Quality Thresholds, Features, and Dosage in Early Care and Education: Secondary Data Analyses of Child Outcomes

Quality Thresholds, Features, and Dosage in Early Care and Education:

Secondary Data Analyses of Child Outcomes

Margaret Burchinal, University of North Carolina-Chapel Hill,

Martha Zaslow, Child Trends, and

Louisa Tarullo, Mathematica Policy Research

Table of Contents

ACKNOWLEDGMENTS

Research reported in this publication was supported primarily by the U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation: Contract Number HHSP23320095642WC awarded to Mathematica Policy Research and partners Child Trends and the University of North Carolina-Chapel Hill (UNC-CH). Additional support came from the Institute for Education Sciences–funded postdoctoral fellowship R305B100028 award to UNC-CH, the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under Award Number P01HD065704, and from the NC state pre-kindergarten evaluation based on funding awarded to Ellen Peisner-Feinberg at UNC-CH from the NC Department of Public Instruction and the NC Department of Health and Human Services. The content is solely the responsibility of the authors and does not represent the official views of these funders. We thank Nina Forestieri for her assistance with manuscript preparation and the researchers, programs, and families that participated in the studies included in the meta-analyses.

ABSTRACT

This monograph addresses the hypotheses that preschool children benefit most strongly when early care and education (ECE) is at or above a threshold of quality, has specific quality features, and/or is of longer duration. These issues are pivotal in recent policies designed to improve the quality of ECE, especially for children from low-income families. Evidence of quality thresholds in which child care quality has stronger impacts in settings with moderate to high levels of quality than in settings with low quality would inform policy initiatives in which monetary incentives or consequences are allocated to ECE settings based on their level of quality. Evidence that specific features of quality, such as quality of teacher-child interactions and of literacy and mathematics instruction, are predictors of gains in child outcomes could help inform quality improvement efforts. Evidence that more time spent in center-based ECE or in instruction in specific content areas predict larger gains among preschoolers could be useful in designing public preschool programs such as Head Start or prekindergarten.

Methods: Secondary data analyses of eight large studies of preschool children in center care were conducted. Analyses focused on quality thresholds, and quality features examined the extent to which three types of quality measures predicted gains in children’s language, literacy, mathematics, and social skills. The measures comprised (1) global quality measures that provide an overall or global rating of quality, focusing on interactions as well as on physical features of the environment, activities, and routines; (2) interaction-specific measures that focus in depth on the quality of interactions between teachers and children with respect to instruction and emotional support; and (3) domain-specific measures that focus on the quality of instruction and stimulation in specific content areas such as early language and literacy. The goal was to provide replicated analyses with data from several projects in order to address each question. Multilevel analyses that controlled for entry skills were conducted, and results were combined by using meta-analysis, nonlinear and nonparametric analyses, and propensity score analyses.

Findings: *With respect to thresholds*, the analyses suggest that higher quality instruction is related to larger gains in language and literacy outcomes, but only in higher quality classrooms. Results point to stronger associations between quality and child outcomes in higher versus lower quality classrooms for measures of the instructional quality of teacher-child interactions and of the quality of specific activities thought to promote early literacy, such as teaching phonemic skills and book reading. In addition, the items focusing on quality of interactions on the global measure also predicted acquisition of language and social skills in higher but not in lower quality classrooms.

With respect to quality features, interaction-specific and especially domain-specific measures of quality remained significant predictors of child outcomes, whereas global measures of quality were never significant positive predictors, when both global and more specific measures of quality were included simultaneously in analyses. There is thus consistent evidence that more specific measures of quality are better predictors of child outcomes.

With respect to dosage, several approaches were used in operationalizing both the cumulative and current dosage of children's exposure to ECE. Propensity score analyses that included baseline scores on outcomes to control for selection into larger dosages suggested that children with two as opposed to one year of Head Start had stronger vocabulary and literacy skills both immediately upon exit from Head Start and at the end of kindergarten. Fewer absences and more observed time spent on instruction were associated with stronger gains in literacy and mathematics skills. Finally, findings revealed that more time spent on instruction in classrooms with higher overall quality was particularly important to the development of mathematics skills. No other replicated evidence of quality by quantity interactions emerged.

I. INTRODUCTION AND LITERATURE REVIEW

Introduction

This monograph focuses on the nature of the relationship of children’s development to the levels of quality in early care and education (ECE), features of quality, and the extent of children’s exposure to ECE. Much previous work has focused on whether the quality and extent of children’s exposure to ECE are related to their development. This monograph builds on this existing body of research and goes beyond it, focusing not on whether but instead on how quality and extent of exposure are related to children’s development. The nature of the association between quality and child outcomes is examined, with analyses focusing on the issues of quality thresholds, features, and dosage. The monograph examines the following research questions: (1) For quality thresholds, is there evidence that the quality of ECE is a stronger predictor of child outcomes in higher rather than in lower quality classrooms? (2) For quality features, do measures of specific aspects of the ECE experience provide stronger predictions of child outcome than measures of the global quality of the environment? (3) For dosage, does more time in ECE, defined in different ways (two versus one year, more cumulative hours with fewer absences, and more time spent on instruction), provide a stronger prediction of child outcomes?

These questions have become increasingly salient in light of recent demographic changes. Maternal employment and regular use of ECE have become normative for families with young children both in the United States and across Organization for Economic Co-operation and Development (OECD) nations. Though there is variation across countries and by age of youngest child (with increases in maternal employment as the youngest child grows older), the OECD average rate of maternal employment exceeds 50 percent for those with children younger than age 3 as well as those between age 3 and 5 (OECD, 2013). Data for the United States reflect the overall pattern for OECD countries, with a majority of mothers with young children now in the labor force, including 64.8 percent of those with children under 6 years, and 57 percent with infants less than one year of age working or looking for work (Bureau of Labor Statistics, April 26, 2013). Data from the spring 2011 Survey of Income and Program Participation indicate that 61 percent of children under age 5 were participating in ECE on a regular basis (Laughlin, 2013). As maternal employment and participation in ECE have become normative, there has been a growing focus on the quality of ECE. The allocation of public funds could be informed by a better understanding of the ways in which children respond to different levels and features of quality and different amounts of time in center-based ECE.

In the United States, policy approaches have focused on raising minimum standards of quality, increasing parents' access to information about quality, integrating different types of ECE into a more coordinated system, and implementing both supports and requirements for increasing quality. For example:

- Regarding minimum standards, in May 2013, the U.S. Department of Health and Human Services announced proposed changes to child care regulations that would, for the first time, require child care providers accepting funds for child care subsidies through the Child Care and Development Fund to receive on-site monitoring, undergo comprehensive background checks, comply with state and local fire, health, and building codes, and receive health and safety training (Office of Child Care, 2013).
- Regarding greater access to information on quality as well as incentives for improving quality, as of 2014 (Build Inititative and Child Trends, 2014), 38 states and localities had implemented Quality Rating and Improvement Systems. These systems provide parents with readily interpretable summary ratings of quality to inform their choices of ECE and provide

ECE settings with information, supports, and resources that can guide their efforts to improve quality (Tout, Zaslow, Halle, & Forry, 2009).

- In a joint initiative, the U.S. Departments of Education and Health and Human Services have awarded Race to the Top Early Learning Challenge grants to 26 states with the aim of fostering quality improvement efforts across types of ECE programs, moving toward more integrated early childhood systems and increasing access to high quality ECE, especially for low-income families (Zaslow, Crosby, & Smith, in press).
- As an example of quality requirements, in December 2011, the Office of Head Start implemented the Designation Renewal System, which requires Head Start Programs to compete for their grants if they do not meet a set of seven conditions, one of which involves demonstrating specific levels of observed quality on the Classroom Assessment Scoring System (CLASS).

These quality initiatives rest on assumptions that children gain academic and social skills more rapidly in higher rather than in lower quality programs. Understanding how quality and child outcomes are related could help target limited resources for quality improvement.

The growing policy emphasis on quality is based on the research evidence relating ECE quality to children’s development. Two types of studies support the conclusion that quality and child outcomes are related: evaluations of ECE programs and studies considering associations of quality and child outcomes when quality varies spontaneously. Evaluations of ECE programs seeking to support young children’s development and school readiness show positive effects on development, with the evidence pointing to moderate to large effect sizes, especially on cognitive outcomes for more intensive programs (Camilli, Vargas, Ryan, & Barnett, 2010; Duncan & Magnuson, 2013; Karoly, Kilburn, & Cannon, 2005). Studies, examining the correlations of quality and child outcomes in data from several ECE settings involving a wide range in quality show modest associations between quality and child outcomes. Given the growing policy emphasis and these findings, there is a need for greater focus on the nature of the association between ECE quality and early development.

Policies such as the development of state Quality Rating and Improvement Systems are based on identifying higher and lower quality ECE settings. It is important that developers of these systems have some confidence in their determination of the cut-points that divide higher and lower quality programs, especially in the case of high-stakes consequences. To date, most descriptive analyses relating observed quality to child outcomes rest on the assumption of a linear relationship; that is, increments in quality are related to parallel increments in child outcomes across the full range of quality. If there is a threshold of quality needed to produce moderate to strong positive child outcomes, such associations would not necessarily be captured in analyses based on the assumption of a linear relationship and an examination of the associations of quality and child outcomes in data sets using the full range of ECE quality. Rather, such analyses would need to be structured to consider associations between quality and child outcomes that are not linear or that consider the possibility of differences in the relation of quality and child outcomes in different quality ranges.

It is noteworthy that a number of the ECE programs for which evaluations showed moderate to strong positive effects involved participation for several years. Abecedarian and Perry Preschool are good examples, with Abecedarian spanning infancy to age 5 and about 75 percent of the children participating in the Perry Preschool Project at both age 3 and 4 (Campbell, Pungello, Miller-Johnson, Burchinal, & Ramey, 2001; Campbell, Ramey, Pungello, Sparling, & Miller-Johnson, 2002; Farnworth, Schweinhart, & Berrueta-Clement, 1985; Schweinhart et al., 2005). Point-in-time assessments of the relationship between quality and child outcomes in descriptive analyses do not generally take into account extent of exposure.

Research on continuity of care and children’s exposure to several ECE settings raises the possibility that, for many children, exposure to any one ECE setting may be of limited duration (Tran & Weinraub, 2006; Weber, 2006).

Finally, the limited documentation of quality, particularly in the early and most widely cited evaluation studies, poses problems. Perhaps the ECE programs that showed moderate to strong effects have such quality features as intentional engagement of children in activities and interactions aimed at building vocabulary or more explicit guidance of children to develop behavioral and cognitive self-regulation. Our current measures of quality tend to focus on overall or global quality (as in the set of Early Childhood Environment Ratings Scales; Harms, Clifford, & Cryer, 1998; Harms, Cryer, & Clifford, 2007) or on the quality of certain types of interactions (such as emotional support and instructional support in CLASS) (Pianta, LaParo, & Hamre, 2004; Pianta, Mashburn, Downer, Hamre, & Justice, 2008). It is in only a small number of studies that the quality of stimulation in specific domains (such as for language and literacy development) is considered in any detail (Zaslow, Martinez-Beck, Tout, & Halle, 2011). With features of quality poorly specified, it is difficult to know what underlies the stronger outcomes in specific aspects of children’s development in particular program evaluations.

The purpose of this monograph is to provide a focused examination of the issues of thresholds, of features of quality, and of dosage in ECE. More specifically, we ask whether stronger associations of quality and child outcomes are identified in descriptive analyses when quality is above (or below) a specified threshold, whether there are specific features of quality that underlie stronger child outcomes, and whether children show stronger effects with greater exposure to center-based ECE, particularly care of higher quality. Each of these issues is examined through secondary analyses of several ECE data sets, focused primarily on preschool children from low-income families, making replication of patterns across data sets a priority.

Literature Review

We turn now to a review of existing research that can inform the study of quality thresholds, features, and dosage. We first summarize contrasting findings from program evaluations as well as from descriptive studies examining the associations of quality and child outcomes. We then discuss substantive and methodological issues from previous research as related to thresholds of quality, features, and dosage that should be taken into account in new analyses.

Contrasting Findings in Evaluation and Descriptive Studies

Evaluation studies have supported the overall conclusion that high quality ECE experiences promote children's cognitive and social development. In a meta-analysis focusing on the evaluations of 20 early childhood programs that involved experimental or quasi-experimental designs, Karoly, Kilburn, and Cannon (2005) found evidence of significant effects in approximately two-thirds of the programs. They also found larger effects that tended to be in the moderate to large range on cognitive outcomes for programs that were more intensive and that focused on improving school readiness. A more recent meta-analysis reported moderate effect sizes in experimental studies with center-based interventions (Camilli, Vargas, Ryan, & Barnett, 2010).

A contrasting approach to considering the effects of quality has involved examining naturally occurring patterns of association between quality and child outcomes in large ECE data sets that include care of widely varying quality. These analyses consider the strength of the association between quality and child outcomes by using an assumption of linear relations (i.e., that each increment in quality is associated with an equivalent increment in scores on child outcome measures across the full range of quality). Results of these descriptive analyses point to only very modest, albeit consistent, associations between quality and child outcomes, in analyses designed to reduce but not eliminate potential biases.

To examine the strength and consistency of associations between quality and child outcomes in descriptive studies, Burchinal and colleagues (Burchinal et al., 2009; Burchinal, Kainz, & Cai, 2011) conducted both a meta-analysis summarizing the findings from published studies and coordinated secondary analyses with data from five large studies of early childhood care and education. For the meta-analysis, a literature review first identified studies that had undergone peer review, involved examination of the association between quality and child outcomes by using common measures of quality, included at least 10 center-based early childhood classrooms, and focused on preschool-age children between the ages of 3 and 5 across all income levels. The inclusion criteria were met by 20 projects (some with several published papers or reports) reporting on 97 associations between measures of quality and child outcomes. For the meta-analysis, associations between the measures of quality and child outcomes were converted to partial correlations. The magnitude of the effects was modest, with partial correlations ranging from 0.05 to 0.17. Comparisons of the effect sizes by type of outcome indicated stronger associations for language outcomes than for those related to social and emotional development across all ages.

Follow-up analyses were conducted by Burchinal, Kainz, and Cai (2011) with data from five data sets that included a large number of low-income children. The analyses found partial correlations between measures of quality and fall-to-spring gains in child outcomes, controlling for site, maternal education, ethnicity, and gender, that ranged from 0 to 0.23, with most less than 0.10. The average partial correlation was again slightly stronger for language (0.06) than for social and emotional outcomes (0.02). Slightly stronger associations were found when the researchers selected items from the quality measures that reflected a more specific aspect of quality and when the aspect of quality considered and the specific child outcomes were conceptually aligned (for example, quality items focused on use of language during caregiver-child interactions and the outcome of interest involved language development). Similarly, Keys et al. (2013) conducted parallel analyses across five large ECE study data sets and found only very modest associations between quality and child outcomes (i.e., $r_p = 0.04$).

The modest associations between quality and child outcomes found in the meta-analysis and secondary data analysis with data from descriptive studies raise important questions about how we are examining the associations between quality and child outcomes.

Thresholds of Quality

One possibility suggested by the juxtaposition of very modest effect sizes in natural history studies of settings (with low to high levels of ECE quality) and large effect sizes in causal evaluation studies (with presumably high levels of ECE quality) is that children may respond to ECE settings differently in different ranges of quality. Children may begin to benefit only when quality is at or above a certain (higher) level of quality, a pattern that might be suggested from the evaluation studies presumably involving programs with quality in a high range. However, another possibility is that increments in quality matter only up to a certain level, at which point further increments in quality no longer make a difference to child outcomes. Weak associations would come from a lack of consideration of the possibility of different patterns of association in different quality ranges.

To what extent does existing research provide evidence for either of these possible patterns? There has been an evolution in the way studies have considered whether a threshold of quality is present. Beginning with studies conducted in the 1990s, researchers have contrasted outcomes for children participating in ECE settings above or below a designated cut-point. However, it is only since 2009 that analyses have explicitly considered the possibility that the relationship between quality and child outcomes follows a nonlinear pattern or differs in strength in different segments of the quality range.

Average or Proportion Scores on Child Outcomes in Different Ranges of Quality

An analytic approach that emerged in the 1990s and that continues to be used with some frequency asks whether *average scores* on a continuous child outcome measure or the *proportion* of children showing a favorable as opposed to unfavorable score on a categorical child outcome measure differs above and below a cut-point in quality. In these studies, ranges of quality have been set according to either (1) the labels assigned to different quality ratings (for example, differentiating settings according to whether a summary rating on an observational measure of quality was at the level of ratings of “good” or higher) or (2) the distribution of quality scores (for example, using a median split on scores on an observational measure of quality).

As one example of a study conducted in the 1990s taking this approach and clearly labeled as focusing on thresholds of quality (titled “Thresholds of quality: Implications for the social development of children in center-based child care”), Howes, Phillips, and Whitebook (1992) used as their threshold the ratings above or below good or very good, respectively, on two summary scores based on the Early Childhood Environment Rating Scale (ECERS) and Infant/Toddler Environment Rating Scale (ITERS) developed by factor analysis: Appropriate Caregiving and Developmentally Appropriate Activities. Data from three samples were used: two from longitudinal studies of children in California who had entered care as infants and data from the Atlanta site of the National Child Care Staffing Study (Howes, Phillips, & Whitebook, 1992). Findings indicate that children from settings rated good or higher on appropriate caregiving differed from those in settings rated lower in terms of the proportion of children showing secure attachment. In addition, children from settings rated very good on developmentally appropriate activities were more likely than those from settings rated lower in quality to be classified as both peer- and adult-oriented rather than solitary.

A subsequent study conducted by Vermeer et al. (2010) provides an example of research in which the pattern of child outcomes was contrasted for children participating in settings above and below a cut-point based on the distribution of quality scores. The research group examined the average pattern of change during the course of the day in cortisol production, as research suggests that rising cortisol during the day is a marker of stress. Using a median split for quality on the ECERS-R for center-based care settings in the Netherlands as well as in Basque Country, the researchers found a rise in cortisol during the day when ECE quality was below the median, but a decline in cortisol when quality was above the median.

Another subsequent study in which average child outcome scores were examined for different portions of the distribution of quality scores comes from research by Watamura, Phillips, Morrissey, McCartney, and Bub (2011), analyzing data from the National Institute of Child Health and Human Development's Study of Early Child Care and Youth Development (NICHD SECCYD). In these analyses, the distributions of ECE quality using the Observational Record of the Caregiving Environment (NICHD Early Child Care Research Network (ECCRN), 1996) and of the quality of the home environment based on the Home Observation for Measurement of the Environment (Caldwell & Bradley, 1984) as well as on observations of mother-child interaction were divided into thirds. Children whose home environments were in the top third of the distribution during their preschool years were protected from the effects of poor quality ECE, defined as care in the bottom third of the distribution, in terms of average scores on measures of social and emotional development. However, children from home environments in the bottom third of the distribution were responsive to ECE in the bottom third as well as to the protective role of high quality ECE (observed to be in the highest third of quality).

Testing for Nonlinear Relations and for Differences in the Strength of Associations Between Quality and Child Outcomes in Different Ranges of Quality

Beginning in 2009, researchers have focused not only on mean scores (or percentage of children showing favorable categorical scores) in different ranges of quality but also on the possibilities of nonlinear associations between quality and child outcomes and differences in the strength of the association between quality and child outcomes in different quality ranges. Nonlinear patterns of association might follow a pattern of improvements in child outcomes in relation to improvements in quality up to only a certain level of quality, with no improvements in child outcomes thereafter. Alternatively, improvements in child outcomes might be found only after quality had reached a certain level. Examining the strength of the association in different ranges of quality allows for a test of whether the strength of the relationship differs above and below a threshold of quality.

The search for thresholds has involved two methodological approaches. The first approach sets the threshold a priori while the second approach uses the data to identify the threshold. Most analyses to date have used cut-points based on professional recommendations and then tested those values for evidence of thresholds. Such an a priori approach provides a means for examining the professional recommendations by testing the same thresholds across child outcomes and studies in order to test the strength of findings. The results may be easily used to create or examine policies that rely on those cut-points for program evaluation or communication with parents. For example, Quality Rating and Improvement Systems often include observational measures of quality and assign points according to whether an ECE program scores below or above selected cut-points. In another approach involving tiered reimbursement, ECE programs may collect higher subsidy rates if the rating for their program exceeds a certain level. Similarly, the Head Start Designation Renewal System used one quality measure—CLASS—as part of its monitoring system, and programs that on average are below a cut-point on CLASS are required to compete for the renewal of their contract. Evidence that there is a threshold at or near the points used by these systems would provide important policy information and could be used to refine the systems. Linear spline or piecewise regressions have typically been used in these studies.

In contrast, the second approach involves examining each bivariate association between a quality measure and a child outcome to see if there is any evidence of thresholds when nonlinear and often nonparametric models are fit. By letting the data speak, this empirical approach provides the greatest opportunity to identify the best threshold for a given data set for the selected quality and outcome variables. This approach is much more useful for thinking about where the thresholds might be instead of evaluating a widely used cut-point. Many types of analytic models have been used, including nonlinear polynomial models (e.g., quadratic models), high-order spline models, and nonparametric nonlinear models.

Both approaches have their advantages and disadvantages. The advantage of the a priori approach is its clarity and policy relevance. By fitting the same model to all data, it is easy to evaluate the evidence for thresholds for a given quality variable across outcomes and across data sets. Findings may be easily translated into policy and practice if they provide support for professional recommendations and should be useful in refining recommendations if the findings do not support them. The disadvantage is that an optimal threshold is not identified. In contrast, the advantage of the empirical approach is the identification of the optimal threshold for a given bivariate association for a particular data set. This approach yields a different cut-point from each analysis across outcomes and across data sets because it is trying to find the optimal regression curves and cut-points for that specific data set and those specific variables. The distribution of the cut-points as determined across analyses of different child outcomes and in different data sets lends itself to examination, but there is no one statistical test that can determine the extent to which a given cut-point is supported across all of the analyses. Thus, the empirical approach should provide greater clarity about probable cut-points for specific quality variables and child outcomes, but it offers no clear answer across quality variables and child outcomes or about the cut-points currently used in policy and practice.

Recent studies taking the first or a priori approach to setting thresholds use piecewise or spline regressions in which the linear associations between quality and outcomes are allowed to differ in lower and higher quality classrooms when the quality ranges are set on conceptual grounds, for example, according to the labels for ratings used in observational measures of quality or previous research with the measure of quality. One slope is determined for the classrooms falling in the lower quality range and another in the higher quality range, and the difference between classrooms is tested. Burchinal, Vandergrift, Pianta, and Mashburn (2010) used this approach in a study analyzing data from an 11-state prekindergarten study. They found that academic outcomes were more strongly related to instructional support on CLASS when classrooms were in the moderate to high quality range than in the low quality range, according to the descriptions of ratings used for CLASS. Further, social outcomes were more strongly related to emotional support on CLASS when classrooms were in the high quality range rather than in the moderate or low quality range.

A recent study by Weiland, Ulvestad, Sachs, and Yoshikawa (2013) illustrates the use of the second or empirical approach. The study asks whether there are linear or nonlinear associations between quality and child outcomes for children participating in prekindergarten programs in Boston public schools. The measures of quality included the Interactions factor on the ECERS-

R; the CLASS Emotional Support, Instructional Support, and Classroom Organization scores; and the Early Language and Literacy Classroom Observation (ELLCO) rating of Classroom Literacy Activities (Smith, Brady, & Anastasopoulos, 2008). Child outcomes included measures of receptive vocabulary, cognitive inhibitory control, and working memory. Cut-points were set by empirical examination of the inflection point where a significant quadratic relationship was found in the analyses. Using inflection point spline knot analyses, CLASS summary scores for Emotional Support and Classroom Organization were positively associated with the measure of cognitive inhibitory control only in the higher range of quality, whereas CLASS Instructional Support was found to be negatively related to the measure of cognitive inhibitory control in the lower quality range, but positively related in the higher quality range. These findings were replicated in a recent study of preschool programs in rural low-income regions (Burchinal, Vernon-Feagans, Vitiello, & Greenberg, 2014).Thus, in the research to date, we see some evidence of thresholds when either an a priori or an empirical approach has been used. In these studies, we see some evidence of an upper quality threshold, such that the relationship between quality and child outcomes is stronger in higher versus lower quality ranges. However, the pattern has not been found consistently across either quality measures or child outcomes. Accordingly, in new work reported in this monograph, we will prioritize the conduct of parallel analyses across several data sets and, using meta-analysis to summarize the patterns, in data sets focusing on children from low-income families. We will use both approaches noted above, conducting spline analyses by using a priori cut-points to set thresholds and seeking to identify inflection points with the use of empirical methods. Based on the research to date, we expect to find a stronger relationship between quality and child outcomes in higher quality ranges. It is important to examine not only whether we see such associations but also the consistency of the pattern across data sets, measures of quality, and child outcomes.

Quality Features

A key question that has emerged in the research on quality is whether there is a stronger association between quality and child outcomes when the association is examined between a child outcome in a specific domain (such as expressive vocabulary) and features of quality that involve stimulation specific to that domain (such as introduction of and encouragement to use new words). Earlier, we noted that Burchinal et al. (2011) found some evidence of slightly stronger associations between quality and child outcomes when analyses involved quality measures that could be considered more closely "conceptually aligned" with specific child outcomes. NICHD ECCRN and Duncan (2003) also considered whether more tightly aligned measures showed stronger associations. The group found evidence supporting this prediction for cognitive and language stimulation and outcomes, but not for other domains.

