Toddlers in Early Head Start: A Portrait of 3-Year-Olds, Their Families, and the Programs Serving Them

Volume II: Technical Appendices

April 2015

Baby FACES 2009

OPRE Report 2015-28

This page is left blank for double-sided printing.

Toddlers in Early Head Start: A Portrait of 3-Year-Olds, Their Families, and the Programs Serving Them

Volume II: Technical Appendices

OPRE Report 2015-28

April 2015

Submitted to:
Amy Madigan, *Project Officer*
Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Submitted by:
Cheri A. Vogel
Pia Caronongan
Yange Xue
Jaime Thomas
Eileen Bandel
Nikki Aikens
Kimberly Boller
Lauren Murphy
Mathematica Policy Research

Project Director:
Cheri A. Vogel
Mathematica Policy Research
P.O. Box 2393
Princeton, NJ 08543-2393

Contract Number: HHSP23320072914YC
Mathematica Reference Number: 06432.150

Suggested citation:
Vogel, C. A., P. Caronongan, Y. Xue, J. Thomas, E. Bandel, N. Aikens, K. Boller, and L. Murphy (2015). Toddlers in Early Head Start: A Portrait of 3-Year-Olds, Their Families, and the Programs Serving Them, Technical Appendices. OPRE Report 2015-28. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

MATHEMATICA Policy Research

baby FACES

OPRE

This page is left blank for double-sided printing.

CONTENTS

This page is left blank for double-sided printing.

APPENDIX A. SAMPLE AND WEIGHTS

Sample and Attrition

In spring 2009, we selected 1,217 children into the Baby FACES sample. From this sample, 109 children were ineligible for the study,[1] 132 eligible children's parents did not consent to participate in the study. Therefore, there were 976 children in the study at baseline. By spring 2010, we obtained consent for six additional children (four in the 1-year-old Cohort and two in the Newborn Cohort), but consent was rescinded for three 1-year-old Cohort children who were part of the baseline sample. Children who left the Early Head Start program from which they were sampled were considered no longer eligible at follow-up, and this was by far the main driver of sample attrition over time. Table A.1 presents the baseline and follow-up sample sizes by year for those with parental consent and who were eligible for the study at that time. Table A.2 presents the same sample sizes, but organized by the child's age. There were 971 children who had parental consent at the end of the study and who were eligible for at least one round of data collection.

Table A.1. Eligible and Consented Sample Sizes at Baseline and Follow-Up by Year

Cohort	2009	2010	2011	2012
Newborn	194	140	100	85
1-Year-Old	782	602	469	n.a.
Combined	976	742	569	85

n.a. = not applicable.

Table A.2. Eligible and Consented Sample Sizes at Baseline and Follow-Up by Age

Cohort	Age 0	Age 1	Age 2	Age 3
Newborn	194	140	100	85
1-Year-Old	n.a.	782	602	469
Combined	194	922	702	554

n.a. = not applicable.

Table A.3 shows the exit patterns for children over the course of the study, regardless of consent status. Most of the children who exited had consent, but not all. A handful of children who remained in the program and eligible for the study had changes in consent status (gaining and losing consent) over time. Note that children who left the program during the month of or prior to the field visit ("recent exiters") were considered eligible for data collection and part of the study population for that visit. For example, there were 21 children who left the program the month of the 2010 field visit (or the month before that), who were considered eligible for 2010 data collection. However, the 249 children who left between the 2009 field visit and more than a month before the 2010 field visit were considered ineligible for 2010 data collection.

[1] Children were ineligible (either at baseline or later) if they were not actually enrolled in the program or if their birth date falls outside the specified windows. The Newborn Cohort included pregnant women whose due dates were no more than two months after the week of the spring 2009 site visit and babies whose birth dates were no more than two months before the spring 2009 site visit. The 1-year-old Cohort included children who were 10 to 15 months of age at the time of the spring 2009 site visit.

Table A.3. Exiters and Recent Exiters over Time, Regardless of Consent Status

	Newborn Cohort	1-Year-Old Cohort	Combined
Sampled	**253**	**964**	**1,217**
Ineligible for study	29	82	111
Eligible 2009: Still in Early Head Start	**224**	**882**	**1,106**
Left just before 2009	1	0	1
Left after 2009	57	192	249
Eligible 2010: Still in Early Head Start	**166**	**690**	**856**
Left just before 2010	5	16	21
Left after 2010	34	118	152
Eligible 2011: Still in Early Head Start	**127**	**556**	**683**
Left just before 2011	8	n.a	8
Left after 2011	7	n.a	7
Eligible 2012: Still in Early Head Start	**112**	**n.a**	**112**

n.a. = not applicable.

Child-Level Weights

After the spring 2010 data collection and processing, we consulted with the Administration for Children and Families (ACF) before constructing this *second* round of child-level weights.[2] We decided to focus primarily on age-specific weights, rather than year-specific weights (2009, 2010, etc.), and to focus each weight either on child-level data (such as staff child reports and child assessments) or staff-level data (such as staff interviews and observations). Staff can be either teachers or home visitors, and observations can be conducted either in a classroom or during a home visit. Table A.4 specifies data-collection times for both cohorts at each age.

Table A.4. Data Collection Points by Age of Sample

Cohort	Age 0	Age 1	Age 2	Age 3
Newborn	Spring 2009	Spring 2010	Spring 2011	Spring 2012
1-year-old	n.a.	Spring 2009	Spring 2010	Spring 2011

n.a. = not applicable.

We constructed both cross-sectional and longitudinal weights. The weighting steps for each involved first defining the eligible population for that particular weight, adjusting for parental consent (and for some weights, adjusting for the presence of a parent interview in the same step), then adjusting for various combinations of completed instruments as defined for each weight. These combinations were designated based on analytical needs. We balanced several competing issues when determining how many, and which, weights to construct. Calculating weights for every variable and every combination of variables would be an overwhelming task and too difficult to make use of during analysis, so we tried to be parsimonious in the number of weights constructed. If a weight had too stringent a definition of "complete"—for example, if we required that five different instruments be complete before assigning a positive weight—it would result in a lot of missing data for children with some but not all of those instruments. On the other hand, if the definition of "complete" is too lax—for example, if we only required that at least one of the five instruments be

[2] Program-level weights were constructed in 2010 the same way as baseline, accounting for each program's probability of selection, study eligibility, and participation. These weights can be used for analysis at the program level of the 89 programs.

complete—it would result in a large amount of missing data for those with positive weights. Omitting cases with missing data from the analyses would increase the risk of biased estimates.

