THE RELATIONSHIP BETWEEN MEASURES OF INFANT-TODDLER CHILD CARE QUALITY AND CHILDREN'S DEVELOPMENTAL OUTCOMES: AN ANALYSIS OF DATA FROM THE EARLY HEAD START RESEARCH AND EVALUATION PROJECT

A Thesis
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Master of Public Policy
in Public Policy

By

Christopher H. Jones, B.A.

Washington, DC
April 10, 2015

THE RELATIONSHIP BETWEEN MEASURES OF INFANT-TODDLER CHILD CARE QUALITY AND CHILDREN'S DEVELOPMENTAL OUTCOMES: AN ANALYSIS OF DATA FROM THE EARLY HEAD START RESEARCH AND EVALUATION PROJECT

Christopher H. Jones, B.A.

Thesis Advisor: Donna Ruane Morrison, Ph.D.

ABSTRACT

Several first-rate early care and education programs have led to substantive, long-term gains in outcomes for children from disadvantaged families. Yet in other settings, widely used instruments for rating the quality of care show only modest associations with child outcomes. Using data from the Early Head Start Research and Evaluation Project, the present study extends previous work by using multiple measures of child care quality and by examining additional cognitive, academic, and social-emotional outcomes from the evaluation's long-term follow-up, when participating children were in grade 5. Measures of quality of care had significant, favorable relationships with a measure of language development at 36 months and an index of success on cognitive/academic outcomes at grade 5. Relationships with the other outcomes were not significant, but several were favorable with effect sizes that were small but on par with findings from other research. Alternative measures of quality did not generally show stronger relationships with outcomes, except for a subscale of one measure, which focused on the physical features and learning activities in the care setting. Additional specifications of the relationship between quality and outcomes did not find consistent evidence for an alternative to a linear model. Results should be interpreted with caution due to potential selection effects. Future research should explore the extent to which the findings from this study and others occur because there is only a small relationship between quality and outcomes, compared to the extent to which they occur because existing instruments do not accurately measure quality.

TABLE OF CONTENTS

Chapter One: Introduction 1

Chapter Two: Prior Literature 4

Chapter Three: Data and Methods 13

Chapter Four: Results 30

Chapter Five: Discussion and Conclusion 46

Tables 56

Appendix 85

References 90

CHAPTER ONE

INTRODUCTION

The first five years of a child's life, and their experiences during these years, are crucial for their development (Shonkoff & Phillips, 2000). Research shows that gaps in achievement emerge as early as preschool among disadvantaged, at-risk children and they continue to grow as children progress through school. At the same time, a majority of young children receive some early care and education outside the home or from people other than their parents (Laughlin, 2013). This has led to a larger focus from policymakers on early care and education as means for giving vulnerable children access to enrichment experiences. For example, several states have instituted or expanded public pre-kindergarten programs to achieve this goal. In addition to supporting access to early care and its affordability, a wide variety of programs and policies are targeted towards improving the *quality* of care provided in early care and education settings so they can support children's development (Boller, Tarrant, & Schaack, 2014).

Defining what constitutes quality in the context of early education and care and devising a way to quantify those dimensions is a difficult undertaking. More challenging still, is determining whether care rated as high-quality based on these assessment measures leads to improved child outcomes. As discussed in more detail in the next chapter, several programs designed to provide a high-quality educational experience have shown large, long-term benefits to children, yet other programs show relatively weak or short-lived impacts on child outcomes. Furthermore, other types of studies that identify high-quality care based on favorable ratings on quality assessment measures find positive but weak effects on child outcomes. While this may not be surprising if the quality assessment ratings were based on proxies of important dimensions of care – for example, if the teacher-staff ratio were used as an indicator of the level of administrative support that teachers enjoy – it also applies to direct observations of aspects of the overall classroom environment, such as the activities available to children or interactions

between caregivers and children. While there are several potential reasons for this weak relationship, policies that emphasize improving quality using these measures may not improve child outcomes as much as policymakers expect. Further research on quality measures and their relationships with child outcomes is needed to understand why these patterns have been observed.

This study aims to address this issue by conducting additional analyses of the relationship between observational measures of quality and child outcomes using data from the Early Head Start Research and Evaluation Project (EHSREP). While primarily an experimental evaluation of Early Head Start, one of the original evaluation's reports included a brief nonexperimental analysis of the relationship between child care quality and key short-term developmental outcomes (ACF, 2004). Since that nonexperimental analysis, the EHSREP has published a follow-up to the experimental evaluation based on data collected when the children were in fifth grade (Vogel et al., 2010).

The current study will extend the original nonexperimental analysis using data from the grade 5 follow-up to examine whether there are effects of quality of care when longer-term outcomes of the evaluation are examined. It will also use different measures of quality that the original study collected to explore whether the relationship between quality of care and child outcomes changes when different measures of quality are used. Finally, it will use alternative specifications of the relationship between quality of care and outcomes to see if models that better fit the relationship can be found.

Results from this study will provide additional insight into the use of observational measures to assess the quality of early care and education. Finding an effect on long-term outcomes would suggest these measures are capturing an important element of quality with persistent effects on outcomes. Finding different effects using different quality measures would provide evidence that some measures are better suited to capturing the aspects of quality that matter most in improving

child outcomes. Findings from additional analyses can provide insight about the nature of the relationship between quality measures and outcomes, such as whether it is linear. This study can guide the use and development of measures of quality of care by researchers and policymakers to use in supporting better child outcomes.

Chapter 2 reviews previous research on measuring quality in early care and education settings and how measures of quality are related to child outcomes. Chapter 3 describes the EHSREP data and the methodology used by the present analysis. Chapter 4 examines the results of the analysis, which are also presented in a series of tables located after the main text. Chapter 5 discusses the interpretation and implications of the results, and concludes. A brief appendix contains some additional details not included in the main text.

CHAPTER TWO

PRIOR LITERATURE

Over the past few decades, the relationship between quality in early care and education settings and subsequent child developmental outcomes has been the subject of extensive study. This chapter reviews this literature, beginning with evaluations of interventions to provide high quality care, moving to efforts to measure quality of care in various settings and assess their effect on children's development, and then examining several specific aspects of these studies. While patterns regarding the overall relationship between quality of care and outcomes have emerged, there are many inconsistencies and unanswered questions, not only about the relationship between quality and outcomes, but about the measures of quality themselves.

Evaluations of early care and education interventions. Much of the early care and education literature examines the effect of programs designed to provide high-quality experiences to the children and families who participate in them. Research on three programs serving at-risk children – the Perry Preschool program, the Carolina Abecedarian Project, and the Chicago Child-Parent Centers – have found large beneficial impacts for participants not only on short-term developmental outcomes but also important long-term adult outcomes such as educational attainment, employment, and earnings (Schweinhart et al., 2005; Campbell et al., 2002; Reynolds et al., 2007). Other programs, such as the Infant Health and Development Program and evaluations of Head Start and Early Head Start, have found overall impacts that were not sustained over time, although in some cases certain impacts (such as for sub-groups) continued (McCormick et al., 2006; Puma et al., 2012; Vogel et al., 2010). Recently, as several states have instituted or expanded pre-kindergarten programs, initial evaluations have found impacts on short-term outcomes, although these range in magnitude (Gormley et al., 2005; Wong et al., 2008; Weiland & Yoshikawa, 2013).

These studies provide evidence that high-quality programs can improve outcomes for disadvantaged children, although several factors should be kept in mind when interpreting these results. For example, the Perry and Abecedarian programs served a small number of participants at a high per-participant cost, which might make them difficult to scale up as larger programs (although the Chicago Child-Parent Centers operated on a larger scale). It may be potentially more difficult for more recent programs to show large effects because the general use of early care and education services has increased over time, giving members of control or comparison groups more options when they are not able to participate in the program. Methodologically, these evaluations tend to provide strong evidence because they use random assignment or quasi-experimental methods that provide a relatively high degree of confidence that the results observed were caused by the program studied and not from other factors. Overall, these evaluations demonstrate the potential for high-quality early care and education to benefit children.

Measuring quality and estimating its effect on child outcomes. While the interventions described above were designed to provide high-quality services, and positive evaluation results suggest this was the case, it is difficult to quantify the actual quality of the care provided. Additionally, outside of specific interventions, children are receiving care from a wide variety of settings. As a result, another area of the literature seeks to define and measure quality in early care and education settings, and then study its relationship with child outcomes.

One approach is to measure quality using structural features of the program, which can range from the ratio of children to caregivers or the educational attainment of the caregivers, to meeting recommended standards in various program areas. Conceptually, these are predicted to lead to higher quality care. For example, a lower child-caregiver ratio means caregivers can devote more time and attention to interacting with each child. Caregivers with higher educational

attainment may have more skills and knowledge about how to appropriately provide care. These are relatively easy to measure, but also represent less direct measures of quality.

Another approach measures quality based on the environment and processes directly experienced by children. Examples include caregiver-child interactions and use of learning activities. These aim to more directly assess the quality of care based on what developmental research has identified as supporting child development. For example, caregivers who talk with children more frequently stimulate more active thinking in the child, as well as provide reassurance during stressful situations. However, these are more difficult to measure and require direct observation of the child care environment.

Several observational measures of the quality of out-of-home care settings have been developed. The most frequently used are a set of rating scales developed to assess the global quality of the care environment: the Early Childhood Environment Rating Scale and its revised version (ECERS and ECERS-R; Harms & Clifford, 1980; Harms, Clifford, & Cryer, 1998), the Infant Toddler Environment Rating Scale and its revised version (ITERS and ITERS-R; Harms, Cryer, & Clifford, 1990; Harms, Cryer, & Clifford, 2003), and the Family Day Care Rating Scale and its revised version (FDCRS and FCCERS-R; Harms & Clifford, 1989; Harms, Cryer & Clifford, 2007). The ECERS is intended for use in center-based settings for children ages 2.5 to 5, the ITERS is intended for younger children in center-based settings, and the FDCRS is intended for children in family child care settings. Each scale contains around 30 to 40 items describing various aspects of environment in the setting, which are then combined into several subscales and an overall scale. The items and subscales are selected because they theoretically represent aspects of high-quality care. They differ depending on the scale; for example, the ITERS subscales are furnishings and display for children, personal care routines, listening and talking, learning activities, interaction, program structure, and adult needs. In all of them, a trained observer visits the setting and records a rating for each item during the visit.

Other measures have been developed that focus on different aspects of quality than the environment rating scales. For example, the Arnett Caregiver Interaction Scale characterizes the nature of the interaction between the caregiver and the children under their care (Arnett, 1989). The Classroom Assessment Scoring System (CLASS) was developed to specifically assess the quality of interaction between caregivers and children, especially regarding support for instruction and learning for pre-academic skills (Pianta, LaParo, & Hamre, 2008). These generally follow the same measurement process: a trained observer records data on several standardized items during a visit to the setting, and data on the items are combined into an overall score.

These measures of quality have been widely used in many different studies. While structural features of care settings can affect child outcomes, this tends to occur indirectly through changes in the environment experienced by children (Love, Schochet, & Meckstroth, 1996; Zaslow et al., 2010). Structural features do not seem to have a direct relationship with child outcomes (Blau 1999; Mashburn et al., 2008). More studies have focused on relationships between quality as measured by observations of care settings and child outcomes; these tend to find relationships that are positive but weak in magnitude (Burchinal, Kainz, & Cai, 2011; Keys et al., 2013; Weiland et al., 2013). In particular, Burchinal, Kainz, and Cai (2011) conducted a meta-analysis of findings from 20 published studies and also analyzed data on low-income children from four existing studies. They found partial correlations ranged from 0.05 to 0.17 from the meta-analysis and were mostly less than 0.10 in the secondary analysis of data.

Methodology. A key challenge to interpreting results of most studies of the relationship between quality of care and child outcomes is that they use an observational methodology. Unlike programs, where it is easier to use random assignment or other quasi-experimental means to compare groups who do and do not receive the intervention, it would be impractical to manipulate the quality of early care and education settings as conceptualized by the measures

discussed previously. As a result, researchers cannot control for unobservable factors that may be related to both child outcomes and the quality of care (Duncan & Gibson-Davis, 2006). However, some studies have used techniques like propensity score matching to account for some of these factors (Li et al., 2011; Ruzek et al., 2014). Other approaches include models that examine effects on a change in outcomes or adjust for a baseline version of the outcome (NICHD & Duncan, 2003).

Long-term outcomes. One specific question of interest is whether there is a relationship between quality of care and longer-term outcomes. This might not occur, given the weak relationships in the short term and because many intervening factors may also affect children's development after their early years. On the other hand, benefits of better care in the earliest years could initially be small or unobserved, but then accumulate over time as initial benefits allow children to learn more and turn these gains into larger benefits. The long-term effects of some of the interventions discussed earlier suggest this is possible. As with most research, long-term outcomes are rarely studied because suitable data is not available. However, the NICHD Study of Early Child Care and Youth Development (SECCYD) tracked children through at least age 15. The study found quality of care had small but statistically significant associations with scores on a vocabulary test (although not other cognitive-academic outcomes and not social-emotional outcomes) at sixth grade as well as with a construct of cognitive-academic achievement and with externalizing behaviors at age 15 (Belsky et al., 2007; Vandell et al., 2010). The size of these effects (effect sizes around 0.09) was similar to the initial effects found at age 4 and a half, suggesting that some effects may endure well beyond the period when children receive early care.

Alignment of quality and outcome measures. Several researchers have theorized that stronger relationships between measures of quality and child outcomes may be observed if the measures being used are more closely aligned to the domain of children's development being

assessed. For example, a connection between a measure of the quality of language instruction and language-related outcomes may be readily observed, while a measure of the quality of emotional support provided by the caregiver may be more strongly associated with children's social-emotional outcomes. Reviews have found evidence of these alignments in several studies, although not always consistently (Zaslow et al., 2010; Burchinal, Kainz, & Cai, 2011). On the outcomes side, measures of quality tend to show stronger effects when academic, cognitive, and language outcomes are used, compared to social-emotional outcomes (Burchinal et al., 2008; Burchinal, Kainz, & Cai, 2011; Keys et al., 2013).

Regarding the instruments mentioned earlier, the CLASS, which was designed to more directly assess interactions and support for learning, is more predictive of child outcomes than the ECERS, which provides a more global measure of quality (Mashburn et al., 2008; Sabol et al., 2013). Within each assessment tool, there is evidence that specific subscales or items may show a stronger relationship with child outcomes. Burchinal, Kainz, & Cai (2011) found that items from the ECERS interaction and program structure subscales showed stronger correlations with outcomes compared to items from the other subscales, while for the CLASS, items on productivity, teacher sensitivity, negative climate, and positive climate showed stronger correlations with outcomes.

Given these possibilities, research has also revealed which aspects of quality tend to be more or less present in child care settings. For example, using the CLASS, the quality of teacher-child interactions around emotional support tends to be higher than the quality of interactions around instructional support for learning in many settings (Burchinal et al., 2008; Mashburn et al., 2008). However, some programs have been able to increase the quality of instructional support to higher levels (Weiland et al., 2013).

Another, more concerning possibility is that existing observational measures do not adequately represent the quality of care that is relevant for improving child outcomes. Several

studies have used factor analysis and reached different conclusions about whether the ECERS measures one or more underlying quality factors, but none have found that the seven subscales actually measures separate quality factors (Cassidy et al., 2005). Studies using psychometric techniques to analyze the ECERS have found issues that may affect its validity as a measure of the quality of care (Gordon et al., 2013). Similarly, an examination of the Arnett Caregiver Interaction Scale found that values tended to cluster on the positive end of the scale, making it difficult to distinguish between different levels of quality according to items on the scale (Colwell et al., 2013).

Alternative specifications. Research has investigated whether alternatives to a straightforward linear specification better represent the relationship between quality and outcomes. For example, the effect of quality may differ depending on the amount of time children spend in the setting. A review by Zaslow et al. (2010) found a pattern of findings in several studies that higher dosages of high-quality care were associated with positive outcomes.

The effect of quality may also differ depending on the level of quality; for example, changes in quality may not have an effect when they are at low levels, but could have effects at higher levels. Some studies only find evidence of a linear effect (Peisner-Feinberg et al., 2001). Others find evidence of differential effects, usually that there are larger effects at higher levels of quality (Burchinal et al., 2010). Vandell et al. (2010) found larger quality effects at higher levels of quality in their analysis of the SECCYD data at age 15, but noted that all previous analyses of this study had not found any nonlinear effects. Burchinal, Kainz, & Cai (2011) found similar evidence, although again this was not consistent across all the relationships they examined.

Quality may also have different effects for different sub-groups of children based on their characteristics. In particular, quality may be more important for disadvantaged children because it compensates for the challenges and lower-quality home environments they face. Or, it may be less important if they simply face too many other challenges for the care setting to make enough

of a difference. Some studies have found evidence of larger effects for more disadvantaged children (Peisner-Feinberg et al., 2001). However, other studies have not found evidence of moderation by socioeconomic status (Vandell et al., 2010; Keys et al., 2013; Ruzek et al., 2014).

Infants and toddlers vs. preschool-age children. Most of the research on the effects of child care quality has focused on preschool-aged children (age 3 to 5) instead of infants and toddlers (birth to age 3), since by this age the large majority of children are in out-of-home care settings. Infants and toddlers are increasingly receiving care in out-of-home settings as well, however, leading to a need for more research on this specific age group.

Findings from a meta-analysis showed effects tend to be larger for younger children (age 2 and 3) than older children (age 4) (Burchinal, Kainz, & Cai, 2011). Ruzek et al. (2014) studied the effect of quality of care (using the ITERS and FDCRS) at age 2 for children in the Early Childhood Longitudinal Survey-Birth Cohort and also found effect sizes larger than typically found. However, an analysis of the SECCYD looking at quality of care during the infant and toddler years compared to preschool-age years found that quality of care at both ages was important; high-quality infant toddler care did not affect long-term outcomes without high-quality preschool, while the effect of high-quality preschool was enhanced by high-quality infant toddler care (Li et al., 2011). This suggests that attention should be paid to care for both age groups, and that efforts should not be concentrated in one area at the expense of the other.

Use in quality rating and improvement systems. Evaluating the extent to which commonly used measures truly reflect quality has become more urgent given their increasing use in state quality rating and improvement systems (QRIS). Under these systems, providers are assigned overall quality ratings (often as a number of stars) based on a combination of their characteristics on multiple dimensions. These systems are designed to improve quality by making ratings available for parents to use in selecting child care, and to providers to use in quality improvement (Schaack et al., 2012).

An evaluation of a pilot QRIS in one state found few links between overall QRIS ratings and child outcomes or between individual QRIS components and outcomes (Zellman et al., 2008). Another analysis used data from a study of pre-kindergarten programs to assign QRIS ratings based on state systems, and found these ratings were not associated with child outcomes (Sabol et al., 2013). While many QRIS use observational measures of quality, they also tend to primarily employ structural features as components, which may explain the lack of a strong connection to child outcomes.

Early Head Start Research and Evaluation Project. The present study involves using data from the Early Head Start Research and Evaluation Project, which will be discussed in more detail in the next chapter. This randomized evaluation found small, positive impacts of Early Head Start on a wide variety of child and family outcomes when children turned 3, although most of these impacts did not continue when the children were in fifth grade (ACF, 2002; Vogel et al., 2010). A supplemental report focused on the impacts on child care use and quality; however, the authors also conducted a basic analysis of the relationship between quality of care and child outcomes, finding some small effects on cognitive and language outcomes but not on a social-emotional outcome (Love et al., 2003; ACF, 2004). The Early Head Start data was also used as part of the analysis by Keys et al. (2013), although they specifically studied the effect of quality of care in pre-kindergarten, after the program group had completed Early Head Start, on outcomes at age 5.

CHAPTER THREE

DATA AND METHODS

This chapter describes the source of data used for the present study, including the variables used or constructed to use as outcomes, quality of care and other key explanatory factors, and other covariates in the analysis. It then presents the specific sample used and the methods employed in the analysis, including the primary model and the variations used to meet the aims of the study. The data and methods used will provide evidence to help answer these questions, but there are several important considerations and cautions that must be kept in mind.

Data:

This study uses data from the Early Head Start Research and Evaluation Project (EHSREP), a randomized evaluation of the Early Head Start program conducted from 1996 to 2000 (ACF, 2002). The evaluation included 17 Early Head Start programs selected from the first groups of programs funded in 1995 and 1996. These programs were purposively, not randomly, selected; while they are not nationally representative of Early Head Start programs, they consist of programs from all regions of the U.S. and from both urban and rural areas. Within each site, families who applied to the program and were eligible both for Early Head Start and the study were randomly assigned to the program or control groups. The program group was eligible to receive Early Head Start services; the control group was not eligible to receive these services, but was otherwise able to receive early care and education services from other sources. A total of 3,001 families were assigned to one of the two groups. A condition for study eligibility was the mother was pregnant or the child was under 12 months of age, and families could receive services until the child turned 3, so families could receive services for at least two years.

According to the Early Head Start guidelines, grantees could choose one of several approaches to providing services. Of the 17 programs, 4 programs chose a center-based approach, where services were primarily provided through out-of-home care, although some

home visits were included. Another 7 programs gave a home-based approach, where services were primarily delivered through home visits and activities with parents in the child's home. The remaining 6 programs employed a mixed approach that combined center and home-based services, either by providing some of both to all families, or providing mostly one type of service to some families and the other service type to other families.