The possibility that there are stronger associations between aligned features of quality and child outcomes would have implications for practice as well as for measurement and analysis. Such findings would suggest that, if we are aiming to strengthen specific school readiness outcomes, it may not be sufficient to improve global quality. Targeted efforts aimed at strengthening specific features of quality, such as language and literacy practices or support for self-regulation, may be needed to strengthen related outcomes.

In reviewing previous research for further evidence that more tightly aligned measures of quality and child outcomes show stronger and/or more consistent association, it is necessary to differentiate among groups of quality measures that vary in the extent to which they focus on specific quality features. We see three "waves" of quality measurement in previous research, with each successive wave focusing on more specific features of quality. We label the waves of quality measurement as involving (1) global measures; (2) "interaction-specific" measures; and (3) "domain-specific" measures.

The first category of quality measures (for example, ECERS; Harms, Clifford, & Cryer, 1998; Harms, Cryer, & Clifford, 2007) provides summary scores looking broadly across different features of quality, including not only caregiver-child interactions but also physical features of the care setting (such as appropriateness of furniture and space for children; availability of play and learning materials), structuring of activities, and features of the environment important for the caregivers. Interaction-specific measures take a major step toward greater specificity by separating different aspects of interactions. A key example is the CLASS (Pianta, La Paro, & Hamre, 2004), which separates Emotional Support and Instructional Support (as well as Classroom Organization). These CLASS summary scores, however, are limited in the extent to which they go the further step of focusing on interactions involving specific content. Examples of domain-specific measures include the Classroom Observation of Early Mathematics (Clements & Sarama, 2008) and the Early Literacy Observation Tool (Grehan & Smith, 2004).

In keeping with their history of emergence, global measures of quality have seen much more widespread use, both in research examining associations of quality and child outcomes and for policy and practice purposes. However, interaction-specific measures are gaining greater use over time. For example, the prekindergarten version of CLASS is now used for the monitoring of quality throughout Head Start programs (Office of Head Start, 2011). Domain-specific measures are most recent in development and least widely used (Zaslow et al., 2011).

It is important to ask if prediction of child outcomes is stronger as we contrast measures from the three categories of quality measurement and move toward measures with greater specificity. This requires analyses of data from studies that include measures of quality from more than one category. For our purposes, it is also important to consider analyses that look across measures of quality and child outcomes as well as to examine associations to those considered conceptually aligned. Whereas the latter approach has the strengths of being hypothesis-driven and limiting the number of associations examined in the present context (see, for example, the analyses by Moiduddin, Aikens, Tarullo, West, & Xue, 2012), the more exploratory analyses are of greater relevance, as they can inform the issue of whether more closely aligned measures of quality and child outcomes are more consistently or strongly related than measures of quality and child outcomes less closely aligned.

There is only a small set of studies examining this set of issues. In one such study, Mashburn et al. (2008) intentionally contrasted patterns of prediction of child outcomes from the first two waves of quality measurement: global and interaction-specific measures. Using data from a study of state-funded prekindergarten programs from 11 states, the research group found that the summary score for the CLASS Instructional Support measure was significantly related to all of the examined measures of child academic and language skills, whereas the CLASS Emotional Support score was related to both improved social competence and fewer behavior problems. Thus, aligned measures were related. However, the global measure of quality—the ECERS-R—predicted only a single outcome related to oral and written language. The authors concluded that focusing on teacher-child interactions provided a stronger basis than global measures for measuring quality, stating that "the measure of pre-k quality that was most consistently and strongly related to measures of children's development was dimensions of teacher-child interaction directly experienced in classrooms" (p. 743).

The study by Weiland et al. (2013) begins to take the next step, looking at patterns of prediction when both interaction-specific and domain-specific measures are considered. Using data from the study of public prekindergarten in Boston, the research group examined the prediction of measures of child language development (using a measure of receptive vocabulary), cognitive development (considering a measure of working memory), and executive function (examining inhibitory control) of interaction-specific measures of quality (the CLASS dimension scores) and a domain-specific measure of language and literacy interaction (the ELLCO Toolkit; Smith, Brady, & Anastasopoulos, 2008). Findings indicate that the interaction-specific scores from CLASS consistently predicted the measure of children's executive function. However, counter to predictions, the ELLCO rating of literacy activities did not predict children's receptive vocabulary. Rather, it also predicted children's executive function. The authors note that there was no evidence of better prediction with aligned measures of quality or of better prediction from the domain-specific measure of quality in general. However, rather than dismissing the possibility that domain-specific measures may provide a better basis for prediction of child outcomes, the authors call for consideration of a wider range of domain-specific measures.

Another key issue emerging in the literature encompasses the issues of quality features and thresholds considered simultaneously. Here, analyses explore whether quality thresholds are detected especially when the measures of quality focus on features of quality in specific domains. As noted, Weiland et al. (2013) did find evidence of thresholds in relating the interaction-specific summary scores from the CLASS and scores on a measure of children's executive function. But there was no evidence for thresholds of quality in the domain-specific measure—the ELLCO rating of literacy activities—examined in relation to the development of expressive vocabulary or the other outcomes considered.

In sum, we have limited research aimed explicitly at contrasting the consistency and strength of prediction of child outcomes from global, interaction-specific, and domain-specific measures of quality. Careful attention to selection factors is needed in analysis, because families who place their young children in higher quality programs also provide many other advantages to their children (Burchinal, Magnuson, Powell, & Hong, 2014).The research to date is constrained, especially by the limited examination of domain-specific measures of quality. The work conducted thus far does suggest that interaction-specific measures of quality are more consistently related to child outcomes than are global measures of quality. However, further work is needed to test for replication of this pattern. In addition, research must take the further step of examining a range of domain-specific measures and asking not only whether the measures show more consistent or stronger associations with child outcomes but also whether there is greater evidence of thresholds when using such measures. To build on the existing research, we make it a priority to examine these issues in new analyses.

Dosage

Studies looking at child outcomes in light of the extent of participation in ECE have most often focused on whether and how children's development varies with more hours or days of either current or cumulative participation in center-based ECE or ECE overall. However, a small set of studies also considers the issue of child outcomes in light of greater participation in ECE that is of higher quality, thus jointly considering the issue of thresholds and dosage. A recurrent set of methodological issues appears in the research on dosage, again pointing to the need to attend carefully to selection factors. Children who either spend more time in center-based ECE settings or in higher quality ECE may have different initial characteristics that help to explain patterns of association of greater exposure and child outcomes. Only by controlling for such factors can we obtain a clear picture of the association between dosage and outcomes.

Extent of Participation in Center-Based Care

There is an accumulation of evidence that greater participation in center-based ECE is associated with stronger cognitive outcomes. However, results for behavioral outcomes are mixed, with some but not all studies pointing to more problematic outcomes with more extensive participation in centers. Some of the recent studies suggest that ratio or group size in center-based ECE settings may help explain more problematic behavioral outcomes when they do occur.

For example, in analyses focusing on the first three years of life, the NICHD Study of Early Child Care and Youth Development (NICHD ECCRN, 2000) found, after controlling for a host of family and child characteristics, that “the longer children were in centers, beginning at age 6 months, the better they performed on cognitive and language measures” (p. 976). Loeb, Fuller, Kagan, and Carrol (2004) contrasted the development of low-income children in center-based versus home-based ECE provided by family, friends, and neighbors at about 2.5 and 4 years. Children in center-based care at both time points had the strongest scores on the Bracken measure of cognitive school readiness and measures of emergent literacy (such as familiarity with books), whereas children who transitioned into center-based care from home-based care between the two time points still scored higher than those who experienced only home-based care. A study by Sylva, Stein, Leach, Barnes, and Malmberg (2011) in a large and demographically diverse sample of infants and toddlers in England found that those who were in group “nursery” care (center-based care) at 18 months had higher scores on measures of cognitive development than children in other types of care.

Some studies examining more extensive participation in center-based ECE report increases in problematic social behaviors, whereas other studies do not show such a pattern. Thus, for example, the NICHD Study of Early Child Care and Youth Development (SECCYD) reported both stronger academics but also increased problem behavior when children in the United States participated in center-based ECE (NICHD ECCRN and Duncan, 2003). Various explanations for the apparent negative effect of center care on social development have included exposure to larger number of children overall or relative to the number of adults (McCartney et al., 2010; Yamauchi & Leigh, 2011). In contrast, the studies in both the U.K. with toddlers (Sylva et al.,

2011) and U.S. with low-income children (Votruba-Drzal, Colely, Chase-Lansdale, 2004) reported more time in center care was related to both stronger cognitive development and stronger emotion regulation scores.

Extent of Overall Participation in Nonmaternal Care

Several analyses of the data collected by the NICHD ECCRN have considered cumulative exposure to ECE, controlling for both type and quality. A pattern of greater cumulative exposure to nonmaternal care and less positive social outcomes was first documented for young children, with more recent analyses showing that the pattern is sustained into adolescence.

For example, an early study by the NICHD ECCRN (1998) reported that spending more hours in care during the first two years of life was associated with less social competence and more behavioral problems. Extending this approach over a longer time period, the NICHD ECCRN (2006) found that the overall extent of participation in ECE was associated with higher levels of caregiver-reported problem behaviors at 36 and 54 months and more caregiver-child conflict at 54 months, yet also with stronger social skills at 24 months. Examining outcomes in high school, Vandell et al. (2010) reported that adolescents who spent more hours in ECE during their first 4.5 years themselves report slightly more risk-taking and impulsive behaviors than adolescents who had spent fewer hours in ECE.

McCartney et al. (2010) reported that “there is an effect of child care hours on externalizing behavior at all levels of quality. The association is multiplicative such that the child care hours effect is smallest in high-quality care and largest in low-quality care.” Further, the “number of hours spent in early child care predicted externalizing scores, controlling for concurrent child care hours as well as selection factors. . . ” (p. 10). Most recently, Belsky and Pluess (2012) reported that self-reported impulsivity at age 15 was predicted by greater exposure to any nonmaternal care in the early years, taking into account background characteristics as well as early difficult temperament.

It is important to note that two NICHD ECCRN analyses (NICHD ECCRN, 2000; NICHD ECCRN & Duncan, 2003) have found no corresponding association between overall cumulative participation in ECE and cognitive outcomes. Further, going beyond the NICHD SECCYD, Votruba-Drzal, Coley, and Chase-Lansdale (2004) found that the extent of overall current participation in ECE at age 3 among a very low-income sample was associated with a diminished likelihood of scoring in the clinical range for behavior problems and with higher scores on a quantitative skills measure. There is a need for further work that looks at overall extent of participation, net of type and quality of care, in a wider range of samples.

Extent of Participation in Care of High Quality

As noted, McCartney et al. (2010) found the pattern of unfavorable associations between more extensive cumulative participation in ECE and children's behavioral outcomes to be weaker when care was of higher quality. This finding raises the possibility that extent and quality of care may show patterns of interaction. Earlier, we noted the possibility that care of particularly high quality, when experienced for longer periods, may help explain the overall stronger pattern of effects in evaluations of ECE programs intentionally focused on strengthening school readiness. Do we see a broader pattern of evidence that greater exposure to higher quality ECE maximizes favorable outcomes and minimizes negative outcomes?

In the study of very low-income children noted above, Votruba-Drzal, Coley, and Chase-Lansdale (2004) also found that children currently participating in high quality care, but not those participating in low quality care, showed a steep decline in both internalizing and externalizing behavior problems as hours of participation in care increased. In addition, as hours in low quality care increased, children's externalizing behavior problems increased. The authors conclude “that extensive hours of care in high-quality arrangements may be protective for children’s socio-emotional functioning, whereas long hours of care in low-quality settings may be particularly detrimental for children’s rates of externalizing behavior problems” (p. 307).

Extending the focus to cognitive outcomes, Dearing, McCartney, and Taylor (2009) measured care quality at 6, 15, 24, 36, and 54 months in the NICHD SECCYD and found that spells in high quality care could reduce the gap in achievement measures associated with the income-to-needs ratio. Children who were in high quality care at three or more of the time points during early childhood demonstrated no association between their income-to-needs ratio and their outcomes on broad mathematics, broad reading, and letter-word identification measures. Even one spell of high quality care in early childhood had statistically significant impacts on the mathematics scores of low-income children.

Whereas the studies summarized above consider current or cumulative exposure to care of varying quality, other studies have looked at extent of participation in programs assumed to be relatively high in quality on the basis of program goals or program monitoring. For example, Hill, Brooks-Gunn, and Waldfogel (2003) found that number of days of exposure to a center-based early intervention program for low-birth-weight babies was related to outcomes on assessments of intelligence at age 8. Children who attended over 400 of the possible 500 days of care demonstrated a 7 to 10 point increase on the Wechsler Intelligence Scale for children’s full and verbal scores at age 8. Low-birth-weight babies who attended over 350 days of the program also showed significant improvement at age 8, although the results were not as large as those experienced by the group that attended over 400 days. Similarly, Hubbs-Tait et al. (2002) looked at the effects of days of attendance at Head Start, which they describe as a program required to meet program performance standards and therefore falling in a higher quality range. They found that, for Head Start children at highest sociodemographic risk, as days of Head Start during a year increased, so did sociability and receptive vocabulary scores.

In summary, studies of dosage conducted to date point most consistently to better cognitive outcomes with greater exposure to center-based care. However, at the same time, they raise the possibility that greater cumulative nonmaternal care, and possibly more exposure to center-based care, may be associated with less positive behavioral outcomes. There is some indication that the pattern of unfavorable behavioral outcomes with a larger dose of nonmaternal care may reflect group contexts involving larger group size. At the same time, we see some indications that a larger dosage of higher quality care or sustained exposure to care in programs with early intervention goals or guided by performance standards may result in more favorable outcomes across both cognitive and behavioral domains, especially for children in low-income families.

In building on the existing literature, it is particularly important to go further in addressing selection effects. Although studies to date systematically control for characteristics that would predict both more participation in ECE and better developmental outcomes (such as more highly educated parents), there may be further important variables that are difficult to document with the available data. Approaches such as propensity score matching would help to account more rigorously for such selection effects. Further, whereas the research to date considers cumulative and current hours of participation, more fine-grained measures of dosage could be extremely informative. For example, rather than looking at quality and cumulative or current hours of participation, research could directly consider a child’s extent of exposure to instructional interactions within an observed sample or time.

In summary, this monograph addresses three interrelated questions about ECE experiences as follows:

For quality thresholds, is ECE quality a stronger predictor of child outcomes in higher versus lower quality classrooms and when using both a priori and empirical approaches to setting thresholds?

For quality features, do measures of specific aspects of the ECE experience—through interaction-specific or domain-specific measures of quality—provide stronger predictions of child outcomes than do measures of the global quality of the environment?

For dosage, does more time in ECE, defined in different ways (two versus one year, more cumulative hours with fewer absences, and more time spent on instruction) provide stronger predictions of child outcomes? Does ECE quality moderate dosage effects?

We address these questions iteratively to build an integrated set of evidence. First, we test for quality thresholds and use those thresholds in subsequent analyses when indicated by the results. We then ask which of the various types of quality measures provide the strongest prediction of child outcomes, using nonlinear analyses for each quality measure if indicated. Next, we ask about the dosage in analyses that account for ECE quality and ask whether quality moderates dosage effects. Our goal is to conduct analyses that are as rigorous as possible by controlling for demographic factors and the child's entry skills and by using methods such as propensity score matching when appropriate. Finally, to provide evidence that is as comprehensive as possible, we systematically look at issues of replication of findings across studies, using meta-analysis where appropriate.

II. METHODS

To address the issues regarding quality thresholds, features, and dosage identified in the literature review, secondary data analyses were conducted using data from eight large-scale ECE research projects. Analyses were conducted in parallel across as many data sets as possible that met our criteria for inclusion in order to identify patterns that were replicated across data sets and across types of center-based ECE (including Head Start, prekindergarten and community-based child care).

All studies included in these analyses focused primarily on children from low-income families. Most studies included only children enrolled in Head Start or public prekindergarten programs, but a few included children in community-based child care. Head Start primarily serves children from low-income families—although the program requires 10 percent of enrollment to comprise children with special needs, regardless of family income. Most state prekindergarten programs restrict service to or favor the recruitment of children from low-income families. In contrast, community-based child care tends to serve families from a wide variety of socioeconomic backgrounds.

Studies

To be included in the analyses, we required a project to have baseline and endpoint assessments of children in their preschool years and to include direct assessments of classroom quality using a widely used measure of global quality, teacher-child interactions, or quality of instruction. About half of the projects had specified both a global quality measure and a measure of either teacher-child interactions or the quality of domain-specific instruction.

Eight studies were included in any of our analyses: three studies of Head Start preschool classrooms, three prekindergarten studies, one multisite set of curricula studies that were largely tested in Head Start or prekindergarten programs, and one follow-up study of Early Head Start in which children experienced a variety of center-based care as preschoolers. Each is briefly described below.

FACES 2006

The Head Start Family and Child Experiences Survey (FACES) was first launched in 1997 as a periodic longitudinal study of Head Start program performance. Successive nationally representative samples of Head Start children, their families, classrooms, and programs provide descriptive information on the population served; staff qualifications, credentials, and attitudes; Head Start classroom practices and quality measures; and child and family outcomes. The FACES data come from a battery of child assessments administered in fall and spring across several developmental domains; interviews with children's parents and teachers about the child in fall and spring; interviews with children's parents, teachers, and program managers about their backgrounds in fall; and direct observations of classroom quality in spring.

The FACES 2006 sample includes 60 Head Start grantees, 135 centers, 410 classrooms, 365 teachers, and 3,315 children who entered Head Start at age 3 or 4 in fall 2006 (West, Tarullo, Aiken, & Hulsey, 2008). Classroom observations were conducted in a representative sample of 335 classrooms attended by 3- and 4-year-old children enrolled in their first year of Head Start, and children were assessed in fall 2006 at entrance into Head Start, through one or two years of program participation, with follow-up in the spring of kindergarten. Approximately two-thirds (63%) of children in the sample were age 3 at enrollment, and the others were age 4 or older. Boys slightly outnumbered girls, a pattern more pronounced among 4-year-old children (54% versus 47%, respectively). Just over a third of children were Hispanic; another third were African American; and one-quarter were European American. On average, children were 3.9 years of age (SD = 0.5) at fall assessment (mean age = 3.5 years for the 3-year-old cohort and 4.5 years for the 4-year-old cohort).

FACES 2009

The FACES 2009 sample includes 60 Head Start grantees, 129 centers, 486 classrooms, 439 teachers, and 3,349 children who entered Head Start at age 3 or 4 in fall 2009 (Moiduddin, Aikens, Tarullo, West, & Xue, 2012). Sixty-one percent of children in the sample were age 3, and the others were age 4 or older. The sample was nearly evenly divided between boys and girls. More than a third of children (36%) were Hispanic/Latino; another third (33%) were African American; and 23% were European American. On average, children were age 4 (SD = 0.6) at fall assessment (mean age = 3.6 years for the 3-year-old cohort and 4.5 for the 4-year-old cohort).

The Head Start Impact Study (HSIS)

The Head Start Impact Study evaluated the effectiveness of Head Start in improving children's outcomes (Puma, Bell, Cook, Heid, & Lopez, 2005). The evaluation involved a nationally representative cluster sample of 84 Head Start grantee/delegate agencies with waiting lists for preschoolers. Children whose families enrolled them in participating Head Start centers in the 2002–2003 school year of Head Start were randomly assigned to either a Head Start center or a control group not eligible to enroll, resulting in a sample of 4,442 newly eligible 3- and 4-year-old children. Overall, the 3-year-old cohort was 4.1 years and the 4-year-old cohort was 5.0 years in fall.

Half of the sample included in the analyses was male (49% of the 3-year-old children and 52% of the 4-year-old children). About one-third was African American (37% of the 3-year-old children and 23 % of the 4-year-old children), Hispanic (34% of the 3-year-old children and 43% of the 4-year-old children), or European American/other (29% of the 3-year-old children and 34% of the 4-year-old children). Over one-third of the mothers in both cohorts had less than a high school education, and fewer than half were married. Data used here come from fall, winter, and spring of the first year of the evaluation. Children were administered language and academic tests individually by trained research assistants in fall and spring. The quality of the child care setting was assessed in winter for children in child care. In spring, parents and teachers were asked to complete questionnaires describing their backgrounds and, in spring, questionnaires describing the child's socioemotional skills. Caregivers were not asked to rate children's socioemotional skills in fall.

In the analyses of ECE quality, for two reasons, we restricted the sample to children enrolled in Head Start regardless of treatment status. First, only center-based ECE was included in any study. Second, the proportion of classrooms with classroom quality data was much higher for children in Head Start than for children in other center-based programs, and about 25% of the control group attended Head Start at other programs and about 20% of the treatment group did not attend Head Start; therefore, treatment/control groups and Head Start attendance were not completely aligned. In contrast, all children were included in analyses of dosage.

The National Center for Early Development and Learning (NCEDL)

The National Center for Early Development and Learning 11-State Pre-Kindergarten Evaluation was conducted in 6 states in 2001 and 5 additional states in 2003 (Howes et al., 2008). States were selected because they had widely used public prekindergarten programs that had been in existence for at least five years. Representative samples of prekindergarten programs within selected regions in the 11 states were recruited (for details, see http://www.fpg.unc.edu/~ncedl or Howes et al., 2008). The 11 states served approximately 80% of children in the United States who attended state prekindergarten programs in the study years 2001–2003. The design involved random sampling of 40 prekindergarten sites in each of 6 states (for a total of 240 classrooms) during the 2001–2002 school year (78% recruitment rate) and 100 sites within each of the additional 5 states (total of 500 classrooms) during 2003–2004 (77% recruitment rate).

In both studies, within each prekindergarten site, one classroom was randomly selected to participate. Four children were randomly selected per classroom (from among children whose parents returned signed consent forms, were scheduled to enter kindergarten the next year, spoke English or Spanish at home, and did not have an Individualized Education Plan). About three-fourths of the prekindergarten programs were targeted to low-income children, resulting in economic diversity in the sample. Among the recruited children, 24% were African American, 24% were Latino/Hispanic, and 55% were European American; 49% were male; and 55% had family incomes that qualified them for free or reduced-price lunch in school (income/federal poverty threshold < 1.8). Children on average were 4.6 years of age (SD = 0.3) in fall of prekindergarten. For all children, children and their classroom experiences were assessed in fall and spring of the prekindergarten year and in fall and spring of the kindergarten year for the first cohort of children.

North Carolina Prekindergarten (NC-PK)

The North Carolina Prekindergarten program, formerly know as More-at-Four, (for full details, see Peisner-Feinberg & Schaaf, 2007, 2008) is a state-funded initiative providing a classroom-based educational program for at-risk 4-year-old children, with the aim of helping them be more successful when they enter kindergarten. The program first targets at-risk “unserved” children (those not already served in a preschool program) and then “underserved” children (those in a program but not receiving child care subsidies and/or those in lower quality settings). NC-PK provides funding for prekindergarten classrooms at a variety of sites, including public schools, Head Start centers, and community child care centers (both for-profit and nonprofit). Local sites are expected to meet program guidelines and standards around curriculum, training and education levels for teachers and administrators, class size and student-teacher ratios, North Carolina child care licensing levels, and provision of other program services. Children are eligible for NC-PK based on family income (up to 75% of state median income or up to 300% of federal poverty status) and other risk factors (limited English proficiency, identified disability, chronic health condition, and developmental/educational need).

The evaluation sample represents a random sampling of classrooms and involves examination of the quality and gains in child outcomes during the child's year in NC-PK. The evaluation sample includes data from three waves, and includes 99 classrooms and 514 children from the 2003–2004 academic year (see Peisner-Feinberg & Schaaf, 2007), 57 classrooms and 478 children from the 2005–2006 academic year (see Peisner-Feinberg & Schaaf, 2007), and 50 classrooms and 321 children from the 2007–2008 academic year (see Peisner-Feinberg & Schaaf, 2008). On average, children were 4.5 years of age (range = 4.0–5.1 years) at at the time of enrollment, 37% of children were African American, 33% were European American, and 25% were Latino/Hispanic; and 51% were male.

My Teaching Partner (MTP)

My Teaching Partner evaluated models of teacher professional development in prekindergarten classrooms (see Pianta, Mashburn, Downer, Hamre, & Justice, 2008). MTP was an NICHD-funded Interagency School Readiness Consortium Intervention grant testing the impacts of two models of teacher professional development and support—a video library and teaching consultation—on the quality of prekindergarten teachers' interactions with children and on children's development of literacy, language, and social skills. School districts were randomly assigned to one of the two treatment conditions or to a comparison condition, and all prekindergarten teachers in those districts were enrolled in the study for two consecutive academic years. Four children were randomly selected from each prekindergarten teacher's classroom during each of the two academic years. Because teachers in the comparison condition did not participate in observations of the quality of classroom interactions, they and their children were not included in the current study.