Cross-sectional weights. To create child-level cross-sectional weights, we ran stepwise logistic regressions separately by cohort. We first created a weight that was positive for those children with parental consent and a completed parent interview, using input variables that were program characteristics[3] (see Table A.5 for list of weights[4]). The pool of independent variables comprised the program's size stratum, service approach (home, center, or mixed), urbanicity (metropolitan statistical area [MSA] versus non-MSA), and U.S. census region. We used the inverse of the propensity score as the weighting adjustment, and applied it to the child base weight, which accounted for program selection probability; program eligibility and participation; and child (sibling) selection probability. By applying this adjustment and creating weight W1P, children with parental consent and a parent interview reflect all eligible children at age 1. The other age 1 cross-sectional weights (W1C and W1S) started with weight W1P, and adjust further for the completion of other instruments as described in the table below. But because these adjustments are only applied to children with a completed parent interview, and the parent interview generates sociodemographic data, these weighting adjustments had a larger pool of variables from which select important predictors of completion. Weights comparable to W1P, W1C, and W1S weights were constructed at age 2 and age 3 in a similar manner.[5]

Table A.5. Child-Level Weights – Cross-Sectional (Age-Specific)

Weight Name	Age	Number with Positive Weight	Sum of Weights	What Constitutes a Complete
W1P	1	839	5943.26	Any parent interview through age 1 (weight use as basis for other weights in this table)
W1C	1	798	5943.26	Any parent interview and an age 1 Staff-Child Report
W1S	1	825	5943.26	Any parent interview and an age 1 staff interview or observation
W2P	2	671	4804.45	Any parent interview through age 2
W2C	2	658	4804.45	Any parent interview and an age 2 child assessment or Staff-Child Report
W2S	2	656	4804.45	Any parent interview and an age 2 staff interview or observation
W3P	3	537	3853.13	Any parent interview through age 3
W3C	3	523	3853.13	Any parent interview and an age 3 child assessment or Staff-Child Report
W3S	3	528	3853.13	Any parent interview and an age 3 staff interview or observation

Note: As an example, to analyze the 1-year old Cohort's vocabulary proficiency at baseline, use weight W1C, which is positive if the child had both a parent interview and a Staff-Child Report in spring 2009 (data on children's vocabulary were collected from the parent interview and Staff-Child Reports).

Longitudinal Weights. We also constructed a set of longitudinal weights targeted for specific planned analyses of Baby FACES data over time. Each one requires eligibility over the appropriate age range, and parental consent as of the end of the study. As with the weights previously described, we constructed the weights separately by cohort, adjusted for consent, then adjusted for completion status as described for each weight presented in Table A.6.

[3] We did not have a lot of information about children without parental consent or without parent interviews. Using the Early Head Start program itself as the weighting cell was problematic due to the relatively small number of children per program.

[4] Some interim weights that were created for the 1-year-old Cohort before we had comparable data for the Newborn Cohort are excluded from Table A.5.

[5] For the age 2 and 3 weights, there was negligible nonresponse among those with a parent interview, so a simple ratio adjustment was used for this final adjustment, rather than a model-generated propensity score.

Table A.6. Child-Level Weights – Longitudinal

Weight Name	Range	Number with Positive Weight	Sum of Weights	What Constitutes a Complete
WL_P2	Age 1-2	500	4804.45	Parent interview at age 1 and age 2
WL_P3	Age 1-3	331	3853.13	Parent interview at ages 1, 2, and 3
WL_PANY	Age 1-3	835	5968.70	Parent interview at age 1, 2, or 3
WL_PANY0	Age 0-3	895	6211.04	Parent interview at age 0, 1, 2, or 3
WL_CANY	Age 1-3	894	5968.70	Staff-Child Report at age 1, 2, or 3
WL_CANY0	Age 0-3	954	6211.04	Staff-Child Report at age 0, 1, 2, or 3
The following weights require the presence of Early Head Start experience data[a] at age 2 or age 3				
WX_A3	Age 3	490	3853.13	Child assessment at age 3 + experience data
WX_R3	Age 3	527	3853.13	Staff-Child Report at age 3 + experience data
WX_V32	Age 3	452	3853.13	Two Bags assessment at age 3 + experience data
WX_P3S	Age 3	464	3853.13	Parent self-administered questionnaire at age 3 + experience data
WX_P3	Age 3	409	3853.13	Parent interview at age 3 + experience data
WX_COGOUT	Age 3	365	3853.13	Cognitive stimulation outcome[b] at age 3 + experience data
WX_P3SA3	Age 3	460	3853.13	Parent self-administered questionnaire + child assessment at age 3 + experience data

Note: As an example, to analyze changes in child socioeconomic status over time, use weight WL_P3, which is positive if the child was eligible (still in the program) throughout the study, had parental consent as of the end of the study, and had a parent interview at ages 1, 2, and 3 (where socioeconomic information was collected).

[a]To be considered as having experience data, the child must have had: (1) a Staff-Child Report at age 2 or age 3; or (2) Family Services Tracking (program services) data between ages 1 and 2 or between ages 2 and 3; or classroom or home visit observation data at age 2 or age 3. There are 831 children meeting these criteria.
[b]Cognitive stimulation outcome: parent interview + parent self-administered questionnaire + child assessment + Two Bags assessment + video data.

We created a few additional weights for analyzing families who exited the program early and for the Family Services Tracking (FST) (program services) data (Table A.7). Because analysis of exiting behavior and FST data spans the year between the first two data collection points (spring 2009 through spring 2010), we required that the children have parental consent throughout this entire period. Out of 1,106 eligible children (224 in the Newborn Cohort and 882 in the 1-year-old Cohort), 973 (194 in the Newborn Cohort and 779 in the 1-year-old Cohort) met the consent criteria. Children for whom we obtained consent after baseline and those who rescinded consent after baseline were excluded from the sample and the reference population for these weights. For these weights, we did not require a parent interview; therefore, other than study cohort, the variables available to use as weighting covariates were limited to those at the program level (program size, service type, MSA status, and census region).

Exit Weights. To compare the characteristics of children and families who exited the Early Head Start program during the first study year with characteristics of those who remained, we created a child-level exit weight. This weight adjusted for the program's probability of selection and its participation, and whether the child had parental consent. Unlike other weights that involve spring 2010 data, we did not exclude children who left the program, as they are of key interest for this analysis. We constructed this weight separately by cohort using stepwise regression to find which of the four program-level variables predicted consent. We then used a logistic regression model to estimate a consent propensity score; we used the inverse as a weighting adjustment. The consent-adjusted weight for the 973 consented children sums to 6,215 (1,157 for the Newborn Cohort and 5,058 for the 1-year-old Cohort).

Among the 244 children who had exited the program (but not during the month of or preceding the spring 2010 data collection visit[6]), we attempted to collect a parent exit interview. This interview aimed to collect information on where the child went after leaving the Early Head Start program, and why he or she left. We constructed a weight to adjust for nonresponse (during the first year) to this interview among those who exited. To construct this weight, we started with the consent-adjusted weight described above and excluded those who had not exited the program or had recently exited. We adjusted this weight separately by cohort. We used the stepwise logistic regression procedure described above, except that the dependent variable was whether we obtained a parent exit interview. The adjusted weight for the 128 children with a completed exit interview (33 from the Newborn Cohort and 95 from the 1-year-old Cohort) sums to 1,354 (314 for the Newborn Cohort and 1,040 for the 1-year-old Cohort).