The project collected a variety of data during the period in which families were eligible to receive services. Baseline data, including information on family characteristics, was available from family application and enrollment forms. Parent Services Interviews (PSIs) were conducted 6, 15, and 26 months after random assignment to collect information on families' receipt of the types of services provided by Early Head Start (from both Early Head Start and other sources), their economic status and participation in education and training programs, and their health and health of their children. An additional exit interview was conducted with program group families when their children were 36 months old to collect information on use of Early Head Start services. Interviews with parents were also conducted when the children were 14, 24, and 36 months old; these Parent Interviews (PIs) focused on the child's development and how the family was functioning. At these ages, the study also conducted direct observations and other assessments of children, their behavior, their development, their interaction with parents, and their home environment. For children receiving out-of-home child care, the study also conducted observations of this care. Other data collected, such as from interviews with fathers or on mothers' risk of depression, was obtained in subsets of sites.

The project followed up with families twice: in the spring before the child entered kindergarten, and when the child was in fifth grade. As with previous data collection, the project interviewed parents and conducted observations and assessments of children.

Data from EHSREP is available in two formats: 1) as a set of restricted-use files only available to qualified researchers; and 2) as a public-use file available through the Child Care

and Early Education Research Connections website. This study uses the public-use file, which contains 2,977 records, one for each family who participated in the evaluation, excluding 24 families where the focus child miscarried, died, or was adopted during the study period. Although the public-use file does not contain all of the data collected for EHSREP, it includes source and constructed data used for the original and follow-up impact evaluations, which can be adapted for this study's analysis.

Variables:

Outcome variables: cognitive and social-emotional development. Measures from two phases of the EHSREP are used as outcomes. First, this study uses the measures in the original analysis (ACF, 2004). Details on these measures are drawn from the impact evaluation reports (ACF, 2001 and ACF, 2002). Specifically:

Children were assessed on their cognitive, language, and social development using the Mental Development Index from the Bayley Scales of Infant Development (BSID-MDI; Bayley, 1993). The EHSREP interviewer asked the child to perform several simple tasks (such as building a wall out of blocks or sorting objects by color) and assessed them on their understanding of concepts in these tasks, all using a standardized protocol. Raw scores were converted into a standardized score based on a national sample of children the same age, where 100 is the mean score and 15 the standard deviation of the score. This score from the assessments given at 24 months and 36 months were used.

Two measures of language and vocabulary were used. At 24 months, the child's parent was asked to complete forms from the MacArthur Communicative Development Inventory (CDI; Fenson et al., 2000), which measures language development. One form asked parents to indicate which of 100 words commonly spoken by young children (such as "kitty" or "up") the child has spoken out loud. The number of these words spoken is used as the vocabulary production score. At 36 months, children were assessed on their receptive vocabulary using the Peabody Picture-

Vocabulary Test-Third Edition (PPVT-III; Dunn & Dunn, 1997). The EHSREP interviewer provided an increasingly difficult series of four-picture sets and, for each set, asked the child to select which one of the four pictures matches the word spoken by the interviewer. As with the BSID-MDI, raw scores are converted to a standardized score, adjusted for age, with a mean of 100 and standard deviation of 15.

Finally, children's social-emotional development was assessed using the Aggressive Behavior subscale from the Child Behavior Checklist (CBCL), which is part of the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach & Rescorla, 2000). The full CBCL for children ages 1.5 to 5 contains 99 items; parents were asked a subset of these items, including the 19 items that make up the aggressive behavior subscale, such as "angry moods" or "physically attacks people". For each item, the parent reports whether this is "not true", "somewhat or sometimes true", or "very true or often true" of the child. The score consists of one point for each item reported as somewhat/sometimes true and two points for each item reported as very/often true, so the total score ranges from 0 to 38. The scores taken from assessments at 24 and 36 months are used.

Outcome measures for this study are also taken from the study's long-term follow-up, when children were in the fifth grade. Details of these measures are drawn from the follow-up evaluation report (Vogel et al., 2010). Children took two direct assessments of cognitive ability and two assessments of academic achievement. This study uses these measures because they make up the academic success and ability success indices used by the EHSREP follow-up study. First, children were assessed again on the PPVT-III, which produced the same type of standardized, age-adjusted score. They were also assessed on the Matrix Reasoning subtest from the Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV; Wechsler, 2003). Children are given a series of abstract patterns and, for each set, asked to select the option that completes the matrix. A raw score corresponding to the number of correct answers was

calculated. Children were also given subsets of the fifth grade reading and mathematics assessments from the Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K; Pollack et al., 2005). Each test included questions from multiple content areas, such as initial understanding and critical stance for reading, and number sense and properties for math. For the math assessment, the 18-question routing form (which in the ECLS-K was given first and the results used determined the difficulty level of the rest of the assessment) was given to the child, and a raw score of the number of items correct out of 18 was used. For the reading assessment, scale scores were developed using Item Response Theory to estimate how the child would have performed on the full 186-question assessment, so the score used could theoretically range from 0 to 186. Finally, the academic success index is used. This index is equal to 1 if a child reaches a threshold on all four cognitive and academic assessments: equal to or greater than 100 on the PPVT-III, 10 on the WISC-IV, 9.6 on ECLS-K math, and 50 on ECLS-K reading. The EHSREP follow-up study broke this out as two separate indices, but this analysis combines them for simplicity and to match the approach for the social-emotional measures.

Several social-emotional outcomes were also measured at grade 5. Five in particular were used for this study because they formed the social-emotional risk and success indices used by the EHSREP follow-up study. Parents were given the complete version of the CBCL for children ages 6 to 18 (Achenbach & Rescorla, 2001). Scores from three of the subscales were used, for Internalizing Behaviors (example items: “too fearful or anxious” or “underactive, slow moving, or lacks energy”), Externalizing Behaviors (“breaks rules at home, school, or elsewhere” or “cruelty, bullying, or meanness to others”; this includes the Aggressive Behavior subscale), and Attention Problems (“fails to finish things he/she starts” or “poor school work”). The three scales have 32, 35, and 10 items, respectively, so the score on each could range from 0 to 64, 70, and 20. Two measures were also used based on children’s self-reported experiences. Children were asked if they had engaged in a series of delinquent behaviors such as “taken or stolen something

from a store without paying for it" or "cheated on a school test." The items were created by the EHSREP study team, drawn from the NICHD Study of Early Child Care and Youth Development, or drawn from another study (Loeber et al., 1991). The score for this measure is the count of behaviors reported out of the 14 asked about. Finally, a measure of bullying from the Child Development Supplement, Wave 2 of the Panel Study of Income Dynamics (PSID-CDS2) was used. Children were asked to report on four bullying-related items, such as "how often have kids in your school or neighborhood taken your things, like your money or lunch, without asking," where responses ranged from a scale of 1 ("never") to 4 ("many times"). Each item's scores were combined, so the total score ranges from 4 to 16. The final measure is the social-emotional success index used by the EHSREP follow-up study, which is composed of the individual measures just discussed. The index is set to 1 if a child reaches a threshold on all five outcomes: for the CBCL internalizing and externalizing behavior scales, a T score (a standardized version of the raw score) of less than 60, and a T score of less than 65 for the attention problems scale; less than 8 on the bullying measure, and less than 3 delinquent behaviors reported.

Key explanatory variables: quality of care. The quality of care children experienced was measured in four ways. As described in the original quality analysis (ACF, 2004), trained observers interviewed caregivers and observed the child care setting during a single visit lasting two to three hours in the morning. Data from this visit was used to construct all four measures.

First, observers recorded data on items from the environment rating scales. As previously discussed, these assess the global quality of the child care environment. Different scales are used depending on care setting and child age: for EHSREP children in center-based settings, the Infant Toddler Environment Rating Scale (ITERS; Harms, Cryer, & Clifford, 1990) was used for the 14 month and 24 month observations, and Early Childhood Environment Rating Scale-Revised (ECERS-R; Harms, Clifford, & Cryer, 1998) was used for the 36 month observations. A

third scale, the Family Day Care Rating Scale, was used for children in home-based settings at 14, 24, and 36 months, but it is not used for this study because home-based settings are not included in this study's sample. For each of 31 items on the ITERS and 39 items on the ECERS-R, the observer recorded a quality score ranging from 1 to 7, where 1 is considered "inadequate" quality, 3 is "minimal" quality, 5 is "good" quality, and 7 is "excellent" quality. Examples of items include "room arrangement for play", "encouraging children to communicate", and "staff interaction and cooperation". The items are grouped into several subscales. For the ITERS, there are seven subscales, which are: furnishings and display for children; personal care routines; listening and talking; learning activities; interaction; program structure; and adult needs. The seven ECERS-R subscales are similar: space and furnishings, personal care routines, language and reasoning, activities, interaction, program structure, and parents and staff. Items are similar between the ITERS and ECERS-R, although there are differences, many due to the different age range

Scores for each subscale are calculated by averaging the scores for each item in the subscale, and the overall average score is calculated by averaging the scores for each item in the full scale. Finally, an average of the overall score across time periods is used for the study – this is the average over 14 and 24 months for the 24-month outcomes and the average over 14, 24, and 36 months for the 36-month and grade 5 outcomes. For example, for a child only observed at 24 months (and missing scores for the other periods, for any reason), the overall score at that time would be used. For a child observed at all three points, the average of the overall scores across the three time periods are used. More details are found in the appendix.

Several studies have used factor analysis to determine whether the seven subscales represent different factors as part of the overall scale. These studies have identified different number of factors and the items they are composed of. However, some have found two factors, 1) Teaching and Interactions and 2) Provisions for Learning (Pianta et al., 2005). Because the individual

items in Teaching and Interactions are all found in the language and reasoning subscale and interaction subscale (in the ECERS-R) this study uses the average score of these two subscales or their equivalents in ITERS as a teaching score. Similarly, the items in the Provisions for Learning factor come from the space and furnishings subscale and activities subscale, this study averages those two subscale scores to use as a provisions score. As with the overall score, the average of this score across the three time periods (excluding periods where the score is missing) is used as the explanatory variable.

Second, the EHSREP observers also used the Arnett Caregiver Interaction Scale (CIS; Arnett, 1989), which measures the quality of the interactions between a caregiver and all of the children under their care. Similar to the environment rating scales, the Arnett scale consists of 26 items. For each item, the observer assigns a score from 1 to 4, where 1 means the item is "not at all true" of the caregiver, 2 is "somewhat true," 3 is "quite a bit true", and 4 is "very much true." For example, items include "encourages the children to try new experience" or "finds fault easily with children." The overall score is the average of the 26 item scores, so it ranges from 1 to 4. The Arnett CIS was collected across all time periods (14, 24, and 36 months) and settings (center and home-based settings). Again, the average of the overall score across each available time period is used as the measure for this study.

Third, observers followed a study-developed instrument called the Child-Caregiver Observation System (C-COS; Boller et al., 1998), which assesses the details of interactions between the caregiver and a specific child, in this case the child under their care who was participating in the study. During a five-minute period, the observer conducted 10 cycles with the first 20 seconds spent looking for the presence of specific actions or behaviors on the part of the caregiver and the child, followed by 10 seconds to record the results. Six of these periods were conducted throughout the overall visit. The EHSREP study constructed counts of the 20-second observation periods during which four behaviors occurred: the caregiver talked to the child at all;

the caregiver initiated talk to the child; the caregiver responded to the child; and the child exhibited a negative behavior. As a result, the score for each item could range from 0 to 60. The C-COS observations were conducted during the 24- and 36-month observations, but not the 14-month observations. They were conducted in both center- and home-based settings. For 24-month outcomes, the score at 24 months was used for this study, while for 36-month and grade 5 outcomes, the average score over 24 and 36 months was used. More details can be found in the appendix.

Finally, the observer recorded the number of children and adults present at six points in time throughout the observation period. These numbers were used to calculate child-adult ratios, and the average ratio across the six observations was used as a measure of quality. A higher number indicates more children per adult. These were collected during all data collection periods (14, 24, and 36 months) and in all settings. The average ratio during each observed period is used for this study.

Intensity of care. During the Parent Services Interviews (PSIs) and Parent Interviews (PIs), parents were asked about information on the care arrangements their child was in, if any. This included the average number of hours per week the child spent in each arrangement. The PIs were conducted at 14, 24, and 36 months, and only asked for information on average weekly hours in care at the time of the interview. The PSIs were scheduled based on the length of time since random assignment, but they asked parents for this information over the time period since random assignment or the last interview, so the EHSREP study used this history to construct the average weekly hours in center care at the times when the child was 14, 24, and 36 months old. Because not all children had turned 36 months by the time the final PSI was conducted, a version of this measure incorporating information from the 36-month PI was created. The public-use file includes these measures, but not the measures based on hours from the PIs. As with the quality

measures, an average of the average number of hours in center care at each time period (14, 24, and 36 months) was used.

Covariates. Several variables measuring child and family characteristics are used as covariates in this analysis. Early Head Start applicants filled out detailed application and enrollment forms at the time they were randomly assigned, which the EHSREP evaluation used to construct several measures. The first set of covariates includes those used the original quality analysis: child gender, maternal race/ethnicity, maternal education, maternal living arrangements/marital status, and maternal age at child's birth.

The original analysis included child age at assessment and an indicator for whether the site was urban; unfortunately, these variables cannot be included because they are not available in the public-use data. The original analysis may have also used slightly different constructs for maternal education and marital status. Other potentially important variables that are not included in the EHSREP public-use file (in order to protect the confidentiality of study participants) include identifiers for each site, although a variable is available indicating whether the site's Early Head Start program used a center-based, home-based, or mixed approach.

A second set of covariates that were used (along with the first set) in the EHSREP impact evaluation are also used in this analysis: these include maternal English-language proficiency, maternal employment or schooling status, number of other children in the household, household income as a proportion of the federal poverty level, participation in welfare programs, self-reported lack of adequate resources to pay for important goods, previous participation by the mother in another child development program such as Head Start, number of moves in the previous year, date of random assignment, age of child at random assignment (including unborn), and for children who had been born at random assignment, whether they were low birthweight, born early, evaluated about their health and development, or identified as at-risk in a number of categories. Most of these measures are categorical, so indicator variables were used to note if the

family or child met the characteristic. Specific measures are listed in Table 1, discussed in the next chapter.

Finally, two other relevant covariates were used in one alternative model specification. During the EHSREP's follow-up at pre-kindergarten, it observed the quality of child care settings using the ECERS-R. This serves as a similar measure of quality as the environment rating scales used during the main EHSREP study. During the fifth grade, a small number of characteristics of the child's school were collected, including the proportion of students eligible for free or reduced-prince lunch (obtained from publicly available U.S. Department of Education school-level statistics). This can be seen a rough proxy for quality of school attended, that could account for differences in fifth grade outcomes.

Analytic Sample:

The study follows the original quality analysis (ACF, 2004) and restricts the analysis sample to children who were randomly assigned to the program group (that is, to receive Early Head Start services), and who were in a center-based care setting that was observed to collect quality measures at least once at 14 months, 24 months, or 36 months. While this limits the size of the sample, it provides several analytic benefits. First, because Early Head Start consisted of a range of services, depending on the site, the children and families in the program group received services not available (or at least not in the same intended intensity) to the control group, such as home visits, health screenings, or parent services such as assistance with employment. Restricting the sample to within the program group avoids the need to account for the differences in services provided between the two groups.

Second, control group families could enroll in other early care in the local community, introducing the risk of selection effects. It is possible that families that are able to find higher-quality care, whether through extra effort, greater resources, or other factors, are also able to help their children in other ways that are difficult to measure. What appears to be an effect of quality

on child outcomes may be simply be the influence of family-level characteristics correlated with both factors. On the other hand, most child care for the EHS treatment group was either provided directly through centers operated by the program, or through referrals to and paying for care at other community centers. The larger role of the EHS program in determining the quality of the child care families received means selection effects may be less likely.

Finally, only a relatively small proportion of the settings where children were receiving care were observed. As laid out in the original analysis, the EHSREP evaluation attempted to observe the care settings of all children whose parents reported (during the PI) they were in at least one arrangement for at least 10 hours per week over the past two weeks. However, the study was not able to conduct observations for many of these settings, because they had to secure permission from the parent, find the provider, and obtain the provider's permission to observe the setting. Slightly more than half of all eligible settings were observed, but a much higher proportion of center-based settings were observed (about 70 percent in each time period) than were home-based settings (about 30 percent in each time period). Furthermore, within center care settings, more program group children's settings were observed than control group children's settings, which is understandable given that many of the program group children's settings were Early Head Start centers. Using only center care settings for program group children uses the group with the highest proportion of data available.

Overall, this sample is much smaller than the EHSREP evaluation sample. Even before the reasons discussed above, parents who did not participate in a particular PI could not be asked about their child's out-of-home care in the first place. Many of those who did respond were not using any out-of-home care or not consistently enough to for the setting to be eligible. As a result, only 465 of the children were in the program group and observed in a center setting at least once during the study, and not all of these also have outcome data (the number of those who do have both ranges from 297 to 337 for the 24-month outcomes and 335 to 409 for the 36-

month outcomes). The sample for the grade 5 outcomes is smaller because not all the children could be reached for the follow-up, although the difference is small (the range is from 294 to 309 for the grade 5 outcomes), partly because some children without short-term outcomes actually have data on grade 5 outcomes.

Analytic Approach:

Using this sample, this study primarily estimates the effect of the quality of care on children's development using multivariate Ordinary Least Squares (OLS) regression. By controlling for other factors that may influence children's development, primarily child and family characteristics, the models estimated here attempt to isolate the contribution of quality of care to these outcomes. Initially, this study follows the approach from the original analysis (ACF, 2004) by using the child outcomes at 24 and 36 months that were previously discussed, explained by quality as measured by the mean ITERS score over 14 and 24 months and mean ITERS/ECERS-R score at 14, 24, and 36 months, respectively. Next, it adds the previously mentioned child outcomes from the EHSREP's grade 5 follow-up to see what relationships, if any, exist over a longer timeframe. Logistic regression is used for the grade 5 success index variables, which only have values of 0 and 1. All models use Huber-White standard errors to deal with potential heteroskedasticity in the error term.

The primary model adds child-adult ratio, average weekly hours of center care, and the first set of child and family characteristics discussed previously. A small number of observations with quality and outcome data have missing data for the child and family characteristics to be used as covariates. Because the number of observations with missing data is small, and in order to maximize the sample with quality and outcome data, this missing data was imputed for each covariate by regressing the covariate on several of the other variables with child and family characteristics. Details on this imputation are in the appendix.

Ideally this analysis would not have needed to handle missing data on the other important explanatory variables, child-adult ratio and average weekly hours in center care. However, as mentioned earlier, data on hours in center care from the PIs were not available in the public-use file. Because data from the PIs on hours in care were used to determine which children had eligible child care arrangements, all observations with quality data should also have data on hours. However, the available data only comes from the PSIs, which were conducted at separate times. Many of the families whose child has quality data did not respond to the PSIs, and this information is sometimes inaccurate – many children with data from the ITERS or ECERS-R (which were only used in center settings) at a given time period are listed as not being in center care or having any hours in center care based on the PSI. As a result, the primary model sets missing values on hours to a constant and includes an indicator variable set to 1 if the hours had been missing. This preserves as much of the size of the sample as possible. Missing values for child-adult ratio were not imputed; however, only a tiny number of observations (2) were lost because they had data for outcomes and for other measures of quality, but not for child-adult ratio.

Next, the different measures of quality (discussed above) are used in the place of the mean ITERS/ECERS-R overall score. If these measures capture different aspects of quality, and those aspects of quality are more or less strongly linked to child outcomes, then these models may have different results from the primary models. One possibility is that measures of quality that more directly assess how the caregiver interacts with children will show a stronger connection than measures that relate to the child care environment but are less direct. For example, the ITERS/ECERS-R teaching score seems to be more focused on direct caregiver-child interaction than the learning provisions score. An example of a very indirect measure of quality is child-adult ratio. The Arnett CIS and C-COS counts also focus more on the quality of interaction, with the C-COS having another advantage in that it specifically concerns the child in the study, not all

children in the care setting. Another possibility is that different measures might link to different domains. For example, the Arnett CIS, which assesses caregiver behaviors such as harshness and detachment, might be expected to be linked more strongly to outcomes in the social-emotional domain than the cognitive domain.

One issue with using the alternative measures of quality is that they do not completely overlap within the sample. The ITERS/ECERS-R subscales line up perfectly with the overall score, and the Arnett CIS score is close, but the C-COS counts have a different sample, largely because they were not conducted during the first time period (14 months). More details are in the appendix.

The next step in this approach is to run several robustness checks using slight differences in the primary model. This starts by running a baseline model without imputed covariates or a missing indicator for hours (which will lower the sample size). Another version adds the missing hours indicator but does not impute covariates. Two other approaches for handling the hours data are tried: imputing hours (details in the appendix), and both imputing hours and using an indicator for the cases that were originally missing data on hours. Finally, versions of the baseline model without hours, and without hours or child-adult ratio (in other words, only using mean ITERS/ECERS-R overall score and the child and family characteristics) are run. This will demonstrate if the results are sensitive to small changes in the model.