Participants in the study included teachers and children who received one of the two models of professional development support. One group of teachers was given access to materials for implementing a year-long, explicit, and comprehensive classroom-based curriculum (see Pianta, Mashburn, Downer, Hamre, & Justice [2008] for a complete description of the materials) as well as access to a web-based video library that provided numerous examples of high quality interactions in prekindergarten classrooms. The other group of teachers was given access to the same materials and the video library, with the addition of a teaching consultancy: direct, regular, and individualized feedback on lessons implemented and consultation to improve implementation. The teaching consultancy was shown to improve classroom quality (Pianta, Mashburn, Downer, Hamre, & Justice, 2008) and some child outcomes (Mashburn, Downer, Hamre, Justice, & Pianta, 2010).

The 157 teachers selected for the secondary analyses met the following two criteria: it was the teacher’s first year participating in the MTP study, and the teacher had at least four videotaped observations collected from his or her classroom during the school year. During the two intervention years, observations of classroom quality and teaching practices were collected every two weeks from the teachers in both treatment groups via submission of 30-minute videotapes featuring their implementation of literacy and language activities. In each of the two intervention years, the four children randomly selected from each prekindergarten teacher’s classroom were administered direct assessments of language and literacy skills in fall and spring. In addition, teachers reported on children’s social relationships and behavioral outcomes at these two time points.

From the 157 classrooms, a total 695 children participated in fall and spring assessments. The average age of children at the beginning of the prekindergarten year was 4.4 years (SD = 0.31); 48% of the children were boys; 48% were African American, 29% were European American, 10% were Hispanic, and 13% were multiracial or another race/ethnicity; the average yearly income of children's families was $26,546 (SD = $19,627); and the average years of maternal education was 12.7 (SD = 2.06).

Preschool Curriculum Evaluation Research (PCER)

The Preschool Curriculum Evaluation Research evaluated selected curricula in Head Start, state prekindergarten, and community child care programs (Preschool Curriculum Evaluation Research Consortium, 2008). It was funded by the Institute of Education Science of the U.S. Department of Education to examine the impacts of a variety of preschool curricula (http://ies.ed.gov/ncer/pubs/20082009/index.asp). The study was conducted by 12 grantees and evaluated the impact of 14 curricula. Within each site, there was random assignment of children or programs to the treatment or control curricula. With few treatment effects observed on classroom quality or child outcomes, the data from the entire sample were included in this study. Data were collected on 2,910 children in 320 preschool classrooms in 210 preschools. The classrooms ranged from being part of a state public prekindergarten (58%) to Head Start (31%) and community-based child care (12%). On average, the children were 4.6 years of age in fall of their last year of preschool. About half (51%) were male. The sample was diverse, with about 33% European American/non-Hispanic, 43% African American, and 16% Hispanic. The children tended to be from low-income families; fewer than half of the parents were married, and 19% of mothers had not completed high school.

Data were collected in fall, winter, and spring of the year before entry to kindergarten. Direct assessments of language and academic skills were collected by trained research assistants in fall and spring. Caregivers rated the child's socioemotional skills in fall and spring. Parents and caregivers provided information about themselves and the programs in fall. In winter, ECE quality was rated in the classroom.

The Follow-Up Study of the Early Head Start Research and Evaluation Project (EHS)

The National Early Head Start Research and Evaluation Project followed children into their preschool settings. The study included 3,001 families living in the areas of 17 program sites. The sites were in diverse communities that reflected the socioeconomic and political context of low-income families in the United States during the late 1990s; within each site families were randomly assigned to either the EHS or control condition. The study reported significant impacts of EHS on parenting and children's cognitive and social skills in the first three years (Love et al., 2005). Children and families were followed after treatment ended at age 3. The current analysis did not examine the impacts of EHS but instead focused on the quality of ECE during the prekindergarten follow-up study as experienced by both treatment and control group children. Seventy-one percent of the sample was followed after participation in Early Head Start, ECE arrangements in the preschool years were observed, and child outcomes were measured during the year before kindergarten entry. Only the 1,043 children in center-based ECE settings during their prekindergarten follow-up year were included in these analyses. Boys and girls were split evenly in the sample. About 37% of the children were European American; 36% were African American; and 24% were Hispanic.

Child outcomes and ECE quality data were collected during spring or summer before kindergarten entry (Vogel et al., 2013). Children were administered language and academic assessments individually by trained research assistants. Caregivers rated the children's socioemotional skills at both time points. The quality of ECE was measured when children were, on average, age 5.3 years (SD = 0.3).

Measures

Classroom Quality Measures

The studies varied in terms of how they measured classroom quality. ECERS (Harms, Clifford, & Cryer, 1998) is widely used as a measure of global quality and was collected in EHS, FACES, NC-PK, NCEDL, PCER, and HSIS. The CLASS (Pianta, La Paro, & Hamre, 2004) measures the quality of teacher-child interactions (here called an interaction-specific measure). The CLASS was administered in FACES, MTP, NC-PK, and NCEDL. Measures designed to describe the quality of instruction in specific content domains included ELLCO (Smith & Dickinson, 2002) in NC-PK and the Teacher Behavior Rating Scale (TBRS) (Landry, Crawford, Gunnewig, & Swank, 2002) in PCER.

Global Quality Measures

Early Childhood Environment Rating Scale-Revised. The primary global quality measure was the Early Childhood Environment Rating Scale-Revised (ECERS-R; Harms, Clifford, & Cryer, 1998). The ECERS is a well-established measure of ECE quality that assesses seven general areas: personal care routines, furnishings and displays for children, language-reasoning experiences, fine and gross motor activities, creative activities, social development, and adult needs. Scores on each of 45 items can range from 1 to 7, with the overall mean score used as a global measure of the developmental appropriateness or quality of the classroom. To be consistent with other research, the adult needs items were not included in the overall classroom quality scores. An overall score from 1 to 2.9 is considered poor quality; scores from 3 to 4.9 are considered medium to good quality; and scores of 5 or greater are considered good to excellent quality. The total scale has good internal consistency (r = .921) (Harms, Clifford, & Cryer, 1998), and, during training, raters must meet a criterion of at least 85 percent agreement within one point on ratings.

In addition to the overall total score, factor analysis of the ECERS-R yielded two factors (Pianta et al., 2005) that we regarded as global quality measures because the items combined a rating of the quality of teacher-child interactions and of the quality of activities available in the environment. Factor 1, labeled Interactions, is a composite of 10 indicators, including staff-child interactions, discipline, supervision, encouraging children to communicate, using language to develop reasoning skills, and informal use of language as well as describing activities and materials (Bryant, 2010). The second factor, labeled Provisions for Learning, is a composite of 12 indicators focused on furnishings, room arrangement, gross motor equipment, free play, group time, and activities in fine motor skills, art, blocks, dramatic play, sand/water, and nature/science. In previous research, only the first factor was found to be related to child outcomes (Howes et al., 2008) and therefore was the only score used in these analyses.

Interaction-Specific Quality Measures

Classroom Assessment Scoring System. The CLASS (Pianta, La Paro, & Hamre, 2004) measures classroom quality as reflected in interactions between children and adults. It includes ratings on 10 dimensions, scored on a scale of 1 to 7 from low to high, which combine into scores on three overarching domains of classroom quality. The first domain, Emotional Support, encompasses four dimensions: Positive Climate (the emotional connection among children and teachers) Negative Climate (expressed negativity such as anger and hostility), Teacher Sensitivity (responsiveness to children’s concerns), and Regard for Student Perspectives (accommodations for children’s points of view). The second domain, Classroom Organization, includes three dimensions: Behavior Management (how effectively behavior is monitored or redirected), Productivity (how well time is organized to maximize learning activities), and Instructional Learning Formats (how well teachers facilitate children’s engagement to maximize learning opportunities). The final domain, Instructional Support, incorporates three dimensions:

Concept Development (how teachers foster higher-order thinking skills), Quality of Feedback (how well teachers extend learning in their responses to children), and Language Modeling (facilitation of language). The scale has demonstrated good inter-rater reliability, ranging from 78.8% to 96.9% agreement within one point, with an average across all items of 87.1% agreement within one point. The domain scores show high internal consistency ($0.83 < \alpha < 0.90$). An earlier version of the CLASS used in NCEDL included two dimensions in the Instructional Support domain (Concept Development and Quality of Feedback). For these analyses, we focus on the two CLASS scales collected across multiple studies, Instructional Support and Emotional Support.

Domain-Specific Quality Measures

Only a few studies collected domain-specific quality measures.

Early Language and Literacy Classroom Observation Toolkit. ELLCO (Smith & Dickinson, 2002) measures the extent to which the classroom provides support for language and literacy development. This observational measure includes three scales: Classroom Observation Scale, Literacy Environment Checklist, and Literacy Activities Rating Scale, each scored on a different metric. The Classroom Observation Scale consists of 14 items across two subscales: General Classroom Environment and Language, Literacy, and Curriculum. Each item is scored on a scale of 1 to 5 , with 1 deficient and 5 exemplary. The Literacy Activities Rating Scale was included in this study. It has a total score ranging from 0 to 13 and contains items describing the quantity and quality of reading and writing activities. For NC-PK, the Literacy Activities Rating Scale met criteria for inter-rater reliability, 81% within one point, and showed moderate internal consistency (alpha = .66).

Teacher Behavior Ratings Scale. TBRS (Landry, Crawford, Gunnewig, & Swank, 2002) was added as a post-intervention measure in the PCER to capture preschool instructional practices. TBRS includes ratings for quality and quantity of teacher instructional practices regarding various aspects of literacy instruction: written expression, print and letter knowledge, phonological awareness, book reading, and oral language use. It also includes a rating for the quantity and quality of instruction in mathematics concepts. The measure contains 11 subscales (total of 50 items) that capture responsive teaching practices, key language and literacy instructional areas, the use of lesson plans and progress monitoring, and classroom structure and organization as well as one subscale (total of 4 items) that captures quality of mathematics instruction. To complete TBRS, observations of classrooms for approximately 1.5 to 2 hours are conducted. The instrument showed good reliability in the development sample (Landry, Crawford, Gunnewig, and Swank, 2002), with inter-rater reliability using generalizability coefficients ranging from 0.80 to 0.98 and high internal consistency. We combined the quality ratings across the four language and literacy areas rated on TBRS to form a literacy quality composite and combined the ratings of quantity in the language and literacy areas to form a literacy quantity composite. Mathematics instruction, in contrast, was rated by using a single item about quantity and a single item about quality. Analyses used the literacy quality, mathematics quality, literacy quantity, and mathematics quantity measures.

Dosage Measures

In defining dosage for the current study and in keeping with the previous research on dosage, we examined a wide variety of measures of both cumulative and current participation. Cumulative participation was measured by number of years of enrollment in Head Start based on data from the FACES 2006 and 2009 studies and HSIS; number of years of high quality care was measured in FACES 2009; and attendance/absence was measured in FACES and NC-PK. Current participation was measured by hours per week of participation in FACES and NCEDL and observed time spent in instructional activities in NCEDL and PCER.

Number of Years of Exposure to Head Start

The FACES 2006 and 2009 and HSIS data included children with one or two years of participation in Head Start. In FACES, children were followed into kindergarten if they entered Head Start at age 3 and stayed for two years and if they entered Head Start at age 4 and stayed for one year. In HSIS, children who entered Head Start at age 3 could have one or two years of Head Start participation. All children were followed into kindergarten, making it possible to find children who entered Head Start at age 3 and had two years of Head Start and to compare them with children who entered at age 4 and had one year to provide a replication of the FACES analyses.

Number of Years of High Quality Care

Only the FACES 2009 and HSIS studies included quality information during each year for children who entered Head Start at age 3. We categorized the children who participated for two years into two groups based on the quality of their care. In FACES 2009, the CLASS was collected in both years. Classrooms in a given year were considered high quality if their CLASS Instructional Support score was greater than 2, CLASS Emotional Support score was greater than 5, and CLASS Classroom Organization score was greater than 4. More than one-third (38%) of the children were in classrooms that met these criteria in both years. In HSIS, classrooms were considered high quality if their ECERS-R total was 4.5 or higher, with about two-thirds of the children in classrooms meeting this criterion in both years.

Attendance/Absences

Three studies collected information on attendance. In the FACES 2006 and 2009 studies, teachers and parents reported on attendance retrospectively at the end of each year of Head Start. In questionnaires, teachers and parents reported days the child had missed Head Start as 1 = never; 2 = 1–5 days; 3 = 6–10 days; 4 = 11–20 days; 5 = more than 20 days. In NC-PK, classrooms were required to provide attendance data for each child enrolled in the program in order to receive reimbursements. Those data were recoded into the five categories used in FACES.

Number of Hours per Week

Three studies collected information on the hours per week that children attended the associated program, based on two different variables. The number of hours per week that each program was offered to a child was available for PCER and NCEDL. We considered this a measure of current participation. Parents reported on the number of hours per week that their child attended Head Start in FACES 2006. We also viewed this number as a measure of current participation.

Time Spent in Particular Instructional Domains

The amount of time spent in literacy or mathematics instruction was collected in two studies. NCEDL used the Snapshot (Ritchie, Weiser, Kraft-Sayre, & Howes, 2001), a time-sampling measure that counted the number of cycles in which activities were observed for the target children. In PCER, trained observers rated the frequency of literacy and mathematics instruction during classroom observations by using the TBRS quantitative scales.

Preschool Child Outcomes

All studies administered individualized direct assessments of children’s language and academic skills with widely used assessment measures and asked teachers to rate children’s social-emotional skills with frequently used questionnaires. Several studies relied on the same measures of language and academic skills, but each used a different measure of social-emotional skills.

In Table 1, we show the primary child outcome measures examined in each study. Almost all of the studies measured language skills with the Peabody Picture Vocabulary Test

(PPVT), mathematics with the Woodcock-Johnson Applied Problems scale (WJ AP), and literacy with the Woodcock-Johnson Letter Word scale (WJ LW). One project used the Test of Preschool Early Literacy (TOPEL) to measure language and literacy skills. Standard scores are reported unless otherwise noted because some of the older studies used earlier versions of these measures, and IRT-based scores were not always available.

The studies used a variety of teacher-rating scales to assess behavior problems and social skills, including the Social Skills Rating Scales (SSRS), the Teacher-Child Relationship Scale (TCRS), the Child Behavior Checklist (CBCL), a measure based on the SSRS, and the Behavior Problem Scale developed for FACES and EHS. We describe these measures below.

Academic Skills

Subtests from the Woodcock-Johnson Tests of Achievement. Five studies administered two subtests from the Woodcock-Johnson Tests of Achievement (Woodcock, McGrew, & Mather, 2001)—the Letter-Word Identification Subtest and the Applied Problems Subtest—as measures of literacy and mathematics skills. The Letter-Word Identification Subtest measures word identification skills. The child is initially asked to identify letters. Further items require the child to read and pronounce written words correctly. The Applied Problems Subtest examines the child's ability to analyze and solve mathematics problems. Five of the studies (FACES, NCEDL, PCER, NC-PK, HSIS) used the Woodcock-Johnson III, whose internal consistency coefficients for the 3- to 5-year-old group range from 0.97 to 0.99 for the Letter-Word Identification Subtest and 0.92 to 0.94 for the Applied Problems Subtest, according to the measure's authors. One study (EHS) that began data collection earlier used the Woodcock-Johnson Revised (Woodcock & Johnson, 1989). The test has internal consistency reliability coefficients of 0.92 for Letter-Word and 0.91 for Applied Problems for 4-year-old children.

MTP administered TOPEL. One subtest, Definitional Vocabulary, measures expressive vocabulary. The other two subtests—Print Knowledge and Phonemic Awareness—measure literacy skills in alphabet knowledge and written language conventions and in elision and sound blending, respectively. Internal consistency reliabilities were 0.85 or higher, and inter-rater reliabilities were 0.96 or higher.

Language Skills

All studies included a measure of receptive vocabulary. The PPVT 3rd edition (Dunn, Dunn & Schlichting, 2005) was administered in NCEDL, PCER, EHS, NC-PK, and HSIS. FACES 2006 and 2009 used the PPVT 4th edition (Dunn & Dunn, 2007). In the PPVT assessment, children are shown a set of four pictures and asked to select the picture that best represents the meaning of a word spoken by the examiner. Internal consistency reliability tends to be high, ranging from 0.92 to 0.98.

Social-Emotional Adjustment

Various measures designed to assess social skills and behavior problems were administered across the studies. The Child Behavior Checklist (CBCL; Achenbach, 1991) was completed by parents and caregivers in EHS. It is a widely used measure of internalizing and externalizing behavior problems. It lists 113 problem behaviors, each of which is rated as true (0), somewhat true (1), or very true (2) of the child. The measure reports high levels of internal consistency and inter-rater reliability of 0.73 to 0.95.

The Social Skills Rating Scale (SSRS; Gresham & Elliott, 1990) was completed in prekindergarten by teachers in fall and spring in PCER and NC-PK. The Social Skills Questionnaire from the SSRS is composed of 38 items describing child behavior, each rated on a three-point scale reflecting how often the child exhibited each behavior. The total score reflects levels of perceived social competence, with internal consistency of 0.90 and test-retest reliability of 0.75 to 0.88) and moderate concurrent and predictive validity to other indices of social competence.

NCEDL used the Teacher-Child Rating Scale (TCRS, Hightower et al., 1986), a behavioral rating scale that assesses children’s social competence and problem behaviors. The Social Competence Scale was computed as the mean of 20 items and had strong internal consistency reliability, with a Cronbach’s alpha of 0.95. The Problem Behavior Scale was computed as the mean of 18 items and likewise had strong internal consistency reliability, with a Cronbach’s alpha of 0.91.

The teacher ratings of children’s social skills and problem behaviors in FACES 2006 and 2009 used items selected from SSRS and the Behavior Problems Index (BPI; Zill, 1990). Teachers were asked to rate each study child in their classroom on a set of items that assess the child’s accomplishments, cooperative classroom behavior, and behavior problems. The internal consistency of the BPI total score ranged from 0.88 to 0.89 in the National Health Interview Survey and the National Longitudinal Study of Youth (NLSY; Berry, Bridges, & Zaslow, 2004).

HSIS did not collect teacher ratings of social-emotional outcomes in fall so we do not include the study in analyses of the association between ECE quality and development of social-emotional adjustment.

In Tables 3 through 5, we present descriptive statistics for the quality, dosage, and child outcome measures used across the various studies.

III. TESTING FOR QUALITY THRESHOLDS AND FEATURES

In this chapter, we report on the analyses focusing on both quality thresholds and quality features. First, we address questions about quality thresholds, using two analytic approaches. The analyses ask whether there is evidence suggesting thresholds in the association between a specific quality measure and a specific child outcome. Second, we extend these analyses to ask whether each child outcome is more strongly related to global quality measures or to quality measures that measure teacher-child interactions or quality of instruction in a given content area. The research to date provides the basis for the articulation of two hypotheses related to quality thresholds and features: (1) the quality of ECE is a stronger predictor of residualized gains in child outcomes in classrooms with higher quality than in classrooms with lower quality and (2) more specific measures of quality are stronger predictors of residualized gains in child outcomes than are global measures. We turn now to analyses intended to address these hypotheses by using data from several data sets.

Overview of Analytic Approach

Two sets of analyses address the hypothesis about quality thresholds, and one set of analyses examines the hypothesis about quality features. The first approach uses meta-analysis to test thresholds set a priori. As discussed in the literature review, thresholds in these analyses are set according to the labels for ratings used in observational measures of quality (for example, ratings of “good” or higher) within the contrasts of the distributions of the quality variables across projects. This approach uses the same thresholds in all projects, fitting parallel analyses across projects in analyses that involved multiple imputations to account for missing data and then combining relevant coefficients from those analyses in a single analysis using meta-analysis. Using data from each project, piecewise regression multilevel analyses estimated separate slopes for higher and lower ECE quality classrooms and tested whether the slopes were different. The slope coefficients were then combined across projects and compared by using meta-analytic techniques. The models included the child's fall score as a covariate in predicting spring scores as well as selected child and family demographic characteristics to account, in part, for child and family differences among children who attended ECE of different quality. The multilevel models accounted for the nesting of children in classrooms or a center. The meta-analyses tested whether results support thresholds in replicated analyses across projects that control the type 1 error rate, allowing us to draw a single conclusion from each set of analyses. We report the results only from meta-analysis in the analyses of the a priori thresholds.

The second set of analyses designed to examine quality thresholds was conducted in an attempt to identify cut-points empirically when the first set of analyses suggested the possible existence of thresholds. We used two nonlinear empirical approaches: LOESS models (LOcal regrESSion) and b-spline models. The LOESS models estimate a function based on locally weighted scatterplot smoothing. These nonparametric, nonlinear regression methods combine multiple regression models in a *k*-nearest-neighbor-based meta-mode. The plots of the estimated LOESS functions overlaid on the data were created and examined visually to determine evidence of thresholds in the smoothed plots of residualized gains in child outcomes and observed ECE quality. The b-spline approach is also a nonlinear regression approach that involves penalized smoothed b-spline to identify the function that best describes the association of quality and child outcomes within a given data set and may include covariates and provide statistical tests for nonlinear associations. We are seeking convergent evidence of thresholds within and across the two approaches as applied to several data sets with different quality measures and child outcomes

The final set of analyses in this chapter involved examining quality features: testing whether global quality measures, in comparison with either interaction-specific or domain-specific quality measures, provide stronger prediction of child outcomes. The analyses were based on the piecewise regression analysis, including the global and specific quality measure for each outcome. The analyses used a piecewise predictor for a given quality measure, if indicated in the analyses described above, and a linear predictor if not so indicated. The global and specific quality variables were combined by using the same analytic strategy involving covariates, multilevel models, and multiple imputation.

In the following sections, we first provide more detail about the analytic approach and then present the results from the analyses based on each approach.

Testing for Thresholds

Description of the Meta-Analytic Approaches Testing for A Priori Thresholds

As noted, the first analytic approach to testing for thresholds calls for fitting parallel analyses involving each quality measure and child outcome by using the data from each project and then combining results by using a meta-analysis. Using the data from all projects, piecewise models test for the same thresholds defined a priori for a given quality measure in the analyses of all outcomes. Hierarchical linear models (HLM) accounted for the nesting of children in classrooms. The analyses of spring scores included the child's fall score on the outcome as a covariate to account, in part, for selection bias. The analyses also included the following as covariates: maternal education or poverty; child gender, race, and home language; and amount of time between fall and spring assessments. In addition, covariates included the site in multiple-site studies(e.g., EHS, PCER, NCEDL) and treatment in studies with treatment and control groups (e.g., EHS, PCER). The coefficients describing the association between classroom quality and child outcomes were then combined in a meta-analysis.

The basic model is depicted in Figure 1, with the lines showing the expected piecewise association between a measure of ECE quality and a measure of children's outcomes. The piecewise association is represented by allowing the slope for quality to vary in the lower and higher quality ranges based on previous research and an a priori division of the quality range derived from the developer's labeling of quality ranges.

The thresholds were defined by using the developer's guidelines to the extent possible. The goal was to use the cut-points that defined higher quality in the developer's guidelines, but we adjusted the cut-points downward by 0.25 increments if fewer than 5% of the children in any projects were in classrooms that were above or below the recommended cut-point. Instead, we used the cut-point from the project with the lowest value and applied that cut-point to all projects that collected that quality measure. Such an approach resulted in testing the same thresholds in all analyses for a given quality measure. In particular, the following cut-points were used for each quality measure:

*4.5: ECERS total and ECERS Interaction Score (range 1–7)

* 5.0: CLASS Emotional Support and Classroom Management score (range 1–7)

* 2.75: CLASS Instructional Support (range 1–7)

* 2: TBRS literacy and mathematics quality scales (range 1–3)

* 4: ELLCO Literacy Scale (range 1–5)

We began by testing for quality thresholds by using a piecewise model that tests whether the association between quality and the child outcome is stronger in higher versus lower quality classrooms based on the threshold defined above for each measure of quality. We used random-intercepts HLMs to account for nesting of children in classrooms. The piecewise regression model estimates linear slopes to describe the association between ECE quality and child outcome, with separate slopes estimated for lower quality classrooms and higher quality classrooms. This model follows:Spring Child Outcome$_{ij}$ = $B_0 + B_1$ Quality$_j$ +

B_2 Quality$_j$ * High Quality Classroom$_j$ + B_3 Fall Outcome$_{ij}$ +

B_4 Gender$_{ijj}$ + B_5 Race$_{ij}$ + B_6 Time between Fall and Spring Assessments$_{ij}$ +

B_7 Home Language$_{ij}$ + B_8 Maternal Education/Poverty$_{ij}$ +

B_9 Site/Treatment Group$_{ij}$ + e_{ij} + u_j

In this model, we are interested in $B_{1,}$ which describes the quality slope in lower quality classrooms and B_2 which describes the difference between the quality slopes in higher and lower quality classrooms (i.e., $B_1 + B_2$ is the estimated slope for the higher quality classrooms). The models account for nesting of children in classrooms by including both e_{ij} as the level 1 residual (i.e., describing within-classroom variability) and u_j representing the level 2 residual (i.e., describing between-classroom variability). In addition, $B_3 – B_9$ are the coefficients for the selected covariates.

Multiple imputations were conducted within each study for each analysis to account for missing data. Ten imputation data sets were generated for each project in which missing data were imputed from regression analyses of the other variables, with random error added to preserve the degree of variability. Analyses were conducted separately for each data set, and the results were then combined in a manner that took into account variation within and between imputation data sets (Rubin, 1976, 1987; Schafer & Graham, 2002).