Table A.7. Child-Level Weights – Consent, Exiter, and FST

Weight Name	Range	Number with Positive Weight	Sum of Weights	What Constitutes a Complete
W0910CSNT	2009-2010	973	6215.54	Eligible at baseline and parental consent in 2009 and 2010
W10EXIT	2009-2010	128	1354.00	Met criteria for W0910CSNT – and – exited between 2009 and 2010 and completed parent exit interview
W10FST	2009-2010	830	6215.54	Met criteria for W0910CSNT – and – have any FST data between 2009 and 2010
WL_FANY	Age 1-2 or age 2-3	786	5968.70	Parental consent as of end of study and any FST data between ages 1 and 2 OR between ages 2 and 3
WL_F1	Age 1-2	765	5968.70	Parental consent as of end of study and any FST data between ages 1 and 2
WL_F2	Age 2-3	516	4804.45	Parental consent as of end of study and any FST data between ages 2 and 3

Note: As an example, to compare first-year exiters to non-exiters, use weight W9010CSNT, which is positive if the child was eligible at baseline and had parental consent during that time period.

Family Services Tracking (FST) weights. We constructed a weight for use with the first year (2009-2010) of FST data. We also constructed an FST weight for age 1 (meaning we had FST data between ages 1 and 2) and for age 2 (meaning we had FST data between ages 2 and 3). Again, we started with the consent-adjusted weight. Then, using the same stepwise procedures described above, we adjusted for whether we received any FST data for the child over the course of the year. The sum of the FST weights is 6,215 children (1,157 for the Newborn Cohort and 5,058 for the 1-year-old Cohort).

[6] There were 20 children who exited during the month of or preceding the site visit. We do not consider these children to have exited early for any of the weights.

This page is left blank for double-sided printing.

APPENDIX B. DATA COLLECTION

This appendix details the process we followed to maintain relationships with study programs, manage samples, collect spring 2011 and 2012 data, and prepare the data for analysis.

Baby FACES Coordinators Maintained Relationships with Study Programs and Families

During the summer and fall of 2010, Baby FACES coordinators (BFCs) contacted on-site coordinators (OSCs) periodically by telephone and email. The purpose of these communications was generally to encourage participation in Family Services Tracking (FST), but it also reminded OSCs of the upcoming spring data collection. We also asked each program to describe their procedures for transitioning children out of Early Head Start.

In early December 2010, we mailed holiday cards to all parents of children in the study. Shortly after, we sent cards to all program directors, OSCs, teachers, and home visitors. Parents received an additional insert listing our toll-free number, encouraging them to call in and provide new contact information. The holiday cards reminded all participants of their involvement in Baby FACES and helped with locating families who may have changed addresses. The post office returned several cards with forwarding addresses that allowed us to update Mathematica's sample management system (SMS) before spring 2011 data collection. The same procedures were used prior to the final 2012 data collection with remaining Newborn Cohort study families.

BFCs Collected Updated Sample Information and Confirmed Site Visit Weeks

To prepare for 2011 data collection, BFCs attempted to obtain updated information on all study participants. We designed a roster confirmation spreadsheet that contained name and contact information of each study child and parent at the program, the name of each child's teacher or home visitor, each child's service option, and an exit date if the family had left the program. We also attempted to gather information about each child's transition plans. BFCs emailed these spreadsheets to OSCs in January and February of 2011 and asked OSCs to (1) review and update the address and phone number of each parent; (2) confirm each study child's service option (e.g., home- or center-based); (3) confirm each study child's teacher or home visitor; and (4) determine if and when families left the program and if their leaving was part of a planned transition out of Early Head Start or if it was a premature departure. Further, the BFCs proposed a week for the upcoming site visit and asked the OSCs to confirm the visit. We attempted to maintain the same visit week from the baseline 2009 visit.[7] Once OSCs returned the spreadsheets, BFCs added to the SMS new phone numbers and addresses and updated teacher and home visitor information, if necessary; this information was used to complete parent interviews and conduct in-home child assessments.

While the same pre-field procedures were followed for spring 2012 data collection, we decided to compress the overall length of the field period given that there were only 84 Newborn Cohort children still enrolled at 49 programs. Week 1 of the field period (the first week in March) corresponded to the same time period in past years, however, all visits were completed by mid-April 2012, approximately 8 weeks sooner than in 2011. Whenever possible, sites visited early in the field period in 2011 were also scheduled early in the 2012 field period. A second consideration in 2012,

[7] We were able to confirm the same visit week for 74 sites and scheduled 10 visits one week before or one week after the week of the 2009 visit. Three visits were scheduled within two weeks of the 2009 site visit week, and one site was scheduled four weeks after the 2009 site visit week. One site did not have a 2011 site visit because all study children had exited the program (although we still conducted the program director interview).

was grouping visits to programs based on geographic proximity to limit the costs associated with traveling to the sites. Since most programs only had one study child remaining in 2012 it was both feasible and beneficial to visit nearby programs during the same week. Despite the constraints of a shorter field period and our coupling of nearby programs, we were still able to visit more than half the programs within a week of their 2011 visit week. Of the 49 programs, 15 were visits the same week in 2012 as in 2011, with an additional 11 programs visited a week before or after their 2011 visit week.

Another important component of data collection preparation was obtaining renewed approval from institutional review boards (IRBs). Six programs had required local IRB or school board approval during the first round of data collection. Beginning in February 2011, the BFCs assigned to these programs worked with the OSCs to obtain any necessary IRB renewals. Only four programs needing local IRB approval remained for the 2012 data collection.

Training and Quality Assurance

Spring 2011 data collection included in-home child assessments and video-recorded interactions with both the 1-year-old Cohort and the Newborn Cohort. To prepare for training and data collection, Mathematica staff trained and certified a set of "gold standard" service quality observers, and hired and trained field assessors. Spring 2012 data collection followed a similar procedure.

Mathematica Survey Staff Trained Field Assessors and Observers

For 10 days in early February 2011, we trained staff to conduct on-site data collection. All field staff were returning members of the team who conducted the 2010 Baby FACES data collection and whose work in past rounds was considered satisfactory. Training on the in-home child assessment took place during the first 5 days, and training on classroom and home visit observations took place during the second 5 days. A subset of field interviewers was selected to conduct 2 year old child assessments for the Newborn Cohort in 2011 and received an additional day of review. Table B.1 presents the 2011 training agenda. We trained a total of 22 field staff (9 bilingual) to administer the in-home child assessments to 3 year olds. Of those, 11 (5 bilingual) were chosen to conduct 2 year old child assessments. We also trained 17 field staff (6 bilingual) to conduct classroom and home visit observations as well as in-home child assessments. Field interviewers attended between 5 and 10 days of training depending on whether they were selected to conduct 2 year old child assessments, were conducting observations, or were completing administrations in Spanish. Prior to the training, we mailed field staff a package of preparation materials. The mailing included the field training manual; a DVD of a child assessment; and practice exercises for field staff to complete.

