Piasta, S. B., & Wagner, R. K. (2010). Developing early literacy skills: A meta-analysis of alphabet learning and instruction. *Reading Research Quarterly, 45*(1), 8-38.

*Pietrangelo, D. J. (1999). *Outcomes of an enhanced literacy curriculum on the emergent literacy skills of Head Start preschoolers.* Unpublished dissertation, State University of New York, Albany.

Ramey, C. T., Campbell, F. A., Burchinal, M., Skinner, M. L., Gardner, D. M., & Ramey, S. L. (2000). Persistent effects of early childhood education on high-risk children and their mothers. *Applied Developmental Science, 4*(1), 2 - 14.

Scarr, S., & Weinberg, R. A. (1986). The early childhood enterprise: Care and education of the young. *American Psychologist, 41*(10), 1140-1146.

Schickedanz, J., & Dickinson, D. K. (2005). *Opening the World of Learning*. Parsippany, NY: Pearson Education, Inc.

Schweinhart, L. J., & Weikart, D. P. (1997). The High/Scope preschool curriculum comparison study through age 23. *Early childhood research quarterly, 12*(2), 117-143.

Schweinhart, L. J., & Weikart, D. P. (1998). Why curriculum matters in early childhood education. *Educational Leadership, 55*, 57-60.

Schweinhart, L. J., Weikart, D. P., & Larner, M. B. (1986). Consequences of three preschool curriculum models through age 15. *Early Childhood Research Quarterly, 1*, 15-45.

Scott-Little, C., Lesko, J., Martella, J., & Milburn, P. (2007). Early learning standards: Results from a national survey to document trends in state-level policies and practices. *Early Childhood Research and Practice, 9*(1), 1-30.

Shaul, M. S., Ward-Zukerman, B., Edmondson, S., Moy, L., Moriarty, C., & Picyk, E. (2003). *Head Start: Curriculum use and individual child assessment in cognitive and language development*. Washington, DC: U.S. General Accounting Office.

Smith, M. W., Dickinson, D. K., & Sangeorge, A. (2002). *The Early Language and Literacy in the Classroom Observation* (ELLCO). Baltimore: Brookes.

Songer, N. B., & Gotwals, A. W. (2005). Fidelity of implementation in three sequential curricular units. Paper presented at the American Educational Research Association.

Spira, E. G., Bracken, S. S., & Fischel, J. E. (2005). Predicting improvement after first-grade reading difficulties: the effects of oral language, emergent literacy, and behavior skills. *Developmental Psychology, 41*, 225-234.

SPSS Statistics 17, Rel. 17.0.0. 2008. Chicago: SPSS Inc.

Starkey, P., Klein, A., & Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childhood Research Quarterly, 19*(1), 99-120.

Tarullo, L., West, J., Aikens, N., & Hulsey, L. (2008). *Beginning Head Start: Children, families and programs in Fall 2006*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families.

The Consortium for Longitudinal Studies. (1983). *As the twig is bent: Lasting effects of preschool programs*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (October, 2009). *Early Childhood Education. Retrieved from* http://ies.ed.gov/ncee/wwc/reports/Topic.aspx?tid=13

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2009). *WWC intervention report: Curiosity Corner.* Retrieved from http://ies.ed.gov/ncee/wwc/reports/early_ed/curious/

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2007). *WWC intervention report: Literacy Express.* Retrieved from http://ies.ed.gov/ncee/wwc/reports/early_ed/lit_express/

U.S. Department of Education, Office of Educational Research and Improvement. (2002, January). *Preschool Curriculum Evaluation Research Grants Preapplication Meeting Summary*. Retrieved on June 27, 2009, from http://www.ed.gov/offices/OERI/pcer_materials/index.html.

U.S. Department of Health and Human Services, Administration for Children and Families (2005). *Head Start Impact Study: First year findings*. Washington, DC.

U.S. Department of Health and Human Services, Administration for Children and Families (2010). *Head Start Impact Study Final Report*. Washington, DC.

Vinovskis, M. A. (1993). Early childhood education: Then and now. *Daedalus, 122*(1), 151-176.

Waltz, J., Addis, M. E., Koerner, K., Jacobson, N. S. (1993) Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology*, *61* (4), 620-630.

Wheeler, J. J., Baggett, B. A., Fox, J., & Blevins, L. (2006). Treatment integrity: A review of intervention studies conducted with children with autism. *Focus on Autism and Other Developmental Disabilities*, *21*(1), 45-54.

Whitehurst, G. (2009). Don't Forget Curriculum [Electronic Version]. *Brown Center letters on education*. Retrieved April 2, 2010 from http://www.brookings.edu/papers/2009/1014_curriculum_whitehurst.aspx.

*Whitehurst, G. J., & et al. (1994). Outcomes of an emergent literacy intervention in Head Start. *Journal of Educational Psychology, 86*(4), 542-555.

Wiese, M.R.R. (1992). A critical review of parenting training research. *Psychology in the Schools, 29*, 229-236.

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.