The results from the analyses were combined across projects by using Hedge’s meta-analysis program (Borenstein, Hedges, Higgins, & Rothstein, 2005). Separate meta-analyses were conducted for each quality measure as a predictor of each outcome in each project. The meta-analysis combines the regression coefficients across projects for a given quality measure and child outcome, taking sample size and sample variability into account. We used the fixed effect findings because we did not regard our set of studies as representing a random selection of ECE studies or classrooms. In these analyses, the regression coefficients for quality were computed; from the coefficients, we computed effect sizes based on Hedge's suggested formula for effect sizes for the project analyses as follows:

d = B sd(quality)/sd(outcome)

Results of Meta-Analyses of A Priori Thresholds

The results from piecewise regressions are shown in Tables 5 through 7. Each cell in the tables presents the results from one of the regressions. In the tables, we show the results from the models in which ECE quality is entered as a piecewise predictor. The first coefficient describes the association between the quality measure and outcome in lower quality classrooms, and the second coefficient describes the association between the quality measure and outcome in higher quality classrooms. The row below the coefficients lists the results of comparison of the two coefficients, showing the direction of the difference when the coefficients were reliably different. The final column shows the coefficients from meta-analysis, listing the combined coefficients for higher and lower quality classrooms in the upper row and indicating if the coefficients differ reliably in the lower row.

In the first row of Table 5, we present the results from the analyses predicting PPVT language scores from the ECERS total score. Whereas the results from meta-analysis shown in the final column do not support thresholds, the results displayed in the first row illustrate our analysis strategy. As shown, the ECERS total was a stronger predictor in higher rather than in lower quality classrooms of residualized gains in PPVT scores in FACES 2006, FACES 2009, and PCER data. In these latter studies, ECERS was a significantly stronger positive predictor in the classrooms with an ECERS Total score of 4.5 or higher (i.e., “higher quality”) than in the classrooms with an ECERS total score of less than 4.5 (i.e., “lower quality”). The final column presents the combined coefficients across the various studies and shows whether the estimated effect sizes in the higher quality and lower quality classrooms reliably differed from zero and whether they reliably differed from each other. Results in the lower row of the final columns did not reveal significant differences in the magnitude of the association between ECERS-R Total and PPVT in higher and lower quality classrooms. Thus, we conclude that there was no evidence of a threshold in the association between the ECERS-R Total score and PPVT scores based on the analysis of the results aggregated across studies.

The subsequent rows of Table 5 show the results from the meta-analysis of the ECERS-R Total score as a predictor of the other child outcomes collected across a number of studies. The meta-analyses did not provide supports for thresholds in the association between the ECERS-R Total and residualized gains in academic or social skills.

Many studies have focused on the ECERS factor scores rather than on the ECERS-R Total score and have indicated that the ECERS-R interaction score is related to child outcomes (see, for example, Howes et al., 2008; Burchinal, Vandergrift, Pianta, & Mashburn, 2010). The ECERS-R Interactions score was examined as a global quality predictor of child outcomes because the items reflect quality of the environment as well as teacher-child interactions. In Table 6, we show the results from these analyses. The meta-analysis comparing the coefficients from the piecewise model indicated that the ECERS-R interaction score was a stronger positive predictor of children's PPVT receptive language scores in higher quality classrooms than in lower quality classrooms. Similarly, evidence for thresholds emerged in the analyses of social competence. Using 4.5 as the cut-point, the analyses indicated very modest but positive effect sizes in the higher quality classrooms and no reliable association in the lower quality classrooms.

The next set of meta-analyses examined the CLASS Instructional Support and Emotional Support domain scores, with the results shown in Table 7. We viewed the CLASS Instructional Support and Emotional Support as interaction-specific quality measures and sought to examine their association with child outcomes in the same domain. Children's social skills and behavior problems were hypothesized as being in the same domain as the CLASS Emotional Support: by providing more frequent and higher quality emotional support, teachers should be promoting the social skills and reducing the behavior problems of the children in their classes. Similarly, we posited that teachers who provided more frequent and higher quality Instructional Support should be promoting the language and academic skills of the children in their classes.

In Table 7, we show the results of the piecewise regression testing whether CLASS scores predicted child outcomes more strongly in higher versus lower quality classrooms. The meta-analysis of the results from the analyses of children's social skills or behavior problems in the four studies with CLASS Emotional Support scores did not support this hypothesis. In contrast, the analysis of the data from the five studies with the CLASS Instructional Support provided evidence suggesting thresholds. The meta-analyses of the piecewise coefficients provided evidence of thresholds for language and literacy skills, both on the WJ Letter-Word across four studies and the TOPEL Phonemic Awareness in the MTP study. Using a score of 2.75 as the cut-point, the CLASS Instructional Support was a stronger positive predictor of the PPVT and WJ Letter-Word scores in higher quality classrooms than in lower quality classrooms (Figure 2). Thus, these analyses of the CLASS provided evidence of thresholds for Instructional Support and language and literacy outcomes, but no evidence of thresholds for Instructional Support and mathematics outcomes or for Emotional Support and either social outcome.

In Table 8, we show the results from the analyses of the domain-specific quality measures. Unfortunately, we were not able to conduct replicated analyses. We found two studies with different domain-specific quality measures and therefore examined the domain-specific measures within each study and could not pursue a meta-analysis.

The ELLCO Literacy Activities scale was collected in the NC-PK evaluation. Higher ELLCO scores were related to higher levels of literacy skills on the WJ Letter-Word in the linear model, but no evidence of thresholds emerged in the analyses.

In contrast, piecewise analyses relating the TBRS Literacy scale to language and literacy outcomes in the PCER data suggested thresholds on both language and literacy (Figure 3). The TBRS literacy scale was a significant and moderately strong predictor of language (PPVT) and literacy (WJ Letter Word) in higher versus lower quality classrooms (i.e., TBRS less than 2.0). No evidence emerged indicating a threshold in the association between the TBRS Numeracy scale and children's mathematics skills. Thus, the analyses provide some evidence of thresholds—replicated child outcomes rather than data sets for one quality measure of literacy instruction and language and literacy outcomes, but not for a measure of mathematics instruction and mathematics skills. The literacy instruction measure described quality of instruction supporting specific reading skills such as phonemic awareness and letter recognition.

Description of Nonlinear Empirical Approaches

The next set of analyses involved fitting LOESS and b-spline models in an attempt to identify cut-points empirically; in other words, we sought to replicate the findings described above by using empirical approaches. The goal was to use analytic methods that identified the functional relations between classroom quality and child outcomes in order to validate and extend the findings based on the piecewise regressions in which we defined the functional relations on a priori grounds. We used two approaches—LOESS and b-spline models. LOESS is both nonparametric and nonlinear and therefore provided the most flexibility in identifying the association between observed quality and child outcomes, whereas the b-spline approach easily allowed for testing nonlinear associations in the presence of covariates.

LOESS

LOESS is a method also known as locally weighted polynomial regression (Cleveland & Devlin, 1988). A low-degree polynomial (often linear) is estimated by fusing predictor variable(s) within a subset of the data. The polynomial is estimated with weighted least squares for each datum, giving more weight to data closer to the value of response that is being estimated and less weight to data further away. The subsets of data used for each weighted least squares fit in LOESS are determined by a nearest-neighbors algorithm. A user-specified input to the procedure called the bandwidth or smoothing parameter determines how much of the data is used to fit each local polynomial. The LOESS fit is complete after regression function values have been computed for each of the data. The degree of the polynomial model and the weights are flexible.

LOESS provides a completely empirical approach to examining thresholds because it does not require the specification of a single function for all data, making it ideal for modeling complex processes for which no theoretical models exist. There are two disadvantages, however: the approach requires large data sets with dense distribution of data, and it does not provide a statistical test of specific parameters that represent characteristics of interest. In our case, there is no specific test that we can use to determine whether a nonlinear relationship exists between observed quality and child outcomes, especially after adjusting for the covariates.

B-Spline Approach

The b-spline approach provides flexibility in modeling the association between quality and child outcomes and does include a test of whether that association is nonlinear, even when covariates are considered. The goal of the b-spline model is to find a function that either interpolates points or fits a smooth curve through them. The model may be viewed as an extension of the piecewise model because it involves estimating more pieces that are forced to be joined and may be defined by nonlinear functions within each piece. Specifically, the general model we used was a cubic spline model with three knots:

$$F\,(x) = B_0 + B_0\,x + B_2\,x^2 + B_3\,x^3 + B_4\,(x - k_1)^3 \;+ B_5\,(x - k_2)^3 \;+ B_3\,(x - k_3)^3\,,$$

where X in our case is the measure of ECE quality and k_1 to k_3 are knots defined by the 25th, 50th, and 75th percentiles in that measure's distribution. Thus, the overall function allows for cubic change but permits the level of the cubic change to differ for the four quartiles of the quality distribution. In addition, the model adds a "penalty" that forces the pieces to join at each knot.

The project data in which statistically significant evidence of thresholds emerged in both the meta-analysis and the univariate analyses were selected for conducting these LOESS and b-spline analyses. The b-spline analyses included the same covariates as linear covariates: site; fall scores on outcomes; type of program if there was some variability; maternal education or poverty; child gender, race, and home language; and amount of time between the fall and spring assessments. The model was tested for an overall nonlinear trend.

Results of Analyses Using LOESS and B-Spline Approaches

Results from the LOESS analyses are discussed below; results from the b-spline analyses appear in Table 9. In Figure 4, we show the LOESS analyses involving the PPVT and ECERS interaction scale. The dark dots are the regression curve predicted from the model that was estimated in the LOESS analysis, and the lighter dots are the observed data. As can be seen, the regression curve on average slopes upward in the higher quality range, but the curve is not linear. There are several upward and downward changes in the regression curve even in the higher quality range. The LOESS results were similar in analyses across NCEDL, PCER, and the 3-year-old cohort of HSIS for the PPVT and ECERS Interaction scale (available on request). In each of these analyses, there was some evidence of somewhat stronger positive slope for the

ECERS Interactions scale as a predictor of PPVT scores in higher rather than lower quality classrooms, but the estimated regression curves were not linear in either the higher or lower range; therefore, no clear cut-point emerged. Similar findings emerged in the LOESS analyses of the social skills and ECERS Interactions in NC-PK and PCER (available on request). Again, a general upward tilt is evident, without a conclusive threshold.

The first rows of Table 9 show the results from the test for nonlinearity in the association between ECERS Interactions and PPVT in b-spline regressions that included all covariates. The results also suggest nonlinearity in associations in data from at least three of the four projects. Thus, the analyses suggest that, as with the a priori piecewise models above, there might be a stronger association between ECERS Interactions and language skills in higher quality classrooms than in lower quality classrooms, but with little conclusive evidence about exactly where the threshold lies or whether the association is linear above or below that threshold.

Next, we examined findings suggesting a threshold in associations between CLASS instructional support and language and literacy skills. The LOESS analyses again suggested a general upward tilt in the higher quality range in the analyses of PPVT in NCEDL and FACES 2009, and the analyses of WJ Letter Word in NC-PK, but also in the very lowest quality range as well in the NCEDL and NC-PK data. No pattern was clearly discernible in the analysis of PPVT in FACES 2006 or WJ Letter-Word in NCEDL. The b-spline analysis indicated nonlinear associations between CLASS Instructional Support with PPVT in FACES 2006 and NCEDL and with WJ Letter-Word in NC-PK. In summary, the analyses provided little clarity regarding whether there was are thresholds, and if so, where the cut-points are.

Finally, we examined findings suggesting a threshold in associations between the TBRS literacy scale and language and literacy outcomes in PCER. The LOESS analyses suggested a slightly upward tilt in the association in the higher range, at least in analyses of literacy scores.

The b-spline analyses also suggested a nonlinear trend in analyses of literacy scores. As before, the analyses do not appear to clarify whether thresholds exist, and if so, where the cut-points are.

In summary, both the LOESS and b-spline analyses provided some indication that the association between ECE quality and residualized gains in child outcomes was stronger at higher levels of quality, but none of the analyses yielded specific cut-points. Furthermore, it was difficult to identify the functional form between quality and outcomes within an analysis or to see consistency across analyses involving the same quality and child outcome measures. The result is not surprising given that the LOESS and b-spline approaches are not designed to find a clear demarcation between two lines with different slopes. Instead, the two approaches maximize the fit of complicated functions to data and thus are likely to allow many points in the regression curves in which the slope may change from an upward to a downward tilt. Such occurs as the local regression curves are estimated and aggregated per the LOESS approach and as the model allows for different quadratic slopes in different regions per the b-spline approach. We conclude that the analyses provide some further support for the meta-analytic evidence suggesting thresholds, but do not provide evidence of the specific location of the thresholds.

Quality Features: Comparison of Global versus More Specific Quality Measures

We built on the analyses of thresholds to examine whether the global quality and interaction-specific or domain-specific quality measures provide a unique prediction of child outcomes when considered together. The primary question involved a contrast between analyses of the predictive ability of the global and more specific quality measures. Linking such analyses to the analyses of thresholds, we also asked whether there was stronger evidence of thresholds in analyses involving interaction-specific or domain-specific measures of quality than for those involving global measures of quality.

Analytic Approach: Analyses Including Global and More Specific Quality Measures

For the analyses, we used the final a priori model from the analyses that included the global quality measure and the analyses that included the specific quality measure. If the spline model indicated reliably different coefficients for higher and lower quality classrooms, then the spline coefficients were estimated for that quality measure. Only a few projects assessed both global and more specific quality assessments. They included the ECERS-R Total and TBRS literacy and numeracy scores in PCER, the ECERS-R Total and ELLCO in NC-PK, and the ECERS-R Total and CLASS Instructional Support in NCEDL, NC-PK, and FACES 2006 and 2009. We did not conduct meta-analyses because of the limited overlap among the projects in terms of use of ECERS and a more specific classroom quality measure.

The analyses were based on the piecewise model analyses that tested for thresholds. For a given quality measure and child outcome in a project's data, we combined the best model for the global quality measure using the ECERS total score with the best model for a given interaction-specific or domain-specific quality measure. The ECERS total score was selected to represent global quality because that is how it is widely used in research, policy, and practice. The two CLASS scores were examined as interaction-specific scores and the ELLCO and TBRS scores as domain-specific scores, respectively. The quality scores were entered into the analyses as piecewise predictors if the analyses of those data described above suggested that slopes were reliably different in higher and lower quality classrooms; otherwise, they were entered into the analyses as linear predictors.

Results of Analyses, Including Global and More Specific Quality Measures

In Table 10, we present the results from the analyses, including both global and more specific quality measures. The domain-specific measure was a positive and significant predictor of residualized gains in mathematics and literacy skills in higher quality classrooms in PCER and

NC-PK and of language skills in higher quality classrooms in FACES. Modest to large effect sizes emerged in some of the analyses that indicated thresholds, with effects sizes for a one SD increase in the domain-specific quality of 0.17 for language and 0.69 for mathematics in higher quality classrooms. The domain-specific quality measures were always positive predictors of residualized gains in these analyses, whereas the global quality measure was never a statistically significant positive predictor and was sometimes a negative predictor when included in the same analysis as the domain-specific predictors.

A similar, albeit weaker, pattern emerged in the analyses of global and interaction-specific quality measures. The CLASS Instructional Support score was a significant positive predictor of residualized gains of mathematics skills in NC-PK and of language, literacy, and mathematics skills in NCEDL. In one case, in NCEDL, a threshold was indicated, and the interaction-specific quality measure was a stronger predictor only in the higher quality classroom. CLASS Emotional Support was also a significant predictor of decreases in behavior problems and increases in social competence in NCEDL. The global measure of quality—the ECERS-R Total—did not reliably contribute as a positive predictor of residualized gains in child outcomes in these analyses when considered along with interaction-specific measures.

Summary of Results for Quality Thresholds and Features

The analyses provide some evidence of quality thresholds and of stronger prediction of child outcomes based on more specific measures of quality. Many analyses were conducted, and for this reason, we discuss and interpret only those findings noted across several sets of analyses.

The meta-analyses suggested thresholds in the association between the acquisition of language skills and social skills and the ECERS-R Interactions factor, but no association with the ECERS-R Total. The meta-analyses indicated that higher global quality using the ECERS-R Interaction factor predicted larger residualized gains in children’s language and social skills in higher quality classrooms than in lower quality classrooms. In addition, the meta-analyses suggested the possibility of thresholds in the quality of instruction and acquisition of academic skills. The meta-analysis of the CLASS Instructional Support domain score suggested that Instructional Support was related to larger gains in language and literacy skills in higher versus lower quality classrooms. Analyses of TBRS in PCER suggested that quality of literacy instruction was related to larger gains in language and literacy skills in higher versus lower quality classrooms. Some further, albeit limited, evidence supporting stronger associations in higher quality classrooms emerged in the LOESS and b-spline analyses that made no a priori assumptions about cut-points. However, in these analyses, no conclusive evidence emerged suggesting where the thresholds lie.

Finally, analyses focused on the prediction of child outcomes when global and either domain-specific or interaction-specific measures of quality were considered simultaneously. In these analyses, the domain-specific quality measures provided significant prediction of residualized gains for at least one outcome in each of the three studies, with a domain-specific and global quality measure considered simultaneously. The interaction-specific quality measure provided significant prediction of residualized gains on at least one outcome in two of the three studies. The global quality measure did not provide significant positive prediction of gains in outcomes in any of the analyses when more specific measures of quality were taken into account. Unfortunately, there was no opportunity to replicate the findings across projects, although findings did replicate across child outcomes within the project data sets. Thus, the findings suggest that classroom quality measures of specific instructional practices within specific content domains (i.e., either literacy or mathematics) and of teacher-child interactions appear to provide stronger prediction of child outcomes than does the global quality measure.

IV. TESTING FOR DOSAGE-OUTCOME ASSOCIATIONS

In this chapter, we turn to the question of whether there is evidence of an association between children's development and the quantity or dosage of ECE across several large studies. As follow–up to the results summarized in the literature review, it is important to control adequately for selection effects in studying effects of dosage. There is also a need to examine different measures of dosage to see if consistent patterns of findings emerge across different measurement approaches. Accordingly, in this chapter, we will summarize analyses by using more rigorous approaches to controlling for selection than those used in previous research and will adopt several approaches to operationalizing dosage. Again, we are seeking replicated findings, as indicated in this section by similar significant findings across projects in analyses of dosage.

Overview of Analytic Approach in Dosage Analyses

The first set of dosage analyses examines whether dosage measured as two years as opposed to one year of participation in Head Start is related to residualized gains in child outcomes based on FACES 2006 and 2009 and HSIS data. Propensity score matching was conducted to provide a more rigorous approach to controlling for selection by accounting for pre-existing differences in measured variables (Shadish, Cook, & Campbell, 2002). Logistic regressions predicted whether children in the 3-year-old cohort stayed in Head Start for one or two years based on a set of family characteristics and children's skills at Head Start entry across domains. The children in the 3-year-old cohort who stayed for two years were matched with children in the 4-year-old cohort who, by definition, had only one year of Head Start, using nearest-neighbor matching with replacement, based on predicted scores from the logistic regressions. The analyses of the matched children compared children with one or two years of Head Start and included the corresponding entry score and other child and family characteristics as covariates.

Representing different approaches to measuring dosage, the next set of dosage analyses considered attendance/absence in ECE, total number of hours per week in ECE, and observed time spent on instruction. Such analyses of dosage are based on regression models that extended the final models from the quality threshold analyses. In the analyses, the dosage variable was added to the final model from the quality threshold analyses for each quality measure. The model then included that quality measure as a piecewise predictor if indicated in the threshold analyses and, if not so indicated, as a linear predictor.

The final set of dosage analyses examined whether more time in higher quality care was related to larger gains in child outcomes. Using data from FACES 2009 and HSIS, we asked whether children who experienced two years of higher quality center-based care showed larger gains in child outcomes than other children with two years of center-based care. The same propensity score analysis strategy described above was employed to compare child outcomes of children with two years of high quality center-based care versus other children with two years of center-based care. Using propensity score matching, we accounted for differences between the groups of children in terms of skills at entry to center-based care and family characteristics. The CLASS was used to determine quality of care in FACES 2009 and ECERS-R to determine quality of care in HSIS. Finally, we examined interactions between quality of care and three further measures of dosage—absences, hours per week of care, and time spent on instruction in specific content areas. Interactions between quality and dosage were tested for the various measures of quality.

Results of Dosage Analyses

One or Two Years of Head Start

Propensity score analyses were first conducted to examine the relationship between the number of years of exposure to Head Start and gains in child outcomes (Shadish, Cook, &

Campbell, 2002). The propensity score analysis identified children who entered Head Start at age 4, experienced one year of Head Start, and had similar family characteristics and standardized entry skill levels as children who entered at age 3 and experienced two years of Head Start. Analyses compared the matched groups on child outcomes to test whether children showed higher skill levels at Head Start exit and in spring of kindergarten if they experienced two years instead of one year of Head Start. We conducted the analyses separately with two cohorts of FACES (FACES 2006 and 2009) and HSIS, allowing us to examine the extent to which findings replicated across cohorts and Head Start samples.

To reduce selection bias, the propensity score analyses balanced confounding covariates in the two groups of interest (in this analysis, the groups participating in Head Start for one versus two years). The analysis focused on identifying the characteristics of the group of 3-year-old children who participated in Head Start for two years and identifying children in the 4-year-old group with similar characteristics who participated in Head Start for a single year.

For analyses with FACES, the following variables were analyzed by using logistic regression to predict which children in the 3-year-old cohort remained in Head Start for two years: child race/ethnicity, gender, disability, household language, family poverty ratio, maternal education, employment, and depressive symptoms, household mobility, neighborhood safety, and child pretest standard scores on PPVT-4 and WJ III Letter-Word Identification, Spelling, and Applied Problems. The coefficients from the analyses were then used to create propensity scores for the 3-year-old cohort with two years of Head Start and the 4-year-old cohort with one year of Head Start. We used multiple imputations to account for missing data and used appropriate weights for the sampling design (Appendix A).

For conducting analyses with HSIS and estimating the propensity score (in other words, the conditional probability of enrollment in Head Start for two years), we included the following covariates that might be related to both enrollment in Head Start for two years and child outcomes: child race/ethnicity; gender; disability; household language; maternal education and depressive symptoms; whether the mother was a recent immigrant to the United States; three family structure variables, including whether both biological parents lived with the child, whether the child's mother was married, and whether the mother was a teenager at the child's birth; and child pretest standard scores on PPVT-3 and WJ III Letter-Word Identification, Spelling, and Applied Problems. The propensity score analysis involved equating on these covariates the children in HSIS who entered Head Start at age 3 and had two years of Head Start with 4-year-old children who had one year of Head Start (Appendix A).

Nearest-Neighbor Matching with Replacement

We used nearest-neighbor matching with replacement to match the two-year group in the 3-year-old cohort with the one-year group in the 4-year-old cohort (Shadish, Cook, & Campbell, 2002). For each child in the two-year group, the potential comparison child in the one-year group with the closest absolute propensity score, or the "nearest neighbor," was selected. Matching with replacement allows some children in the one-year group to be used more than once to match to children in the two-year group. In FACES 2006, the resulting sample after propensity score matching included 809 to 854 children in the two-year group and 377 to 404 children in the one-year group, with the total sample ranging from 1,084 to 1,118 (the sample size varies because of multiple imputation). Approximately half of the one-year group children was matched to children in the two-year group. In FACES 2009, the resulting sample after matching included 736 to 795 children in the two-year group and 391 to 433 children in the one-year group, with the total sample ranging from 1,209 to 1,246. More than half of the one-year group was matched to children in the two-year group. In HSIS, the resulting sample after matching included 714 children in the two-year group and 770 children in the one-year group, with a total sample of 1,484. Close to 85% of the one-year group were matched to children in the two-year group. In Appendix A in Tables 1 and 2, we show the descriptive statistics for those enrolled in Head Start for one year versus two years before and after matching. In all three samples, the children with one and two years of Head Start in the propensity matched samples did not differ significantly, and none of standardized mean difference was |0.10| or greater on the child/family characteristics or baseline scores used in propensity score matching.

Next, multilevel analyses compared the matched children with one or two years of Head Start on each outcome. The multilevel analyses accounted for nesting of children in centers and included the same set of child and family covariates (Table 11). In both FACES 2006 and 2009, children with two years of Head Start scored significantly higher than children with one year on vocabulary skills at exit from Head Start and a year later at the end of kindergarten ($0.10 \leq d \leq 0.17$). Two years compared to one year of Head Start were related to higher mathematics skills in FACES 2009 but not in FACES 2006, whereas two years compared to one year of Head Start were related to higher literacy skills in FACES 2006 but not in FACES 2009. In HSIS, children with two years of Head Start scored significantly higher than children with one year on literacy skills at exit from Head Start and at the end of kindergarten ($0.14 \leq d \leq 0.16$). No evidence emerged suggesting that more years of Head Start were related to social skills or behavior problems.

Sensitivity analyses were conducted. In addition to nearest-neighbor matching, we conducted robustness testing by using two other propensity score approaches: caliper matching and propensity score weighting (Shadish, Cook, & Campbell, 2002). In our case, caliper matching selected all 4-year-old children whose characteristics are sufficiently close in propensity score units to those of a 3-year-old child with two years of Head Start. We carried out caliper matching with replacement so that a potential comparison child in the one-year group may be matched to several children in the two-year group. The findings are similar to the results from nearest-neighbor matching (not reported).

We tried to weight the two-year and one-year groups by using the inverse probability of treatment (the inverse of the propensity score for the two-year group and the inverse of one minus the propensity score for the one-year group). Again, the findings are similar to the results from nearest-neighbor matching (not reported).