Table B.1. Training Agenda—Spring 2011

Day 1	2 year old child assessment review and discussion of changes
	Practice of 2 year old administration and individual feedback sessions on 2010 administration issues
Day 2	Welcome back to all interviewers, discussion of changes for 2011
	PLS-4 review and paired practice
	Individual feedback sessions on 2010 administration issues
Day 3	ECI and Two-Bags review
	Height and weight review
	PPVT-4 training and practice
Day 4	Overview of entire home visit and procedures for other interviews
	Child assessment certification (for bilingual interviewers)
	Practice conducting child assessment
	Bilingual child assessment training
Day 5	Bilingual child assessment training
	Child assessment certification (non-bilingual interviewers)
Day 6	Bilingual certification
Day 7	Review of CLASS-T and reliability
Day 8	CLASS-T in-field reliability observation
Day 9	HOVRS-A review and recertification
Day 10	HOVRS-A recertification

In-home child assessment training. The topics covered during in-home child assessment training included scheduling appointments; a review of setting up and properly using the video camera; measuring height and weight; administering the Preschool Language Skills (PLS-4) assessment and Early Communication Indicator (ECI) and Two Bags Parent-Child Interaction tasks; and a general session called "Working with 3-Year-Olds." In addition, all field staff received one-on-one feedback based on their work in 2010. As in the previous year, all field interviewers were required to conduct full child assessments attempting certification in the following areas: (1) conducting the PLS-4, (2) administering and successfully recording the ECI; and (3) conducting the PPVT-4. The certification assessments were conducted with 3-year-old children during the 2011 training.

Senior staff trained nine Mathematica project staff to be gold standard assessors for child assessment certification. Each staff member received nine hours of formal training on what to look for and how to conduct the certification observations. To measure adherence to the step-by-step protocol instructions, Mathematica staff members who were trained to serve as gold standard assessors for child assessment certification, observed each field staff member during the administration of the child assessment with a 3-year-old, with a certification form that we developed. Certifiers provided trainees with detailed written feedback on their performance. All staff earned certification at the training. Unlike the previous year, field interviewers were not required to complete a post-training video of a child assessment to be certified.

Classroom and home visit observation training. During the spring 2011 training, we reviewed and re-certified 16 field staff observers and 4 quality assurance observers to conduct the

Classroom Assessment Scoring System-Toddler (CLASS-T), and Home Visit Rating Scale-Adapted (HOVRS-A). The project team, in consultation with the developer decided to use the 2010 version of the CLASS-T measure again in 2011 (and 2012) rather than change to the newer version that had been developed in order to preserve our ability to track changes over time. Therefore, only staff previously trained and certified were allowed to be re-certified in 2011. Once again, an author of the CLASS-T came to training to review materials and re-certify all field staff. Observers reviewed the dimensions and coding practice videos then watched and coded three certification videos. Of the 20 observers retrained, 11 certified on the first three videos (meaning they met the 80 percent inter-rater reliability threshold overall and for each dimension on 2 out of the 3 certification videos). Observers who did not pass the first three videos at training were able to re-certify remotely after training[8]. In total, 19 of the 20 observers trained were successfully certified. Observers also conducted CLASS-T observations in classrooms, and all observers were reliable.

These staff were also re-certified on the HOVRS-A. A group review and practice session was held at the field training followed by additional sessions in which observers viewed and coded video clips to achieve certification. By the end of training, all 16 field staff and 4 quality assurance observers were certified by demonstrating 100 percent reliability with the master codes.

Field teams. Before staff conducted any field visits, we created field teams. Field teams consisted of a team leader and one or more field observers and assessors. The team leader was responsible for managing on-site activities, including scheduling classroom and home visit observations, child assessments, and in-person interviews with teachers and home visitors. Team leaders were also the main point of contact with the OSC during the site visit week. If study children at a site were exposed to Spanish in their households, we sent at least one bilingual team member to the program site. Teams and team leaders were reconfigured each week of data collection to accommodate the needs and sizes of the programs visited.

2012 Training. A total of nine field interviewers (four bilingual) were invited to continue as Baby FACES data collection staff in 2012. The nine selected were the strongest interviewers from previous rounds of data collection. No new field staff was hired for 2012. In February 2012, we trained all staff trained for seven days with bilingual interviewers attending an eighth day. Table B.2 presents the 2012 training agenda. Days 1 through 4 of training focused on reviewing the CLASS-T and HOVRS-A protocol and included all 9 field staff and 2 additional quality assurance observers being re-certified on both measures. As in past years, the developer of the CLASS-T came to the training and conducted the review session using the 2010 CLASS-T version of the instrument. CLASS-T recertification was conducted in classrooms, and all observers were reliable with their fellow observers including at least one observer that was a quality assurance observer in previous rounds of data collection[9]. Days 5 through 7 focused on the child assessment administration and highlighted changes to procedures and refinements to the instruments. In addition, each returning field interviewer had a one-on-one session to review their work in 2011. On the final day of training, each field interviewer conducted an assessment with a 3-year-old and was observed by a gold

[8] Observers received an email with a link to three more videos and were instructed to watch the videos and send their scores to Mathematica who forwarded them to the developer. Five more observers passed on this second certification attempt. The remaining observers were sent a final opportunity to certify on three additional videos. All but one passed.

[9] Due to the small sample size, we no longer needed such a large team of quality assurance observers; consequently, observers who previously acted as gold standards, were no longer needed in this role and were instead considered regular observers for 2012 data collection.

standard assessor to be certified. All field interviewers were certified by the end of training. Bilingual interviewers were also certified on their ability to complete the assessment in Spanish.

Table B.2. Training Agenda—Spring 2012

Day 1	Welcome back, discussion of changes for 2012
	CLASS-T review
Day 2	CLASS-T in-field reliability observations
Day 3	HOVRS-A review and recertification
Day 4	HOVRS-A recertification
Day 5	Child assessment overview, PLS-4 review, PPVT review
	ECI and 2 Bags review, Height and Weight review
	One-on-one feedback based on previous year issues
Day 6	Observe PLS administration
	Basal and ceiling review and quiz
	Paired practice
Day 7	English certification
	Bilingual child assessment review
Day 8	Bilingual paired practice
	Bilingual certification

Prior to the training in early February, each field interviewer was sent a pre-training packet that included a manual, DVD of a child assessment administration, and a worksheet to be completed and returned before training. The worksheet concentrated on changes for 2012 areas that gave interviewers trouble in the past. Results of the worksheet exercise were incorporated into the full training.

Gold Standard Observers and Assessors Conducted Quality Assurance Field Visits

For quality assurance (QA) purposes, we sent trained gold standard observers to monitor each member of the field staff conducting the HOVRS-A and CLASS-T. In 2011, five Mathematica staff members and two employees of Branch Associates served as the gold standard observers on 16 sites visits over 9 weeks. For the HOVRS-A, the gold standard observer accompanied each field observer to monitor a home visitor conducting a home visit. The gold standard and field observers rated the visit independently and discussed their scores immediately afterward. Inter-rater reliability was calculated as having at least 80 percent exact agreement with the gold standard. For the CLASS-T QA, the gold standard observer and up to two observers completed the classroom observations simultaneously, and then discussed each item immediately following the observations. Reliability was calculated where observers were considered reliable if they were within one point of the gold standard observer at least 80 percent of the time.

We also conducted a QA review of the PLS-4 portion of the in-home visit. Interviewers were asked to video record a child assessment visits during their first week or two in the field in 2011. Each assessor recorded a full administration of the PLS-4 during a regularly scheduled in-home child assessment and sent it to Mathematica for review by gold standard assessors. Feedback for those who did not pass the PLS-4 QA was sent to the assessor and he or she was asked to complete a second video for QA review. Videotapes were received over the course of the entire field period. All

but one interviewer passed the QA review of the child assessment administration. The only field interviewer who did not pass was let go from the project and the assessments that he conducted were re-administered by certified staff.