Finally, this analysis runs alternative models using different specifications, while keeping the ITERS/ECERS-R overall score as the measure of quality. Some of these can be considered additional checks for sensitivity, while others test different assumptions about the nature of the relationship between child care quality and outcomes, especially regarding whether the relationship is strictly linear. First, a model is run that has no child and family characteristics (only the quality, child-adult ratio, and hours variables), followed by one that adds the full set of covariates used in the EHSREP impact evaluation.

Next, variables for the quality of care in pre-kindergarten (as measured by the ECERS-R overall score) and school quality in grade 5 (as measured by percent of the school eligible for free or reduced price lunch) are added to the models with grade 5 outcomes. Pre-kindergarten quality is added as a separate variable than quality during Early Head Start, instead of being combined, since these are different experiences and may make separate contributions to child outcomes.

The next alternative adds an interaction term for the mean ITERS/ECERS-R overall score and average weekly center hours, in order to find out whether the effect of quality on outcomes changes depending on the intensity of care. Theoretically, any effects of quality, positive or negative, should be larger when the child spends more time in care and is therefore more exposed to the child care setting.

Another model investigating whether the relationship between quality and outcomes is non-linear adds a squared term for quality. Theoretically, the direction of the squared term could be positive or negative. A positive value would suggest that at lower levels of quality, improvements in quality are not enough to make a difference for children, but at higher levels, the quality of care can make a bigger impact. A negative value suggests the opposite – that there are benefits to raising quality out of lower levels, but once higher levels are reached, outcomes are less sensitive to changes in quality. As discussed in the previous chapter, evidence found so far supports the former theory.

Finally, the continuous measures of quality that have been used are replaced with a trio of mutually exclusive indicators noting whether the quality of care fell into a certain range of scores. Two approaches are used. First, the groups are defined using the ITERS/ECERS-R developers' thresholds for quality categories, which split the 1-7 scale in three evenly spaced groups. The low group is for quality that falls between 1 ("inadequate" quality) and 3 ("minimal" quality), the middle group covers scores between 3 and 5 ("good" quality), while the high group

is for scores between 5 and 7 (“excellent” quality). In the second approach, groups are assigned based on the distribution of scores in the sample. Here, the low group is the bottom quartile of scores, the middle group is the middle two quartiles of scores, and the high group is the top quartile of scores. Because the mean ITERS/ECERS-R scores are relatively high on the scale, these thresholds are higher than the threshold scores of 3 and 5 in the first approach. Specifically, the bottom quartile runs through a score of 4.18, while the top quartile starts at 5.65. The first approach will show whether the developer thresholds are a useful way of dividing up quality, while the second approach provides evidence for whether isolating the children in the highest and lowest relative settings yields a difference in outcomes.

Ideally this analysis would have looked at sub-groups, as the EHSREP evaluations found larger impacts for certain sub-groups, such as for African-American children. Given the already reduced sample size for this analysis, however, sub-group analyses are not conducted.

The fact that this study consists of estimating a large number of models means the results for any single model should be viewed with caution, as a few statistically significant or other noteworthy results may appear by chance. Instead, the patterns that emerge across models are important.

CHAPTER FOUR

RESULTS

Having outlined the data and methods employed for this study, this chapter examines the results of these analyses. The tables containing the full results are located after the main text. First, descriptive statistics of the variables used for the analysis are discussed, followed by a preliminary analysis of simpler bivariate relationships between the variables. Finally, the multivariate section contains the regression results for the short-term outcomes (24 months and 36 months) and longer-term outcomes (grade 5) using the primary model, followed by results using other measures of quality, results of robustness checks, and results from alternative specifications of the model.

Descriptive Statistics:

Table 1 includes descriptive statistics for the key explanatory variables, outcomes, and covariates. Patterned after Table A.6 in the original analysis (ACF, 2004), statistics are presented for the primary analysis sample, children who were observed in center care at least once during the 14 month, 24 month, and 36 month periods. The data in Table 1 generally lines up with that of the original analysis. There are small differences for the child-adult ratios, because unlike this study, the original used samples at each time period to calculate these figures instead of the full sample of children with an observation at any of the three time periods. The CBCL aggressive behavior scale is also different; given that the maximum score is higher than in this analysis and that the EHSREP study asked parents to fill out additional items on the CBCL, the Table A.6 in the original analysis may list statistics for the full questionnaire. The most important difference is the average number of weekly hours in center care. The present study shows a slightly higher number of hours and more variation compared to the original analysis. The most likely explanation is that the original study used data on hours from the PIs, which as discussed in the

previous chapter, are the most accurate source of data on hours in center care around the time the child care observations of quality were conducted.

The information in Table 1 highlights several important points about quality. First, the average quality of care received is relatively high. The average ITERS/ECERS-R overall score across all time periods is 4.85 on the scale of 1 to 7, just slightly under the threshold of 5 corresponding to the developer's definition of "good" quality care. Only 6.7 percent of the children received care that was below the "minimal" quality level of 3 on average during the program, with the rest split between "minimal" to "good" and "good" to "excellent" quality. As the original analysis notes, the Cost, Quality, and Child Outcomes Study in the 1990s found an average score of 3.4 on the ITERS, well below the EHS average (ACF, 2004). A more recent dataset, the Early Childhood Longitudinal Study-Birth (ECLS-B) cohort, showed an average score of 4.16 for its low-income sample (Ruzek et al, 2014), which is still about two-thirds of a standard deviation below the EHS scores.

Turning to the alternative measures of quality, the ITERS/ECERS-R teaching score is higher (and has more variation) than the learning provisions score, illustrating the lack of uniformity that can exist in the components of scores across different assessment measures. The average Arnett CIS score is also high – across all time periods, the mean is 3.38 out of a possible 4. As a percentage of its maximum, the average Arnett score is a more favorable rating of quality of care than the ITERS/ECERS-R scores. The C-COS counts show that the caregiver spoke to the child during about half of the observation periods (29.9 out of 60), with about three quarters of those exchanges initiated by the caregiver. Children exhibited negative behavior during about 10 percent of the observation periods (4.8 out of 60).

The relative disadvantage of this population is readily apparent in the Bayley MDI scores (90.8 at 24 months and 91.9 at 36 months) and the PPVT-III scores (84.5 at 36 months and 94.4 at grade 5), all well below the national mean score of 100. By grade 5, using the EHSREP's

definitions of success, only 14.3 percent of children reached the benchmarks on each cognitive/academic measure, although 43 percent of children reached all of the social-emotional benchmarks. It is important to note that the 24-month and 36-month outcomes have different sample sizes. Response rates are higher for measures based on parent reports (the CDI vocabulary production and CBCL aggressive behavior) than for measures based on direct observations (the Bayley MDI and PPVT-III).

The demographic covariates provide further evidence of the relative disadvantage of this population, although the degree varies somewhat across the measures. For example, 41 percent of mothers had less than a high school education, 40 percent were the only adult in their household, 30 percent were less than 19 years old when their EHS child was born, and 44 percent were not employed or in school. On the other hand, noteworthy shares of the mothers had more than a high school education (30 percent), lived with a spouse (18 percent), or were employed (30 percent). In context, it is useful to remember that the program occurred in the mid- to late-1990s, a period of strong economic growth.

Participation in social safety net programs varied. About a quarter of the families received welfare benefits (Aid to Families with Dependent Children as it transitioned into Temporary Assistance to Needy Families following welfare reform) and half received food stamps. Almost all participated in the Women, Infants, and Children (WIC) nutrition, which is not surprising given that this sample is a low-income population of women who were pregnant or had a young child when enrolling in Early Head Start. Perhaps because of these programs, relatively low percentages of mothers reported not having resources to pay for food or medical care. However, a large percentage reported difficulty with affording other necessities and transportation. Turning to children, a relatively small percent had a low birthweight and/or were born very early, although a third were identified as having some kind of established, bio/medical, or environmental risk that could affect their development.

Bivariate Statistics:

Table 2 shows the pairwise correlations between the various measures of quality used in this analysis. As expected, the ITERS/ECERS-R overall score is tightly linked to the teaching and provisions scores (r = 0.90 or higher), although the two subscores are less strongly correlated with each other (r = 0.78), which suggests they may be capturing different aspects of quality. The Arnett CIS score is highly associated with the ITERS/ECERS-R scores, but the C-COS scores, surprisingly, are much more weakly correlated with the other quality measures. In fact, C-COS caregiver initiated talk has no correlation with the ITERS/ECERS-R and Arnett CIS measures, although the C-COS responding talk and negative behavior scores show some weaker relationships with other measurement scores. The Arnett CIS and C-COS measures tend to be more strongly correlated with the ITERS/ECERS-R teaching score than the provisions score. This makes sense since those measures are more focused on caregiver interaction with children. As would be expected, the C-COS negative behavior rating is negatively correlated with measures for which a high value is favorable. Similarly, the child-adult ratio (the lower the value, the more beneficial) is negatively correlated with the quality measures, although fairly weakly (r is between -0.07 and -0.24). Finally, hours in center care shows very little or consistent association with the quality measures. With the exception of the C-COS counts, these results line up with expectations that while related, the different measures of quality do not overlap completely. This makes an investigation of their different relationship with child outcomes of interest.

Turning to the outcomes, table 3 shows the results of correlating 24-month, 36-month, and grade 5 cognitive/academic and social-emotional outcomes with the ITERS/ECERS-R overall score. Although nearly all of the correlations run in the expected direction, all are weak, and most are not statistically different from zero. However, three measures – PPVT-III at 36 months, the academic success index at grade 5, and the CBCL externalizing behavior at grade 5 – are

correlated with the ITERS/ECERS-R overall score ($p < 0.10$) with Pearson correlation coefficients of roughly 0.10 in strength. So, although there is limited evidence of a relationship between measures of quality across most of the outcomes, modest effects may be detected in the multivariate results.

Table 4 contains means and sample counts for key variables of interest, outcomes, and covariates for several different samples. This table does not reflect any tests for differences in means. The first column contains statistics for the full treatment group (families and children assigned to receive EHS services), as a reference point. The second column consists of children known to have received center care at some point during the program, based on information gathered from one of two sources. Given that the ITERS/ECERS-R assessment tools involve direct observation of the care environment, children with non-missing overall scores over 14, 24, and 36 months had, by definition, been in center care. While no observations were conducted for the remaining children, they were included based on parental report of center care. The composition of this sample is certainly subject to error because the report of center care is based on the Parent Services Interviews (PSIs). However, it is the best alternative in the absence of the same data from the Parent Interviews (PIs). The EHSREP attempted to collect data on quality of care for all children in this group (and, as with the full sample, to collect data on outcomes). Under a theoretical scenario where this data was obtained for all children, this is the sample that would be used for the analysis. The children in the center care group have slightly higher average scores on some of the outcomes compared to the full treatment group, but these are not large. There appear to be some substantive differences based on family characteristics; for example, mothers using child care had more education and were less likely to live with their spouse. This makes sense that the families choosing out-of-home care might differ from the full group.

The third column consists of families for whom quality (the ITERS/ECERS-R overall score) and outcomes from the main ESHREP evaluation are available; for the variables of interest and

covariates, the 36-month PPVT-III outcome is used as the filter (using other 24-month or 36-month outcomes would yield slightly different samples). The fourth column further restricts the sample to families for whom data on average weekly hours in center care is available at least once during the timeframe. Again, because this came from the PSIs, this is a smaller sample than in column 3. For the outcomes, the sample is defined by the presence of data for quality and for that particular outcome, instead of for the PPVT-III score.

Finally, column 5 consists of cases for whom the quality measures (again, the ITERS/ECERS-R overall score) and grade 5 outcomes are available. For the variables of interest and the covariates, the grade 5 PPVT-III is used; patterns of missing data are more consistent across the grade 5 outcomes than the 24-month and 36-month outcomes, so this filter is very similar to ones that use other grade 5 outcomes. Note that while columns 1 through 4 are progressively smaller sub-groups, column 5 is not a subset of columns 3 or 4. As expected, there are cases with 36-month, but not grade 5 outcomes, as well as vice versa. Overall, these columns show that there are only small differences in quality between the groups. However, there do appear to be small but non-trivial differences in the child and family characteristics. For example, the sample with quality and outcome data tends to have more maternal education but the sample members are less likely to be Hispanic (at least until the grade 5 sample is considered). This underscores the need to consider the various samples carefully when running these analyses; as some differences in effects may be due to compositional differences in the samples.

Table 5 contains the results of regressing a few of the quality measures on the demographic covariates taken from EHS baseline information. Weak or nonexistent relationships between these characteristics and quality of care would provide some evidence that selection effects were not a problem, although this only applies to these observable characteristics. By and large, the associations are indeed weak, 0.2 points or less for the ITERS/ECERS-R, where the standard deviation is about 1 point on a 1-7 scale, and less than a tenth of a point for the Arnett CIS,

which has a standard deviation of about a half point on a 1-4 scale. The largest association is found between families where the mother lived with a spouse and higher quality care, an association of half a point on the ITERS/ECERS-R overall score and 0.15 points on the Arnett score. The C-COS any talk measure is more strongly correlated with the demographic characteristics, particularly race and ethnicity. Overall, the demographic characteristics only explain about 4 to 5 percent of the variation in the ITERS quality measures and less for the Arnett, although they explain 10 percent of the variance in the C-COS count of any talk. This suggests any selection effects (at least observable characteristics) are not large, but these characteristics should still be accounted for in the model.

Another piece of evidence around the extent of selection effects might be gained by examining the relationship between the quality of care during EHS and during the pre-kindergarten follow-up. Theoretically, if the former was determined more by EHS program factors and less by family choices, while the latter reflects much more of the families' control (because EHS had concluded for them), then there should only be a weak relationship. The correlation between mean ITERS/ECERS-R overall score during the program (average over 14, 24, and 36 months) and ECERS-R overall score in pre-kindergarten was 0.23 ($p < 0.001$). While this is a relatively weak relationship, it is significantly different from zero, and it is in the same range as the correlation for the control group ($r = 0.20$, $p = 0.02$), where a higher correlation might be expected. This is highly speculative, as it assumes pre-kindergarten quality involves selection effects, and that there are not other plausible reasons why the two may be linked. However, it does suggest that families whose children received higher quality care during EHS tended to receive higher quality care afterwards. There is no correlation between the quality of care and school poverty (as measured by the percent of fifth grade students who are eligible for free and reduced price lunch ($r = -0.03$, $p = 0.62$); the correlation between this measures was slightly stronger at -0.10 for the control group, but still not statistically significant ($p = 0.22$).

Multivariate Regression Results:

Primary models. Tables 6-A through 6-C present the results of the primary regression model, which uses the mean ITERS/ECERS-R overall score over 14, 24, and 36 months as the key explanatory variable. The model includes the average child-adult ratio, includes average weekly center hours and adds an indicator when hours are missing, and imputes missing values for the demographic covariates.

In Table 6-A, the relationship between quality and the 24-month and 36-month outcomes runs in the expected direction, except for 36-month CBCL aggressive behavior. For most of the outcomes, the effect is not statistically significant, but the effect of quality on the PPVT-III at 36 months is significant ($p = 0.016$). Holding the other variables in the model constant, we expect an increase in the mean ITERS/ECERS-R overall score of 1 point (for example, going from 3 to 4, or from 5.5 to 6.5) to be associated with an increase in the PPVT-III score of nearly 2 points. This is an effect size of a little more than a tenth of a standard deviation ($d = 0.12$). The relationship between quality and the 24-month Bayley MDI is about the same size ($d = 0.10$), but not significant below the 0.10 level.

Turning to weekly child-adult ratio and weekly center hours, each has a significant association with 24-month CBCL aggressive behavior. However, the direction is actually unfavorable for child-adult ratio, as an increase in child-adult ratio (which is negative) leads to an increase in the CBCL score (higher is also negative here). The coefficient for hours seems small at -0.05, but in practical terms it means an increase of 20 hours per week (for example, going from half-day to full-day care) is associated with a 1-point decrease in the CBCL score, which is why the effect size ($d = -0.17$) is relatively large. The indicator for whether hours are missing is also significant, showing that the cases with missing values on hours have a better CBCL score, on average, than cases with data on hours. Otherwise, the relationships between

these variables and the other short-term outcomes are small and insignificant, and in some instances, operate in an unexpected direction.

The covariates generally have the expected relationships with these outcomes, with less maternal education, male children, and African-American mothers being associated with more negative outcomes. Interestingly, teenage mothers tend to have more positive outcomes compared to older mothers. Overall, the models predict the cognitive outcomes more strongly than the social-emotional outcomes.

Continuing to Table 6-B, the effect of quality is not significant for any of the individual grade 5 cognitive/academic outcomes. There are non-trivial effect sizes for the PPVT-III and ECLS-K reading tests (d = 0.07 and 0.08, respectively). Intriguingly, there is a significant effect on the academic success index. This column includes odds ratios, meaning that an increase of 1 point in the mean ITERS/ECERS-R score makes a child 1.66 times more likely to be an academic success. This translates into an average marginal effect of 5 percentage points across the sample; that is, increasing quality by 1 point for everyone in the sample but holding all else constant would lead to an average of 5 percentage points more children being academically successful, with the change dependent on the level of quality and the child's other characteristics.

Child-adult ratio also has a significant effect on the academic success index, although again it operates counterintuitively because a higher ratio (unfavorable) increases the odds of success. Otherwise, there are no significant effects of ratio or hours on the long-term cognitive/academic outcomes, although hours has a relatively large relationship with three of the individual tests, with effect sizes of 0.08 or higher. Of the covariates, maternal education is even more important than it was with the short-term outcomes, and children of African-American mothers have poorer outcomes. Children born when their mothers were 18 or younger continue to surprisingly have better outcomes. Overall, these models tend to account for more of the variation in these

outcomes compared to variation in the short-term outcomes, with an R-squared of 0.25 for the PPVT-III score being the highest.

Finally, at Table 6-C, the relationship between quality and the grade 5 social-emotional outcomes is positive, as expected (higher quality decreases the scores), but none are significant, or have an effect size greater than 0.06 standard deviations. The effect of quality on the social-emotional index is actually negative, although not significant, which runs counter to the other outcomes. Child-adult ratio has a significant, positive effect on CBCL internalizing behavior ($p = 0.030$), but not on any other outcomes, while average weekly hours in center care has almost no relationship with any outcome – the closest is delinquent behavior ($d = -0.10$, but not significant). The most interesting effect involving the covariates is that children of black and Hispanic mothers have significantly better outcomes across the social-emotional measures, which are not expected given that these tend to be more associated with disadvantage. Overall, the models predict less of the variation in the social-emotional outcomes, with R-squared between 0.07 and 0.12.

Alternative measures of quality. Tables 7-A through 7-C show the results when different measures of quality are substituted in for the mean ITERS/ECERS-R overall score, while keeping the other aspects of the model the same. Unlike Tables 6-A to 6-C, only the effect of the quality measure is shown. Of the mean ITERS/ECERS-R subscores, the learning provisions score tends to have stronger relationships with the cognitive and academic outcomes (both short-term and at grade 5) than does the teaching score. For the most part, the level of statistical significance of these relationships does not change, but for the 24-month Bayley MDI and the grade 5 PPVT-III the provisions score is statistically significant (at the 0.10 level) where the overall score was not, and correspondingly the effect size is larger. The teaching score remains significant for the 36-month PPVT-III, but not for the academic success index. For some other measures, the effect size for the teaching score drops and is very close to zero. This pattern does

not hold for the social-emotional outcomes; the effects for the teaching and provisions scores do not seem to change much compared to the overall score.

The mean Arnett CIS score has a stronger effect on the 24-month outcomes compared to the ITERS/ECERS-R overall score, as the relationship with CDI vocabulary production and CBCL aggressive behavior become significant and the magnitudes of the effect sizes increase to 0.12 and 0.10. In contrast, the relationships with the 36-month outcomes are weaker under the Arnett score. Worse, the Arnett score has a pattern of negative effects on grade 5 cognitive/academic outcomes, one of which is significant (ECLS-K math score, $p = 0.024$). Like the ITERS/ECERS-R scores, the Arnett score has essentially no relationships with the grade 5 social-emotional outcomes, and is arguably even weaker.

In general, when the C-COS counts of any caregiver talk appear to have an effect on an outcome, it is actually unfavorable. For example, an additional period with caregiver talk is associated with a 0.16-point reduction in the 36-month Bayley MDI score ($d = -0.16$, $p = 0.005$), although most of the others are not statistically significant. This appears to reflect different contributions of the effects of caregiver-initiated talk and talk responding to the child. The former displays a similar negative pattern, but one that tends to be slightly larger in magnitude and is statistically significant for a couple of additional outcomes (24-month Bayley MDI and ECLS-K reading). The latter, on the other hand, tends to have favorable relationships with outcomes. Especially noteworthy is the stronger connection to CDI vocabulary score ($d = 0.17$). Finally, C-COS counts of negative child behavior also tend to have positive effects on most outcomes. Most effects, positive or negative, are larger and more significant for the short-term outcomes and grade 5 cognitive/academic outcomes; nothing rises to conventional levels of statistical significance for the grade 5 social-emotional outcomes.