In summary, the analyses provide replicated evidence that two years of Head Start appear to have a larger impact on children's academic but not social skills than a single year, both at program exit and one year later in spring of kindergarten.

Attendance/Absence

The next set of analyses examined whether absences predicted residualized gains in child outcomes. Absences were added to the final model from the quality threshold analyses. Again, separate analyses were conducted for each quality measure, using that quality indicator as a piecewise predictor if indicated in the threshold analyses and otherwise as a linear predictor . The analyses included the same set of covariates, including the child's fall score on the outcome.

Two data sets—FACES 2006 and NC-PK—measured absences. FACES 2006 measured the number of days the child was absent based on teacher and parent reports, whereas NC-PK recorded attendance based on a daily teacher report. The data from both projects were recoded into the FACES categories (1 = never; 2 = 1–5 days; 3 = 6–10 days; 4 = 11–20 days; 5 = more than 20 days). The first set of columns in Table 12 gives the results, listing the smallest coefficient for absences from the analyses of that outcome that used different quality measures. We focused on the smallest coefficient for absences from the separate models with different measures of quality because that model provided the most conservative test based on the assumption that the quality variable in that model accounted for the greatest variance in outcomes (all results are available on request). As shown in Table 12, children with more absences according to attendance reports had smaller residualized gains in language in NC-PK, in mathematics according to both teacher and parent absence reports in FACES, and in literacy according to teacher absence reports in FACES. Thus, the findings provide some evidence that children with more absences show lower levels of academic but not social skills, although the specific academic outcomes related to absences varied across studies and informants.

Number of Hours per Week

The next set of analyses used the same strategy to examine associations between hours per week and child outcomes, using data from the two studies in which hours per week were reported. The program director or classroom teacher reported hours of operation in FACES 2006 and NCEDL; hours varied across programs in both studies. We added hours per week of ECE to the final quality threshold model, with results of analyses reported in the middle set of columns in Table 12. Again, separate analyses of each outcome were conducted by using each quality measure, adding hours per week to the final model in the threshold analyses involving HLMs, covariates, and multiple imputations. For a given outcome, we report the coefficient in which quality accounted for the greatest variance from the models involving different quality measures. Hours per week of operation of ECE programs did not emerge as a statistically significant predictor; thus, no evidence of replicated associations emerged from the analyses.

Instructional Time within Specific Content Areas

The same strategy was used to examine instructional time, drawing on data from two studies. For NCEDL, we used the proportion of time observed in instruction on Snapshot; for PCER, we used the rating of time observed in instruction according to TBRS. In both data sets, we examined instruction time in mathematical activities as a predictor of mathematics outcomes and instruction time in literacy activities as a predictor of language and literacy outcomes. Results are in the final set of columns in Table 12. The analyses also used the final model from the threshold analyses, including each quality measure as either a piecewise or linear predictor, the same covariates, HLM analyses, and multiple imputation. The analyses allow us to consider time on instruction in a particular content area, holding constant the observed measure of quality included in the threshold analysis.

As shown in the final columns of Table 12, time spent in mathematics instruction was a consistent predictor of mathematics skills (NCEDL: d = 0.04; PCER: d = 0.07), and time spent in literacy instruction was a consistent predictor of literacy outcomes (NCEDL: d = 0.04–0.05; PCER: d = 0.11), even after accounting for quality of instruction, the child’s skills in the fall, and demographic covariates.

Quality by Dosage Interactions

The next set of analyses tested whether there were interactions between quality and dosage. Using FACES 2009 and HSIS data, we first used the propensity score approach to examine whether children with more years of high quality care had better outcomes. In another set of analyses, we added the interaction between the final quality terms in the quality threshold analysis and hours per week, absence, and time spent in specific instructional domains.

Number of Years of High Quality ECE

The propensity score analyses of the FACES 2009 and HSIS data examined two years of high quality center-based care. The analyses examined whether children’s academic and social skills were higher when children experienced two years of high quality care (as opposed to only one or no years of high quality care) among children with two years of center-based ECE. In both studies, the analyses included the cohort of 3-year-old children with two years of ECE. The children were classified into two groups based on whether they experienced high quality care in both years. Propensity score matching identified matches for children with two years of high quality care on children’s entry skills and family characteristics (child age, gender, race, disability, household language, income-to-needs ratio, maternal education, maternal depression, maternal employment, single-parent family, household mobility, neighborhood safety, as well as language, literacy, mathematics, and social skills based on the fall 3-year-old assessments). Propensity score matching using nearest-neighbor matching eliminated initial differences between the groups on these variables. We attempted to use the same cut-points as in the threshold analyses but were not able to identify a sufficient number of children with high quality care in FACES based on those criteria, especially on the Instructional Support scale. Instead, we used the criteria that Head Start uses for its monitoring system; that is, a low quality classroom was defined with scores of lower than 2 on Instructional Support, 5 on Emotional Support, and 4 on Classroom Organization, with 258 children in the high quality group and 419 in the comparison group after matching. In HSIS, a high quality classroom was defined with a score above 4.5 on the ECERS-R Total score, with 361 children in the high quality group and 177 in the comparison group for the group with fewer than two years of high quality in HSIS.

Results in Table 13 suggest that the outcomes of children with two years of high quality care in Head Start did not differ from those of other children at Head Start exit and spring of kindergarten.

Interactions Between Quality and Dosage

We tested the interaction between the final quality terms in the threshold analysis and absences, hours per week, and time spent in content-specific instruction. Only one of the interactions was observed across studies and therefore meets our goal of obtaining replicated evidence.

First, we examined interactions between absences and quality. Significant interactions between quality and absences emerged in the analysis of FACES 2006, but not in the analyses of the other two studies with absence data. In higher quality care according to the ECERS total score, teacher-reported absence in FACES 2006 was a stronger negative predictor of the residualized gains in language (PPVT-4) and mathematics (WJ III Applied Problems). No interactions were statistically significant in FACES 2009 or NC-PK.

No evidence of interactions emerged in analyses examining hours per week of ECE and quality based on data from NCEDL. Inconsistent findings emerged in analyses examining associations between hours per week of care and quality in relation to child outcomes based on FACES 2006 data. Findings suggested that hours per week was a stronger predictor of PPVT in higher quality care according to the ECERS total and a weaker predictor of ECLS-B mathematics skills in higher quality care according to CLASS instructional support. No evidence supporting interactions emerged in analyses of NCEDL data.

Finally, interactions between time spent in instruction and quality of care were examined. In both NCEDL and PCER, interactions between time spent in mathematics activities and quality of mathematics instruction were statistically significant. Results indicated that more time in mathematics instruction was a stronger predictor of residualized gains in mathematics skills (WJ Applied Problems) when quality of instruction was higher according to the CLASS instructional support in NCEDL and according to the TBRS rating of the quality of mathematics instruction in PCER. No evidence of interactions emerged in analyses of the quality and quantity of literacy instruction.

In summary, there was very limited evidence in the analyses of an interaction of quality and dosage. The only replicated finding pointed to greater gains in mathematics skills when time on mathematical instruction was higher, especially in programs with higher quality instruction.

Summary of Results for Dosage Analyses

The dosage analyses provided evidence that dosage of ECE was related to the acquisition of academic skills. Propensity score analyses indicated that academic skills were enhanced with an additional year in Head Start, but not necessarily with an additional year in higher quality care among children with two years of center-based ECE. Children's academic skills were also stronger when they were in classrooms where more time was spent in instruction. Analyses of absences provided some replicated evidence that children with fewer absences showed larger gains across outcomes in one study, but not across studies. Hours per week of the center-based programs were not reliably or consistently related to child outcomes. Finally, replicated findings of interactions between quality and dosage emerged only in analyses of mathematics outcomes as a function of the quality and quantity of mathematics instruction. Associations between dosage and child outcomes tended to be modest in all of these analyses.

V. DISCUSSION AND CONCLUSIONS

By providing an in-depth examination of thresholds of quality, specificity of quality measurement, dosage operationalized in several ways, and interactions of dosage and quality, the secondary data analyses reported here sought to extend the understanding of how early childhood quality and child outcomes are related. We report findings from analyses that included the child's entry skills and family demographic characteristics as covariates and only report findings that were replicated across studies or across measures as a means to reduce, but not eliminate, potential biases.

Results from the threshold analyses suggested that children experience larger gains in language and literacy skills when the quality of instruction is higher, but only when the quality of instruction is in the moderate to high range. This conclusion was supported by the meta-analysis involving measures of instructional quality based on CLASS Instructional Support and the analysis of the quality of literacy instruction based on the TBRS literacy measure. In the higher, but not lower, quality classrooms, gains in language and literacy outcomes were predicted modestly by the measure of instructional support and moderately by the measure of the quality of literacy instruction. Little to no evidence emerged suggesting thresholds in quality as related to mathematics and behavior problems. Other findings, building on the threshold analyses, indicated that more specific quality instruments (interaction-specific or domain-specific) provide more consistent associations with child outcomes than do global measures of quality.

The dosage analyses extended the quality findings. We expected to find that more time in higher quality programs would be especially advantageous for children. Instead, we found evidence that a greater overall dosage of center-based ECE was related to higher levels of skills when dosage was defined in specific ways. Children with two years of Head Start demonstrated better academic skills at exit from Head Start and at the end of kindergarten than children with only one year of Head Start. Children showed larger gains in academic skills when teachers spent more time in instruction in related content areas. Children with more absences showed smaller gains in academic outcomes. A finding that gains in mathematics skills were larger when children experienced more time in higher quality mathematics instruction was the only result linking stronger outcomes to a larger dose of higher quality.

We turn next to a more detailed summary of the findings for thresholds, features of quality, and dosage, and to a discussion of these findings within the context of policy, practice and research.

Thresholds

The analyses designed to test whether there might be thresholds in the association between ECE quality and child outcomes, using a priori cut-points, provided fairly consistent evidence of thresholds when quality of instruction was examined, especially in relation to children's outcomes in the areas of language and literacy. In the meta-analyses, increments in the interaction-specific measure of instructional support (CLASS Instructional Support) and the domain-specific measures of instruction in language and literacy (TBRS Literacy Quality) were each related to larger gains in children's language and literacy skills in higher rather than in lower quality classrooms. A similar pattern held for the ECERS-R Interaction score (though not for the ECERS-R Total) for children's acquisition of language and for social skills. The ECERS-R Interaction Score was a stronger predictor of these child outcomes in higher rather than in lower quality classrooms.

The analyses provided the most consistent evidence for thresholds for instructional quality. The results suggested that higher instructional quality was related to larger gains in child outcomes in classrooms that were considered moderate to high quality, and was not related in classrooms considered low quality. These findings suggest that instructional quality must reach a threshold before it has an impact on child outcomes. There was no evidence for thresholds in which quality was a stronger predictor in lower than higher quality settings, thus providing no evidence that there is an asymptotic level above which improving quality no longer has an impact on child outcomes. Thus, we can say with confidence that we are seeing evidence suggesting that gains in quality in higher quality classrooms produce the largest gains in child outcomes.

One goal of the analyses was to identify specific cut-points for the thresholds. The plan was to use a priori approaches to identify the most consistent evidence of thresholds and then use the empirical approaches to identify the specific cut-points. The first part of the plan was successful, but the second part was not. Our strategy involved focusing on the meta-analyses to provide evidence using a priori cut-points based on observational measure rating labels and in the ranges now used in evaluations of prekindergarten programs and state Quality Rating and Improvement Systems. The meta-analyses provided replicated evidence suggesting that thresholds might exist for three types of measures related to the instructional quality and quality of the teacher’s interactions with the class. The subsequent empirical analyses tended to suggest that the shape of the associations between classroom quality and child outcomes varied across the continuum of classroom quality, with some suggestion of the pattern of higher thresholds. However, the resulting functions did not provide clear evidence of specific cut-points.

Use of the meta-analytic approach with a priori cut-points as our primary analytic approach allowed us to draw conclusions about whether we were seeing the same pattern of findings across studies. The analyses used the same explicit piecewise models (as noted, based on labels in quality measures and using cut-points currently employed in some policy decisions) in replicated analyses across data from several projects. The meta-analysis combined the findings from these analyses and directly addressed the question about whether the quality measure was a stronger predictor in higher rather than in lower quality classrooms. Thus, these analyses provide fairly strong evidence for thresholds, especially when findings were observed across multiple child outcomes.

The alternative approach focused on the empirical results, first identifying consistency in the smoothed regression lines across studies and then using meta-analysis to test those thresholds indicated in the a priori analyses. These flexible nonlinear methods found the function that optimally describes the association between quality and outcomes for those data; as expected, the functions varied across analyses. Some of the variability was likely attributable to differences in true functional forms and some to error. Based on our experiences, we are convinced that visual examination of the functions is unlikely to yield simple or straightforward conclusions about thresholds that are consistent across data from different projects. Accordingly, the empirical methods made it difficult to answer specific questions with confidence. The methods are nonetheless useful in descriptive analyses designed to generate hypotheses or in sensitivity analyses designed to examine whether alternative functions better describe associations than the a priori relationships tested. Thus, in hindsight, it is not surprising these approaches were not useful in identifying exact cut-points, and it was reassuring that they provided some evidence supporting the results from the a priori analyses.

The findings of the a priori analyses reported here provide more robust evidence than previously available with respect to the possible existence of quality thresholds. The findings replicate findings reported by Burchinal, Kainz, and Cai (2011) and strengthen those findings by analyzing several data sets and conducting meta-analyses to provide omnibus tests across projects. Interestingly, this approach provided stronger evidence for thresholds for academic outcomes than for social outcomes. Only the analyses of ECERS-R Interaction scores—not CLASS Emotional Support scores—provided some evidence similar to results from earlier studies suggesting thresholds in quality in ECE to predict social-emotional outcomes (Burchinal, Vandergrift, Pianta, & Mashburn, 2010; Vandell et al., 2010). Though we found a threshold in the association of quality and social-emotional outcomes by using only one of the quality measures, we found evidence of thresholds across several measures of quality and several academic outcomes.

These findings may be helpful to policy makers and practitioners. These findings are consistent with hypotheses—but do not prove—that children show gains in their language and literacy outcomes when instructional quality improves, but only among classrooms above the quality threshold (labeled "active range"). That is, these findings suggest children in classrooms with lower initial quality will not show gains in these outcomes as instructional quality is improved until quality has improved sufficiently to exceed the threshold. Although improvement in instructional quality may be especially important for lower quality classrooms for many reasons, it is likely to translate into smaller gains on these outcomes than is improvement in instructional quality in higher quality classrooms. This interpretation is consistent with the contrasting evidence from two studies: findings from the Boston Pre-kindergarten study suggested that literacy instruction through a specific curriculum, Opening the World of Learning (OWLS), resulted in large gains in language and literacy skills in a setting where instructional quality was initially good (Weiland et al., 2013). In contrast, the recent experimental study in Tennessee found that the OWLS curriculum was not effective in a setting where instructional quality was initially low (Kaiser, Dickinson, Darrow, Roberts, Freiberg, & Hofer, 2011).

This pattern raises questions about effective means to increase instruction quality in lower quality programs. Although widely used, workshops that provide information to teachers without also providing the opportunity for on-site practice with feedback from a supportive source are unlikely to bring about substantial changes in instructional quality (Zaslow, Tout, Halle, Whittaker, & Lavelle, 2010). In contrast, there are specific professional development approaches that have been shown to be effective in strengthening quality in lower quality classrooms. For example, some professional development work suggests that targeted coaching may be effective (Burchinal et al., 2014). For example, even among teachers with lower quality CLASS scores, explicit training in meeting CLASS dimensions resulted in higher CLASS scores (Hamre, Pianta, Burchinal, & Downer, 2010) and explicit training in literacy instruction practices resulted in substantial gains in children's literacy skills, (Powell et al., 2010). In addition, other approaches have been proposed. For example, tiered approaches help programs establish foundational practices (such as daily routines or health and safety practices) before focusing intentionally on supporting children's social and emotional development, early math skills, or language and literacy development. Such tiered approaches have been explored in specific quality improvement initiatives (Allard Agnamba, 2012), but their effectiveness relative to implementing the same quality improvement approaches focusing on stimulation in particular domains of development (such as coaching to improve language and literacy stimulation) in programs with differing starting levels of quality has not yet been explored.

For practitioners, both teachers and those directing programs, it will be critical to learn from and build on efforts that have improved instructional quality. The threshold findings on CLASS Instructional Support suggest that professional development support should point toward strategies that focus on improving the instructional content in teacher-child interactions by, for example, ensuring extended conversations with children and improving the quality of feedback to children. The threshold findings on TBRS literacy and the dosage findings for TBRS literacy and mathematics suggest that professional development should support those approaches with intentional content-specific instruction that includes structured and unstructured activities designed to teach specific skills such as oral language, phonemic skills, and print knowledge for literacy and numeracy and geometry skills for mathematics.

A hallmark of high quality instruction is that it is enjoyable and engaging and organized in a manner that facilitates the learning and retention of these important language, literacy, and mathematics skills. We are learning, for example, from studies such as the evaluation of prekindergarten in Boston, more about what high instructional quality looks like. Such instructional quality involves engaging activities, small- and large-group instruction, and sequenced presentation of the instructional materials that allows for deep learning. Children are actively engaged in learning activities that teachers organize and sometimes direct. The curricula provide many opportunities to practice the prerequisite skills before the presentation of more complex skills. Such curricula involve some whole-group instruction such as circle time, but much of the instruction takes place in small groups in which teachers differentiate the lessons depending on the children's performance on monitoring tools. The curricula are aimed at helping children learn domain-specific content (for example, helping them extend their vocabularies or learn mathematics concepts) through both structured and unstructured activities that are enjoyable, involve engagement with learning materials, and are marked by responsive child-adult interactions.

An important caveat to the findings involves the exact identification of the thresholds. We caution that, although we lack the evidence to indicate exactly where thresholds occur, results do suggest that the thresholds are probably in the ranges examined in this study with the a priori approach. In other words, although we cannot say conclusively that the threshold for CLASS Instructional Support is 2.75, 3.00, or 3.25, findings from this study and previous studies point to that general range selected primarily based on accepted definitions for low, moderate, and high quality on the selected observational measures of quality. The findings provide some support for cut-points used in Quality Rating and Improvement Systems and in monitoring systems such as the Head Start Designation Renewal System. These systems have also moved forward in setting cut-points, often by using a priori approaches in designating thresholds, given pressing needs and the absence of conclusive evidence on where thresholds should be designated. The research now seeking to validate the distinctions in levels of quality made in Quality Rating and Improvement Systems (Elicker, Langill, Ruprecht, Lewsader, & Anderson, 2011; Sabol, Hong, Pianta, & Burchinal, 2013; Tout et al., 2011; Zellman, Perlman, Le, & Setodju, 2008) will provide one important perspective on thresholds. The results of that research will likely be used in the creation and refinement of programs such as state Quality Rating and Improvement Systems or the Head Start Designation Renewal System. Interaction-specific quality measures are becoming widely used in these monitoring systems as well as in research involving those systems. Such research could also benefit from the more widespread adoption of domain-specific measures of quality within specific content areas as well as from attention to the question of whether the same thresholds are appropriate for groups that vary in terms of socioeconomic status or home language.

Specificity of Quality Measurement

Perhaps the most robust finding from the analyses highlighted in this monograph involves the comparisons of global and more specific observed quality measures. The analyses comparing the predictive ability of global and more specific quality measures suggested that interaction-specific and domain-specific quality measures provided better prediction of children’s outcomes thought to be influenced by those specific quality measures than did global quality measures. The findings are consistent with those from earlier studies that examined either quality measures and children’s outcomes in several data sets (Burchinal, Kainz, & Cai, 2011) or several measures of quality in a single data set (Mashburn et al., 2008).

The findings may hold particular interest for policymakers and practitioners who are undertaking initiatives designed to provide high quality care for children to improve school-readiness skills, especially among children with personal or social risk factors. The finding that interaction-specific quality measures provide modestly better prediction of gains in academic outcomes than do global quality measures is consistent with policies such as the recent Head Start Designation Renewal System (Office of Head Start, 2011) that relies on CLASS as one metric to monitor quality. It is also consistent with a growing number of states that have included CLASS in their Quality Rating and Improvement Systems (Tout et al., 2010) and with the growing reliance on CLASS for measuring quality in descriptive and evaluation studies of ECE (e.g., Moiduddin, Aikens, Tarullo, West, & Xue, 2012; Vitiello, Moas, Henderson, Greenfield, & Munis, 2012; Weiland, Ulvestad, Sachs, & Yoshikawa , 2013).

The finding that domain-specific quality measures provided moderately larger prediction of gains in children's outcomes than did other measures is not widely reflected in either policy or research. It makes sense that children acquire more language and literacy skills when teachers provide higher quality instruction in those areas, but some providers remain anxious about approaches that focus on intentional instruction (Gordon, 2014). There are both encouraging indications and challenges in looking toward greater reliance on domain-specific measures and toward a greater focus on improvement in particular domains of quality within classrooms and programs.

One important indication of greater emphasis on environmental supports within particular domains is the development of curricula for early childhood settings that focus on supporting specific aspects of school readiness. Experimental evaluations of specific curricula have demonstrated moderate to large gains for mathematics curricula (e.g., Clements & Sarama, 2008; Weiland, Ulvestad, Sachs, & Yoshikawa, 2013), literacy curricula (e.g., Bierman et al., 2008;

Farver, Lonigan, & Eppe, 2009; Landry, Anthony, Swank, & Monesque-Bailey, 2009; Neuman & Cunningham, 2009; Powell, Diamond, Burchinal, & Koehler, 2010; Weiland, Ulvestad, Sachs, and Yoshikawa et al., 2013), language curricula (Neuman & Cunningham, 2009; Wasik & Hindman, 2011), and social-emotional skills curricula (e.g., Bierman et al., 2008; Fantuzzo, Gadsden, & McDermott, 2011; Raver et al., 2011). The evidence is growing that greater attention should be paid to the content of instruction when evaluating ECE programs, suggesting that consideration of program quality without consideration of the content of instruction is not sufficient.

As noted, an increasing number of state and local rating systems are using CLASS (Tout et al., 2010). Yet, the transition from using global measures of quality to relying on interaction-specific measures within states or localities has not been an easy one. Another shift to a focus on domain-specific measures of quality within specific content areas would certainly be challenging, even if the evidence suggests that this approach is the most promising for guiding quality improvement efforts that will translate into improvements in child outcomes.

However, we are beginning to see approaches that use an interaction-specific measure, such as the CLASS, along with selected domain-specific measures in the particular content areas on which programs or systems have decided to focus (see, for example, Weiland, Ulvestad, Sachs, & Yoshikawa, 2013 for selection of quality measures in the Boston prekindergarten system). Such complementary use of interaction-specific and *selected* domain-specific quality measures may emerge as an especially important strategy. It may be particularly appropriate in contexts in which programs are called upon to prioritize the content areas they will address per their ongoing monitoring of children's school-readiness assessments. Rather than being overwhelmed by attempting to prioritize and measure all domains of school readiness in depth, a program might, for example, flag expressive language, literacy, and self-regulation as its highest priority areas and combine interaction-specific measures of quality with measures of stimulation specifically in the areas of environmental supports for language, literacy, and social-emotional development to guide its efforts (see Bierman et al., 2008 for such an example).

Dosage

The dosage analyses found evidence—across different measurement approaches—that extent of exposure to ECE appears to contribute to children's development. Our findings indicate that longer exposure was related to children's academic outcomes in analyses examining number of years of Head Start, absences, and instructional time, but not in analyses examining hours of ECE per week and number of years of high quality ECE.

Our analyses indicated that children showed modestly larger gains in literacy and mathematics skills when their teachers spent more time on literacy and mathematics instruction. In light of these findings, the relatively small amounts of time spent in instruction in ECE as reported in FACES 2006 and 2009 and HSIS and in studies of ECE in general (Howes et al., 2008) should be worrisome. The results point to the importance of working to increase time focused on instruction in language, literacy, and mathematics in ECE.

These findings might be useful for policy and practice. It is not surprising that children, on average, learn more mathematics when more time is spent teaching mathematics and that children acquire more literacy skills when more time is spent teaching early reading skills. Furthermore, it is also not surprising that children learn more when the instruction is sequenced to introduce increasingly more difficult concepts and link the new concept to those learned in prior lessons (Clements & Sarama, 2008; Weiland, Ulvestad, Sachs, & Yoshikawa, 2013). There is a growing consensus that intentional teaching, which can include structured teacher-directed lessons, is important to ensure that all children enter kindergarten with the skills needed to succeed in school. This change will require consideration of teacher attitudes about whether there should be intentional instruction for young children (Hamre, 2012) as well as the provision of supports—drawing on approaches such as coaching (Zaslow, Tout, Halle, Whittaker, & Lavelle, 2010)—for improving the implementation of curricula with content-rich, engaging, and sequenced instruction. It will be critical to convince teachers that intentional instruction provides preschool-age children with enhanced opportunities to learn school-readiness skills needed to succeed in school.