For spring 2012, we followed the same procedures sending QA observers to monitor CLASS-T and HOVRS-A observations. In total we were able to complete QA visits on CLASS-T with all 9 field staff and on HOVRS-A with 8 of the 9 field staff (the ninth field interviewer was a bilingual gold standard). In 2012 we decided to conduct in-person QA observations of child assessment administrations, rather than requesting field staff to send in video. Each field interviewer was observed by a gold standard QA person who watched the assessment and completed a certification form (the same one used to recertify staff during the training). The QA observer then gave the field interviewer in-person feedback and determined if a second observation was needed. No field interviewers required a second observation. QA observers followed up with a memo detailing the strengths and weaknesses of the assessment. This memo was given to both the field interview and the project staff. Whenever possible, bilingual interviewers were observed conducting an assessment in Spanish, although this was not always possible.

Three Teams of Video Coders Were Trained to Reliability

We created three teams of coders, each supervised by a gold standard coder; two senior survey researchers supervised the entire coding team. One team coded the Two Bags task using the Two Bags Parent-Child Interaction Rating Scales for the Two Bags Assessment. Another team coded the Two Bags task using the Parenting Interactions with Children: Checklist of Observations Linked to Outcomes (PICCOLO), with the third team coding the ECI task. Each team received extensive training and ongoing reliability checks as described below.

Two Bags. In spring 2011, the Two Bags gold standard coders from the 2010 coding team trained 11 members of the Mathematica coding team. Coders independently coded a video-recorded interaction coded a priori by a Two Bags expert who provided the coding training in 2010 to serve as the certification video. The certification criterion required that coders achieve 92 percent agreement (exact or within one point) with the ratings assigned by the expert across the 12 scales. Two additional certification videos were available to coders who did not certify on the initial certification video.

Of the original team of coders who were trained by the gold standard coders, six members composed the final coding team. Following training and certification, the team leaders worked with the six-member coding team to establish and maintain inter-rater reliability throughout the coding period. Inter-rater reliabilities between the team leaders and coding team members were established on the 12 seven-point scales to a criterion of 80 percent, allowing for a one-point difference in scores. Thereafter, the team conducted weekly inter-rater reliability checks on three to five randomly selected videos.

PICCOLO. For the Baby FACES study in 2011, the PICCOLO gold standard coders from the 2010 coding team 11 members of the Mathematica coding team, two of whom had coded PICCOLO in 2010. Coders independently coded four video interactions coded a priori by the PICCOLO research team to serve as the certification videos. Coders were required to achieve 80 percent exact agreement using a binary scale in which "0" and "1" were collapsed to represent behaviors that were absent or infrequently observed. A score of 2 still indicated behaviors that were "clearly evident" and frequent in their occurrence and/or intensity.

Of the original team of coders who were trained by the gold standard coders, seven members composed the final coding team. Following training and certification, the team leaders worked with the seven-member coding team to establish and maintain inter-rater reliability throughout the coding period. Inter-rater reliabilities between the team leaders and coding team members were established on the binary scale to a criterion of 80 percent. Thereafter, the team conducted weekly inter-rater reliability checks on three to five randomly selected videos.

Early Communication Indicator. Prior to the Baby FACES spring 2011 data collection, two ECI gold standard coders trained nine members of the Mathematica coding team (two coded the ECI in 2010). One of the gold standard coders was a gold standard coder in 2010 and the other gold standard coder was a member of the 2010 coding team. ECI coders were required to become certified on two videos coded a priori by the developer. The certification criterion required that coders achieve 85 percent agreement with the developer ratings on each video. Coders recoded the videos as many times as necessary until meeting the certification criterion.

Of the original team of coders who were trained by the ECI gold standard coders, five members composed the final coding team. Following training and certification, the gold standard coders worked with the five-member coding team to establish and maintain inter-rater reliability throughout the coding period. Inter-rater reliabilities between the team leaders and coding team members were established on the four ECI dimensions to a criterion of 80 percent. Thereafter, the team conducted weekly inter-rater reliability checks on three to five randomly selected videos.

Coding 2012 Videos. Given there were only 84 children remaining in the study in 2012, we decided to limit the coding teams to the team leaders and coders from past rounds. Three teams of two people, one of whom was bilingual, coded all the videos. As in the past, each team of two specialized in their coding—either coding the Two Bags task using the Two Bags Parent-Child Interaction Rating Scales for the Two Bags Assessment or using the PICCOLO; or coding the ECI. We completed reliability coding of 20 percent of the Two Bags and PICCOLO videos and 16 percent of the ECI videos to maintain inter-rater reliability between team members throughout the coding period.

Telephone Interviewers Were Trained to Administer the Parent Survey

In 2011, we followed the same training process for the parent survey as was instituted for the previous two rounds of Baby FACES data collection. Two groups of (daytime and evening) telephone interviewers received eight hours of training for the parent survey in early February 2011. We trained 16 telephone interviewers (four were bilingual in English and Spanish). In addition, six monitors (two of whom were bilingual) and four telephone supervisors participated in the training sessions. Training involved a brief overview of the project and how the parent interview fit into the overall data collection effort, instruction on gaining cooperation and screening of parents, and a question-by-question review of the survey instrument. At the conclusion of the formal training, interviewers were paired up to conduct mock interviews with one another using the Computer-Assisted Telephone Interview (CATI) instrument under the guidance of the trainer and supervisors. During the first weeks of telephone interviewing, each interviewer was monitored and given immediate feedback. Ongoing monitoring of 10 percent of the interviews continued throughout the telephone field period. We monitored bilingual interviewers in both English and Spanish.

For spring 2012, we brought back six telephone interviewers (two daytime and four evening) and trained them together over two days in a four-hour session and two-hour practice session. As in

the past, the training included a brief overview of the study (this time emphasizing how this final round of data collection was different), a question-by-question review, followed by paired practice. Ongoing monitoring of 10 percent of the interviews was conducted throughout the 2012 parent data collection period.

Training on Program Director Survey

In late March 2011, we reassembled the team of researchers that conducted the program director interview in previous years. A new researcher was added to the team replacing a previous team member. The Baby FACES project and survey directors met with the two returning and one new researcher for a four-hour training session. The training built on what the researchers had learned from the past surveys, highlighting areas that sparked questions in prior years and describing the intent of new questions added for this round. In addition, the training included a review of the programs' structure and stressed the importance of gathering enough information from the program director to accurately record the information on the questionnaire. Extensive spreadsheets were created to capture additional information provided during the semi-structured interviews about programs' organization and activities.

We did not conduct interviews with program directors in 2012.

Interviews, Observations, and Assessments

Mathematica Conducted Exit/Matriculation Interviews with Families Leaving Early Head Start

The Baby FACES research team periodically conducted exit interviews to better understand why families leave Early Head Start and where they go once they leave. There were two versions of exit interview. One for those who we describe elsewhere in the report as "early exiters" and a "matriculation" interview administered to parents of 1-year-old Cohort children who remained enrolled until their child was 3. We conducted the first round of the exit interviews in fall 2009 and a second round at the same time as the parent interview during spring 2010.