Robustness checks: Tables 8-A through 8-C contain the results of robustness checks, which were conducted using the mean ITERS-ECERS-R overall score and the same set of demographic

variables as covariates. Compared to the primary model, the “baseline” model in the first row does not have imputed values for covariates, nor does it set cases with missing data for average weekly centers hours to zero and add an indicator for their missing status. The second row involves a model where the missing hours indicator is used, but the demographic characteristics are not imputed. The sample sizes demonstrate that the sample is much more affected by missing data on hours compared to missing data for demographic characteristics, as there is a large increase in sample size when the former is accounted for but only a small increase when adding the latter. The baseline model generally shows larger effects of quality than the primary model used, to the point where multiple grade 5 outcomes (PPVT-III, ECLS-K reading, CBCL internalizing behavior, and CBCL externalizing behavior) become significant with p-values less than 0.10 and effect sizes with magnitudes around 0.10 to 0.15. As with the sample sizes, the relationships between quality and outcomes tend to scale down in magnitude when adding the missing hours indicator, and then the imputed covariates. For example, the coefficient for the grade 5 PPVT-III decreases from 2.20 in the baseline model, to 1.27 (and no longer statistically significant) when adding the missing hours indicator, and to 1.00 when imputing the demographic covariates.

The fourth and fifth rows are for models that handle the missing hours data in slightly different ways. The fourth row imputes the missing hours instead of using a missing indicator, while the fifth row both imputes missing hours and uses an indicator to note cases where the hours were originally missing. Both models produce almost identical results to the primary model, indicating that the approaches to handling the missing hours data do not affect the results. The only difference in terms of inference is that under one of the models, the 24-month Bayley MDI crosses the threshold for significance at $p < 0.10$.

The bottom two rows return to the baseline model but drop the hours variable, and then both hours and ratio. The quality effects for the model dropping hours are similar to the model using a

missing hours indicator – these are the same sample but only differ in whether hours is included, and they have very similar results, suggesting that hours was either unrelated to the outcomes or to the level of quality. This makes sense, as Table 2 shows that ITERS/ECERS-R scores were not correlated with hours, and from Table 6-A to 6-C (and results from the other robustness checks, which are not shown), hours are generally not significantly related to outcomes aside from the 24-month CBCL aggressive behavior score. Finally, there is no clear pattern of changes when dropping the child-adult ratio – the effect of quality becomes slightly larger for some outcomes, like 24-month CBCL aggressive behavior and grade 5 PPVT-III, but slightly smaller for others, like 24-month Bayley MDI and 36-month PPVT-III.

Alternative specifications. Tables 9-A, 9-B, and 9-C display results from alternative specifications. These models continue to use imputed values for demographic covariates and an indicator that average weekly hours in center care is missing. After the primary model in the first row of each table, the next two rows alter the composition of the demographic variables. Removing all of them, as in the second row, and leaving only the mean ITERS/ECERS-R overall score, child-adult ratio, and average weekly center hours, the relationship between center quality and outcomes seems to weaken for the cognitive/academic outcomes. For example, the effect size for the 24-month Bayley MDI drops from 0.10 to 0.06, although it was not statistically significant in the first place. The effect of quality on the PPVT-III also decreases, although it remains significant. On the other hand, the associations with quality seem to strengthen for the social-emotional outcomes, and the grade 5 CBCL internalizing and externalizing behavior scores become significant ($p = 0.097$ and 0.072, respectively), with effect sizes increasing in magnitude to 0.09 units of a standard deviation.

Adding the full set of demographic covariates from the EHSREP impact evaluation does not lead to many changes in the estimated effects of quality on the various outcomes. The standard errors increase in some cases, but only by small amounts, suggesting the presence of the large

number of additional variables is offset by the their ability to explain more of the variance (for example, the adjusted R-squared for the model using grade 5 PPVT-III as an outcome increases from 0.215 to 0.259 when expanding the covariates). The most noteworthy change is that the quality effect for the 24-month Bayley MDI becomes larger and statistically significant (d = 0.16).

Adding the ECERS-R overall score from pre-kindergarten and school free and reduced price lunch rate in grade 5 to the models with grade 5 outcomes reduces the effect of quality from the EHS program to the point where it is actually slightly negative (but not significant) for three of the four cognitive/academic outcomes, although the effect on the academic success index remains significant. The individual social-emotional outcomes remain in the expected directions, and the social-emotional success index changes to the expected positive direction, but none are statistically significant. Notably, the sample sizes are much smaller, reflecting the fact that many cases with EHS quality and outcome data do not also have pre-kindergarten quality data, whether because the child was not in an eligible care setting or was not observed by the EHSREP follow-up evaluation.

When adding the interaction term for ITERS/ECERS-R overall score and average weekly center hours, the quality effects remain significant for the outcomes that had significant quality effects in the primary model (36-month PPVT-III and grade 5 academic success index), as indicated by an F-test of joint significance for the main quality term and the interaction term. However, the interaction terms are not significant, and the quality effect on 36-month PPVT-III runs in the opposite direction as expected, by weakening as the number of hours increases. On the other hand, the grade 5 PPVT-III relationship with quality, which was positive but not significant originally, is now significant (for the joint F-test, p = 0.078, and for the interaction term, p = 0.086). For most of the other outcomes, the direction of the quality effect is as

expected, with stronger effects at higher numbers of hours, but remains not significantly different from zero.

Similarly, adding the squared term for ITERS-ECERS-R overall score, to allow the effect of quality to be nonlinear, the quality effects for 36-month PPVT-III and grade 5 academic success index are still significant through the F-test of joint significance of the linear term and squared term. However, in each case the squared term is not significantly different from zero, and as with the interaction with hours, the direction of the quality effects for the 36-month PPVT-III are the opposite of expected, as the estimated effect is smaller at higher levels of quality. For many but not all of the other outcomes, the relationship between quality and the outcome is larger at higher levels of quality, although not significant.

The final rows of each table show the results from two models where quality is modeled by dividing the sample into three groups based on the mean ITERS/ECERS-R overall score, first using the developer thresholds, and then by using the distribution of scores in this sample. The significant relationships between quality and outcomes in the primary model (PPVT-III at 36 months and the academic success index at grade 5) remain significant here. For the 36-month PPVT-III there is only a significant contrast between the low ("inadequate" to "minimal" quality) and high ("good" to "excellent" quality) groups for the developer thresholds (the middle/high contrast is not shown in the table, but $p = 0.168$). However, for the sample distribution, both the low/high (bottom quartile vs. top quartile) and middle/high (the middle two quartiles vs. the top quartile; not shown, $p = 0.022$) contrasts are significant, with only the low/middle contrast not significant. For the academic success index, the developer threshold model was not run because there was no variation in the low group – no children experiencing inadequate to minimal quality (which was a relatively small group) had outcomes meeting the definition for academic success. In the sample distribution model, all three contrasts were significant (for middle/high, not shown, $p = 0.076$).

Note the effect sizes are larger than in other models, although this is mostly an artifact of the new method that was used to calculate them. Because the quality variables are now indicators, the effect size in these models is simply the coefficient for quality divided by the standard deviation of the outcome. This gives the standardized effect in moving from one quality group to another, which could be a change of from 2 to 4 points in the ITERS/ECERS-R overall score. This is a much larger change than moving up a standard deviation in quality, which is about 1 point.

In addition, for the ECLS-K math score, the effect of quality in the developer threshold now becomes significant, judging by an F-test of joint significance of the coefficients for the two quality groups. However, the adjusted mean for the high group ("good" to "excellent" quality) is actually slightly lower than the middle group ("minimal" to "good" quality). In the sample distribution model there are also significant low/high (bottom quartile vs. top quartile) contrasts for the grade 5 PPVT and ECLS-K reading scores, but the overall F-tests reveal a lack of significance ($p = 0.103$ and 0.170, respectively). For other outcomes, especially the social-emotional outcomes, these categorical quality models do not find significant relationships.

In general, the sample distribution categories seem to yield stronger quality relationships with outcomes than the developer threshold categories. This might be because the average scores are relatively high on the scale (from Table 1, the mean ITERS/ECERS-R overall score is almost 5, the threshold for "good" quality care), and few participants received care at the low end of the scale. The sample distribution forces more variation into the groups and for a higher quality level to be reached to be in the middle and high groups.

CHAPTER FIVE

DISCUSSION AND CONCLUSION

This chapter discusses the results of the present study in context. After re-stating the aims of the analysis and summarizing the main findings, it goes into the strengths of the data involved and important limitations that apply when interpreting the results. It then considers several other features and implications of the findings, both in general and for specific sets of results, and then concludes.

Summary of Aims and Findings:

This study explored the relationship between measures of child care quality and child outcomes using data on participants in the Early Head Start Research and Evaluation Project (EHSREP). After starting with the 24-month and 36-month outcomes examined by the original analysis, it looked at longer-term outcomes, when children were in the fifth grade. This analysis also looked at different measures of quality to see if they showed a stronger or otherwise different pattern of relationships with the same child outcomes. Finally, different approaches to modeling the effect of quality were tested.

Interpreting results from such a large set of analyses is challenging. To summarize, in the primary model the relationship between quality and outcomes was generally favorable but not statistically significant, although it was effectively zero in a couple of cases. For two outcomes – a measure of vocabulary at 36 months and an index indicating positive results on the four grade cognitive/academic measures at grade 5 – the results were favorable and statistically significant, with effect sizes that are small but could be meaningful. Alternative measures of quality did not generally show stronger relationships with outcomes, other than the measure that focused more on the physical features and learning activities of the care setting than on caregiver-child interactions. Different specifications of the relationship between quality and outcomes did not find consistent evidence for an alternative to the linear approach used in the primary model.

Strengths:

This study took advantage of the EHSREP's extensive collection of child outcomes and use of different quality measures to run a large number of analyses that varied the quality measure, outcome measure, and the specification of the model connecting them. Several features of the EHSREP data strengthen these analyses. Child outcomes are measured through multiple sources, including parent reports, child self-reports, and trained observers. Quality is assessed using different measures that were also collected by trained observers during the same session, making comparisons easier. The data contains a rich set of family and child characteristics that were collected prior to experiencing the quality of care and can serve as baseline covariates. The longitudinal nature of the study means that data was collected at several different points in time during the program, and the follow-up at grade 5 allows for an examination of longer-term relationships.

Cautions:

Internal validity. The primary limitation of this study is its level of internal validity. Using multivariate regression means that the influence of observable factors can be accounted for, but other, unobserved factors could lead to biased results if they are related to both child outcomes used and quality. Selection effects, where families who obtain higher-quality care for their children are likely to also support them in other ways, mean that observed effects may be wrongly attributed to higher quality care (alternatively, if selection effects run in the other direction, it could mean that effects of higher quality care exist, but are not detected). Additionally, because several requirements had to be met for a child care observation to take place, there could be bias resulting from various levels of nonresponse.

However, there are reasons to believe that these issues are less problematic in this study than in other contexts. As the original analysis observes, most children in the program group who received center-based care received it directly through Early Head Start or through another

provider referred to by the program (ACF, 2004). This was a low-income population whose options for quality care outside of Early Head Start were likely more limited. Analysis of this data also provides some evidence that selection effects are limited. The quality of care children experienced did not seem to be strongly linked with their background characteristics. While most analyses used a relatively small set of demographic characteristics, an alternative model that included the full set of covariates from the EHSREP impact evaluation found similar results. Still, the overarching point is that potentially important factors cannot be accounted for in this design because they are not observable, so there is not a high degree of confidence that higher quality care caused the favorable outcomes that were observed.

Multiple comparisons. One risk inherent in making comparisons across a large number of models is encountering "false positives" (i.e., results with a low p-value that actually occurred by chance). Because a probability threshold of 0.10 was used for determining statistical significance in the current study, every tenth result might be expected to reach that threshold. In fact, two of 17 outcomes, or about 12 percent, showed significant effects in the primary model (although these both had p-values under 0.05). This can be demonstrated by applying the Sidak method for multiple comparison tests, which provides an adjusted significance level above which any individual test might not be different from zero, given the original significance level of the collection of tests (Abdi, 2007). Using an overall significance level of 0.10 and the 17 outcomes run under the primary model, the adjusted significance level for any given outcome drops to 0.0062, which none of the results met (the two significant results had p-values of between 0.01 and 0.02). A less strict application considers groups of outcomes together; for example, if asking whether the set of five cognitive/academic outcomes at grade 5 involve significant effects, the adjusted significance level is now 0.0209, which the p-value for the grade 5 index meets. Similarly, considering the three 36-month outcomes a set, the p-value for the 36-month PPVT-III meets the adjusted significance level.

Still, there are reasons to conclude these results are significant and did not simply occur by chance. They are robust to various specifications and variations in the model, including changes to how missing values are treated, which covariates are included, and how quality is specified. In fact, there were some larger effects of quality in other models, although mainly ones that omitted part of the sample. Also, the fact that most of the outcomes with which quality showed the strongest relationships tended to involve language (the two PPVT-III scores and the ECLS-K reading score) could mean that the quality of care received affected this developmental domain more than other domains. It could also occur if outcomes in the language domain are easier to measure than outcomes in other domains, especially the social-emotional domain. Regardless of the reasons, other work has indeed found larger effects of quality of care on academic and language outcomes compared to social-emotional ones (Burchinal, Kainz, & Cai, 2011).

Generalizability. The results of this analysis should also be understood in terms of the population to which they are most applicable. Given the targets and eligibility requirements for Early Head Start, this sample was a low-income, disadvantaged group of families. Importantly, the sample also consisted of those participants who, in conjunction with the program, chose a significant amount of care in center settings. These findings may not be as generalizable to people with stronger preferences for alternatives to center-based care, including staying at home with their child if possible. Again, however, given the fact that child care choices were expanded due to the resources offered by Early Head Start, many families likely took advantage of child care they would not have been able to afford without the program.

Public-use data. One more limitation is worth mentioning: using the public-use file led to a few shortcomings in the analysis, mostly due to data elements that were not available. Two of the covariates used in the original analysis were not available. Data on average hours in center care from the interviews used to determine which children were observed also were not included, so data from a different set of interviews had to be used, and this data was not as complete or

accurate. These differences could have affected the estimates of the relationship between quality and outcomes. However, the results from this study for the 24-month and 36-month outcomes were similar to results from the original analysis, suggesting that these issues did not greatly affect the estimates of the role of quality.

Context and Implications:

Overall, these results should also be examined through the lens of effect sizes. The relatively small size of the sample used in this analysis limits the inferences that can be made about most of these relationships. Many relationships between quality and outcomes had effect sizes in the 0.07 to 0.10 range, which can be substantively important. If these continued to appear with a larger sample, they would have been significant. The fact that many outcomes (although not all) were in a favorable direction and many had non-trivial effect sizes suggests that the small sample played a factor in these results. To place these in context, the EHSREP experimental evaluation found overall impacts in effect sizes of about 0.10 to 0.20 (ACF, 2002). While the follow-up study had enough power to detect differences of this magnitude, the reduction in sample size meant smaller impacts were not detectable, and the follow-up study found non-significant effect sizes as high as 0.06 on the outcomes used in this analysis. Additionally, a meta-analysis of these types of studies found an average effect size of quality on outcomes of 0.11, or an average partial correlation of 0.12 (Burchinal, Kainz, & Cai, 2011). While not reaching conventional levels of statistical significance, several of the effect sizes found in this analysis come close to this level.

Long-term outcomes. Turning to the results of the different aims of the study and starting with the long-term outcomes, finding a significant, positive result for the grade 5 academic success measure is a little unexpected on its face, given that none of its component measures were significantly related with quality. Still, the component scores tended to have favorable relationships with quality even if they were not significant, and it is also possible that these average effects masked different effects amongst students. For example, if higher quality did not

help all the children but provided a strong boost to the ones it did help – enough for them to reach the success threshold on each measure – then this could explain these results. The EHSREP follow-up study found a similar result: there was a significant treatment/control difference on the social-emotional success index, even though the treatment effect on the individual components was not significant (Vogel et al., 2010).

These relationships with long-term outcomes are important because they show how the children had developed approximately seven years after their center-based care experiences during Early Head Start. Although these children still have many more important years ahead of them, outcomes in elementary school can still be powerful predictors of later achievement (Lesnick et al., 2010). While this portion of the analysis did not find significant effects outside of the academic success index, it found suggestive evidence similar to the evidence found when the children were still in Early Head Start, suggesting that observed quality may have at least a small relationship in their later development. On the one hand, this may not be surprising, given how influential the earliest years of life can be, and the results from previously discussed programs showing lasting effects of high-quality early experiences. On the other hand, because of all the other influences that occur after early child care, and the risks and stresses that disadvantaged children encounter, for some evidence of an effect of quality to remain is striking.

Measures of quality of care. Next, the results for different measures of quality can shed light on whether using measures that focus on different aspects of quality or take different measurement approaches makes a difference. Most of the alternative measures – the ITERS-ECERS-R teaching score, the Arnett CIS score, and the C-COS counts of caregiver talk – were more directly focused on how the caregiver interacted with children. To the extent that this interaction is the most important factor for children's development, then these instruments might be expected to show stronger relationships with outcomes than the global quality represented by the ITERS/ECERS-R overall score. However, the ITERS/ECERS-R provisions score, which

focused more on the physical features and available activities in the setting, showed the strongest connection to outcomes. The Arnett CIS score showed favorable effects at the earliest time period (24 months), but these became mixed or even unfavorable at later time periods. It is possible the scale is better suited to capturing quality at earlier ages, but even if true that would limit its utility.

The C-COS counts of caregiver talk actually showed more unfavorable than favorable relationships. This is especially disappointing since the C-COS focused on the specific child in the study, not the setting as a whole, which might be expected to find a more precise measurement of quality as experienced by the child studied. These negative results appear to be driven by the talk that is initiated by the caregiver. Perhaps caregivers spoke up more often because of "negative" situations, for example to ask the child to stop doing something viewed as inappropriate, even if the child had not first spoken to the caregiver. The more favorable connections with talk involving the caregiver's response to the child show that caregiver-child interaction is still important, and suggests that however the caregiver can achieve this, having the child communicate more often can be beneficial. In fact, the two counts involving favorable results (responding talk and negative child behavior) had more to do with direct child behavior. While the count of negative child behavior seems to blur the line between quality of child care and a child outcome, it points to the potential for measuring quality through different, specific actions and behaviors on the part of caregivers and children.

Alternative specifications. Finally, testing alternative specifications could show whether models of the effect of quality on outcomes should depart from an assumption of a linear relationship. It is difficult to make any strong conclusions from these results. Adding the quality of pre-kindergarten child care meant that neither period's quality of care showed a connection with outcomes. However, because of the response rates at each period of data collection, few children had complete data for this model, meaning the results are likely also affected by the

different sample involved. Adding an interaction term with hours or a squared term for quality led to many effects in the expected direction – that the effects of quality would increase with more hours in care and the effects of quality would increase at higher levels of quality, respectively – but these were not statistically significant. Also, for at least one significant finding they ran in the opposite, unexpected direction. The results for hours may been affected by the need to use less accurate and less complete data. The fact that levels of quality in the EHSREP study were relatively high to begin with may have affected the ability to detect a nonlinear relationship. This fact also may have affected results from the models using categorical levels of quality, as there were few cases of children receiving the lowest levels of care. The slightly stronger relationships found using the sample distribution to create groups, which led to higher thresholds, does suggest that effects on outcomes may be larger at higher levels of quality.

Future Research and Conclusion:

Repeating this analysis using the restricted-use data would allow for more insight in several areas, with the primary benefit being access to the data on average hours in center care from the Parent Interviews. Because data from these interviews determined which families were using eligible care arrangements, a more detailed analysis of potential selection effects would use this data to compare characteristics of those with eligible child care arrangements to those who actually had completed child care observations. Having more accurate and complete information on hours would improve the estimates of the effects involving child care intensity.

Other areas of interest using the restricted-use data include access to more detailed quality measures from the ITERS/ECERS-R scales or C-COS observations. For the former, having item-level data would allow for a factor analysis or using the items found by previous studies to form more accurate teaching and provisions subscores. For the latter, it would be interesting to see if children who hear more requests for language or other communication from their caregivers have better outcomes than children who receive more directives or requests to take a certain action.

Finally, having other information available, particularly site identifiers, would allow for more options for covariates to include in a model.

More broadly, research should continue to understand the degree to which these types of results happen because the effects of quality of care are truly small, as opposed to occurring because of difficulties accurately measuring quality. Given the knowledge of the importance of early development, the large effects of several early care programs, and previous research on measures of quality, it seems likely that improvements in quality measurement could lead to finding stronger relationships between quality and child outcomes (Burchinal, Kainz, & Cai, 2011). For example, the CLASS, with a more targeted approach to quality that focuses on interactions between teachers and children, has found larger relationships with outcomes compared to the global quality measures used in this analysis (Mashburn et al., 2008; Sabol et al., 2013). Of course, it is also possible that, especially in large-scale settings or observational-only studies, that the effects of quality will never be as large and long-lasting (at least on their own) as the effects of programs like Perry Preschool or Abecedarian. Changes in the child care environment and challenges in scaling up from small, intensive programs have to be taken into account.