Findings also suggested that children with more absences showed smaller gains on academic outcomes. Other measures of dosage such as attendance have received much attention in public school. For example, attendances and absences have emerged in early elementary school as important predictors of academic achievement (Chang, 2008). School policies encourage high rates of attendance and may sanction families whose children demonstrate low rates of attendance. Just as in our results, other studies are beginning to point to the importance of supporting attendance in ECE programs for later academic outcomes (Connolly, 2012; Ehrlich, 2013). Establishing a causal link for attendance is especially difficult because the family typically must ensure that the child attends school, raising major questions about family selection bias in analyses relating attendance to child outcomes. We used the child’s skill levels at entry to the preschool classroom as a covariate to address concerns about bias but probably did not fully account for potential bias. Our analyses yielded some, albeit limited, replicated evidence indicating that children with more absences had lower academic skills. Even though it is important to keep in mind the need for further rigorous work on attendance as related to selection issues, the findings indicate that addressing family barriers or concerns about program participation during a child’s preschool years may not only help support children’s progress toward school readiness but also help establish patterns of attendance that are important later in schooling.

We had hypothesized that more time in higher quality care would provide the strongest prediction of higher levels of gains in academic and social skills. Only one of the many interactions we examined supported our hypothesis, despite careful analyses in which we used the information from the quality threshold analyses to represent quality. Instead, we found evidence that attendance and time in instruction matter, independent of quality, and some evidence that more time in higher quality mathematics instruction leads to larger gains in mathematics.

One of the most consistent findings in these analyses involved comparing one and two years of Head Start. The propensity score analysis of the FACES 2006 and 2009 and HSIS data provided evidence that children who entered Head Start at age 3 and stayed for two years had modestly higher vocabulary scores at the end of Head Start and kindergarten than children who entered Head Start at age 4 and stayed for one year. Even though some pre-kindergarten programs (such as in the District of Columbia) have been built on an assumption of two years as opposed to one year, such an approach may not be feasible in all contexts. Future work may need to examine one or two years of other programs such as prekindergarten or combinations such as one year of Head Start at age 3 and one year of prekindergarten at age 4, as these combinations may become more typical. It is encouraging and important that further levers for increasing dosage are also supported by the results presented here.

Limitations of the Present Study

Several limitations of the present study must be considered. First, the findings in the study are correlational, not causal. We attempted to adjust for family selection factors by including both the child’s entry-level skills and family characteristics. Our use of propensity score matching involved steps beyond those usually taken to control for selection in analyses focusing on quality or dosage and child outcomes. However, these approaches should reduce, but not eliminate bias, and therefore findings should be interpreted as suggestive, but not definitive evidence of causality.

Second, we conducted many analyses--separate analyses of each outcome for each quality and dosage measure. We used meta-analysis as an attempt to identify patterns across data sets when possible and to focus on replication when meta-analysis was not possible. Nevertheless, findings should be interpreted cautiously—especially findings that are not replicated in independent data sets (e.g., findings regarding thresholds for content-specific quality measures or findings regarding negative associations with absences).

Third, in terms of limitations with existing samples and studies included in our secondary analyses, some of the projects in the secondary data analysis did not have wide variation in quality (e.g., FACES). Further, almost all of the studies measured children for the post-test only six to eight months after they measured them for the pre-test, perhaps not enough time for large changes in child outcomes to occur.

Next Steps for Research

In considering how to build on the findings of the present set of analyses, we conclude that research designed explicitly to address questions about thresholds is needed. The literature review and the secondary data analysis offer consistent evidence that domain-specific and interaction-specific quality measures provide better prediction of child outcomes than do global quality measures. A study designed explicitly to study the issues of thresholds should focus on interventions designed to improve both domain-specific instruction and teacher-child interactions. Similarly, the literature review and secondary data analysis provide evidence that there might be a quality threshold in improving language and literacy skills. The implications for study design suggest contrasting improvements in child outcomes when quality of instruction in language and literacy improves through a quality intervention for classrooms that started with lower versus higher initial levels of quality. Our findings point to the hypothesis that larger gains in child outcomes would occur when domain-specific quality improves in higher rather than in lower quality classrooms. To test such a question, it would be valuable to design a study that ensures a sufficient number of classrooms with initially higher and lower domain-specific quality. The use of the same quality improvement approach in higher and lower quality classrooms might be contrasted with the use of more intensive or tiered quality improvement approaches in lower quality classrooms. It would be valuable to assess the costs that are involved in improving quality in higher and lower quality ranges as well, particularly if the evidence points to the need for more intensive or tiered approaches in the lower quality range.

Finally, the secondary data analysis suggests that children perform better on specific academic outcomes with larger dosages of ECE, measured here as two years of Head Start, greater attendance, and more time spent in instruction. To build on and test further these patterns of dosage, an experimental design would randomly assign children to one of three groups: one year of Head Start or prekindergarten at age 3; one year of Head Start or prekindergarten at age 4; or two years of Head Start or prekindergarten starting at age 3. Research could also seek to implement and rigorously evaluate interventions to address early emerging patterns of chronic absence, examining whether such interventions are effective not only in improving school readiness but also in sustaining effects on attendance in elementary school. Further, consistent attendance is critical in benefitting from interventions aimed at improving instructional quality and the content of that instruction as well as time spent on instruction.

In conclusion, we began the set of analyses with a focus, first, on thresholds, then on quality features, and finally on dosage. We conclude with the strongest sense of certainty about the need for ECE that focuses on intentional instruction involving activities designed to promote specific school-readiness skills and content-rich interactions between teachers and all children.

Greater amounts of interaction with an instructional focus that scaffolds learning for the young child and longer periods of time in ECE settings with standards for quality and monitoring also appear beneficial. Whereas we do see indications of quality thresholds, especially when more specific measures of instructional quality are used and academic skills are considered, more rigorous testing for thresholds will require new research that examines the results of the same or intentionally differentiated quality improvement efforts within different ranges of quality. Such work should build on the interconnectedness we have found between/among quality features, dosage, and thresholds by focusing on quality improvement efforts aimed at increasing the amounts of high quality instructional interactions providing specific content.

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 profiles*. Burlington: University of Vermont, Department of Psychiatry.

Allard Agnamba, L. T. (2012). *Preparation and ongoing support for early childhood instructional coaches: A case study exploration of an instructional coaching program.* Unpublished doctoral dissertation. University of Pennsylvania, Philadelphia.

Berry, D.J., Bridges, L.J., & Zaslow, M.J. (2004). *Early childhood measures profiles.* Washington, DC: Child Trends. Retrieved from http://aspe.hhs.gov/hsp/ecmeasures04/report.pdf

Belsky, J., & Pluess, M. (2012). Differential susceptibility to long-term effects of quality of child care on externalizing behavior in adolescence? *International Journal of Behavioral Development*, *36*(1), 2-10.

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

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

Bryant, D. (2010). *Observational measures of quality in center-based early care and education programs* (OPRE Research-to-Policy Brief #5). Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation, and Child Trends.

Build Initiative and Child Trends. ((2014). A Catalog and Comparison of Quality Rating and Improvement Systems (QRIS). (2014, November 3). Retrieved November 7, 2014, from http://qriscompendium.org/

Burchinal, P., Kainz, K., Cai, K., Tout, K., Zaslow, M., Martinez-Beck, I., & Rathgeb, C. (2009). *Early care and education quality and child outcomes* (OPRE Research-to-Policy Brief #1). Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation, and Child Trends.

Burchinal, M., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. In M. Zaslow (Ed.), *Reasons to take stock and strengthen our measures of quality* (pp. 11-31). Baltimore, MD: Brooks Publishing.

Burchinal, M., Magnuson, K., Powell, D., & Hong, S. S. (2014). Early child care and education and child development. In M. Bornstein, R. Lerner, & T. Leventhal (Eds.) *Handbook of Child Psychology and Developmental Science*. Volume IV, (pp223-267). New York, NY: Wiley Press

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A., (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly. 25,* 166-176.

Burchinal, M., Vernon-Feagans, L., Vitiello, V., Greenberg, M., & the Family Life Project Key Investigators. (2014). Thresholds in the association between child care quality and child outcomes in rural preschool children. Early Childhood Research Quarterly, 29, 41-51.

Bureau of Labor Statistics (April 26, 2013). *Employment Characteristics of Families-2012* (USDL-13-0730). Retrieved from http://www.bls.gov/news.release/pdf/famee.pdf

Caldwell, B., & Bradley, R. (1984). *Home Observation for Measurement of the Environment (HOME) Inventory-Revised Edition.* Little Rock: University of Arkansas.

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*, 579-620.

Campbell, F. A., Pungello, E., Miller-Johnson, S., Burchinal, M., & Ramey, C. T. (2001). The development of cognitive and academic abilities: Growth curves from an early childhood education experiment. *Developmental Psychology, 37*, 231-242.

Campbell, F. A., Ramey C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education: Young adult outcomes from the Abecedarian Project. *Applied Developmental Science, 6*, 42-57.

Chang, H., (2008). Present, engaged and accounted for: The critical importance of addressing chronic absence in the early grades. New York: National Center for Children in Poverty.

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal,45*, 443-494.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, *83*(403), 596-610.

Connolly, F. (2012). Early elementary performance and attendance in Baltimore city school’s pre-kindergarten and kindergarten. Baltimore: Baltimore Education Research Consortium.

Dearing, E., McCartney, K., & Taylor, B. (2009). Does higher-quality early child care promote low-income children’s math and literacy achievement in middle childhood? *Child Development, 80*, 1329-1349.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... & Japel, C. (2007). School readiness and later achievement. *Developmental psychology*, *43*(6), 1428.

Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *The Journal of Economic Perspectives*, *27*(2), 109-132.

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody picture vocabulary test*. Harcourt Test Publishers

Dunn, L. M., Dunn, D. M., & Schlichting, J. E. P. T. (2005). *Peabody picture vocabulary test-III-NL*. Circle Pines, MN: American Guidance Service.

Ehrlich, S. (2013). Preschool attendance in Chicago Public Schools: Relationships with learning outcomes and reasons for absences. University of Chicago Consortium on Chicago School Research.

Elicker, J. G., Langill, C. C., Ruprecht, K. M., Lewsader, J., & Anderson, T. (2011). Final report: Evaluation of paths to QUALITY, Indiana's child care quality rating and improvement system. Purdue University: Department of Human Development and Family Studies.

Fantuzzo, J. W., Gadsden, V. L., & McDermott, P. A. (2011). An integrated curriculum to improve mathematics, language, and literacy for Head Start children. *American Educational Research Journal, 48*, 763-793.

Farnworth, M., Schweinhart, L. J., & Berrueta-Clement, J. R. (1985, Fall). Preschool intervention, school success and delinquency in a high-risk sample of youth. *American Educational Research Journal*, *22*, 445–464.

Farver, J., Lonigan, C., & Eppe, S. (2009). Effective early literacy skill development for young Spanish-speaking English language learners: An experimental study of two methods. *Child Development, 80,* 703-719.

Gordon, R. A. (2014). Assuring quality preschool: Where are we and where do we need to go? Presented in the American Educational Research Association Presidential Session on

Universal Preschool: What Have We Learned, and What Does It Mean for Practice and Policy? (Chair: Rachel A. Gordon, Discussant: Libby Doggett), Philadelphia, PA.

Grehan, A. Smith, L. (2004). Early literacy observation tool. Memphis TN: University of Memphis Center for Research in Educational Policy.

Gresham, F. M., & Elliott, S. N. (1990). *Social skill rating system.* Circle Pines, MN: American Guidance Services.

Hamre, B. (2012). Enhancing teachers' intentional use of effective interactions with children. In R. C. Pianta (Ed.), *Handbook of early childhood education* (pp. 507-532). New York: Guilford Press.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environment rating scale (Rev. ed.).* New York: Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. M. (2006). *The infant/toddler environment rating scale (Rev. ed., updated).* New York: Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. M. (2007). *The child care environment rating scale (Rev. ed.).* New York: Teachers College Press.

Hightower, A. D., Work, W. C., Cowen, E. L., Lotyczewski, B. S., Spinell, A. P., Guare, J. C., & Rohrbeck, C.A. (1986). The teacher-child rating scale: A brief objective measure of elementary children's school problem behaviors and competencies. *School Psychology Review, 15,* 393-409.

Hill, J. L., Brooks-Gunn, J., & Waldfogel, J. (2003). Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology, 39*(4), 730-744.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly, 23,* 27-50.

Howes, C., Phillips, D. A., & Whitebook, M. (1992). Thresholds of quality: Implications for the social development of children in center-based child care. *Child Development, 63*, 449-460.

Hubbs-Tait, L., Culp, A. M., Huey, E., Culp, R., Starost, H., & Hare, C. (2002). Relation of Head Start attendance to children's cognitive and social outcomes: Moderation by family risk. *Early Childhood Research Quarterly, 17*, 539-558.

Kaiser, A., Dickinson, D., Roberts, M., Darrow, C., Freiberg, J., & Hofer, K. (2011). The Effects of Two Language-Focused Preschool Curricula on Children's Achievement through First Grade. *Society for Research on Educational Effectiveness*.

Karoly, L. A., Kilburn, M. R., & Cannon, J. S. (2005). *Early childhood interventions: Proven results, future promise*. Santa Monica, CA: The RAND Corporation.

Keys, T. D., Farkas, G., Burchinal, M. R., Duncan, G. J., Vandell, D. L., Li, W., ... & Howes, C. (2013). Preschool center quality and school readiness: Quality effects and variation by demographic and child characteristics. *Child development*, *84*(4), 1171-1190.

Landry, S. H., Anthony, J. L., Swank, P. R., & Monesque-Bailey, P. (2009). Effectiveness of comprehensive professional development for teachers of at-risk preschoolers. *Journal of Educational Psychology, 101*, 448-465.

Landry, S. H., Crawford, A., Gunnewig, S. B., & Swank, P. R. (2002). *Teacher behavior rating scale*. Unpublished instrument, University of Texas Health Science Center at Houston, Center for Improving the Readiness of Children for Learning and Education.

Laughlin, L. (2013). *Who's minding the kids? Child care arrangements: Spring 2011* (Current

Population Reports, P70-135). Washington, DC: U.S. Census Bureau.

Loeb, S., Fuller, B., Kagan, S. L, & Carrol, B. (2004). Child care in poor communities: Early learning effects of type, quality, and stability. *Child Development, 75*, 47-65.

Love, J. M., Kisker, E. E., Ross, C., Raikes, H., Constantine, J., Boller, K., . . . Vogel, C. (2005). The effectiveness of Early Head Start for 3-year-old children and their parents: Lessons for policy and programs. *Developmental Psychology, 41*(6), 885.

Mashburn, A. J., Downer, J. T., Hamre, B. K., Justice, L. M., & Pianta, R. C. (2010). Consultation for teachers and children's language and literacy development during pre-kindergarten. *Applied Developmental Science, 14*, 179-196.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J., Barbarin, O. A., Bryant, D., . . . Howes, C. (2008). Measures of classroom quality in pre-kindergarten and children's development of academic, language and social skills. *Child Development, 79,* 732-749.

McCartney, K., Burchinal, M., Clarke-Stewart, A., Bub, K. L., Owen, M. T., Belsky, J., & the NICHD Early Child Care Research Network. (2010). Testing a series of causal propositions relating time in child care to children's externalizing behavior. *Developmental Psychology, 46,* 1-17.

Moiduddin, E., Aikens, N., Tarullo, L., West, J., & Xue, Y. (2012). Child Outcomes and Classroom Quality in FACES 2009. OPRE Report 2012-37a. *Administration for Children & Families*.

Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *American Educational Research Journal, 46,* 532-566.

National Institute of Child Health and Human Development (NICHD) Early Child Care Research Network (ECCRN). (1996). Characteristics of infant child care: Factors contributing to positive caregiving. *Early Childhood Research Quarterly, 11*, 269-306.

National Institute of Child Health and Human Development (NICHD) Early Child Care Research Network (ECCRN). (1998). Early child care and self-control, compliance, and problem behaviors at twenty-four and thirty-six months. *Child Development, 69*(4), 1145-1170.

National Institute of Child Health and Human Development (NICHD) Early Child Care Research Network (ECCRN). (2000). The relation of child care to cognitive and language development. *Child Development, 71*, 960-980.

National Institute of Child Health and Human Development (NICHD) Early Child Care Research Network (ECCRN). (2006). The relations of classroom contexts in the early elementary years to children's classroom and social behavior. In A. C. Huston & M. N. Ripke (Eds.). *Developmental contexts in middle childhood: Bridges to adolescence and adulthood* (pp. 217-236). New York, NY: Cambridge University Press.

National Institute of Child Health and Human Development (NICHD) Early Child Care Research Network & Duncan, G. J. (2003). Modeling the impacts of child care quality on children’s preschool cognitive development. *Child Development, 74*(5), 1454-1475.

OECD (2013) *Maternal employment rates: OECD Family Database,* Social Policy Division, Directorate of Employment, Labour and Social Affairs. Retrieved from http://www.oecd.org/els/family/LMF1_2_Maternal_Employment_July2013.pdf

Office of Child Care (2013). *ACF announces proposed changes to child care regulations.* Retrieved from http://www.acf.hhs.gov/programs/occ/news/acf-announces-proposed-changes-to-child-care-regulations

Office of Head Start (2011). *Designation renewal*. Washington, DC: Office of Head Start Early Childhod Learning and Knowledge Center. Retrieved from eclkc.ohs.acf.hhs.gov/hslc/hs/grants/drPeisner-Feinberg, E. S. & Schaaf, J.M. (2007). Evaluation of the North Carolina More at Four Pre-kindergarten Program: Children's Outcomes and Program Quality in the Fifth Year. Chapel Hill, NC: FPG Child Development Institute.

Peisner-Feinberg, E. S. & Schaaf, J.M. (2008). Evaluation of the North Carolina More at Four Pre-kindergarten Program: Performance and Progress in the Seventh Year (2007-2008). Chapel Hill, NC: FPG Child Development Institute.

Peisner-Feinberg (2013). *North Carolina Pre-Kindergarten Program Evaluation: Summary of Research 2002-2013.* . Chapel Hill: The University of North Carolina, FPG Child Development Institute. Downloaded on October 15, 2014 from http://fpg.unc.edu/sites/fpg.unc.edu/files/resources/reports-and-policy-briefs/Summary of NC Pre-K Evaluation Findings 2005-2014.pdf

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R. M., Early, D. M., & Barbarin, O. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Prediction of observed classroom quality and teacher-child interactions. *Applied Developmental Science, 9*(3), 144-159.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2004). *Classroom assessment scoring system [CLASS] manual: Pre-K.* Baltimore, MD: Brookes Publishing.

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher–child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, *23*, 431-451.

Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on Head Start teachers and children. *Journal of Educational Psychology, 102,* 299-312.

Preschool Curriculum Evaluation Research Consortium (PCER). (2008). *Effects of preschool curriculum programs on school readiness: Report from the preschool curriculum evaluation research initiative.* Washington DC: National Center for Education Research.

Puma, M., Bell, S., Cook, R., Heid, C., & Lopez, M. (2005). Head Start Impact Study: First Year Findings. *Administration for Children & Families*.

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism. *Child Development, 82*, 362-378.

Ritchie, S., Weiser, B, Kraft-Sayre, M. & Howes, C. (2001). *Emergent academics snapshot scale* (Unpublished instrument). Los Angeles: University of California.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Sabol, T. J., Soliday Hong, S. L., Pianta, R. C., & Burchinal, M. R. (2013). Can rating pre-k programs predict children's learning? *Science, 23 (341), 845-846.*

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). Lifetime effects: The High/Scope Perry preschool study through age 40. Ypsilanti, MI: High/Scope Press.Available at http://www.highscope.org/Research/PerryProject/perrymain.htm

Shadish, W. R.; Cook, T. D.; Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin. ISBN 0-395-61556-9.

Smith, M. W., Brady, J. P., & Anastasopoulos, L. (2008). *Users guide to the early language and literacy classroom observation tool, pre-k.* Baltimore: Brookes Publishing.

Smith, M., & Dickinson, D. K. (2002). *User's guide for the early language and literacy classroom observation toolkit.* Boston, MA: Brookes Publishing.

Sylva, K., Stein, A., Leach, P., Barnes, J., & Malmberg, L. E. (2011). Effects of early child-care on cognition, language, and task-related behaviours at 18 months: An English study. *British Journal of Developmental Psychology*, *29*(1), 18-45.

Tout, K., Starr, R., Isner, T., Cleveland, J., Albertson-Junkans, L., Soli, M., & Quinn, K. (2011). *Evaluation of parent aware: Minnesota's quality rating and improvement system pilot.* Minneapolis, MN: Child Trends.

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). *Issues for the next decade of quality rating and improvement systems* (OPRE Issue Brief #3). Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Tran, H. & Weinraub, M. (2006). Child care effects in context: Stability and multiplicity in nonmaternal child care arrangements during the first 15 months of life. *Developmental Psychology, 42*, 566-82.

US Departments of Education and Health and Human Services, (2104) http://www.ed.gov/news/press-releases/six-states-awarded-race-top-early-learning-challenge-rtt-elc-grants-build-statew

Vandell, D. L., Belsky, J., Burchinal, M., Steinberg, L., Vandergrift, N., & NICHD Early Child Care Research Network. (2010). Do effects of early child care extend to age 15 years? *Child Development, 81*, 737–756.

Vermeer, H. J., Groeneveld, M. G., Larrea, I., van IJnendoorn, M. H., Barandiaran, A., & Linting, M. (2010). Child care quality and children's cortisol in Basque Country and the Netherlands. *Journal of Applied Developmental Psychology, 31*, 339-347.

Vitiello, V. E., Moas, O., Henderson, H. A., Greenfield, D. B., & Munis, P. M. (2012) Goodness of fit between children and classrooms: Effects of child temperament and preschool classroom quality on achievement trajectories. *Early Education & Development*, *23*, 302-322.

Vogel, C., Brooks-Gunn, J., Martin, A., & Klute, M.M.. (2013). Impacts of Early Head Start participation on child and parent outcomes at ages 2, 3, and 5. In *What Makes a Difference: Early Head Start Evaluation Findings in a Developmental Context*, edited by J. M. Love, R. Chazan-Cohen, H. Raikes, and J. Brooks-Gunn. Boston: Wiley.

Votruba-Drzal, E., Coley, R. L., & Chase-Lansdale, P. L. (2004). Child care and low-income children's development: Direct and moderated effects. *Child Development, 75,* 296-312.

Wasik, B. A., & Hindman, A. H. (2011). Improving vocabulary and pre-literacy skills of at-risk preschoolers through teacher professional development. *Journal of Educational Psychology, 103*, 455-469.

Watamura, S. E., Phillips, D. A., Morrissey, T. W., McCartney, K., & Bub, K. (2011). Double jeopardy: Poorer social-emotional outcomes for children in the NICHD SECCYD experiencing home and child-care environments that confer risk. *Child Development, 82*(1), 48-65.

Weber, R. (2006). *Measurement of child care arrangement stability: A review and case study using Oregon Child Care Subsidy Data.* (Doctoral dissertation). Oregon State University.

Weiland, C., Ulvestad, K., Sachs, J., & Yoshikawa, H. (2013). Associations between classroom quality and children's vocabulary and executive function skills in an urban public kindergarten program. *Early Childhood Research Quarterly, 28,* 199-209.

West, J., Tarullo, L., Aikens, N., & Hulsey, L. (2008). Study Design and Data Tables for FACES 2006 Baseline Report. *Washington, DC: US Department of Health and Human Services*.

Woodcock, R. W., & Johnson, M. B. (1989). *WJ-R tests of cognitive ability.* Itasca, IL: Riverside.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III: Tests of achievement.* Itasca. IL: Riverside Publishing.

Yamauchi, C., & Leigh, A. (2011). Which children benefit from non-parental care? *Economics of Education Review, 30,* 1468-1490.

Zaslow, M., Tout, K., Halle, T., Whittaker, J. V., & Lavelle, B. (2010). *Toward the identification of features of effective professional development for early childhood educators.* Washington, DC; U.S. Department of Education, Office of Planning, Evaluaiton and Policy Development, Policy and Program Studies Service. Available at http://www.ed.gov/about/offices/list/opepd/ppss/reports.html

Zaslow, M., Martinez-Beck, I., Tout, K., & Halle, T. (Eds.). (2011). *Quality measurement in early childhood settings*. Baltimore, MD: Brookes Publishing.

Zaslow, M., Crosby, D.A., & Smith, N. (in press). Issues of quality and access emerging from the changing early childhood policy context: Toward the next generation of research. In E.T. Gershoff, R.S. Mistry, & D.A. Crosby (Eds.), *Societal Contexts of Child*

Development: Pathways of Influence and Implications for Policy and Practice. New York, NY: Oxford University Press.

Zellman, G. L., Perlman, M., Le, V., & Setodju, C. M. (2008). *Assessing the validity of the Qualistar early learning quality rating and improvement system as a tool for improving child care quality* (MG-650-QEL). Santa Monica, CA: RAND Corporation.

Zill, N. (1990). *Behavior problems index based on parent report*. Child Trends.