Spring 2011 Exit/Matriculation Interview. In spring 2011 we expanded the third round of the exit interview to include questions about transitioning out of Early Head Start. The 20-minute interview was administered to parents who indicated they had left the program. It collected (1) the child's exit date from the perspective of the parent; (2) the reasons for leaving the program; (3) satisfaction with the program; (4) child health; (5) current child care arrangements; and (6) annual household income and sources of income support. For parents who indicated in the survey that their child had transitioned out of Early Head Start, we asked additional questions about how the program prepared and assisted parents with the transition. We offered parents $20 to complete the exit interview. We incorporated the exit interview items as a separate module of the parent interview so all telephone interviewers trained on the parent interview received training on the exit items as part of the parent interview training.

The third round of exit interviews began in February 2011 at the same time as the parent interview and concluded in mid-June when we closed out the parent interview. It contained new exits we learned about from (1) roster spreadsheets being returned to BFCs prior to spring data collection; (2) staff entering exit dates into the FST; (3) parents informing interviewers during the

parent interview; and (4) field staff notification during site visits. The total sample interviewed was 169 parents.

We operationally decided to continue treating families who recently stopped receiving Early Head Start services as "still enrolled" for the purposes of data collection if they exited near the time of their site visit week. We based our definition of "near the time of the visit" on the fact that we received exit dates from multiple sources (roster confirmation spreadsheets, OSCs, field staff, and parent interviews), making it difficult to assign an exact "exit" date. Parents and programs were often able to provide only an exit month. Therefore, our eligible data collection window included cases with exit dates in the month of the site visit or the month before the site visit. For example, if the first day of the site visit was scheduled for March 15, then any exit date from February 1 to March 14 was considered inside the data collection window. Any exit date before February would be considered outside the data collection; these parents would instead be contacted to complete the exit interview.

Recently exiting families considered eligible for regular data collection were treated the same as families still receiving Early Head Start services from the study program. We attempted the in-home child assessment if the family was still in the program area and still called to complete the regular parent interview. We also asked teachers and home visitors to fill out a staff-child rating (SCR) and complete the teacher/home visitor interview. If a child was in center-based care before leaving the program, we conducted a CLASS-T observation of the child's last classroom. If a child was in home-based care and he or she was the only study child on the home visitor's caseload, we were unable to complete a HOVRS-A with the home visitor. However, we attempted observations if the home visitor served other study families.

Fall 2011 Matriculation Interview. Beginning in August 2011, we contacted all 1-year-old Cohort parents with whom we had attempted or completed an interview during the spring 2011 data collection period (that is, who remained enrolled in the program roughly through age 3). We presumed that by fall 2011, all of our 1-year-old Cohort children would have left the Early Head Start program and could answer question about the transition process. Parents were called between August and October 2011 and asked the same series of questions asked of those who exited or transitioned out of the program earlier. In the rare cases where we learned during our call that the child had not yet transitioned out of Early Head Start, we determined when the transition was scheduled and postponed the interview until after that date. As in the past, parents who completed the exit interview were sent $20 for their time participating. We trained 13 telephone interviewers, 10 of which worked on the parent interview in the spring. The total sample for the fall matriculation interview was 463 and of those 337 completed interviews (73 percent).

On-Site Data Collection Consisted of Classroom and Home Visitor Observations, Teacher/Home Visitor Interviews, and In-Home Child Assessments

This section describes in detail the various components of the data collection effort in spring 2011. The spring 2012 data collection was identical in that the same elements were collected, albeit on a smaller group of children. Because the 1-year-olds Cohort had left the study by spring 2012, only 3 year olds were included in this final round of data collection.

Teams of field assessors visited 88 sites over 15 weeks from March 2011 through mid-June 2011.[10] Several weeks before the site visits, BFCs and Survey Operations Center (SOC) field supervisors developed data collection plans (DCPs) for each site. The DCPs serve as the team leader's guide to collecting all required instruments. Each DCP included (1) the addresses and contact information of program centers and OSCs; (2) a listing of each study child including his/her teacher or home visitor, service type, cohort, and whether he or she should be assessed in Spanish; and (3) a listing of teachers/home visitors, including a checklist of the required instruments.

Classroom observations. We conducted observations of all teachers using the CLASS-T to observe all teachers providing services to children in both the 1-year-old Cohort and the Newborn Cohort (at ages 3 and 2, respectively in spring 2011). The CLASS-T collects information on classroom climate (both positive and negative), teacher sensitivity, regard for child perspectives, behavior guidance, facilitation of learning and development, quality of feedback, and language modeling.

We conducted classroom visits similar to the 2009 and 2010 data collections. We scheduled the classroom observations in the morning, when children are most active. During each observation the observers completed two counts of children and adults (spaced at least an hour apart) and completed a post visit rating consisting of three items; there was no interaction with the children or the teachers. Classroom observations generally lasted two to three hours. We gave a gift bag of classroom supplies worth $25 to the teacher in each observed classroom.

Home visitor observations. Using the HOVRS-A, observers conducted a home visit observation of each home visitor serving at least one child in the study, the same as in the 2009 and 2010 data collections. This observation tool focuses on the quality and nature of aspects of the home visit interaction, including home visitor responsiveness to the family, the relationship between the home visitor and the parent, and the engagement of the parent and the child during the home visit. During these observations, observers did not interfere with or react to home visit activities or interactions. Each HOVRS-A observation lasted an average of 90 minutes. We gave each observed home visitor a gift bag of supplies that was identical to those we gave the teachers.

Teacher/home visitor interview and staff-child report. We again interviewed each teacher and home visitor providing services to at least one study child. The in-person interview was nearly identical to the spring 2010 interview. It lasted approximately 30 minutes and focused on their background, training, services they provide to families, and expectations the programs place on them. Minor changes to the interview included some modifications to age-specific questions (since the children were now 2- and 3- years old).

The field assessor conducted the interview and recorded the teacher/home visitor's responses on a paper questionnaire for later data entry. The two instruments were nearly identical, with a few questions modified to reflect services provided in the home versus in a classroom.

At the start of the week of the site visit, we again distributed SCR forms to each teacher and home visitor of study children and instructed them to complete this self-administered questionnaire (SAQ) about each of the study children. In 2011, there were four versions of the SCR: one each for home visitors and teachers of children in the Newborn and 1-year old Cohorts. (That is, there was a

[10] Assessors visited between 1 and 9 sites during each week of data collection (there was one week with no site visits scheduled).

home visitor form for Newborn Cohort children and another for 1-year-old Cohort children, and so on.) Teachers were asked to report on two additional questions specific to child language exposure in the classroom. All teachers and home visitors were asked to report on the child's social skills, language development, and parent-staff relationships. The key difference between the Newborn and 1-year old Cohorts' SCRs was the version of the language measure (CDI) used; teachers and home visitors of the Newborn Cohort were asked to report on a different set of words than those of the 1-year old Cohort.

After the visit, the team leader collected and reviewed all completed instruments and sent the documents to Mathematica for receipt and review. Whenever possible, we collected completed SCRs before the end of the visit week. The forms took about 15 minutes to complete for each child. Teachers and home visitors received $5 for each completed form. If the SCRs were not delivered or completed by the end of the site visit, teachers and home visitors were given prepaid business reply envelopes and were asked to mail the documents to the SOC.