Even if this is true, and out-of-home care has a small impact compared to family characteristics, quality of care will remain important for policymakers. The related area of primary and secondary (K-12) education is instructive; many programs find effect sizes that seem small, but it is more important to compare them to benchmarks based on empirical knowledge of the field (Hill et al, 2008; Lipsey et al., 2012). Improving the quality of early care could remain important, for example, depending on the relative costs and benefits of the improvements (Duncan and Gibson-Davis, 2006).

Clearly, quality is important, but to what extent remains not fully known, nor does the ideal approach to observe and measure it. More research in this area will inform policymakers and can

help determine the best way to use early care and education policies and programs to protect our nation’s most disadvantaged children.

TABLES

The following terms and abbreviations appear throughout the tables:

ITERS – Infant Toddler Environment Rating Scale

ECERS-R – Early Childhood Environment Rating Scale-Revised

Arnett CIS – Arnett Caregiver Interaction Scale

C-COS – Child-Caregiver Observation System

Bayley MDI – Bayley Scales of Infant Development, Mental Development Index

CDI Vocabulary – MacArthur Communicative Development Inventory, Vocabulary Production

CBCL – Child Behavior Checklist

PPVT-III: Peabody Picture Vocabulary Test-Third Edition

WISC-IV – Wechsler Intelligence Scale for Children-Fourth Edition

ECLS-K – Early Childhood Longitudinal Study-Kindergarten Class of 1998-99

The source of data for all tables is the Early Head Start Research and Evaluation Project Public-Use File. Only data from the EHS program group (those randomly assigned to receive Early Head Start services) is used.

Table 1. Descriptive statistics for Early Head Start program group children observed in center care at least once during 14, 24, and 36 months

	N	Mean	Standard Deviation	Min	Max
Infant-Toddler Child Care Quality					
ITERS overall score (14 months)	273	4.71	1.12	1.5	6.8
ITERS overall score (24 months)	289	4.95	1.08	1.65	6.76
ECERS-R overall score (36 months)	316	4.96	1.11	1.24	6.82
Mean ITERS overall score (14 and 24 months)	372	4.80	1.10	1.58	6.76
Mean ITERS teaching score (14 and 24 months)	372	5.23	1.39	1	7
Mean ITERS provisions score (14 and 24 months)	372	4.55	1.05	1.53	6.75
Mean ITERS/ECERS-R overall score (14, 24, and 36 months)	465	4.85	1.07	1.24	6.79
Mean ITERS/ECERS-R teaching score (14, 24, and 36 months)	465	5.15	1.31	1	7
Mean ITERS/ECERS-R provisions score (14, 24, and 36 months)	465	4.59	1.02	1.25	6.75
Mean Arnett CIS overall score (14 and 24 months)	370	3.37	0.43	1.31	4
Mean Arnett CIS overall score (14, 24, and 36 months)	461	3.38	0.44	1.31	4
C-COS count of any caregiver talk (24 months)	273	33.6	11.8	8	60
C-COS count of caregiver initiated talk (24 mo.)	273	25.7	11.0	0	55
C-COS count of caregiver-responding talk (24 mo.)	273	8.5	9.5	0	48
C-COS count of negative child behavior (24 mo.)	273	5.7	6.4	0	50
Mean C-COS count of any caregiver talk (24 and 36 months)	385	29.9	11.2	2	58
Mean C-COS count of caregiver-initiated talk (24 and 36 months)	385	22.4	9.8	0	55.5
Mean C-COS count of caregiver-responding talk (24 and 36 months)	385	8.0	7.7	0	46
Mean C-COS count of negative child behavior (24 to 36 months)	385	4.8	4.9	0	36

	N	Mean	Standard Deviation	Min	Max
Child-adult ratio (14 months)	287	2.92	1.35	0.5	10.83
Child-adult ratio (24 months)	309	3.52	1.60	1	11.58
Child-adult ratio (36 months)	319	5.47	2.64	0.78	14.83
Mean child-adult ratio (14 and 24 months)	389	3.30	1.42	0.93	11.58
Mean child-adult ratio (14, 24, and 36 months)	463	4.13	1.83	0.78	13.1
Mean ITERS overall score is inadequate to minimal (14 and 24 months)	372	0.081		0	1
Mean ITERS overall score is minimal to good (14 and 24 months)	372	0.476		0	1
Mean ITERS overall score is good to excellent (14 and 24 months)	372	0.444		0	1
Mean ITERS/ECERS-R overall score is inadequate to minimal (14, 24, and 36 months)	465	0.067		0	1
Mean ITERS/ECERS-R overall score is minimal to good (14, 24, and 36 months)	465	0.465		0	1
Mean ITERS/ECERS-R overall score is good to excellent (14, 24, and 36 months)	465	0.469		0	1
Mean ITERS overall score is in bottom quartile (14 and 24 months)	372	0.25		0	1
Mean ITERS overall score is in middle quartiles (14 and 24 months)	372	0.495		0	1
Mean ITERS overall score is in top quartile (14 and 24 months)	372	0.251		0	1
Mean ITERS/ECERS-R overall score is in bottom quartile (14, 24, and 36 months)	465	0.245		0	1
Mean ITERS/ECERS-R overall score is in middle two quartiles (14, 24, and 36 months)	465	0.503		0	1
Mean ITERS/ECERS-R overall score is in top quartile (14, 24, and 36 months)	465	0.252		0	1

	N	Mean	Standard Deviation	Min	Max
Child Care Intensity					
Average weekly center hours (14 months)	328	28.3	25.7	0	130
Average weekly center hours (24 months)	286	25.2	22.1	0	85
Average weekly center hours (36 months)	318	29.9	15.9	0	60
Mean of average weekly center hours (14 and 24 months)	336	26.5	22.2	0	105
Mean of average weekly center hours (14, 24, and 36 months)	369	28.0	17.5	0	86.7
Child Outcomes					
24-Month Bayley MDI	366	90.8	13.0	49	118
24-Month CDI Vocabulary	415	57.1	23.2	3	100
24-Month CBCL Aggressive Behavior	418	12.4	6.7	0	34
36-Month Bayley MDI	356	91.9	11.8	51	121
36-Month PPVT-III	335	84.5	15.0	40	125
36-Month CBCL Aggressive Behavior	409	10.7	6.3	0	36
Grade 5 PPVT-III	294	94.4	14.7	55	133
Grade 5 WISC-IV Matrix Reasoning	296	8.3	3.3	1	19
Grade 5 ECLS-K mathematics	296	8.0	4.5	0	17
Grade 5 ECLS-K reading	299	127.3	28.9	31.8	174.5
Grade 5 academic success index	294	0.143		0	1
Grade 5 CBCL Internalizing Behavior	309	5.9	6.0	0	30.2
Grade 5 CBCL Externalizing Behavior	309	8.3	8.0	0	42
Grade 5 CBCL Attention Problems	309	4.1	4.0	0	16
Grade 5 self-reported delinquent behavior	298	1.4	1.7	0	10
Grade 5 self-reported bullying by peers	296	6.9	2.8	4	16
Grade 5 social-emotional success index	293	0.430		0	1

	N	Mean	Standard Deviation	Min	Max
Family and Child Characteristics (subset used in primary model)					
Child gender is male	465	0.499		0	1
Mother race/ethnicity is White	461	0.330		0	1
Mother race/ethnicity is African-American	461	0.416		0	1
Mother race/ethnicity is Hispanic	461	0.226		0	1
Mother has other race/ethnicity	461	0.028		0	1
Mother education is less than high school	451	0.412		0	1
Mother education is high school or GED	451	0.284		0	1
Mother education is more than high school	451	0.304		0	1
Mother lives with a spouse	465	0.181		0	1
Mother lives with other adults	465	0.422		0	1
Mother lives with no other adults	465	0.398		0	1
Mother was less than 19 years when child born	463	0.300		0	1
Family and Child Characteristics (additional set used in alternative model)					
Mother was less than 20 years old when child born	463	0.389		0	1
Mother was 20 to 24 years old when child born	463	0.272		0	1
Mother was 25 years old or older when child born	463	0.339		0	1
Mother primary language is English	459	0.847		0	1
Mother primary language not English but speaks English well	459	0.081		0	1
Mother primary language not English and does not speak English well	459	0.072		0	1
Mother is employed	453	0.296		0	1
Mother is in school or training	453	0.263		0	1
Mother is unemployed or not in labor force	453	0.442		0	1

	N	Mean	Standard Deviation	Min	Max
Number of other children aged 0 to 5 in household	465	0.4	0.6	0	4
Number of other children aged 6 to 17 in household	465	0.6	0.9	0	4
Household income is less than 33 percent of poverty level	465	0.228		0	1
Household income is 33 to 67 percent of poverty level	465	0.284		0	1
Household income is 67 to 99 percent of poverty level	465	0.202		0	1
Household income is 100 percent or more of poverty level	465	0.114		0	1
Household income is missing	465	0.172		0	1
Household receives AFDC/TANF	450	0.273		0	1
Household receives food stamps	445	0.474		0	1
Household receives WIC	445	0.861		0	1
Household receives SSI	445	0.083		0	1
Household has inadequate resources for food	463	0.039		0	1
Household has inadequate resources for housing	461	0.113		0	1
Household has inadequate money to buy necessities	454	0.185		0	1
Household has inadequate resources for medical needs	441	0.093		0	1
Household has inadequate resources for transportation	449	0.174		0	1
Mother previously enrolled in Head Start or another child development program	446	0.164		0	1
Number of times household has moved in previous year	436	0.8	1.1	0	4
Random assignment was before 10/96	465	0.277		0	1
Random assignment was between 10/96 and 6/97	465	0.262		0	1
Random assignment was after 6/97	465	0.460		0	1

	N	Mean	Standard Deviation	Min	Max
Child was unborn at random assignment	465	0.183		0	1
Child was age 0 to under 5 months at random assignment	465	0.338		0	1
Child was age 5 months or older at random assignment	465	0.480		0	1
Child had low birthweight (less than 2500 grams)	442	0.072		0	1
Child born early (more than 3 weeks)	460	0.115		0	1
Child was evaluated about health and development	461	0.046		0	1
Child had established, bio/medical, or environmental risk	465	0.338		0	1
Preschool Child Care Quality and Grade 5 School Characteristics					
ECERS-R overall score (pre-kindergarten)	234	5.07	1.27	1.55	7
Arnett CIS overall score (pre-kindergarten)	227	3.35	0.54	1.04	4
Percent of students at child's school eligible for free or reduced price lunch (grade 5)	287	61.5	23.5	1.29	99.4

Note: Sample includes all children in the EHS program group who had at least one ITERS or ECERS-R overall score during 14, 24, or 36 months. N = 465, smaller counts for individual items indicate missing data. Statistics do not include imputed values.

AFDC/TANF – Aid to Families with Dependent Children/Temporary Assistance for Needy Families.
WIC – Special Supplemental Nutrition Program for Women, Infants, and Children.
SSI – Supplemental Security Income.

Table 2. Correlations between measures of quality and intensity

	ITERS/ ECERS-R overall score	ITERS/ ECERS-R teaching score	ITERS/ ECERS-R provisions score	Arnett CIS score	C-COS count of any talk	C-COS count of initiated talk	C-COS count of respond. talk	C-COS count of negative behavior	Child-adult ratio	Hours in center care
ITERS/ECERS-R overall score	1									
ITERS/ECERS-R teaching score	0.90*	1								
ITERS/ECERS-R provisions score	0.95*	0.78*	1							
Arnett CIS score	0.71*	0.80*	0.61*	1						
C-COS count of any talk	0.05	0.15*	-0.02	0.17*	1					
C-COS count of initiated talk	<0.01	0.01	-0.01	0.02	0.74*	1				
C-COS count of responding talk	0.08	0.20*	-0.01	0.21*	0.49*	-0.21*	1			
C-COS count of negative behavior	-0.31*	-0.35*	-0.26*	-0.32*	-0.12*	-0.01	-0.17*	1		
Child-adult ratio	-0.15*	-0.21*	-0.10*	-0.22*	-0.23*	-0.07*	-0.24*	0.13*	1	
Hours in center care	-0.01	0.02	-0.05	-0.06	0.18*	0.16*	0.06	0.02	0.07	1

Note: Each cell lists the pairwise correlation between the two measures for all observations, regardless of whether any of the child outcome variables were present. N ranges from 315 to 467, depending on the pair.

* $p < 0.05$

Table 3. Correlations between quality (ITERS/ECERS-R overall score) and outcomes

	24-Months			36-Months		
	Bayley MDI	CDI Vocabulary	CBCL Aggressive Behavior	Bayley MDI	PPVT-III	CBCL Aggressive Behavior
ITERS/ECERS-R overall score	0.051	0.021	-0.089	0.025	0.099*	-0.014
	Grade 5 Cognitive/Academic					
	PPVT-III	WISC-IV Matrix Reasoning	ECLS-K math	ECLS-K reading		Academic success index
ITERS/ECERS-R overall score	0.059	-0.008	0.009	0.078		0.106*
	Grade 5 Social-Emotional					
	CBCL Internal-izing Behavior	CBCL External-izing Behavior	CBCL Attention Problems	Self-reported delinquent behavior	Self-reported bullying by peers	Social-emotional success index
ITERS/ECERS-R overall score	-0.075	-0.100*	-0.043	-0.053	-0.068	0.006

Note: The mean ITERS overall score over 14 and 24 months is used for the 24-month outcomes, and the mean ITERS/ECERS-R overall score over 14, 24, and 36 months is used for the 36-month and grade 5 outcomes.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4. Characteristics of sample with and without quality data

	EHS program group	Eligible center care setting	Quality and 36-month outcomes	Quality and 36-month outcomes and hours	Quality and grade 5 outcomes
Key Explanatory Variables					
Mean ITERS/ECERS-R overall score (14, 24, and 36 months)	4.85 (1.07) n = 465	4.85 (1.07) n = 465	4.82 (1.00 n = 335	4.81 (1.04) n = 280	4.84 (1.02) n = 294
Mean ITERS/ECERS-R teaching score (14, 24, and 36 months)	5.15 (1.31) n = 465	5.15 (1.31) n = 465	5.11 (1.25) n = 335	5.10 (1.29) n = 280	5.17 (1.26) n = 294
Mean ITERS/ECERS-R provisions score (14, 24, and 36 months)	4.59 (1.02) n = 465	4.59 (1.02) n = 465	4.55 (0.96) n = 335	4.53 (0.99) n = 280	4.57 (1.00) n = 294
Mean Arnett CIS overall score (14, 24, and 36 months)	3.38 (0.44) n = 463	3.38 (0.44) n = 463	3.39 (0.42) n = 331	3.39 (0.44) n = 276	3.41 (0.41) n = 291
Mean C-COS count of any talk (24 and 36 months)	29.9 (11.2) n = 387	29.9 (11.2) n = 387	30.2 (10.9) n = 292	30.9 (11.1) n = 243	29.9 (11.0) n = 251
Mean child-adult ratio (14, 24, and 36 months)	3.99 (1.92) n = 589	4.12 (1.84) n = 498	4.12 (1.75) n = 333	4.13 (1.74) n = 278	4.18 (1.89) n = 292
Mean of average weekly center hours (14, 24, and 36 months)	13.4 (17.8) n = 1029	26.6 (16.8) n = 518	28.5 (17.0) n = 280	28.5 (17.0) n = 280	28.2 (17.5) n = 247
Child Outcomes					
24-Month Bayley MDI	89.8 (13.8) n = 930	90.8 (12.9) n = 461	91.6 (13.0) n = 297	91.0 (13.5) n = 221	n/a
24-Month CDI Vocabulary	55.9 (23.3) n = 1076	57.2 (22.9) n = 538	57.4 (23.5) n = 337	56.6 (23.5) n = 246	n/a
24-Month CBCL Aggressive Behavior	12.3 (6.8) n = 1090	12.4 (6.6) n = 540	12.2 (6.5) n = 340	12.4 (6.4) n = 246	n/a

	EHS program group	Eligible center care setting	Quality and 36-month outcomes	Quality and 36-month outcomes and hours	Quality and grade 5 outcomes
36-Month Bayley MDI	91.2 (12.4) n = 879	91.7 (11.9) n = 452	91.9 (11.8) n = 356	92.3 (11.7) n = 303	n/a
36-Month PPVT-III	83.6 (14.9) n = 752	84.2 (14.7) n = 417	84.5 (15.0) n = 335	84.6 (14.7) n = 280	n/a
36-Month CBCL Aggressive Behavior	10.9 (6.5) n = 1069	10.7 (6.4) n = 545	10.7 (6.3) n = 409	10.6 (6.4) n = 346	n/a
Grade 5 PPVT-III	94.1 (15.3) n = 796	94.8 (14.5) n = 379	n/a	n/a	94.4 (14.7) n = 294
Grade 5 WISC-IV Matrix Reasoning	8.5 (3.3) n = 803	8.4 (3.2) n = 381	n/a	n/a	8.3 (3.3) n = 296
Grade 5 ECLS-K mathematics	8.2 (4.6) n = 803	8.1 (4.6) n = 381	n/a	n/a	8.0 (4.5) n = 296
Grade 5 ECLS-K reading	128.1 (28.1) n = 805	128.0 (27.9) n = 384	n/a	n/a	127.3 (28.9) n = 299
Grade 5 academic success index	0.155 n = 800	0.135 n = 379	n/a	n/a	0.143 n = 294
Grade 5 CBCL Internalizing Behavior	5.7 (5.7) n = 835	5.8 (6.0) n = 397	n/a	n/a	5.9 (6.0) n = 309
Grade 5 CBCL Externalizing Behavior	7.8 (7.8) n = 835	8.2 (7.9) n = 397	n/a	n/a	8.3 (8.0) n = 309
Grade 5 CBCL Attention Problems	4.0 (3.7) n = 835	4.1 (3.9) n = 397	n/a	n/a	4.1 (4.0) n = 309

	EHS program group	Eligible center care setting	Quality and 36-month outcomes	Quality and 36-month outcomes and hours	Quality and grade 5 outcomes
Grade 5 self-reported delinquent behavior	1.5 (1.8) n = 802	1.5 (1.7) n = 382	n/a	n/a	1.4 (1.7) n = 298
Grade 5 self-reported bullying by peers	6.7 (2.7) n = 801	6.8 (2.7) n = 381	n/a	n/a	6.9 (2.8) n = 296
Grade 5 social-emotional success index	0.454 n = 793	0.440 n = 377	n/a	n/a	0.430 n = 293
Family and Child Characteristics (subset from primary model)					
Child gender is male	0.516 n = 1490	0.504 n = 619	0.493 n = 335	0.493 n = 280	0.514 n = 294
Mother race/ethnicity is White	0.371 n = 1483	0.340 n = 614	0.381 n = 331	0.388 n = 276	0.336 n = 292
Mother race/ethnicity is African-American	0.343 n = 1483	0.415 n = 614	0.417 n = 331	0.446 n = 276	0.401 n = 292
Mother race/ethnicity is Hispanic	0.239 n = 1483	0.210 n = 614	0.172 n = 331	0.145 n = 276	0.233 n = 292
Mother has other race/ethnicity	0.047 n = 1483	0.034 n = 614	0.030 n = 331	0.022 n = 276	0.031 n = 292
Mother education is less than high school	0.477 n =1454	0.416 n = 601	0.362 n = 323	0.351 n = 268	0.378 n = 286
Mother education is high school or GED	0.273 n = 1454	0.283 n = 601	0.307 n = 323	0.332 n = 268	0.308 n = 286
Mother education is more than high school	0.250 n = 1454	0.301 n = 601	0.331 n = 323	0.317 n = 268	0.315 n = 286
Mother lives with a spouse	0.250 n = 1503	0.179 n = 619	0.149 n = 335	0.154 n = 280	0.180 n = 294
Mother lives with other adults	0.385 n = 1503	0.415 n = 619	0.439 n = 335	0.432 n = 280	0.412 n = 294
Mother lives with no other adults	0.365 n = 1503	0.405 n = 619	0.412 n = 335	0.414 n = 280	0.408 n = 294
Mother was less than 19 years when child born	0.290 n = 1503	0.300 n = 617	0.294 n = 333	0.301 n = 279	0.291 n = 292

	EHS program group	Eligible center care setting	Quality and 36-month outcomes	Quality and 36-month outcomes and hours	Quality and grade 5 outcomes
N	1503	619	335	280	294

Note: Each cell presents the mean in the first row, followed by the standard deviation in parentheses, followed by the sample size (n). Standard deviations are not reported for indicator variables.

"EHS program group" consists of all families randomly assigned to treatment status for the EHS evaluation. From this group, "eligible child care setting" consists of all families whose child was in center care at some point during the program, defined as having a non-missing ITERS/ECERS-R overall score at 14, 24, or 36 months, or the parent reported the child was in center care at least once over that time but no observations were conducted.

From this group, "quality and 36-month outcomes" consists of all families where quality data (the mean ITERS/ECERS-R overall score at 14, 24, and 36 months) and short-term outcome data are present. From this group, "quality and 36-month outcomes and hours" consists of all families who also have non-missing intensity data (average weekly center hours at 14, 24, and 36 months).

From the "eligible center care setting" group, "quality and grade 5 outcomes" consists of all families where quality data and grade 5 outcomes are present.

For the rows with explanatory variables and family and child characteristics, the outcome used in the 36-month columns is the PPVT-III, while for the grade 5 column it is the grade 5 PPVT-III. For the rows with child outcomes, the outcome used is the outcome of that row.