Table 1 Descriptive Statistics by Project: Demographic Characteristics

	EHS		**FACES**		**FACES**		**NC-PK**		**NCEDL**		**PCERS**		**HSIS**		**HSIS**		**MTP**	
			2006		**2009**								**3yr Cohort**		**4yr Cohort**			
	N	**M(sd)**	**N**	**M(sd)**	**N**	**M(sd)**	**N**	**M(sd)**	**N**	**M(sd)**	**N**	**M(sd)**	**N**	**M(sd)**	**N**	**M(sd)**	**N**	**M(sd)**
		Prop		**Prop**		**Prop**		**Prop**		**Prop**		**Prop**		**Prop**		**Prop**		**Prop**
Gender	1043		2501		2381		1313		2966		2900		1355		1004		679	
Male		.50		.51		.50		.51		.49		.51		.48		.52		.48
Female		.50		.49		.50		.49		.51		.49		.52		.48		.52
Race/ethnicity	1043		2498		2381		1310		2898		2640		1355		1004		677	
African American		.36		.34		.33		.37		.18		.43		.37		.41		.48
White/nonHispanic		.37		.20		.23		.33		.41		.33		.30		.09		.29
Latino/Hispanic		.24		.37		.40		.25		.26		.16		.34		.50		.10
Other		.03		.08		.04				.15		.08						.13
Age (fall)	1043	5.3 (.45)	2459	4.4 (.56)	2280	3.9 (.55)	1304	4.53 (.29)	2757	5.05 (.32)	2850	4.56 (.32)	1270	4.12 (.33)	940	5.00 (.74)	681	4.38 (.31)
Mother's Ed	1043		2361		2229				2885		2410		1355		1004		673	
Less than HS		.46		.37		.30				.19		.19		.35		.42		.16
High school		.30		.32		.34				.40		.33		.35		.31		.26
More than HS		.24		.31		.29				.41		.48		.30		.27		.58
Poverty	1043	.64	2448	..75	2301	.62	1313	.74	2750	.58								
Home Lang	1043		2501		2381		1207		2888		2410		1355		1004		682	
English		.80		.71		.70		.78		.73		.75		.73		.62		.81
Not English		.20		.29		.30		.22		.27		.25		.27		.48		.19
Head Start	1043	.52	2501	1.0	2381	1.0	1311	.12	2983	.15	2910	.30	1355	1.0	1004	1.0	695	.00

Table 2 Descriptive Statistics by Project: Child outcomes

		EHS		FACES 2006		FACES 2009		NC-PK[a]		NCEDL		PCERS		HSIS 3yr Cohort		HSIS 4yr Cohort	
Child Outcome		**36m**	**PK**	**Fall**	**Spring**	**Fall**	**Spring**	**Fall**	**Spring**	**Fall**	**Spring**	**Fall**	**Spring**	**Fall**	**Spring**	**Fall**	**Spring**
Language	N	1424	1674	3004	2817	2024	2192	1273	1165	2298	2435	2840	2690	1196	1143	911	858
PPVT	M	83.0	91.5	80.9	83.2	84.6	85.9	84.5	89.1	94.0	96.3	88.4	92.8	91.3	92.8	90.5	92.0
	(sd)	(15.6)	(15.2)	(18.8)	(17.2)	(15.91)	(16.5)	(19.7)	(18.2)	(15.0)	(14.3)	(15.8)	(15.2)	(7.2)	(8.3)	(9.9)	(10.4)
Math	N		1755	2408	2483	1597	1944	1266	1160	2273	2435	2750	2640	930	1136	595	857
WJ-AP	M		88.3	89.5	90.0	89.4	89.9	94.1	98.0	98.4	99.1	93.7	95.8	90.4	89.3	90.9	88.5
	(sd)		(20.1)	(17.4)	(14.6)	(14.6)	(15.1)	(14.4)	(12.0)	(13.7)	(12.9)	(15.2)	(14.0)	(17.9)	(18.4)	(14.9)	(16.1)
Literacy	N		1755	2468	2600	1962	2032	320	291	1485	1583	2780	2650	1196	1147	911	859
WJ-LW	M		89.3	93.7	98.5	94.8	101.2	93.4	96.5	101.2	102.9	98.7	103.0	88.1	94.3	89.1	93.8
	(sd)		(13.7)	(17.1)	(16.9)	(18.4)	(17.3)	(12.2)	(12.3)	(16.0)	(14.1)	(16.1)	(13.9)	(21.3)	(19.3)	(16.5)	(15.3)
Behavior	N	2031	2049	3155	2782	2320	2309	1274	1134	2308	2345	2790	2600				
Problems	M	11.1	10.9	7.07	6.41	4.59	4.26	98.7	98.6	1.50	1.50	100.5	100.6				
	(sd)	(6.47)	(5.6)	(6.31)	(6.22)	(4.53)	(4.55)	(12.7)	(12.7)	(.53)	(.54)	(13.6)	(13.5)				
	Test	CBCL Aggr	CBCL Aggr	BPI	BPI	BPI	BPI	SSRS	SSRS	TCRS	TCRS	SSRS	SSRS				
Social	N		2060	3154	2781	2319	2307	1257	1128	2307	2352	2730	2600				
Skills	M		12.0	15.4	17.4	15.31	17.35	100.7	108.8	3.49	3.65	100.7	106.5				
	(sd)		(1.9)	(4.77)	(4.62)	(4.84)	(4.54)	(15.6)	(14.9)	(.76)	(.77)	(15.9)	(15.0)				
	Test		SSRS	SSRS	SSRS	SSRS	SSRS	SSRS	SSRS	TCRS	TCRS	SSRS	SSRS				

Table 2 continued

		MTP	
Child Outcome		**fall**	**spring**
Literacy TOPEL Blending	N	631	597
	M	15.5	17.2
	(sd)	(3.11)	(2.87)
Behavior Problems	N	603	607
	M	1.52	1.50
	(sd)	(.56)	(.55)
	Test	TCRS	TCRS
Social Skills	N	605	607
	M	3.66	3.95
	(sd)	(.92)	(.92)
	Test	TCRS	TCRS

Note [a] NCPK added the WJ LW in the most recent wave of data collection

EHS=Early Head Start Research and Evaluation Project; FACES = Family and Child Experiences Survey-2006; NC-PK= More-at-Four NC Pre-kindergarten Evaluation; NCEDL= NCEDL 11 State Pre-kindergarten Study; PCERS=Preschool Curriculum Evaluation Research Study; MTP=MyTeachingPartner.

ASBI = Adaptive Social Behavior Index; BPI = Behavior Problem Inventory; CBCL = Child Behavior Checklist; PLS = Preschool Language Scale, PPVT = Peabody Picture Vocabulary Test, RLS = Reynell Developmental Language Scales, SSRS= Social Skills Rating Scale; TCRS=Teacher-Child Rating Scale; WJ-AP = Woodcock Johnson Applied Problems, WJLW = Woodcock Johnson Letter-Word Identification

Table 3 Descriptive Statistics by Project: Preschool Classroom Quality[a]

Quality Measures	Cut point		EHS	FACES 2006	FACES 2009	NC-PK[a]	NCEDL	PCERS	MTP	HSIS 3yr Cohort	HSIS 4-yr Cohort
Global Quality											
ECERS total	4.5	N	984	335	369	206	705	320		175	102
		M (sd)	5.25 (1.14)	3.56 (.58)	4.27 (.78)	4.79 (.92)	3.84 (.82)	3.99 (.99)		5.2 (0.9)	5.4 (0.8)
		% high	75%	6%	41%	65%	22%	33%		53%	46%
ECERS-Interactions Factor	4.5	N		335	369	206	705	320		175	102
		M (sd)		4.03 (.92)	4.68 (.98)	5.56(1.10)	4.72 (1.18)	4.64 (1.26)		5.7 (1.1)	5.8 (0.9)
		% high		27%	61%	83%	60%	58%		54%	46%
Teacher-child Interactions Quality											
CLASS Emotional Support	5.0	N			369	50	694		157		
		M (sd)			5.29 (.50)	5.77 (.85)	5.56 (.68)		5.33 (.64)		
		% high			76%	62%	79%		76%		
CLASS Instructional Support	2.75	N		333	369	50	694		157		
		M (sd)		1.89 (0.55)	2.24 (.65)	3.06 (.96)	2.06 (.84)		2.86 (.50)		
		% high		7%	20%	62%	21%		61%		
Domain-specific Instructional Quality											
ELLCO Literacy	4.0	N				107					
		M (sd)				3.54 (.59)					
		% high				48%					
TBRS – Literacy	2.0	N						300			
		M (sd)						1.29 (.60)			
		% high						12%			
TBRS-Numeracy	2.0	N						300			
		M (sd)						1.00 (.68)			
		% high						8%			

Note

[a] NC-PK collected the ECERS in all three waves of data collection and the CLASS in the most recent wave

EHS=Early Head Start Research and Evaluation Project; FACES = Family and Child Experiences Survey-2006; NC-PK= More-at-Four NC Pre-kindergarten Evaluation; NCEDL= NCEDL 11 State Pre-kindergarten Study; PCERS=Preschool Curriculum Evaluation Research Study.

CLASS= Classroom Assessment Scoring System; ECERS-R= Early Childhood Environment Rating Scale-Revised; ELLCO= Early Language and Literacy Classroom Observation Tool- Literacy Activities; TBRS= Teacher Behavior Rating Scale

[a]Quality measures presented at the classroom level

Table 4. Descriptive Statistics for Dosage Measures by Project

		FACES 2006	NC-PK	NCEDL	PCER
Teacher reported absence	N	240			
	M (sd)	2.81 (.96)			
Parent-reported absence	N	232			
	M (sd)	2.60 (.86)			
Class records absence	N		1306		
	M (sd)		4.05 (1.01)		
Hour per week	N	244		2519	
	M (sd)	24.86 (9.19)		23.08(12.6)	
Proportion time in Math Activities (Snapshot)	N			2225	
	M (sd)			.07 (.06)	
Proportion time in Reading Activities (Snapshot)	N			2225	
	M (sd)			.10 (.07)	
Rating – time on math (TBRS)	N				300
	M (sd)				1.00 (0.68)
Rating – time on reading (TBRS)	N				300
	M (sd)				5.74 (1.86)

Note: FACES = Family and Child Experiences Survey-2006; NC-PK= More-at-Four NC Pre-kindergarten Evaluation; NCEDL= NCEDL 11 State Pre-kindergarten Study; PCERS=Preschool Curriculum Evaluation Research Study. SECCYD=NICHD Study of Early Child Care and Youth Development

Table 5.Testing for Quality Thresholds Using ECERS Total across Studies

		EHS		**FACES**		**FACES**		**NC-PK**		**NCEDL**		**PCERS**		**HSIS**		**HSIS**		**Meta-Analysis**[a]	
				2006		**2009**								**3yr Cohort**		**4yr Cohort**			
		Lower	Higher	Lower	Higher	Lower	Higher	Lower	Higher	Lower	Higher	Lower	Higher	Lower	Higher	Lower	Higher	**Lower**	**Higher**
Language PPVT	B	.12	.98	. 27	6.83*	0.28	2.25+	0.19	0.24	-.11	.02	. -.01	2.00*	-.40*	-.07	12	-.20	**.06**	**1.90****
	(se)	(.23)	(.72)	(.41)	(3.33)	(0.23)	(1.30)	(.21)	(.80)	(.16)	(1.14)	(.20)	(.96)	(.15)	(.44)	(.18)	(.46)	**(.23)**	**(1.32)**
	d	.01	.07	.01	.23	.01	.11	.01	.01	-.01	.00	-.01	.10	-.05	-.01	.02	-.03	**.00**	**.08**
	differ?	ns		L<H*		L<H*		ns		ns		L<H*		ns		ns		**ns**	
Math WJ-AP	B (se)	-.11	2.10*	. 47	.-2.89	0.13	1.00	.39*	0.26	.33	1.68	.03	4.22**	-.42	1.42	.08	-1.84^{+}	**.09**	**.97^{+}**
	d	(.29)	(.79)	(.56)	(4.35)	(0.27)	(1.49)	(.18)	(.72)	(.17)	(1.20)	(.22)	(1.61)	(.39)	(1.10)	(.41)	(.98)	**(.30)**	**(1.74)**
	d	-.01	.12	.02	-.11	.01	.05	.02	.03	.02	.11	.02	.11	-.02	.08	.00	-.10	**.01**	**.04**
	differ?	L<H*		ns		ns		ns		ns		L<H*		L<H*		L<H*		**ns**	
Literacy WJ-LW	B	-.12	.06	.80	6.44	-0.01	-0.41	0.67+	0.93	-.16	-.81	.04	1.02	-.03	1.52	23	-.22	**.17**	**1.22**
	(se)	(.23)	(.43)	(.60)	(4.78)	(0.31)	(1.80)	(.36)	(1.22)	(.19)	(1.31)	(.26)	(1.25)	(.42)	(1.19)	(.40)	(.92)	**(.34)**	**(1.83)**
	d	-.01	.01	.03	.22	-.00	-.02	.05	.07	-.01	-.05	-.01	-.05	-.00	.07	.01	-.01	**.01**	**.03**
	differ?	ns		ns		ns		ns		ns		ns		ns		ns		**ns**	
Behavior Problems	B	-.07*	-.23*	-.11	-.58	-0.01	-0.27	-0.00	-0.56	-.01	.03*	54+	.47					**.06**	**-.12**
	(se)	(.03)	(0.10)	(.23)	(1.92)	(0.08)	(0.46)	(.21)	(.77)	(.01)	(.05)	(.29)	(1.42)					**(.15)**	**(.86)**
	d	-.01	-.05	-.01	-.05	-.00	-.05	.00	-.04	-.02	.05	-.01	.03					**-.01**	**-.01**
	differ?	ns		ns		ns		ns		ns		ns						**ns**	
Social Comp	B	06+	.12	.30	.81	-0.18+	-0.12	0.04	2.10*	.02	.01	-.68+	.33					**-.11**	**.43**
	(se)	(.04)	(.11)	(.20)	(1.66)	(0.09)	(0.53)	(.25)	(.93)	(.01)	(.08)	(.36)	(1.70)					**(.16)**	**(.91)**
	d	.04	.07	.04	.10	-.03	-.02	.00	.13	.02	.01	.02	.01					**.01**	**.04^{+}**
	differ?	ns		ns		ns		L<H*		ns		ns						**ns**	

Note: [a] meta-analysis examined t-statistics computed from the regression coefficients and standard errors and results should be interpreted as represented aggregated partial correlation coefficients The effect sizes for each project was computed as d=B sd(quality)/sd(outcome).

+ .05 <= p < .1; *** p<0.001; ** .001 <= p < .01; * .01 <= p < .05.

Covariates include fall score on same outcome, gender, race, time between fall and spring assessments, whether child speaks English at home, and if relevant site, whether program was a Head Start program and whether the program was located in a school

Table 6: Testing for Quality Thresholds Using ECERS Interactions Scores across Studies

ECERS Interaction		**FACES 2006**		**FACES 2009**		**NC-PK**		**NCEDL**		**PCERS**		**HSIS 3yr Cohort**		**HSIS 4yr Cohort**		**Meta Analysis[a]**	
		Lower	Higher	Lower	Higher	Lower	Higher	Lower	Higher	Lower	Higher	Lower	Higher	Lower	Higher	**Lower**	**Higher**
Language PPVT	B	-.13	1.08	-0.19	0.16	.50+	.11	-.08	.80+	.02	.96+	-.21	.45	.28	-.02	**-.02**	**.62***
	(se)	(.21)	(1.03)	(0.23)	(0.79)	(.17)	(.53)	(.14)	(.41)	(.18)	(.51)	(.20)	(.34)	(.23)	(.40)	**(.19)**	**(.61)**
	d	.00	.04	-.01	.01	.03	.01	.00	.05	.00	.06	-.03	.07	.05	-.00	**.00**	**.04**[*]
	differ?	ns		ns		ns		L<H*		L<H*		L<H*		ns		**L<H***	
Math WJ-AP	B	-.12	-.90	-0.23	-0.43	.29	-.43	-.17	.28	.24	-.43	-.29	1.09	.08	-.71	**-.04**	**-.24**
	(se)	(.27)	(1.32)	(0.27)	(0.91)	(.23)	(.47)	(.15)	(.45)	(.19)	(1.03)	(.57)	(.99)	(.53)	(.74)	**(.27)**	**(.87)**
	d	.00	-.04	-.01	-.03	.02	-.03	-.01	.02	.02	-.03	-.02	.07	.01	-.05	**.00**	**-.01**
	differ?	ns		ns		ns		ns		ns		ns		ns		**Ns**	
Literacy WJ-LW	B	-.39	.53	-0.52+	-0.27	-.54	,99*	-.09	.24	.00	.38	-.30	.21	.52	.29	**-.21**	**.30**
	(se)	(.31)	(1.48)	(0.31)	(1.07)	(.48)	(.87)	(.18)	(.50)	(.23)	(.65)	(.54)	(1.00)	(.47)	(.76)	**(.32)**	**(.90)**
	d	-.01	.02	-.03	-.02	-.04	.07	-.01	.01	.00	.03	-.02	.01	.03	.02	**-.01**	**.02**
	differ?	ns		ns		L<H*		ns		ns		ns		ns		**ns**	
Behavior Problem	B	-.23*	-.69	0.09	0.24	38	-.13	.003	-.0005	-.03	-1.27+					**.01**	**-.41**
	(se)	(.12)	(.55)	(0.09)	(0.28)	(.26)	(.53)	(.006)	(.01)	(.27)	(.76)					**(.15)**	**(.43)**
	d	.02	-.06	.02	.05	-.01	-.05	.01	-.001	.00	-.09					**.01**	**-.03**
	differ?	ns		ns		ns		ns		ns						**ns**	
Social Comp	B	.04	.01	-0.18+	-0.02	-.50	1.45*	-.002	.004	.13	2.75**					**-.06**	**.83**
	(se)	(.10)	(.49)	(0.09)	(0.32)	(.32)	(.64)	(.01)	(.02)	(.32)	(.91)					**(.17)**	**(.50)**
	d	.01	.00	-.04	-.00	-.03	.09	.00	.00	.01	.18					**-.01**	**.05**
	dif	ns		ns		L<H*		ns		L<H*						**L<H***	

Note: [a] meta-analysis examined t-statistics computed from the regression coefficients and standard errors and results should be interpreted as represented aggregated partial correlation coefficients The effect sizes for each project was computed as d=B sd(quality)/sd(outcome). + .05 <= p < .1; *** p<0.001; ** .001 <= p < .01; * .01 <= p < .05.

Covariates include fall score on same outcome, gender, race, time between fall and spring assessments, whether child speaks English at home, and if relevant site, whether program was a Head Start program and whether the program was located in a school

Table 7: Testing for Quality Thresholds Using CLASS Interaction-specific Quality Scores in Several Studies

		FACES 2006		**FACES 2009**		**NC-PK**		**NCEDL**		**MTP**		**Meta-analysis[a]**	
		Lower	**Higher**	**Lower**	**Higher**	**Lower**	**Higher**	**Lower**	**Higher**	**Lower**	**Higher**	**Lower**	**Higher**
CLASS Emotional Support													
Problem Behavior	B			-0.02	-0.41	-0.23	1.50	.004	-.03	-.01	-.10	**-.05**	**.11**
	(se)			(0.07)	(0.44)	(.53)	(1.75)	(.005)	(.02)	(0.01)	(0.10)	**(.12)**	**(.47)**
	d			-.00	-.05	-.02	.10	.01	-.04	-.01	-.12	**-.00**	**-.01**
	differ?			ns		ns		ns		L>H*		**ns**	
Social Skills	B			0.03	0.51	-0.85	-2.10	-.01	.05^{+}	. 04*	.19*	**-.14**	**.18**
	(se)			(0.08)	(0.50)	(.62)	(2.04)	(.008)	(.03)	(0.01)	(0.14)	**(.14)**	**(.55)**
	d			.00	.06	-.05	-.12	-.01	.04	.03	.15	**-.01**	**.03**
	differ?			ns		ns		L<H*		L<H*		**ns**	
CLASS Instructional Support													
Language: PPVT	B	-.65	2.95	0.19	2.30	-0.78	-0.86	.13	1.44*			**-.20**	**1.76*****
	(se)	(0.50)	(2.03)	(0.48)	(1.53)	(.55)	(.90)	(.27)	(.65)			**(.43)**	**(1.30)**
	d	-.02	.09	.01	.09	-.04	-.05	.01	.08			**-.01**	**.07**
	differ?	L<H*		L<H*		ns		L<H*				**L<H***	
Math: WJ-AP	B	-1.03	-.37	0.00	-0.47	-0.11	0.43	-.32	.07			**-.40**	**-.14**
	(se)	0.64)	(2.61)	(0.57)	(1.78)	(.60)	(.96)	(.29)	(.71)			**(.50)**	**(1.55)**
	d	-.04	-.01	.00	-.02	-.01	.03	-.03	-.02			**-.02**	**-.01**
	differ?	ns		ns		ns		ns				**Ns**	
Literacy: WJ-LW	B	-.88	2.71	0.53	2.34	-1.19^{+}	0.46	-.16	1.35+			**-.33**	**1.88****
	(se)	(0.72)	(2.93)	(0.64)	(1.97)	(.59)	(.95)	(.37)	(.75)			**(.57)**	**(1.70)**
	d	-.03	.09	.02	.09	-.09	.04	-.01	.08			**-.02**	**.08**
	differ?	ns		ns		L<H*		L<H*				**L<H***	
Reading:Topel Phon Awareness	B									-.0.03	1.05		
	(se)									(0.11)	(0.58)		
	d									-.01	.18		
	differ?									L<H*			

Note: [a] meta-analysis examined t-statistics computed from the regression coefficients and standard errors and results should be interpreted as represented aggregated partial correlation coefficients The effect sizes for each project was computed as d=B sd(quality)/sd(outcome).+ .05 <= p < .1; *** p<0.001; ** .001 <= p < .01; * .01 <= p < .05.

Covariates include fall score on same outcome, gender, race, time between fall and spring assessments, whether child speaks English at home, and if relevant site, whether program was a Head Start program and whether the program was located in a school

[a] meta-analysis examined t-statistics computed from the regression coefficients and standard errors and results should be interpreted as represented aggregated partial correlation coefficients The effect sizes for each project was computed as d=B sd(quality)/sd(outcome).

Table 8: Testing for Quality Thresholds Using Domain-specific Quality Measures in Several Studies

		NC-PK		**PCER**			
		ELLCO Literacy		**TBRS Literacy**		**TBRS Numeracy**	
		Lower	**Higher**	**Lower**	**Higher**	**Lower**	**Higher**
Language: PPVT	B(se)	-0.53 (.43)	-6.05^{+} (3.39)	.78(.51)	4.20*(1.98)		
	D	-0.02	-0.18	.03	.17		
	Differ?	ns		L<H*			
Math: WJ-AP	B(se)					1.06*(.53)	3.12(2.17)
	d					.06	.15
	differ?					ns	
Literacy: WJ-LW	B(se)	1.14 (.77)	15.95^{+} (8.04)	1.96**(.59)	13.96***(2.39)		
	d	0.05	0.10	.08	.60		
	differ?	ns		L<H*			

+ .05 <= p < .1; * p<.05; ** p<.01; *** p<0.001

Covariates include fall score on same outcome, gender, race, time between fall and spring assessments, whether child speaks English at home, and if relevant site, whether program was a Head Start program and whether the program was located in a school

Table 9. B-spline analyses: Evidence for nonlinear associations

	FACES 06 **Nonlinear F**	**FACES 09** **Nonlinear F**	**NCEDL** **Nonlinear F**	**NC-PK** **Nonlinear F**	**PCER** **Nonlinear F**	**HSIS-3 year olds** **Nonlinear F**
ECERS Interactions						
PPVT (language)			3.22**		2.43*	2.03+
Social Competence				5.72***	9.42***	
CLASS Instructional Support						
PPVT (language)	2.55*	1.23	2.48*			
WJ-LW (literacy)			0.95	2.82*		
TBRS Literacy						
PPVT (language)					1.78^{+}	
WJ LW (literacy)					8.06***	

Note: $^{+}$ .1 < p < .05; * p < .05; ** p <.01; *** p <.001

Covariates include site, fall score, time since fall testing type of program, maternal education, child sex, race, home language

Table 10. Effect Sizes from Analyses that Include Both Global and Specific Quality Measures

		Global & Domain-Specific Quality Measures				Global & Interactions-Specific Quality Measures							
		PCERS		NC-PK		NC-PK		NCEDL		FACES 06		NCEDL	
		ECERS	TBRS	ECERS	ELLCO Literacy Activities	ECERS	CLASS Instruct Support	ECERS	CLASS Instruct Support	ECERS	CLASS Instruct Support	ECERS	CLASS Emotional Support
PPVT	d-low	.00	.02							.01	-.02		
	d-high	.12	.13							.16	.09		
	d-linear			-.04	.02	-.02	.02	.02	.04**				
WJ-AP	d-linear	.01	.05*	-.11***	.07*	-.05	.04	-.12**	.10***	-.04	.01		
WJ-LW	d-low		.10***						.00				
	d-high		.69***						.09*				
	d-linear	-.07		.03	.13*	.01	.05*	-.03		.03	.01		
Behavior Problems	d-linear											.03	-.04*
Social Competence	d-linear											-.02	.05*

Note: + .05 <= p < .1; *** p<0.001; ** .001 <= p < .01; * .01 <= p < .05. Listed are the effect sizes for the quality measures. The piecewise model was fit if slopes for quality in higher and lower quality classroom were significantly different in prior analyses. Otherwise, the quality measure entered these models as a linear predictor.