Child assessments. During spring 2011, assessors conducted in-home child assessments with each study child—now 2 and 3 years old. The assessors attempted to schedule the assessments with parents a week in advance of the site visit, when possible. Approximately two weeks prior to the target week, we sent parents a letter explaining the visit and included a list of activities the parent could "try out" with the child to prepare for the visit (see Parent SAQ below). Different versions of the letter were sent to parents of 2 year olds and parents of 3 year olds. When parents completed the telephone interview in advance of the site visit, assessors provided them with a toll-free field number they could call to schedule the assessment and obtain more information. If an assessment was not scheduled ahead of the site visit, the team leader worked with the OSC and teacher or home visitor to contact the family and find an appropriate date. Occasionally, depending on the comfort level and/or availability of the parent, the program provided a private setting at a center to conduct the assessment.

We created two different versions of the child assessment record form—the 24-Month Child Assessment Record Form and the 36-Month Child Assessment Record Form. Both provided the assessor a script to follow, step-by-step instructions for administering each item of the PLS-4 (using a separate flip book) and coding pages to record responses and observations. There were two different language versions of the 24-Month and 36-Month Child Assessment Record Forms: an English version and a bilingual version (containing both English and Spanish items). If our records or previously collected data indicated the child is exposed to Spanish in the home, we sent the assessor a bilingual record form. If we did not have this information, the assessor asked the parent questions (during the scheduling call or at the beginning of the visit) to determine which form to use.

The first activity performed with the child was the PLS-4 scale, which is used to determine the child's language comprehension skills. Children are asked to follow directions and perform tasks using manipulatives (such as a box, toys, and blocks) or point to objects found in a picture manual. The PLS-4 took approximately 20–30 minutes to conduct depending on the age of the child (and slightly longer for bilingual children who are asked some questions in both languages).

The PLS-4 was followed by the collection of the child's height and weight. The assessor twice measured the child's height in centimeters and weight in kilograms. If the two measurements differed by more than 0.2 kg or 2 cm, assessors took a third measurement. When measuring height, the assessor asked the child to stand straight up (without interference) against a wall or door. The

assessor then lowered a carpenter's triangle from above the child to where it firmly touched the crown of the child's head. The assessor placed a removable stick-on arrow flag on the wall at the bottom of the triangle to mark the child's height, then used a metal measuring tape to determine the child's height. To record the child's weight, the assessor used a digital scale.

The next activity performed with 3-year old children only, was the PPVT-4, which is used to determine the child's receptive vocabulary skills. Children are asked to point to a picture representing a stimulus word named by the assessor. The PPVT-4 was only conducted in English. Spanish speaking children were asked to do the best they could identifying pictures based on the English words. Non-English speaking children were routed out of the PPVT-4 if they were unable to pass the warm-up items.

Next, the assessor recorded the parent-child interaction task. The purpose of the interaction (Two Bags task) is to assess parent and child behaviors as the pair interact in semi-structured free play. To set up the interaction, the assessor first found an unobtrusive place for recording, placed a yoga mat on the floor, and set up the video camera. Then, the assessor began recording by holding a signboard in front of the camera for 15 seconds (to identify the video). Once set-up was complete, the assessor asked the parent and child to sit on the mat. The assessor then placed bag number 2 on the floor to the parent's left and bag number 1 on top of it. The first bag contained a book entitled *Goodnight Gorilla*, and the second bag contained a set of toy dishes and play food. The assessor, using scripted standardized instructions, informed the parent that the recording would last eight minutes and that he or she may now play with the child. The assessor asked the parent to begin playing with bag number 1 first, then move on to bag number 2 whenever he or she liked. During the recording, the assessor monitored the video camera to verify that the parent and child stayed within the video frame.

After completing the parent-child interaction task, the assessor set up the observer-child task, the ECI. The purpose of this task is to measure the child's expressive communication skills. With the video camera running, the assessor changed places with the parent on the mat, unpacked a Fisher-Price farm set, and interacted with the child for eight minutes. Per the ECI developer training standards, the assessors were trained to follow the child's lead, talk about things of interest to the child, comment on the child's actions, and repeat what the child was saying (particularly if the child spoke softly and the microphone might not pick up their vocalizations). The assessor was instructed to ask questions sparingly. Following the video tasks, the assessor gave the child an age appropriate book as a gift.

The assessor finished the visit by asking the parent a series of questions intended to help researchers understand the context of the visit. The questions ask whether the child's behavior during the visit was typical, whether the child is generally shy or outgoing, and whether the parent thought the child did his or her best. The assessor then collected the self-administered questionnaire that the parent had received at the start of the visit (see Parent SAQ below). Next, the assessor verified that the parent had completed the telephone interview. If the interview had not been completed, the assessor gave the parent our toll-free number and called to set up an appointment with the SOC. The assessor then provided the parent with an appointment card that included the toll-free number and time and date of the appointment. Finally, before leaving the house, the assessor gave the parent a $35 check as a thank you for both completing the telephone interview and their involvement in the child assessment. The visit lasted one and a half hours, on average.

As soon as possible after leaving the home, the assessor completed ratings of the (1) child's behavior during the session (Bayley BRS); (2) their observation of the home (using the Home Observation for Measurement of the Environment, or HOME); and (3) the neighborhood, using items from a study of neighborhoods in Chicago (Ross et al. 2008). The HOME observation measures the quality of stimulation and support available to children in their home environments, and the neighborhood items describe the condition and safety of the neighborhood.

Parent SAQ. Parents were asked to complete a self-administered questionnaire (SAQ) during the child assessment. The SAQ consisted of several measures about child development in a number of areas, including social skills, communication and language development, gross and fine motor development, and personal growth. Two different versions of the SAQ were created—one for parents of 2 year olds and one for parents of 3 year olds. Some of these measures were administered at baseline during the telephone interview. The SAQ was available in either English or Spanish. As previously mentioned, parents received an advance letter that included a list of activities the parent could try out with the child to prepare for the visit. The purpose of this list was to prepare parents to complete the Ages and Stages, Third Edition (ASQ-3) portion of the questionnaire. This measure asks parents whether the child can regularly, sometimes, or not yet do particular activities. Because the PLS-4 and recorded interactions require a substantial amount of time, the intent of the advance letter was to provide parents an opportunity to try some of these activities (which parents may have never tried or seen their children do) before the visit. Some activities from the advance letter for 2 year olds included:

- Let your child drink from a cup and use a spoon
- Let your child kick a soccer-sized ball; watch him or her run, jump, and go up and down stairs
- Let your child practice stacking with a least eight small blocks or containers
- Give your child beads or Cheerios and string to practice beading
- Read books to your child, ask questions about the pictures, and let him or her try to turn the pages
- Let your child scribble and try to make lines with crayons or markers

While the following activities were listed as part of the 3 year olds' advance letter:

- Show you child how a zipper on a coat moves up and down and ask your child to move the zipper
- Let your child catch a large ball with both hands from about five feet away
- Let you child put together a five to seven piece interlocking puzzle
- Show you child how to make a bridge with blocks, boxes, or cans and let your child copy you
- Say three numbers and let your child repeat them
- Let your child use a large spoon to scoop applesauce from a jar to a bowl
- Let your child copy basic shapes and patterns using a pencil or crayon without tracing

After handing the parent the SAQ, the assessor explained that the parent should not expect to see the child do everything listed. If the parent wanted to try out an activity, he or she was told to

circle it. The parent could then come back to the activity at the end of the visit and try it with the child.