Table 5. Regression of quality measures on child and family characteristics from primary model

	ITERS/ ECERS-R overall score	ITERS/ ECERS-R teaching score	ITERS/ ECERS-R provisions score	Arnett CIS score	C-COS count of any talk
Male child	0.08	-0.03	0.11	0.01	2.97***
Black	0.19	0.20	0.20	-0.01	0.58
Hispanic	0.17	0.07	0.31***	-0.02	-5.98***
Other race/ethnicity	-0.17	-0.71*	0.19	-0.10	-9.99***
High school	-0.01	-0.17	0.05	-0.01	0.23
More than high school	0.11	0.01	0.16	0.03	-0.94
Lives with spouse	0.49***	0.68***	0.38***	0.15***	-0.10
Lives with other adult	0.06	0.21	0.02	0.05	2.42**
Less than 19 at child birth	-0.03	0.01	-0.01	0.08	-1.59
Constant	4.56	4.91	4.25	3.30	29.47
R^2	0.04	0.05	0.04	0.02	0.10
N	448	448	448	446	378

Note: Each column presents results from a model regressing the measure of quality on the child and family characteristics from the primary model. Each row contains the coefficient for the variable in the row.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6-A. Primary regression model for 24- and 36-month outcomes

	24-Months			36-Months		
	Bayley MDI	CDI Vocabulary	CBCL Aggressive Behavior	Bayley MDI	PPVT-III	CBCL Aggressive Behavior
ITERS/ECERS-R overall score	1.19 (0.76) d = 0.10	0.13 (1.14) d = 0.01	-0.22 (0.33) d = -0.04	0.47 (0.61) d = 0.04	1.84** (0.76) d = 0.12	0.05 (0.28) d = 0.01
Child-adult ratio	0.35 (0.71) d = 0.03	-0.65 (0.88) d = -0.04	0.46* (0.27) d = 0.10	0.48 (0.34) d = 0.07	0.27 (0.40) d = 0.03	-0.13 (0.15) d = -0.04
Hours in center care	0.02 (0.05) d = 0.04	-0.03 (0.07) d = -0.03	-0.05*** (0.02) d = -0.17	0.02 (0.04) d = 0.03	-0.02 (0.05) d = -0.03	-0.01 (0.02) d = -0.03
Hours are missing	2.03 (2.15)	1.29 (3.72)	-2.27** (1.00)	-2.96 (2.16)	-2.19 (2.51)	0.49 (1.00)
Male child	-1.77 (1.46)	-7.65*** (2.54)	-0.31 (0.69)	-2.16* (1.17)	-4.06*** (1.55)	1.17* (0.62)
Black	-2.58 (1.83)	-5.57* (2.98)	-0.37 (0.81)	-7.30*** (1.45)	-6.60*** (1.73)	-2.41*** (0.77)
Hispanic	4.95** (2.14)	-1.09 (3.84)	0.17 (1.06)	-0.79 (1.68)	-3.68 (2.63)	-2.36*** (0.85)
Other race/ethnicity	-0.78 (2.51)	-1.61 (7.06)	-2.61* (1.37)	-2.44 (4.46)	1.60 (3.61)	-0.78 (1.61)
High school	4.04** (1.90)	0.94 (3.49)	-0.58 (0.95)	3.42** (1.65)	3.51 (2.25)	1.28 (0.87)
More than high school	7.62*** (2.08)	9.17** (3.73)	-0.48 (0.99)	6.17*** (1.65)	11.15*** (2.44)	-0.36 (0.87)
Lives with spouse	-0.11 (2.35)	-0.31 (4.20)	-1.46 (1.02)	2.24 (1.88)	-1.85 (2.60)	-0.36 (0.88)
Lives with other adults	-0.88 (1.78)	-2.24 (2.69)	0.26 (0.83)	-2.30 (1.41)	-0.79 (1.75)	-0.63 (0.77)
Less than 19 at child birth	3.49 (1.98)	8.07** (3.39)	-0.42 (1.03)	4.31*** (1.58)	2.39 (2.13)	0.22 (0.91)
R^2	0.09	0.07	0.05	0.15	0.15	0.06
N	295	335	334	354	333	407

Note: Each column presents results from a model regressing the outcome on the variables in each row. Cells present coefficients, standard errors in parentheses, and for key variables, effect sizes (d). For key variables, means over 14 and 24 months are used for 24-month outcomes, and means over 14, 24, and 36 months are used for 36-month outcomes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6-B. Primary regression model for grade 5 cognitive/academic outcomes

	PPVT-III	WISC-IV Matrix Reasoning	ECLS-K math	ECLS-K reading	Academic success index
ITERS/ECERS-R overall score	1.00 (0.82) d = 0.07	0.01 (0.19) d = 0.00	0.16 (0.25) d = 0.04	2.23 (1.67) d = 0.08	x1.67** (0.33) ame= 0.05
Child-adult ratio	-0.45 (0.43) d = -0.06	-0.02 (0.11) d = -0.01	0.06 (0.15) d = 0.02	-0.74 (0.88) d = -0.05	x1.19* (0.11) ame= 0.02
Hours in center care	0.06 (0.05) d = 0.08	0.02 (0.01) d = 0.09	0.01 (0.01) d = 0.04	0.16 (0.11) d = 0.11	x1.00 (0.01) ame= 0.00
Hours are missing	0.20 (2.92)	-0.50 (0.63)	-0.68 (0.88)	-5.83 (5.57)	x0.56 (0.36)
Male child	2.11 (1.50)	0.01 (0.37)	0.59 (0.50)	-2.08 (3.16)	x1.55 (0.57)
Black	-10.86*** (1.84)	-0.69 (0.46)	-2.49*** (0.61)	-3.28 (3.87)	x0.27*** (0.12)
Hispanic	-4.85** (2.35)	-0.09 (0.51)	-0.55 (0.73)	-1.21 (4.52)	x0.34** (0.18)
Other race/ethnicity	-5.28 (5.43)	0.99 (1.04)	2.76** (1.19)	11.94 (7.27)	x0.42 (0.40)
High school	3.90** (1.93)	1.05** (0.54)	1.19* (0.68)	9.66** (4.27)	x2.69 (1.64)
More than high school	12.21*** (2.37)	2.21*** (0.56)	2.82*** (0.72)	21.28*** (4.85)	x9.54*** (5.66)
Lives with spouse	0.91 (2.17)	0.89 (0.59)	1.23 (0.77)	5.40 (4.84)	x2.62** (1.27)
Lives with other adults	-1.72 (2.07)	0.19 (0.43)	-0.65 (0.61)	-4.95 (4.05)	x1.81 (0.94)
Less than 19 at child birth	1.93 (2.08)	0.74 (0.52)	1.80*** (0.67)	10.05** (4.19)	x1.89 (1.11)
R^2	0.25	0.09	0.15	0.13	0.16
N	292	294	294	297	292

Note: Each column presents results from a model regressing the outcome on the variables in each row. Cells present coefficients, standard errors in parentheses, and for key variables, effect sizes (d). The success index uses a logistic regression model; cells present odds ratios (x) and their standard errors, and average marginal effects (ame). For key variables, means over 14, 24, and 36 months are used for all outcomes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6-C. Primary regression model for grade 5 social-emotional outcomes

	CBCL Internal-izing Behavior	CBCL External-izing Behavior	CBCL Attention Problems	Self-reported delinquent behavior	Self-reported bullying by peers	Social-emotional success index
ITERS/ECERS-R overall score	-0.37 (0.30) d = -0.06	-0.43 (0.40) d = -0.05	-0.00 (0.21) d = -0.00	-0.04 (0.09) d = -0.02	-0.12 (0.15) d = -0.04	x0.94 (0.12) ame= -0.01
Child-adult ratio	-0.33** (0.15) d = -0.10	-0.07 (0.21) d = -0.02	-0.02 (0.10) d = -0.01	-0.03 (0.05) d = -0.03	0.13 (0.10) d = 0.09	x0.99 (0.06) ame= 0.00
Hours in center care	0.00 (0.02) d = -0.01	0.00 (0.03) d = 0.01	-0.00 (0.01) d = -0.02	0.00 (0.01) d = 0.03	-0.02 (0.01) d = -0.10	x0.99 (0.01) ame= 0.00
Hours are missing	-0.22 (1.04)	1.48 (1.55)	0.75 (0.74)	0.01 (0.33)	-0.58 (0.52)	x0.43** (0.18)
Male child	0.05 (0.67)	2.29*** (0.88)	1.27*** (0.45)	0.99*** (0.19)	-0.45 (0.33)	x0.71 (0.18)
Black	-3.91*** (0.81)	-4.67*** (1.13)	-2.19*** (0.56)	-0.70*** (0.24)	-1.21*** (0.39)	x2.94*** (0.92)
Hispanic	-1.80* (1.04)	-3.81*** 1.31	-2.19*** (0.61)	-0.51* (0.31)	-0.71 (0.48)	x2.64*** (0.97)
Other race/ethnicity	-2.74 (1.71)	-1.48 (2.27)	-1.14 (1.49)	-0.55 (0.43)	-0.35 1.05	x3.39* (2.50)
High school	-0.23 (0.92)	-0.26 (1.14)	-0.03 0.57	-0.16 (0.27)	-0.45 (0.41)	x1.37 (0.46)
More than high school	0.92 (1.08)	0.04 (1.23)	0.03 (0.63)	-0.31 (0.29)	-0.70 (0.46)	x1.45 (0.52)
Lives with spouse	-1.94* (1.08)	-2.61** (1.27)	-0.68 (0.66)	-0.19 (0.27)	-0.70 (0.44)	x1.71 (0.64)
Lives with other adults	-1.55* (0.81)	-1.29 (1.06)	-0.24 (0.53)	-0.02 (0.23)	-0.39 (0.43)	x1.95** (0.59)
Less than 19 at child birth	-0.22 (0.91)	-0.65 (1.14)	-0.31 (0.56)	-0.10 (0.28)	-0.05 (0.50)	x1.01 (0.35)
R^2	0.12	0.12	0.11	0.12	0.07	0.07
N	307	307	307	296	294	291

Note: Each column presents results from a model regressing the outcome on the variables in each row. Cells present coefficients, standard errors in parentheses, and for key variables, effect sizes (d). The success index uses a logistic regression model; cells present odds ratios (x) and their standard errors, and average marginal effects (ame). For key variables, means over 14, 24, and 36 months are used for all outcomes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7-A. Alternative quality measures for 24- and 36-month outcomes

	24-Months			36-Months		
	Bayley MDI	CDI Vocabulary	CBCL Aggressive Behavior	Bayley MDI	PPVT-III	CBCL Aggressive Behavior
ITERS/ECERS-R overall score	1.19 (0.76) d = 0.10 N = 295	0.13 (1.14) d = 0.01 N = 335	-0.22 (0.33) d = -0.04 N = 334	0.47 (0.61) d = 0.04 N = 354	1.84** (0.76) d = 0.12 N = 333	0.05 (0.28) d = 0.01 N = 407
ITERS/ECERS-R teaching score	0.76 (0.64) d = 0.08 N = 295	0.08 (0.99) d = 0.00 N = 335	-0.20 (0.27) d = -0.04 N = 334	0.10 (0.51) d = 0.01 N = 354	1.35** (0.64) d = 0.11 N = 333	-0.10 (0.24) d = -0.02 N = 407
ITERS/ECERS-R provisions score	1.28* (0.78) d = 0.10 N = 295	0.11 (1.20) d = 0.00 N = 335	-0.23 (0.35) d = -0.04 N = 334	0.94 (0.61) d = 0.08 N = 354	2.21*** (0.81) d = 0.14 N = 333	0.02 (0.31) d = 0.00 N = 407
Arnett CIS overall score	2.53 (2.04) d = 0.08 N = 295	6.43** (3.00) d = 0.12 N = 336	-1.44* (0.84) d = -0.10 N = 335	-1.31 (1.65) d = -0.05 N = 353	2.65 (2.15) d = 0.07 N = 332	-0.21 (0.74) d = 0.01 N = 406
C-COS count of any caregiver talk	-0.09 (0.08) d = -0.08 N = 239	0.09 (0.13) d = 0.05 N = 264	-0.02 (0.04) d = -0.04 N = 261	-0.16*** (0.06) d = -0.16 N = 311	-0.09 (0.08) d = -0.07 N = 292	-0.02 (0.04) d = -0.03 N = 352
C-COS count of caregiver-initiated talk	-0.17** (0.08) d = -0.15 N = 239	-0.21 (0.14) d = -0.10 N = 264	-0.07 (0.04) d = -0.11 N = 261	-0.18*** (0.06) d = -0.16 N = 311	-0.05 (0.09) d = -0.03 N = 292	-0.04 (0.04) d = -0.07 N = 352
C-COS count of caregiver-responding talk	0.15* (0.09) d = 0.10 N = 239	0.45*** (0.16) d = 0.17 N = 264	0.05 (0.06) d = 0.07 N = 261	0.02 (0.09) d = 0.01 N = 311	-0.09 (0.13) d = -0.04 N = 292	0.02 (0.05) d = (0.02) N = 352
C-COS count of child negative behavior	-0.08 (0.13 d = -0.04 N = 239	-0.19 (0.19) d = -0.05 N = 264	0.14** (0.07) d = 0.14 N = 261	-0.24 (0.17) d = -0.10 N = 311	-0.20 (0.19) d = -0.06 N = 292	0.18*** (0.06) d = 0.14 N = 352

Note: Each cell presents the estimate of the partial effect of the quality measure in the row on the outcome in the column, using the primary model from Table 6-A and replacing the ITERS/ECERS-R overall score with the quality measure in the row. Cells present coefficients, standard errors in parentheses, effect sizes (d), and model sample sizes (N). For ITERS/ECERS-R and Arnett scores, means over 14 and 24 months are used for 24-month outcomes, and means over 14, 24, and 36 months are used for 36-month outcomes. For C-COS counts, counts from 24 months are used for 24-month outcomes and means over 24 and 36 months are used for 36-month outcomes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7-B. Alternative quality measures for grade 5 cognitive/academic outcomes

	PPVT-III	WISC-IV Matrix Reasoning	ECLS-K math	ECLS-K reading	Academic success index
ITERS/ECERS-R overall score	1.00 (0.82) d = 0.07 N = 292	0.01 (0.19) d = 0.00 N = 294	0.16 (0.25) d = 0.04 N = 294	2.23 (1.67) d = 0.08 N = 297	x1.67** (0.33) ame= 0.05 N = 292
ITERS/ECERS-R teaching score	0.29 (0.68) d = 0.02 N = 292	-0.01 (0.16) d = -0.00 N = 294	-0.05 (0.21) d = -0.01 N = 294	1.08 (1.48) d = 0.05 N = 297	x1.30 (0.22) ame= 0.03 N = 292
ITERS/ECERS-R provisions score	1.63* (0.85) d = 0.11 N = 292	0.14 (0.19) d = 0.04 N = 294	0.29 (0.26) d = 0.06 N = 294	2.85 (1.76) d = 0.10 N = 297	x1.80** (0.42) ame= 0.06 N = 292
Arnett CIS overall score	-0.13 (2.20) d = -0.00 N = 292	-0.25 (0.49) d = -0.03 N = 294	-1.51** (0.67) d = -0.13 N = 294	-3.87 (5.27) d = -0.05 N = 297	x0.66 (0.31) ame= -0.04 N = 292
C-COS count of any caregiver talk	-0.08 (0.08) d = -0.06 N = 251	0.00 (0.02) d = 0.00 N = 253	-0.02 (0.03) d = -0.06 N = 253	-0.13 (0.16) d = -0.05 N = 255	x0.97* (0.02) ame= -0.00 N = 252
C-COS count of caregiver-initiated talk	-0.14 (0.09) d = -0.08 N = 251	-0.03 (0.02) d = -0.07 N = 253	-0.03 (0.03) d = -0.05 N = 253	-0.35* (0.20) d = -0.11 N = 255	x0.97 (0.02) ame= -0.00 N = 252
C-COS count of caregiver-responding talk	0.05 (0.10) d = 0.03 N = 251	0.04* (0.02) d = 0.10 N = 253	-0.00 (0.03) d = -0.00 N = 253	0.30 (0.23) d = 0.08 N = 255	x0.99 (0.03) ame= -0.00 N = 252
C-COS count of child negative behavior	-0.31 (0.19) d = -0.10 N = 251	-0.12*** (0.04) d = -0.17 N = 253	-0.13** (0.05) d = -0.13 N = 253	-0.77** (0.39) d = -0.12 N = 255	x0.99 (0.04) ame= -0.00 N = 252

Note: Each cell presents the estimate of the partial effect of the quality measure in the row on the outcome in the column, using the primary model from Table 6-B and replacing the ITERS/ECERS-R overall score with the quality measure in the row. Cells present coefficients, standard errors in parentheses, effect sizes (d), and model sample sizes (N). The success index uses a logistic regression model; cells present odds ratios (x) and their standard errors, and average marginal effects (ame). For ITERS/ECERS-R and Arnett scores, means over 14, 24, and 36 months are used for all outcomes. For C-COS counts, means over 24 and 36 months are used for all outcomes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7-C. Alternative quality measures for grade 5 social-emotional outcomes

	CBCL Internal-izing Behavior	CBCL External-izing Behavior	CBCL Attention Problems	Self-reported delinquent behavior	Self-reported bullying by peers	Social-emotional success index
ITERS/ECERS-R overall score	-0.37 (0.30) d = -0.06 N = 307	-0.43 (0.40) d = -0.05 N = 307	-0.00 (0.21) d = -0.00 N = 307	-0.04 (0.09) d = -0.02 N = 296	-0.12 (0.15) d = -0.04 N = 294	x0.94 (0.12) ame= -0.01 N = 291
ITERS/ECERS-R teaching score	-0.27 (0.25) d = -0.06 N = 307	-0.53 (0.34) d = -0.08 N = 307	-0.19 (0.18) d = -0.06 N = 307	-0.08 (0.07) d = -0.06 N = 296	-0.02 (0.12) d = -0.01 N = 294	x1.03 (0.11) ame= 0.01 N = 291
ITERS/ECERS-R provisions score	-0.37 (0.30) d = -0.06 N = 307	-0.50 (0.42) d = -0.06 N = 3-7	0.09 (0.21) d = 0.02 N = 307	-0.02 (0.10) d = -0.01 N = 296	-0.20 (0.16) d = -0.07 N = 294	x1.00 (0.13) ame= -0.00 N = 291
Arnett CIS overall score	0.01 (0.74) d = 0.00 N = 306	-0.25 (1.02) d = -0.01 N = 306	0.13 (0.56) d = 0.01 N = 306	0.10 (0.22) d = 0.02 N = 296	0.28 (0.36) d = 0.04 N = 294	x1.06 (0.33) ame= 0.01 N = 291
C-COS count of any caregiver talk	0.01 (0.03) d = 0.01 N = 262	0.05 (0.05) d = 0.07 N = 262	0.02 (0.02) d = 0.05 N = 262	0.00 (0.01) d = 0.02 N = 254	0.02 (0.02) d = 0.09 N = 253	x0.99 (0.01) ame= -0.00 N = 250
C-COS count of caregiver-initiated talk	0.01 (0.03) d = 0.02 N = 262	0.01 (0.05) d = 0.01 N = 262	0.03 (0.02) d = 0.07 N = 262	0.00 (0.01) d = 0.01 N = 254	0.01 (0.02) d = 0.04 N = 253	x1.00 (0.02) ame= -0.00 N = 250
C-COS count of caregiver-responding talk	-0.02 (0.04) d = -0.02 N = 262	0.08 (0.07) d = 0.07 N = 262	-0.02 (0.03) d = -0.03 N = 262	0.00 (0.01) d = 0.02 N = 254	0.02 (0.02) d = 0.06 N = 253	x0.99 (0.02) ame= -0.00 N = 250
C-COS count of child negative behavior	0.07 (0.07) d = 0.05 N = 262	0.12 (0.11) d = 0.07 N = 262	0.04 (0.05) d = 0.04 N = 262	-0.01 (0.02) d = -0.01 N = 254	0.02 (0.04) d = 0.03 N = 253	x0.97 (0.03) ame= -0.01 N = 250

Note: Each cell presents the estimate of the partial effect of the quality measure in the row on the outcome in the column, using the primary model from Table 6-C and replacing the ITERS/ECERS-R overall score with the quality measure in the row. Cells present coefficients, standard errors in parentheses, effect sizes (d), and model sample sizes (N). The success index uses a logistic regression model; cells present odds ratios (x) and their standard errors, and average marginal effects (ame). For ITERS/ECERS-R and Arnett scores, means over 14, 24, and 36 months are used for all outcomes. For C-COS counts, means over 24 and 36 months are used for all outcomes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8-A. Robustness checks for 24- and 36-month outcomes