Table 11. Dosage: Comparing children with one or two years of Head Start in propensity-score matched samples: FACES 2006, FACES 2009, and HSIS

	FACES 2006				**FACES 2009**				**HSIS**			
	Head Start Exit		**Spring Kindergarten**		**Head Start Exit**		**Spring Kindergarten**		**Head Start Exit**		**Spring Kindergarten**	
	B (se)	d	B (se)	d	B (se)	d	B (se)	d	B (se)	d	B (se)	d
Language: PPVT[a]	3.61*** (0.96)	0.16	2.77** (0.85)	0.15	3.49*** (.85)	0.17	1.88* (0.92)	0.10	0.07 (0.55)	0.00	1.15+ (0.63)	0.08
Literacy: WJ Letter Word[a]	6.00** (1.87)	0.14	2.74 (1.96)	0.08	3.61* (1.48)	0.11	3.94* (1.79)	0.10	2.36** (0.85)	0.16	2.06* (1.05)	0.14
Math: WJ Applied Problems[a]	1.49 (1.25)	0.06	1.64 (1.53)	0.07	0.30 (0.48)	0.08	-0.02 (0.49)	-0.00	0.28 (0.76)	0.02	1.04 (1.02)	0.07
Social skills[b]	0.44 (0.31)	0.08	0.03 (0.37)	0.01	-0.22 (0.36)	-0.03	-0.34 (0.47)	-0.04				
Behavior problems[b]	-0.37 (0.38)	-0.06	-0.21 (0.59)	-0.02	-0.22 (0.36)	-0.03	-0.34 (0.47)	-0.04	-0.22 (0.65)	-0.01	-0.39 (0.61)	-0.03

Note: $^{+}$ .05 < p < .1; * p<.05; ** p<.01; *** p<.001

Estimates are weighted. Covariates include corresponding pretest score, child age, gender, race/ethnicity, household language, poverty ratio, maternal education, employment, and depressive symptoms, household mobility, and neighborhood safety. Missing data were handled using multiple imputation (N = 10). Children in Head Start for two years were matched with those in Head Start for one year based on propensity scores estimated from baseline characteristics and scores (nearest neighbor matching with replacement).

[a] Controlled for baseline (fall of first year) standard scores on corresponding measures.

[b] Did not control for baseline scores.

Table 12. Dosage: Associations between child outcomes and absences, hours/week of care, and instruction time

		Absences			Hours/Week		Time in Reading Instruction		Time in Math Instruction	
		FACES 2006- teacher report	**FACES 2006- parent report**	**NC-PK- class records**	**FACES 2006**	**NCEDL**	**NCEDL**	**PCER**	**NCEDL**	**PCER**
Language: PPVT	B	-.44+	-.27	-.35*	-.00	-0.01				
	(se)	(.25)	(.28)	(.21)	(.03)	(0.02)				
	d	-.03	-.01	-.07	-.00	-.01				
Math WJ AP	B	-1.16***	-.89*	-.53+	-.00	-0.01			8.21*	1.35**
	(se)	(.32)	(.35)	(.30)	(.04)	(0.02)			(3.26)	(0.51)
	D	-.07	-.05	-.04	-.00	-.01			.04	.07
Literacy: WJ LW	B	-1.01**	-.38	-.38	09+	-0.02	7.48*	.80***		
	(se)	(.35)	(.39)	(.48)	(.05)	(0.02)	(2.81)	(.17)		
	D	-.06	-.02	-.03	.05	-.01	.04	.11		
Social Skills	B	-.11	-.01	-.14	-.02+	-0.02				
	(se)	(.08)	(.08)	(.33)	(.01)	(0.01)				
	d	--.02	-.00	-.01	-.04	-.03				
Behavior Problems	B	.23*	.067	-.64+	.03+	0.01				
	(se)	(.10)	(.10)	(.34)	(.02)	(0.01)				
	d	.04	.01	-.05	.04	.02				

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

The coefficient reported in the table is the coefficient for absences in the model in which the quality variable accounted for the most variance. Separate analyses examined each quality measure as a covariates – using the best model from the threshold analysis for that study using that outcome and quality measure. In no case was a different inference draw in the analyses that involved the other quality variables. The other covariates include fall score on same outcome, gender, race, time between fall and spring assessments, whether child speaks English at home, and if relevant site, whether program was a Head Start program and whether the program was located in a school.

Quality measures included ECERS-R Total and Interaction in FACES, NC-PK, NCEDL, and PCER; CLASS Instructional Support in FACES 2006 & 2009, NC-PK, and NCEDL, CLASS Emotional Support in NC-PK & NCEDL, and TBRS in PCER

Coefficients reported analyses involving all classroom quality analyses – reporting the largest and smallest coefficients from those analyses.

Control variables include gender, race, maternal education, whether or not below poverty line, Head Start, whether program was located in a public school, state and fall assessment

Table 13. Testing dosage threshold: Comparing FACES 2009 children with 2 years of high quality care and less than 2 years of high quality care using propensity score matching[ab]

	FACES 2009[a]		**HSIS**[b]	
	Head Start Exit B (se)	**Spring Kindergarten B (se)**	**Head Start Exit B (se)**	**Spring Kindergarten B (se)**
Language: PPVT	0.95 (1.51)	0.67 (1.60)	0.37 (0.85)	0.55 (0.066)
Literacy WJ Letter Word	0.41 (1.55)	1.53 (1.50)	0.02 (1.15)	0.80 (1.16)
Math WJ Applied Problems	0.06 (1.55)	-0.27 (1.56)	-0.92 (1.04)	2.18^{+} (1.27)
Behavior Problems	-0.44 (0.46)	0.00 (0.78)		
Social skills	0.76 (0.51)	-0.08 (0.70)		

$^{+}$ .1 < p < .05
Matched on: child age, gender, race, disability, household language, income to needs ratio, maternal education, maternal depression, maternal employment, single-parent family, household mobility, neighborhood safety, and baseline scores.

[a] FACES: High quality classroom is defined having scores of greater than 2 on Instructional Support, 5 on Emotional Support, and 4 on Classroom Organization. Sample sizes range n=250-258 for the group with 2 years of high quality and n=410-419 for the group with less than 2 years of high quality

[b] HSIS: High quality classroom is defined as having score above 4.5 on ECERS-R total score. There were n=361 with two years of high quality and n=177 for the group with less than 2 years of high quality in HSIS.

Table 14. Summary of Findings

	Language	Literacy	Math	Social skills	Problem Behaviors
Quality Thresholds					
ECERS-R Total	ns	ns	ns	ns	ns
ECES-R Interactions	L<H	ns	ns	L<H	ns
CLASS Emotional Support				ns	ns
CLASS Instructional Support	L<H	L<H	ns		
ELLCO Literacy Activities	ns	ns			
TBRS – Literacy	L<H[a]	L<H[a]			
TBRS- Math			ns		
Quality Feature					
Global v **I**nteraction-Specific	I>G	I>G	I>G	I>G	I>G
Global v **D**omain-Specific	D>G	D>G	D>G		
Dosage					
1 v 2 years of Head Start	2>1	2 >1	2>1[a]	ns	ns
Absences	ns	Neg[a]	Neg[a]		
Hours/Week	ns	ns	ns	ns	Ns
Instruction time in content area		Pos	Pos		
1 v 2 years high quality care	ns	ns	ns	ns	Ns
Dose x quality interactions			Instruction quality x time		

Note: L<H indicates significant difference in slopes from higher and lower quality classrooms in meta-analysis
I>G indicates that the interaction-specific quality measure contributed more than global quality measure
D>G indicates that the doamin-specific quality measure contributed more than global quality measure
2 >1 indicates children with 2 years of HS had higher scores than children with one year
Neg indicates a significant negative association
Pos indicates a significant positive association

[a] finding was significant in a single study, but not in more than one study

Figure Captions

Figure 1. Hypothesized threshold in effects of child care quality on child outcomes

Figure 2. Effect sizes from meta-analysis of academic outcomes using Class Instructional Support in higher and lower quality classroom. Note: * indicates that quality coefficients for high and low quality classrooms are reliably different in the meta-analysis.

Figure 3. Effect sizes from academic outcomes using TBRS Instructional Quality in higher and lower quality classroom in the PCER. Note: * indicates that quality coefficients for high and low quality classrooms are reliably different in the meta-analysis.

Figure 4. LOESS plot: Looking for threshold in ECERS Interactions scores for the 3-year-old Cohort in HSIS.

Figure 1. Hypothesized threshold in effects of child care quality on child outcomes

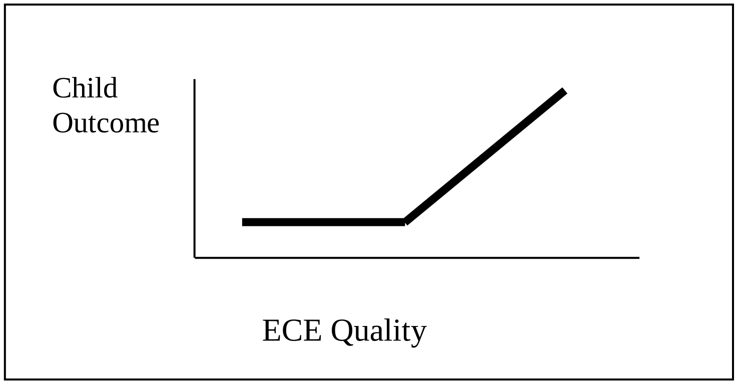
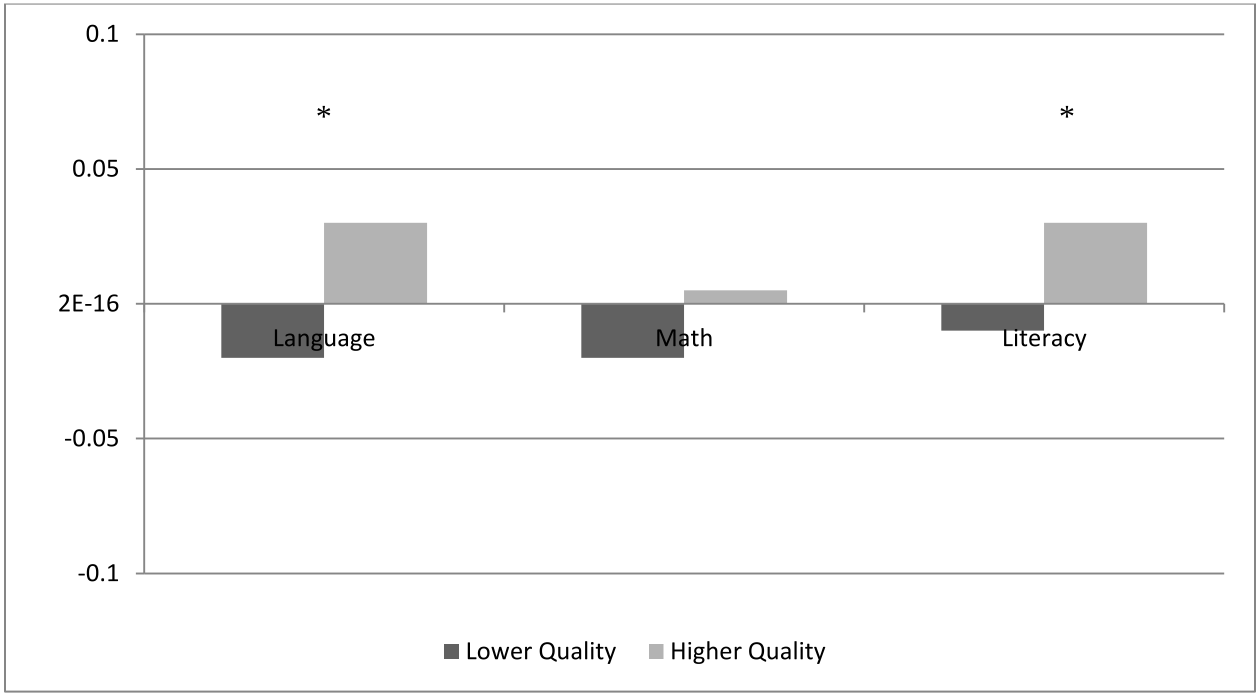
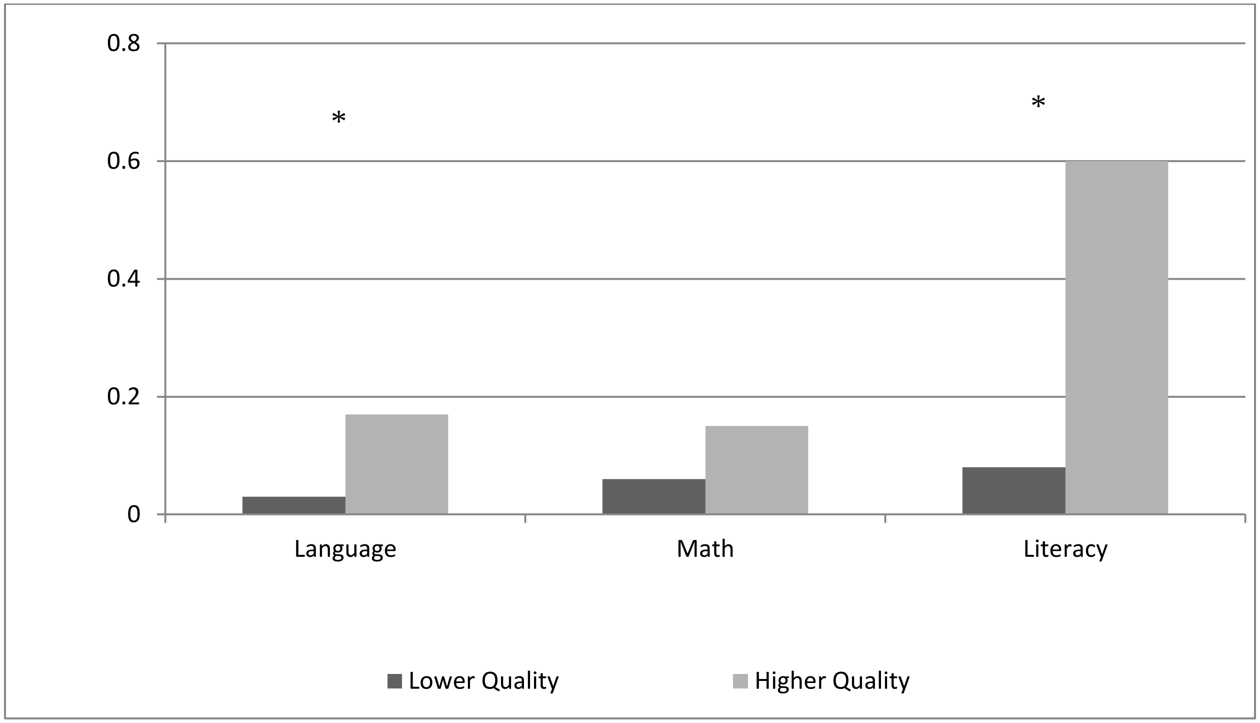


Figure 2. Effect sizes from meta-analysis of academic outcomes using Class Instructional Support in higher and lower quality classroom



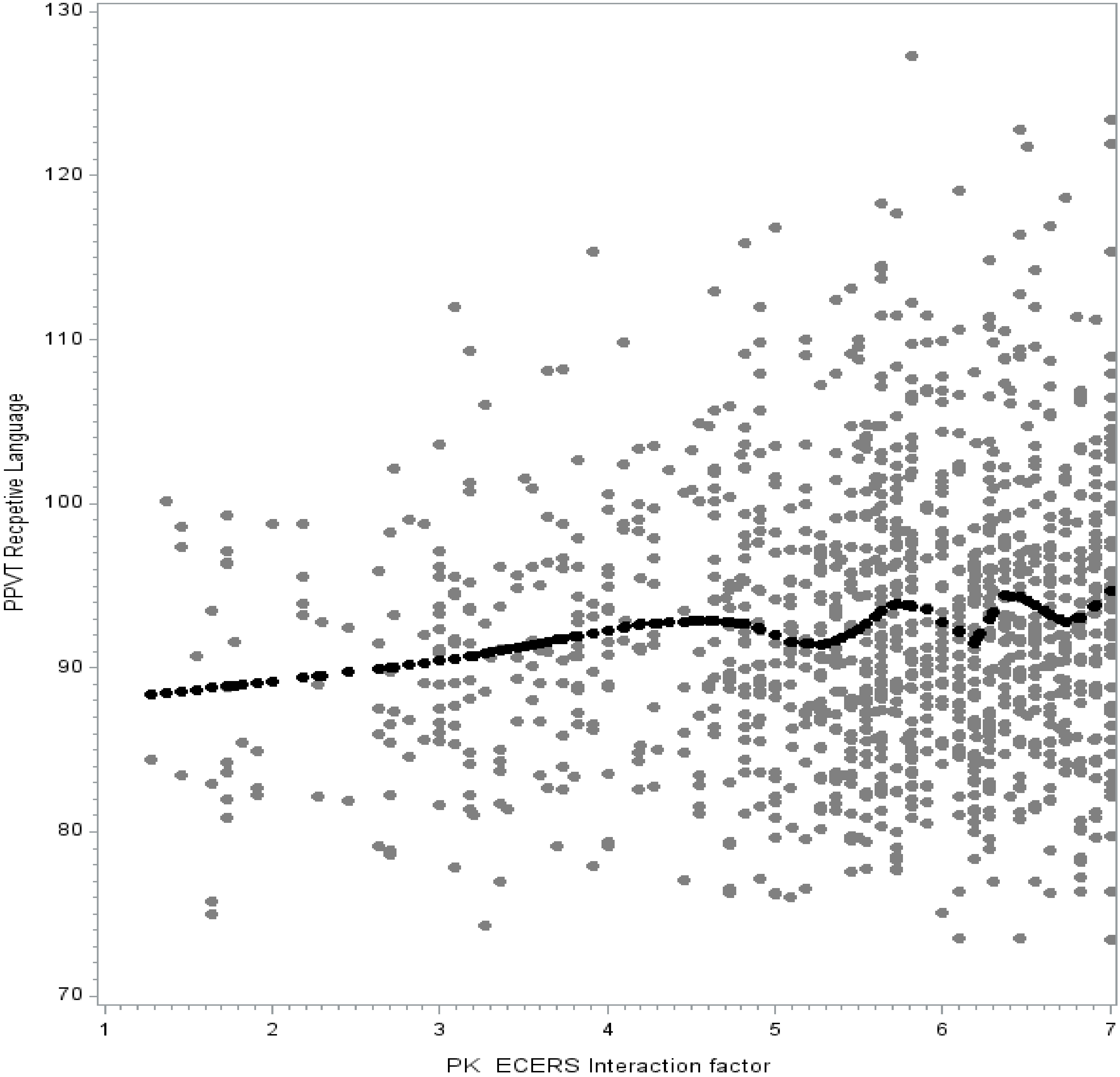
* quality coefficients reliably different from each other in meta-analysis

Figure 3. Effect sizes from academic outcomes using TBRS Instructional Quality in higher and lower quality classroom in the PCER



* quality coefficients reliably different in higher and lower quality classrooms

Figure 4. LOESS plot: Looking for threshold in ECERS Interactions scores for the 3-year-old Cohort in HSIS.



Appendix A

Table 1a. Comparing Child and Family Characteristics in Unmatched and Matched FACES Samples

		FACES 2006							FACES 2009						
		Before Matching			After Matching			Not Matched	Before Matching			After Matching			Not Matched
		2 years in Head Start	1 year in Head Start	t-test	2 years in Head Start	1 year in Head Start	t-test	1 year in Head Start	2 years in Head Start	1 year in Head Start	t-test	2 years in Head Start	1 year in Head Start	t-test	1 year in Head Start
Gender (Boy)	%	51.38	53.46		51.05	56.32		51.62	51.38	47.96		51.31	47.78		47.64
Race/Ethnicity															
European-Am	%	22.17	27.32	*	22.42	24.79		26.93	22.77	22.77		22.80	20.15		24.43
Black	%	43.54	28.44	***	42.83	46.60		20.32	43.87	35.22	***	43.79	42.22		30.00
Hispanic	%	25.19	35.00	***	25.62	20.14		43.42	24.55	34.61	***	24.59	28.91		38.82
Asian	%	1.04	1.71		1.06	0.81		2.14	1.76	1.86		1.76	2.25		1.67
Other	%	7.99	7.53		8.01	8.14		6.50	7.05	5.45		7.06	6.06		4.99
Child with a disability diagnosis	%	5.74	4.75		5.70	5.61		4..20	3.24	2.62		3.24	3.66		2.17
Household language (not English)		13.71	28.02	***	13.93	12.41		37.91	14.82	22.62	***	14.84	17.31		26.62
Poverty ratio		2.74	2.78		2.74	2.71		2.84	2.58	2.56		2.57	2.56		2.55
Maternal education															
< high school	%	30.78	35.70	*	30.78	30.03		39.47	25.64	33.74	***	25.68	27.11		38.07
High school/GED	%	32.76	34.22		32.41	37.50		30.97	39.77	36.92		39.81	40.70		34.90
Some college	%	29.96	24.54	*	30.42	27.94		23.76	28.34	23.98		28.27	25.58		22.66
BA+	%	6.82	6.28		6.73	5.89		6.33	6.59	5.58		6.58	6.72		4.46
Maternal employment	%	56.89	50.86	*	56.75	57.35		47.31	52.80	52.02		52.81	50.95		52.15
Maternal depressive symptoms	%	5.76	5.09	*	5.67	6.08		4.49	5.05	4.93		5.06	5.00		4.89
Single parent	%	48.62	48.33		48.89	51.22		46.85	54.20	50.71		54.23	50.18		51.42
Household mobility	%	22.85	23.44		22.72	19.37		24.74	40.53	50.32**		40.59	37.00		60.04

		FACES 2006							FACES 2009						
		Before Matching			After Matching			Not Matched	Before Matching			After Matching			Not Matched
		2 years in Head Start	1 year in Head Start	t-test	2 years in Head Start	1 year in Head Start	t-test	1 year in Head Start	2 years in Head Start	1 year in Head Start	t-test	2 years in Head Start	1 year in Head Start	t-test	1 year in Head Start
Neighborhood safety	M	12.90	12.52		12.52	14.04		12.10	10.48	10.78		10.49	9.49		11.67
Pretest scores															
PPVT-4 standard score	M	85.55	82.87	***	85.54	85.47		81.08	86.69	85.52		86.70	85.49		85.48
WJ III Letter Word Identification	M	93.55	92.22		93.52	94.28		91.31	96.64	94.21**		96.57	96.00		93.09
WJ III Spelling	M	98.35	90.45	***	97.82	96.50		87.54	95.18	94.52		95.15	94.51		94.45
WJ III Applied Problems	M	92.20	83.99	***	91.69	90.63		80.55	87.26	87.13		87.26	86.09		87.71
Sample size		822-868	759-810		809-854	377-404		383-406	737-796	741-809		736-795	391-433		349-376

Note: Children enrolled in Head Start for two years were matched to those enrolled in Head Start for one year based on propensity scores using nearest neighbor matching with replacement.

*p < .05; **p < .01; ***p < .001.

Table 1b. Descriptive statistics for the matched and unmatched child samples: HSIS

		HSIS						
		Before Matching			After Matching			Not Matched
		2 years in Head Start	1 year in Head Start	t-test	2 years in Head Start	1 year in Head Start	t-test	1 year in Head Start
Gender (Boy)	%	48.57	50.97		48.88	50.84		48.49
Race/Ethnicity								
European American	%	31.16	35.14		29.69	29.55		33.48
Black	%	35.78	19.69	***	36.28	36.56		32.92
Hispanic	%	33.06	45.17	***	34.03	33.89		33.59
Child Disability	%	14.56	13.64		14.57	13.17		10.94
Household language (not English)	%	26.80	37.71	***	24.93	24.09		26.28
Maternal education								
Less than high school	%	33.61	42.21	***	33.61	31.23		36.89
High school/GED	%	34.56	31.40		34.31	35.71		33.42
Beyond high school	%	31.84	26.38	*	32.07	33.05		29.69
Maternal depressive symptoms	%	19.86	21.75		20.44	21.57		19.25
Mother teenaged at birth	%	13.61	17.63	*	14.01	13.59		16.02*
Mother married	%	44.08	45.69		42.72	43.42		45.65
Both biological parents in HH	%	50.34	52.77		49.16	49.30		49.55
Mother recent immigrant	%	14.83	23.81	***	14.15	14.43		17.13*
Urban location	%	82.86	85.07		82.35	81.93		82.92
Pretest scores								
PPVT-3	M	91.40	90.85		91.40	91.78		91.83
WJ III Letter Word ID	M	91.52	92.28		91.52	91.94		91.20
WJ III Applied Problems s	M	90.11	90.02		90.11	89.67		89.69
Sample size	N	807	918		777	735		141

Mother teenaged at birth	13.61	17.63*		14.01	13.59	16.02*
Mother married	44.08	45.69		42.72	43.42	45.65
Biological parents live together	50.34	52.77		49.16	49.30	49.55
Mother recent immigrant	14.83	23.81***		14.15	14.43	17.13*
Urban location	82.86	85.07		82.35	81.93	82.92
Pretest scores						
PPVT-3 standard score	91.40	90.85		91.40	91.78	91.83
WJ III Letter Word Identification standard score	91.52	92.28		91.52	91.94	91.20
WJ III Applied Problems standard score	90.11	90.02		90.11	89.67	89.69
Sample size	807	918		777	735	141

Note: Children enrolled in Head Start for two years were matched to those enrolled in Head Start for one year based on propensity scores using nearest neighbor matching with replacement.

*p < .05; **p < .01; ***p < .001.