	24-Months			36-Months		
	Bayley MDI	CDI Vocabulary	CBCL Aggressive Behavior	Bayley MDI	PPVT-III	CBCL Aggressive Behavior
Baseline	1.17 (0.93) d = 0.10 N = 211	0.00 (1.35) d = 0.00 N = 235	-0.39 (0.37) d = -0.07 N = 235	0.12 (0.65) d = 0.01 N = 287	2.02** (0.87) d = 0.14 N = 264	-0.08 (0.33) d = -0.01 N = 329
Indicator for missing hours	1.11 (0.80) d = 0.09 N = 286	-0.18 (1.19) d = -0.01 N = 324	-0.29 (0.34) d = -0.05 N = 323	0.26 (0.63) d = 0.02 N = 339	1.94** (0.82) d = 0.13 N = 318	0.08 (0.30) d = 0.01 N – 392
Indicator for missing hours, impute covariates (PRIMARY)	1.19 (0.76) d = 0.10 N = 295	0.13 (1.14) d = 0.01 N = 335	-0.22 (0.33) d = -0.04 N = 334	0.47 (0.61) d = 0.04 N = 354	1.84** (0.76) d = 0.12 N = 333	0.05 (0.28) d = 0.01 N = 407
Impute hours, impute covariates	1.25* (0.75) d = 0.10 N = 295	0.20 (1.14) d = 0.01 N = 335	-0.28 (0.33) d = -0.05 N = 334	0.49 (0.61) d = 0.04 N = 354	1.81** (0.76) d = 0.12 N = 333	0.05 (0.28) d = 0.01 N = 407
Impute hours, indicator for missing hours, impute covariates	1.20 (0.75) d = 0.10 N = 295	0.12 (1.14) d = 0.01 N = 335	-0.22 (0.33) d = -0.04 N = 334	0.48 (0.61) d = 0.04 N = 354	1.83** (0.76) d = 0.12 N = 333	0.05 (0.28) d = 0.01 N = 407
No hours	1.16 (0.79) d = 0.10 N = 286	-0.15 (1.20) d = -0.01 N = 324	-0.35 (0.34) d = -0.06 N = 323	0.29 (0.63) d = 0.02 N = 339	1.93** (0.81) d = 0.13 N = 318	0.08 (0.29) d = 0.01 N = 392
No ratio or hours	0.91 (0.79) d = 0.07 N = 288	0.13 (1.09) d = 0.01 N = 326	-0.55* (0.32) d = -0.09 N = 325	0.10 (0.64) d = 0.01 N = 341	1.65** (0.82) d = 0.11 N = 320	0.07 (0.29) d = 0.01 N = 394

Note: Each cell presents the estimate of the partial effect of the ITERS/ECERS-R overall score on the outcome in the column, modifying the primary model from Table 6-A as indicated in each row. Details of these modifications are discussed in chapters 3 and 4. The third row is the primary model. Cells present coefficients, standard errors in parentheses, effect sizes (d), and model sample sizes (N). The mean ITERS/ECERS-R overall score over 14 and 24 months is used for 24-month outcomes, and mean score over 14, 24, and 36 months is used for 36-month outcomes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8-B. Robustness checks for grade 5 cognitive/academic outcomes

	PPVT-III	WISC-IV Matrix Reasoning	ECLS-K math	ECLS-K reading	Academic success index
Baseline	2.20** (0.88) d = 0.15 N = 235	0.12 (0.22) d = 0.04 N = 236	0.16 (0.28) d = 0.04 N = 235	3.54** (1.80) d = 0.12 N = 238	x1.89*** (0.46) ame= 0.07 N = 227
Indicator for missing hours	1.27 (0.87) d = 0.09 N = 281	0.10 (0.20) d = 0.03 N = 283	0.09 (0.26) d = 0.02 N = 283	2.11 (1.77) d = 0.07 N = 286	x1.69** (0.35) ame= 0.06 N = 281
Indicator for missing hours, impute covariates (PRIMARY)	1.00 (0.82) d = 0.07 N = 292	0.01 (0.19) d = 0.00 N = 294	0.16 (0.25) d = 0.04 N = 294	2.23 (1.67) d = 0.08 N = 297	x1.67** (0.33) ame= 0.05 N = 292
Impute hours, impute covariates	0.98 (0.81) d = 0.07 N = 292	0.01 (0.19) d = 0.00 N = 294	0.16 (0.25) d = 0.04 N = 294	2.18 (1.67) d = 0.08 N = 297	x1.69** (0.35) ame= 0.06 N = 292
Impute hours, indicator for missing hours, impute covariates	0.99 (0.81) d = 0.07 N = 292	0.01 (0.19) d = 0.00 N = 294	0.16 (0.25) d = 0.04 N = 294	2.21 (1.67) d = 0.08 N = 297	x1.67** (0.34) ame= 0.05 N = 292
No hours	1.26 (0.86) d = 0.09 N = 281	0.10 (0.20) d = 0.03 N = 283	0.09 (0.26) d = 0.02 N = 283	2.11 (1.77) d = 0.07 N = 286	x1.71** (0.35) ame= 0.06 N = 281
No ratio or hours	1.42* (0.83) d = 0.10 N = 283	0.09 (0.19) d = 0.03 N = 285	0.08 (0.26) d = 0.02 N = 285	2.26 (1.73) d = 0.08 N = 288	x1.66** (0.36) ame= 0.06 N = 283

Note: Each cell presents the estimate of the partial effect of the ITERS/ECERS-R overall score on the outcome in the column, modifying the primary model from Table 6-B as indicated in each row. Details of these modifications are discussed in chapters 3 and 4. The third row is the primary model. Cells present coefficients, standard errors in parentheses, effect sizes (d), and model sample sizes (N). The success index uses a logistic regression model; cells present odds ratios (x) and their standard errors, and average marginal effects (ame). The mean ITERS/ECERS-R overall score over 14, 24, and 26 is used for all outcomes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8-C. Robustness checks for grade 5 social-emotional outcomes

	CBCL Internal-izing Behavior	CBCL External-izing Behavior	CBCL Attention Problems	Self-reported delinquent behavior	Self-reported bullying by peers	Social-emotional success index
Baseline	-0.63* (0.34 d = -0.10 N = 246	-0.86* (0.46) d = -0.11 N = 246	-0.17 (0.23) d = -0.04 N = 246	-0.05 (0.10) d = -0.03 N = 237	-0.16 (0.16) d = -0.06 N = 235	x1.02 (0.15) ame= 0.00 N = 232
Indicator for missing hours	-0.40 (0.31) d = -0.07 N = 296	-0.40 (0.42) d = -0.05 N = 296	0.01 (0.22) d = 0.00 N = 296	-0.05 (0.09) d = -0.03 N = 285	-0.04 (0.15) d = -0.02 N = 283	x0.93 (0.12) ame= -0.02 N = 280
Indicator for missing hours, impute covariates (PRIMARY)	-0.37 (0.30) d = -0.06 N = 307	-0.43 (0.40) d = -0.05 N = 307	-0.00 (0.21) d = -0.00 N = 307	-0.04 (0.09) d = -0.02 N = 296	-0.12 (0.15) d = -0.04 N = 294	x0.94 (0.12) ame= -0.01 N = 291
Impute hours, impute covariates	-0.37 (0.30) d = -0.06 N = 307	-0.42 (0.40) d = -0.05 N = 307	-0.00 (0.21) d = -0.00 N = 307	-0.04 (0.09) d = -0.02 N = 296	-0.12 (0.15) d = -0.04 N = 294	x0.95 (0.12) ame= -0.01 N = 291
Impute hours, indicator for missing hours, impute covariates	-0.37 (0.30) d = -0.06 N = 307	-0.43 (0.40) d = -0.05 N = 307	-0.00 (0.21) d = -0.00 N = 307	-0.04 (0.09) d = -0.02 N = 296	-0.12 (0.15) d = -0.04 N = 294	x0.95 (0.12) ame= -0.01 N = 291
No hours	-0.40 (0.31) d = -0.07 N = 296	-0.40 (0.42) d = -0.05 N = 296	0.01 (0.22) d = 0.00 N = 296	-0.05 (0.09) d = -0.03 N = 285	-0.04 (0.15) d = -0.01 N = 283	x0.93 (0.12) ame= -0.02 N = 280
No ratio or hours	-0.32 (0.31) d = -0.05 N = 298	-0.46 (0.41) d = -0.06 N = 298	0.00 (0.22) d = 0.00 N = 298	-0.05 (0.09) d = -0.03 N = 287	-0.05 (0.14) d = -0.02 N = 285	x0.95 (0.12) ame= -0.01 N = 282

Note: Each cell presents the estimate of the partial effect of the ITERS/ECERS-R overall score on the outcome in the column, modifying the primary model from Table 6-C as indicated in each row. Details of these modifications are discussed in chapters 3 and 4. The third row is the primary model. Cells present coefficients, standard errors in parentheses, effect sizes (d), and model sample sizes (N). The success index uses a logistic regression model; cells present odds ratios (x) and their standard errors, and average marginal effects (ame). The mean ITERS/ECERS-R overall score over 14, 24, and 26 is used for all outcomes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9-A. Alternative specifications for 24- and 36-month outcomes

	24-Months			36-Months		
	Bayley MDI	CDI Vocabulary	CBCL Aggressive Behavior	Bayley MDI	PPVT-III	CBCL Aggressive Behavior
Subset of child and family covariates (PRIMARY)	1.19 (0.76) d = 0.10 N = 295	0.13 (1.14) d = 0.01 N = 335	-0.22 (0.33) d = -0.04 N = 334	0.47 (0.61) d = 0.04 N = 354	1.84** (0.76) d = 0.12 N = 333	0.05 (0.28) d = 0.01 N = 407
No child and family covariates	0.73 (0.74) d = 0.06 N = 295	0.05 (1.20) d = 0.00 N = 335	-0.27 (0.33) d = -0.05 N = 334	0.39 (0.62) d = 0.03 N = 354	1.66** (0.80) d = 0.11 N = 333	-0.11 (0.29) d = -0.02 N = 407
Full set of child and family covariates	1.92** (0.82) d = 0.16 N = 295	0.22 (1.16) d = 0.01 N = 335	-0.18 (0.35) d = -0.03 N = 334	0.42 (0.64) d = 0.04 N = 354	1.50* (0.79) d = 0.10 N = 333	-0.23 (0.31) d = -0.04 N = 407
Add pre-k quality and grade 5 school poverty	n/a	n/a	n/a	n/a	n/a	n/a
Interact quality and hours: quality	0.33 (1.04)	-0.79 (1.64)	-0.33 (0.48)	1.19 (1.00)	1.97* (1.00)	-0.29 (0.43)
Interact quality and hours: quality x hours	0.04 (0.03)	0.04 (0.05)	0.01 (0.02)	-0.03 (0.03)	-0.01 (0.05)	0.02 (0.02)
Interact quality and hours: joint effect	β_{20} = 1.10 β_{40} = 1.87 p = 0.184	β_{20} = 0.08 β_{40} = 0.94 p = 0.689	β_{20} = -0.22 β_{40} = -0.12 p = 0.773	β_{20} = 0.56 β_{40} = -0.07 p = 0.490	β_{20} = 1.85 β_{40} = 1.73 p = 0.045 **	β_{20} = 0.02 β_{40} = 0.33 p = 0.599
Add squared term: quality	-1.71 (5.12)	-1.42 (7.85)	-1.28 (2.24)	-5.44 (4.40)	3.05 (6.25)	0.77 (2.02)
Add squared term: quality-squared	0.31 (0.55)	0.17 (0.86)	0.11 (0.24)	0.63 (0.46)	-0.13 (0.67)	-0.08 (0.22)
Add squared term: joint effect	β_4 = 0.78 β_6 = 2.03 p = 0.247	β_4 = -0.08 β_6 = 0.59 p = 0.977	β_4 = -0.36 β_6 = 0.10 p = 0.724	β_4 = -0.38 β_6 = 2.14 p = 0.262	β_4 = 2.00 β_6 = 1.47 p = 0.058 *	β_4 = 0.15 β_6 = -0.16 p = 0.921

	24-Months			36-Months		
	Bayley MDI	CDI Vocabulary	CBCL Aggressive Behavior	Bayley MDI	PPVT-III	CBCL Aggressive Behavior
Categorical quality, developer thresholds: minimal to good	1.31 (3.77) d = 0.10 N = 295	-5.87 (4.75) d = -0.25 N = 335	-0.46 (1.56) d = -0.07 N = 334	-3.32 (3.10) d = -0.28 N = 354	4.88 (3.42) d = 0.33 N = 333	-0.72 (1.31) d = -0.11 N = 407
Categorical quality, developer thresholds: good to excellent	2.95 (3.68) d = 0.23 N = 295	-4.86 (4.51) d = -0.21 N = 335	-0.44 (1.51) d = -0.07 N = 334	-1.48 (3.00) d = -0.12 N = 354	7.14** (3.30) d = 0.48 N = 333	-0.50 (1.27) d = -0.08 N = 407
Categorical quality, developer thresholds: joint effect	p = 0.500	p = 0.463	p = 0.955	p = 0.260	p = 0.059 *	p = 0.848
Categorical quality, sample distribution: middle quartiles	2.49 (2.18) d = 0.19 N = 295	2.01 (3.30) d = 0.09 N = 335	-1.10 (0.98) d = -0.17 N = 334	1.42 (1.61) d = 0.12 N = 354	1.48 (2.09) d = 0.10 N = 333	0.75 (0.84) d = 0.12 N = 407
Categorical quality, sample distribution: top quartile	3.60 (2.39) d = 0.28 N = 295	0.40 (3.67) d = 0.02 N = 335	-1.27 (1.09) d = -0.20 N = 334	2.77 (1.72) d = 0.23 N = 354	5.74*** (2.19) d = 0.38 N = 333	0.65 (0.88) d = 0.10 N = 407
Categorical quality, sample distribution: joint effect	p = 0.320	p = 0.795	p = 0.443	p = 0.268	p = 0.016 **	p = 0.654

Note: Cells generally present the estimate of the partial effect of the ITERS/ECERS-R overall score on the outcome in the column, using alternative specifications to the primary model from Table 6-A as indicated in each row. Details of these specifications are discussed in chapters 3 and 4. The first row is the primary model. The mean ITERS/ECERS-R overall score over 14 and 24 months is used for 24-month outcomes, and mean score over 14, 24, and 36 months is used for 36-month outcomes.

For the interaction model, the joint effect row presents estimates for the marginal effect of quality on the outcome at 20 and 40 average weekly hours of center care. For the squared-term model, the joint effect row presents estimates for the marginal effect of quality on the outcome when the mean ITERS/ECERS-R score is 4 points and 6 points. In models with two quality variables, the p-value is from the F-test of joint significance of the two variables.

Cells present coefficients, standard errors in parentheses, effect sizes (d), and model sample sizes (N). For the categorical quality models, the effect size is the expected change in the outcome in standard deviations from being in the category (for other models, where quality is continuous, the effect size is based on a change of one standard deviation in the quality measure).

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9-B. Alternative specifications for grade 5 cognitive/academic outcomes

	PPVT-III	WISC-IV Matrix Reasoning	ECLS-K math	ECLS-K reading	Academic success index
Subset of child and family covariates (PRIMARY)	1.00 (0.82) d = 0.07 N = 292	0.01 (0.19) d = 0.00 N = 294	0.16 (0.25) d = 0.04 N = 294	2.23 (1.67) d = 0.08 N = 297	x1.67** (0.33) ame= 0.05 N = 292
No child and family covariates	0.65 (0.86) d = 0.05 N = 292	-0.02 (0.19) d = -0.01 N = 294	0.04 (0.26) d = 0.01 N = 294	2.03 (1.75) d = 0.07 N = 297	x1.39** (0.23) ame= 0.04 N = 292
Full set of child and family covariates	1.31 (0.84) d = 0.09 N = 292	0.03 (0.20) d = 0.01 N = 294	0.14 (0.28) d = 0.03 N = 294	1.73 (1.79) d = 0.06 N = 297	x1.68** (0.45) ame= 0.05 N =292
Add pre-k quality and grade 5 school poverty	-0.09 (1.20) d = -0.01 N = 160	-0.17 (0.27) d = -0.05 N = 161	-0.08 (0.37) d = -0.02 N = 161	1.01 (2.28) d = 0.03 N = 162	x1.70** (0.39) ame= 0.06 N = 161
Interact quality and hours: quality	-0.61 (1.31)	-0.22 (0.28)	-0.12 (0.40)	1.19 (2.91)	0.30 (0.32)
Interact quality and hours: quality x hours	0.08* (0.05)	0.01 (0.01)	0.01 (0.01)	0.05 (0.10)	0.01 (0.01)
Interact quality and hours: joint effect	β_{20} = 0.97 β_{40} = 2.54 p = 0.078 *	β_{20} = 0.01 β_{40} = 0.24 p = 0.525	β_{20} = 0.15 β_{40} = 0.43 p = 0.439	β_{20} = 2.21 β_{40} = 3.24 p = 0.274	β_{20} = x1.63 β_{40} = x1.98 p = 0.015 **
Add squared term: quality	-0.33 (6.09)	-0.01 (1.28)	0.74 (1.72)	-5.22 (12.34)	0.25 (1.45)
Add squared term: quality-squared	0.14 (0.64)	0.00 (0.14)	-0.06 (0.19)	0.81 (1.27)	0.03 (0.15)
Add squared term: joint effect	β_4 = 0.82 β_6 = 1.39 p = 0.447	β_4 = 0.01 β_6 = 0.02 p = 0.998	β_4 = 0.24 β_6 = -0.01 p = 0.744	β_4 = 1.23 β_6 = 4.45 p = 0.243	β_4 = x1.59 β_6 = x1.77 p = 0.028 **

	PPVT-III	WISC-IV Matrix Reasoning	ECLS-K math	ECLS-K reading	Academic success index
Categorical quality, developer thresholds: minimal to good	5.08 (3.83) d = 0.35 N = 292	0.77 (0.75) d = 0.23 N = 294	1.95** (0.91) d = 0.43 N = 294	0.88 (8.11) d = 0.03 N = 297	n/a
Categorical quality, developer thresholds: good to excellent	4.83 (3.84) d = 0.33 N = 292	0.29 (0.73) d = 0.09 N = 294	1.61* (0.87) d = 0.36 N = 294	5.84 (8.03) d = 0.20 N = 297	n/a
Categorical quality, developer thresholds: joint effect	p = 0.412	p = 0.445	p = 0.100 *	p = 0.324	n/a
Categorical quality, sample distribution: middle quartiles	0.56 (2.07) d = 0.04 N = 292	0.28 (0.49) d = 0.09 N = 294	0.20 (0.70) d = 0.05 N = 294	2.71 (4.82) d = 0.09 N = 297	x3.03** (1.69) ame= 0.12 N = 292
Categorical quality, sample distribution: top quartile	4.29* (2.35) d = 0.29 N = 292	0.58 (0.60) d = 0.17 N = 294	0.42 (0.79) d = 0.09 N = 294	8.42* (5.00) d = 0.29 N = 297	x6.31*** (3.71) ame= 0.19 N = 292
Categorical quality, sample distribution: joint effect	p = 0.103	p = 0.631	p = 0.867	p = 0.170	p = 0.007 ***

Note: Cells generally present the estimate of the partial effect of the ITERS/ECERS-R overall score on the outcome in the column, using alternative specifications to the primary model from Table 6-B as indicated in each row. Details of these specifications are discussed in chapters 3 and 4. The first row is the primary model. The mean ITERS/ECERS-R overall score over 14, 24, and 26 is used for all outcomes.

For the interaction model, the joint effect row presents estimates for the marginal effect of quality on the outcome at 20 and 40 average weekly hours of center care. For the squared-term model, the joint effect row presents estimates for the marginal effect of quality on the outcome when the mean ITERS/ECERS-R score is 4 points and 6 points. In models with two quality variables, the p-value is from the F-test of joint significance of the two variables.

Cells present coefficients, standard errors in parentheses, effect sizes (d), and model sample sizes (N). For the categorical quality models, the effect size is the expected change in the outcome in standard deviations from being in the category (for other models, where quality is continuous, the effect size is based on a change of one standard deviation in the quality measure). The success index uses a logistic regression model; most cells present odds ratios (x) and their standard errors, and average marginal effects (ame). The first categorical model could not be estimated because there was no variation in the success index for the low quality category.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9-C. Alternative specifications for grade 5 social-emotional outcomes

	CBCL Internal-izing Behavior	CBCL External-izing Behavior	CBCL Attention Problems	Self-reported delinquent behavior	Self-reported bullying by peers	Social-emotional success index
Subset of child and family covariates (PRIMARY)	-0.37 (0.30) d = -0.06 N = 307	-0.43 (0.40) d = -0.05 N = 307	-0.00 (0.21) d = -0.00 N = 307	-0.04 (0.09) d = -0.02 N = 296	-0.12 (0.15) d = -0.04 N = 294	x0.94 (0.12) ame= -0.01 N = 291
No child and family covariates	-0.52* (0.31) d = -0.09 N = 307	-0.74* (0.41) d = -0.09 N = 307	-0.16 (0.22) d = -0.04 N = 307	-0.09 (0.09) d = -0.05 N = 296	-0.17 (0.15) d = -0.06 N = 294	x0.99 (0.12) ame= -0.00 N = 291
Full set of child and family covariates	-0.50 (0.31) d = -0.08 N = 307	-0.43 (0.41) d = -0.05 N = 307	-0.02 (0.23) d = -0.01 N = 307	-0.01 (0.09) d = -0.01 N = 296	-0.10 (0.15) d = -0.04 N = 294	x0.94 (0.13) ame= -0.01 N = 291
Add pre-k quality and grade 5 school poverty	-0.46 (0.46) d = -0.07 N = 165	-0.28 (0.55) d = -0.04 N = 165	-0.03 (0.27) d = -0.01 N = 165	-0.02 (0.13) d = -0.01 N = 162	-0.06 (0.21) d = -0.02 N = 161	x1.02 (0.19) ame= 0.00 N = 161
Interact quality and hours: quality	-0.11 (0.45)	-0.12 (0.60)	0.28 (0.33)	0.01 (0.14)	-0.08 (0.22)	-0.17 (0.18)
Interact quality and hours: quality x hours	-0.01 (0.02)	-0.02 (0.02)	-0.01 (0.01)	-0.00 (0.01)	-0.00 (0.01)	0.01 (0.01)
Interact quality and hours: joint effect	β_{20} = -0.36 β_{40} = -0.61 p = 0.352	β_{20} = -0.42 β_{40} = -0.73 p = 0.466	β_{20} = 0.00 β_{40} = -0.27 p = 0.528	β_{20} = -0.04 β_{40} = -0.09 p = 0.855	β_{20} = -0.12 β_{40} = -0.16 p = 0.698	β_{20} = x0.94 β_{40} = x1.05 p = 0.653
Add squared term: quality	0.02 (2.30)	-1.78 (2.72)	0.06 (1.62)	-0.26 (0.53)	1.52 (0.99)	1.01 (0.89)
Add squared term: quality-squared	-0.04 (0.24)	0.15 (0.29)	-0.01 (0.17)	0.02 (0.06)	-0.18* (0.11)	-0.11 (0.10)
Add squared term: joint effect	β_4 = -0.31 β_6 = -0.48 p = 0.443	β_4 = -0.61 β_6 = -0.03 p = 0.497	β_4 = 0.01 β_6 = -0.02 p = 0.999	β_4 = -0.07 β_6 = 0.03 p = 0.780	β_4 = 0.10 β_6 = -0.61 p = 0.187	β_4 = x1.10 β_6 = x0.69 p = 0.457

	CBCL Internal-izing Behavior	CBCL External-izing Behavior	CBCL Attention Problems	Self-reported delinquent behavior	Self-reported bullying by peers	Social-emotional success index
Categorical quality, developer thresholds: minimal to good	-1.47 (1.54) d = -0.24 N = 307	-1.58 (1.70) d = -0.20 N = 307	-0.78 (1.05) d = -0.20 N = 307	0.23 (0.32) d = 0.13 N = 296	0.99* (0.53) d = 0.35 N = 294	x1.39 (0.78) ame= 0.07 N = 291
Categorical quality, developer thresholds: good to excellent	-2.05 (1.54) d = -0.34 N = 307	-2.14 (1.65) d = -0.27 N = 307	-0.63 (1.03) d = -0.16 N = 307	0.07 (0.31) d = 0.04 N = 296	0.52 (0.50) d = 0.19 N = 294	x1.16 (0.64) ame= 0.03 N = 291
Categorical quality, developer thresholds: joint effect	p = 0.351	p = 0.412	p = 0.752	p = 0.674	p = 0.155	p = 0.746
Categorical quality, sample distribution: middle quartiles	-0.06 (0.80) d = -0.01 N = 307	-1.13 (1.23) d = -0.14 N = 307	0.25 (0.57) d = 0.06 N = 307	-0.25 (0.25) d = -0.14 N = 296	-0.10 (0.45) d = -0.04 N = 294	x1.10 (0.35) ame= 0.02 N = 291
Categorical quality, sample distribution: top quartile	-0.86 (0.83) d = -0.14 N = 307	-1.21 (1.27) d = -0.15 N = 307	-0.02 (0.60) d = -0.01 N = 307	-0.17 (0.27) d = -0.10 N = 296	-0.62 (0.47) d = -0.22 N = 294	x1.03 (0.39) ame= 0.01 N = 291
Categorical quality, sample distribution: joint effect	p = 0.479	p = 0.598	p = 0.843	p = 0.626	p = 0.312	p = 0.952

Note: Cells generally present the estimate of the partial effect of the ITERS/ECERS-R overall score on the outcome in the column, using alternative specifications to the primary model from Table 6-C as indicated in each row. Details of these specifications are discussed in chapters 3 and 4. The first row is the primary model. The mean ITERS/ECERS-R overall score over 14, 24, and 26 is used for all outcomes.

For the interaction model, the joint effect row presents estimates for the marginal effect of quality on the outcome at 20 and 40 average weekly hours of center care. For the squared-term model, the joint effect row presents estimates for the marginal effect of quality on the outcome when the mean ITERS/ECERS-R score is 4 points and 6 points. In models with two quality variables, the p-value is from the F-test of joint significance of the two variables.

Cells present coefficients, standard errors in parentheses, effect sizes (d), and model sample sizes (N). For the categorical quality models, the effect size is the expected change in the outcome in standard deviations from being in the category (for other models, where quality is continuous, the effect size is based on a change of one standard deviation in the quality measure). The success index uses a logistic regression model; most cells present odds ratios (x) and their standard errors, and average marginal effects (ame).

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

APPENDIX

This appendix contains additional details from this analysis not reported in the main text.

Key Explanatory Variables: Quality of Care.

ITERS/ECERS-R. For each item in the environment rating scales such as the ITERS or ECERS-R, there are multiple indicators describing the features corresponding to each labeled quality level (1, 3, 5, and 7). The observer evaluates whether the indicator has been met and uses the results of the indicators to determine the score for that item. An individual item score must be an integer from 1 to 7. Subscales have different numbers of items, and the total score is an item-level average, not a subscale average. The EHSREP public-use file includes the overall average score and the score for each subscale, but not item-level scores. The EHSREP evaluation states they used 33 items from the ITERS, omitting three items from the adult needs subscale. However, the full ITERS had 35 items, and based on the values for the subscale scores in the data, it appears the evaluation used 31 items, omitting three from the adult needs section (as these scores in the data are only integers, suggesting there was only one item) and one from the program structure section (these scores are divisible by three – 5.33, 5.67, and so on, instead of divisible by four, the number of items in the full subscale, suggesting there were only three items). Similarly, the evaluation notes they used 39 items from the ECERS-R by omitting four items from the parents and staff section. The ECERS-R does have 43 items; however, based on the data, it appears three items were omitted from the parents and staff section; and the fourth item not included was in the program structure section.

Using an item-level average of the two subscale components of the teaching score and the learning provisions score would better align with the method used to calculate the overall score. It is possible to construct this average by weighting the average subscale scores by their number of items. However, some observations may not have had scores for all their items, so this method may have introduced some error. This study uses subscale averages instead. Within the teaching

score, the breakdown for the ITERS is 2 items in listening and talking vs. 3 items in interaction, and for the ECERS-R it is 4 items in language and reasoning vs. 5 items in interaction. Within the provisions score, the breakdown for the ITERS is 5 items in furnishings and display vs. 8 items in activities, and for the ECERS-R it is 8 items in space and furnishings vs. 10 items in activities. So, the largest split is slightly above 60/40.

Arnett CIS. Before calculating the overall score, scores for items representing negative behavior are reversed so "not at all true" is now 4 and "very much true" is 1. The Arnett CIS has four subscales corresponding to caregiver characteristics: sensitivity, harshness, detachment, and permissiveness. The EHSREP study performed factor analyses for each time period (at 14, 24, and 36 months), but did not find any consistent set of factors, so only the overall score was analyzed and included in the public-use file.

C-COS. During the C-COS observation periods, several pieces of detailed information were recorded about: the type of caregiver talk, who the child spoke to, what the child was doing, how the child was behaving, and which caregivers were interacting with the child. For example, if the caregiver spoke to the child, it was classified as requesting language or some other form of communication from the child, requesting the child perform a certain action, reading to the child, or another type of talking (including singing). The first three measures used by the EHSREP evaluation report (and included in the public-use file) only note the periods during which any caregiver talk occurred, and then whether the period involved caregiver-initiated talk or a response to the child. Finally, a period of negative child behavior meant that the child either was wandering around/unoccupied, was upset/crying, was being hit/bit/bothered by another child, or was hitting/biting/bothering another child. The observer also recorded general assessments of the quality of the caregiver and child's behaviors during each five-minute period, but the EHSREP study did not report on these or include them in the public-use file.

Imputed Variables:

Imputed covariates. Imputation was done using Stata's *impute* command. Because the command only accepts up to 30 explanatory variables, a subset of the full set of child and family characteristics were used. These are: child gender, maternal race/ethnicity, maternal education, maternal living arrangements/marital status, maternal age at birth (specifically the indicators for being younger than 20, between 20 and 24, and 25 or older), maternal English-language proficiency, maternal employment or schooling status, number of other children age 0 to 5 and age 6 to 17 in the household, poverty to income ratio, participation in AFDC/TANF and food stamps, having inadequate resources to buy necessities, medical needs, and transportation, the number of moves in the previous year, child age at random assignment, whether child was low birthweight, and whether child had a risk identified.

Because almost all the child and family characteristics are indicator variables, most predictions were calculated using a linear probability model, so the very small number of imputed values that did not fall between 0 and 1 were bottom- or top-coded at those values. When a set of mutually exclusive indicator variables (for example, the maternal race/ethnicity) variables had imputed values that summed to more or less than 1, values were re-scaled so the total was equal to 1.

Imputed hours. Hours were imputed for each time period (14, 24, and 36 months) before their mean was taken. First, values above 80 hours per week were topcoded at 80, partly because the original analysis' descriptive statistics show this was the maximum, and partly because larger numbers do not seem plausible (they might be for any care, but for center care, an arrangement lasting more than 16 hours/5 days a week or about 11-12 hours/7 days a week seems very unlikely). Next, cases where the average weekly hours were reported as 0 even though an ITERS or ECERS-R score was present were set to missing, under the assumption the parent report from

the Parent Services Interviews (PSIs) was incorrect because an observation had been conducted, and that the values should be imputed.

Other data from the PSIs listed the average weekly hours of center care from the start of the program up until the time of the interview. This was used, along with the number of months between each PSI, to calculate an average over the period since the last interview. While the three PSIs were held at different times from the three PIs, the first PSIs were held in the same general timeframe as the first PIs, and so on. If available, the average over the period between PSIs was imputed as the 14, 24, or 36 month value. However, this was only used for a few cases as most hours data was missing because the parent did not respond to the PSI at all.

The main step, which filled in most of the missing data, was to impute missing values using the same regression process as for the child and family characteristics. The same covariates were used in the imputation model for hours, except the resources to buy necessities/medical care/transportation were removed, and indicators for the site's EHS type (center-based, mixed, or home-based model) were added. Imputed values below 0 were set to 0.

The final step was to take the mean of the imputed hours over each time period, in order to produce the measures to be used in the regression models. All steps were conducted using only EHS program group participants.

Samples for Quality Measures:

A small number of children had observations on the ITERS/ECERS-R but not the Arnett CIS (6, 5, and 9 children at 14, 24, and 36 months). However, there were more missing observations on the C-COS for children with ITERS/ECERS-R – 21 and 28 cases at 24 and 36 months. Because the indicator for receiving center care came from the PSIs, it had the same challenges of missing and incorrect data. For example, 29, 44, and 30 children had a non-missing ITERS/ECERS-R score at 14, 24, and 36 months, respectively, but according to their family's PSI, were not in center care at that time. As a result, presence of an ITERS/ECERS-R score was

taken as more reliable than the PSI report. Unlike the ITERS/ECERS-R, where the instrument reflected the care setting (because FDCRS scores reflect home-based settings), the Arnett CIS and C-COS were given in all settings. This study included an Arnett or C-COS result if: 1) an ITERS/ECERS-R score was present, or 2) if it was missing, but the FDCRS score was also missing and the PSI reported the child was in center care at the time. Very few children had Arnett scores but neither an ITERS/ECERS-R or FDCRS score, so this latter criteria only applied to a few additional observations – 3, 0, and 2 children at 14, 24, and 36 months. Similarly, 0 and 2 children had C-COS scores but no ITERS/ECERS-R score, and were reported as being in center care, at 24 and 36 months.

REFERENCES

Abdi, H. (2007). "Bonferroni and Šidák corrections for multiple comparisons". In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics.* Thousand Oaks, CA: Sage.

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms and profiles.* Burlington, VT: University of Vermont Department of Psychiatry.

Achenbach, T. M., & Rescorla, L. A. (2001*). Manual for the ASEBA school-age forms & profiles.* Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Administration for Children and Families. (2001). *Building their futures: How Early Head Start programs are enhancing the lives of infants and toddlers of low-income families.* Volumes I–II. Washington, DC: U.S. Department of Health and Human Services.

Administration for Children and Families (2002). *Making a difference in the lives of infants and toddlers and their families: The impacts of Early Head Start.* Volumes I-III. Washington, DC: U.S. Department of Health and Human Services.

Administration for Children and Families (2004). *The role of Early Head Start programs in addressing the child care needs of low-income families with infants and toddlers: Influences on child care use and quality.* Washington, DC: U.S. Department of Health and Human Services.

Arnett, J. (1989). Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology, 10*(4), 541-552. doi:http://dx.doi.org/10.1016/0193-3973(89)90026-9

Bayley, N. (1993). *Bayley Scales of Infant Development, second edition: Manual.* New York: The Psychological Corporation, Harcourt Brace & Company.

Belsky, J., Vandell, D. L., Burchinal, M., Clarke-Stewart, K. A., McCartney, K., & Owen, M. T. (2007). Are there long-term effects of early child care? *Child Development, 78*(2), 681-701.

Boller, K., & Sprachman, S., & Early Head Start Research Consortium (1998). *The Child-Caregiver Observation System instructor's manual*. Princeton, NJ: Mathematica Policy Research.

Boller, K., Tarrant, K. & Schaack, D.D. (2014). *Early care and education quality improvement: A typology of intervention approaches.* OPRE Research Report #2014-36. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Retrieved from http://www.acf.hhs.gov/programs/opre/resource/early-care-and-education-quality-improvement-a-typology-of-intervention-approaches

Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher–child interactions and instruction. *Applied Development Science, 12*(3), 140-153.

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, I. Martinez-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 11-31). Baltimore, MD: Brookes.

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*(2), 166-176.

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*(1), 42-57.

Cassidy, D. J., Hestenes, L. L., Hegde, A., Hestenes, S., & Mims, S. (2005). Measurement of quality in preschool child care classrooms: An exploratory and confirmatory factor analysis

of the Early Childhood Environment Rating Scale-Revised. *Early Childhood Research Quarterly, 20*(3), 345-360.

Colwell, N., Gordon, R. A., Fujimoto, K., Kaestner, R., & Korenman, S. (2013). New evidence on the validity of the Arnett Caregiver Interaction Scale: Results from the Early Childhood Longitudinal Study-Birth Cohort. *Early Childhood Research Quarterly, 28*(2), 218-233. doi:http://dx.doi.org/10.1016/j.ecresq.2012.12.004

Duncan, G. J., & Gibson-Davis, C. M. (2006). Connecting child care quality to child outcomes: Drawing policy lessons from nonexperimental data. *Evaluation Review, 30*(5), 611-630.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test–third edition.* Circle Pines, MN: American Guidance Service.

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short-form versions of the MacArthur communicative development inventories. *Applied Psycholinguistics, 21*(01), 95-116.

Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS-R with implications for measures of child care quality and relations to child development. *Developmental Psychology, 49*(1), 146-160.

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

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

Harms, T., & Clifford, R. M. (1989). *Family Day Care Rating Scale.* New York, NY: Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. M. (1990). *Infant/Toddler Environment Rating Scale*. New York, NY: Teachers College Press.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early Childhood Environment Rating Scale, Revised Edition.* New York, NY: Teachers College Press

Harms, T., Cryer, D., & Clifford, R. M. (2003). *Infant/Toddler Environment Rating Scale-Revised.* New York, NY: Teachers College Press

Harms, T., Cryer, D., & Clifford, R. M. (2007). *Family Child Care Environment Rating Scale Revised Edition (FCCERS-R).* New York, NY: Teachers College Press.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172-177.

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.

Laughlin, L. (2013). *Who's minding the kids? Child care arrangements: Spring 2011.* Washington, DC: U.S. Department of Commerce, Bureau of the Census. Retrieved from http://www.census.gov/prod/2013pubs/p70-135.pdf

Lesnick, J., Goerge, R., Smithgall, C., & Gwynne J. (2010). *Reading on grade level in third grade: How is it related to high school performance and college enrollment?* Chicago, IL: Chapin Hall at the University of Chicago. Retrieved from http://www.chapinhall.org/sites/default/files/Reading_on_Grade_Level_111710.pdf

Li, W., Farkas, G., Duncan, G. J., Burchinal, M. R., Vandell, D. L., Ruzek, E. A., & Dang, T. T. (2011). Which combination of high quality infant-toddler and preschool care best promotes school readiness? *Society for Research on Educational Effectiveness.* Retrieved from the ERIC database. (ED517847)

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms.* NCSER 2013-3000. Washington, DC: National

Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Loeber, R., Stouthamer-Loeber, M., Van Kammen, W., & Farrington, D. P. (1991). Initiation, escalation and desistance in juvenile offending and their correlates. *J. Crim. L. & Criminology, 82*, 36.

Love, J. M., Harrison, L., Sagi-Schwartz, A., Van IJzendoorn, M. H., Ross, C., Ungerer, J. A., . . . Brooks-Gunn, J. (2003). Child care quality matters: How conclusions may vary with context. *Child Development, 74*(4), 1021-1033.

Love, J. M., Schochet, P. A., & Meckstroth, A. L. (1996). *Are they in any real danger? What research does—and doesn't—tell us about child care quality and children's well-being.* Princeton, NJ: Mathematica Policy Research.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., . . . Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*(3), 732-749.

McCormick, M. C., Brooks-Gunn, J., Buka, S. L., Goldman, J., Yu, J., Salganik, M., . . . Bernbaum, J. C. (2006). Early intervention in low birth weight premature infants: Results at 18 years of age for the infant health and development program. *Pediatrics, 117*(3), 771-780.

National Institute of Child Health and Human Development 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. doi:10.1111/1467-8624.00617

Pollack, J. M., Atkins-Burnett, S., Najarian, M., and Rock, D.A. (2005). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS–K), Psychometric report for the fifth grade.* NCES 2006–036. U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., & Yazejian, N. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development, 72*(5), 1534-1553. doi:10.1111/1467-8624.00364

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

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system.* Baltimore, MD: Brookes.

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., . . . Downer, J. (2012). *Third grade follow-up to the Head Start impact study: Final report.* OPRE Report #2012-45. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Reynolds, A. J., Temple, J. A., Ou, S., Robertson, D. L., Mersky, J. P., Topitzes, J. W., & Niles, M. D. (2007). Effects of a school-based, early childhood intervention on adult health and well-being: A 19-year follow-up of low-income families. *Archives of Pediatrics & Adolescent Medicine, 161*(8), 730-739.

Ruzek, E., Burchinal, M., Farkas, G., & Duncan, G. J. (2014). The quality of toddler child care and cognitive skills at 24 months: Propensity score analysis results from the ECLS-B. *Early Childhood Research Quarterly, 29*(1), 12-21.

Sabol, T., Hong, S. S., Pianta, R., & Burchinal, M. (2013). Can rating pre-K programs predict children's learning? *Science, 341*(6148), 845-846.

Schaack, D., Tarrant, T., Boller, K., & Tout, K. (2012). Quality rating and improvement systems: Alternative approaches to understanding their impact on the early learning system. In S.L.

Kagan and K. Kaurez (Eds.), *Early childhood systems: Transforming early learning* (pp.71-86). New York, NY: Teachers College Press.

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.

Shonkoff, J. P. & Phillips, D. A. (Eds.) (2000). *From neurons to neighborhoods: The science of early childhood development.* Washington, DC: National Academy Press.

Vandell, D. L., Belsky, J., Burchinal, M., Steinberg, L., & Vandergrift, N. (2010). Do effects of early child care extend to age 15 years? Results from the NICHD study of early child care and youth development. *Child Development, 81*(3), 737-756.

Vogel, C. A., Xue, Y., Moiduddin, E. M., Carlson, B. L., & Kisker, E. E. (2010). *Early Head Start children in grade 5: Long-term follow-up of the Early Head Start Research and Evaluation Project study sample.* OPRE Report #2011-8. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Wechsler, D. (2003). *WISC-IV Wechsler Intelligence Scale for Children.* San Antonio, TX: The Psychological Corporation.

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 prekindergarten program. *Early Childhood Research Quarterly, 28*(2), 199-209.

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development, 84*(6), 2112-2130. doi:10.1111/cdev.12099

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management, 27*(1), 122-154.

Zaslow, M., Anderson, R., Redd, Z., Wessel, J., Tarullo, L., & Burchinal, M. (2010). *Quality dosage, thresholds, and features in early childhood settings: A review of the literature.* OPRE 2011-5. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Zellman, G. L., Perlman, M., Le, V., & Setodji, C. M. (2008). *Assessing the validity of the Qualistar early learning quality rating and improvement system as a tool for improving child-care quality.* Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/monographs/MG650.html