Mathematica Conducted Parent Interviews and Program Director Interviews by Telephone

Parent interviews. Parents of each study child were asked to complete a telephone interview in 2011 and 2012 around the time of the on-site data collection effort. Mathematica telephone interviewers conducted the interview at the SOC in Princeton, New Jersey. The interview was programmed and administered using CATI, thereby allowing the individual path of each interview to be determined based on the responses given to previous questions or preloaded information. The interview was conducted in Spanish when necessary. The parent interview had 20 sections, although not all parents were asked questions in every section. The parent interview instrument, along with all other study instruments, will be available on the Administration for Children and Families (ACF) website. As mentioned earlier, a new module for the exit interview was added to the parent interview in spring 2011.

We conducted parent interviews from mid-February through June of each year. We attempted to conduct the parent interviews before the field site visits. We began calling parents at least one month prior to their site visits, when possible. Parents were sent an advance postcard informing them of the upcoming interview and reminding them of the upcoming Baby FACES data collection. In 2011 we released four large groups of sample to the CATI system between February and April and temporarily stopped calling parents during their site visit week and the week prior to allow field staff to schedule appointments for the in-home child assessments. Due to the much smaller sample size in 2012, there was only one sample release.

In the screener portion of the interview, parents were asked to confirm whether they were still receiving services from the study program. If the family had left the program, the parent was asked for an exit date. The CATI program was designed to calculate whether the exit date was inside the eligible data collection window for families to receive the parent interview. If it was, the interviewer proceeded to complete the full parent interview. In 2011, if the exit date was outside the eligible data collection window, the CATI program took the interviewer to the exit module within the program. In 2012, the CATI program would end the interview since we did not conduct exit interviews that round.

While calling parents, interviewers identified many incorrect or nonworking telephone numbers. We generally put these cases on hold and waited for field staff to locate the parents for the in-home child assessment. BFCs also played a role, by contacting OSCs and asking them for updated phone numbers. In some instances, the programs provided private space at a center where parents could call in and complete the interview using the center's phone. As a final attempt, SOC locators attempted to find telephone numbers through directory assistance and online sources.

The average length of the parent interview was approximately 45 minutes; Spanish interviews lasted 60 minutes, on average. Parents received $35 for completing this round of data collection.[11] Those who completed only the exit interview questions (and therefore had no home visit) received a $20 thank you check.

[11] Parents were either given a check at the conclusion of the in-home child assessment or mailed a check at the conclusion of the parent interview if they did not complete the in-home assessment.

Program director interviews. To learn about program practices, policies, and overall enrollment, we conducted interviews with program directors in spring 2011. The program director interview was broad in scope and asked directors about the entire program. We gathered program-level information through an hour-long telephone interview. While the process for conducting the interviews was similar to previous rounds in 2009 and 2010 we eliminated the self-administered portion of the interview and revised the telephone questionnaire to include new questions incorporating key aspects of questions previously gathered through the SAQ. The interview focused on program structure, involvement with community partners, approaches to serving DLLs, implementation, and program goals.

We conducted interviews from April to July 2011. Each program director was mailed an advance letter and about a week after the mailing, researchers began calling to schedule the telephone interviews at the program director's convenience.

Upon completion, each researcher reviewed his or her own interviews, entered verbatim comments into a spreadsheet, and determined whether a call-back was needed. Because the program director interview resembled a semi-structured executive interview, the interviewers recorded extensive additional information on spreadsheets to capture data that went beyond the questionnaire form and could facilitate better understanding of program activities.

Response Rates

Most Parents Completed the Parent Interview

We completed telephone interviews with 425 parents (76 percent) during the 2011 data collection. A few parents partially completed the interviews: 7 parents (1 percent) completed the household composition and household languages sections, and another 13 parents (2 percent) completed at least half of the interview. We completed five first-time interviews, meaning we had not previously completed an interview with any parent. In addition, 14 interviews were conducted with new parents, meaning someone other than 2009 or 2010 respondent completed the 2011 parent interview.

A high percentage of parents from both cohorts completed the interview: a total of 87 percent of Newborn Cohort parents and 78 percent of 1-year-old Cohort parents. We completed 104 interviews in Spanish (16 from the Newborn Cohort and 88 from the 1-year-old Cohort).

We were unable to complete the interview with 114 parents (21 percent). Of parents not completing the interview, most (64 percent) were parents with working telephone numbers for whom we continuously reached voicemail systems, failed to receive an answer, or could not set up a suitable appointment time. We were unable to locate a working telephone number for 18 percent of this group, and 18 percent refused to complete the interview. In two instances, the interview could not be completed because of a language barrier.

In 2012, we interviewed the same parents for the fourth year in a row. We completed interviews with 60 parents (71 percent) and one parent partially completed the interview. Of parents not completing the interview, most (70 percent) were parents with working telephone numbers for whom we could not reach. We were unable to locate a working telephone number for 13 percent of this group, and 17 percent refused to complete the interview.

Assessors Achieved a High Response Rate for the Child Assessment

In spring 2011, we conducted a large percentage of in-home child assessments and video-recorded interactions with children and their parents. Assessors completed child assessments with 503 children (90 percent). We received 142 bilingual versions of the Child Assessment Record Form. A small number of parents (3 percent) refused the entire in-home visit. We were unable to locate two families. The remaining parents typically were unable to overcome scheduling conflicts to participate.

From the SAQs we distributed during the in-home assessment, we collected or received 481 (out of the 503 visits and 559 total cases) for a completion rate of 86 percent. In some instances the parent did not complete the SAQ in the home or the child assessment took place at an Early Head Start center without the parent. In addition, we received 465 (83 percent) codeable Two Bags videos and 473 (85 percent) working ECI videos. See Table B.3 for completion rates.

We were able to achieve equally high completion rates in spring 2012 with our Newborn Cohort. Assessors completed assessments with 76 children (90.5 percent). We received 22 bilingual versions of the Child Assessment Record Form. Further, we collected or received 70 SAQs (83 percent), 68 (81 percent) codeable Two Bags videos and 73 (86.9 percent) ECI videos.

Teachers and Home Visitors Continued to Participate in High Numbers

Response rates in 2011 remained high for teacher and home visitor instruments. We observed 231 classrooms of teachers with study children, achieving a response rate of 99 percent on the CLASS-T. We observed 139 home visits for a response rate of 84 percent on the HOVRS-A. In addition, we completed 232 teacher interviews and 174 home visitor interviews, for response rates of 99 percent for each. Teachers and home visitors were again receptive to completing the SCRs. We received a total of 538 SCRs from 319 teachers and 219 home visitors for a 96 percent completion rate.

Cooperation and completion rates with teachers and home visitors of our Newborn Cohort children continued to be very high in spring 2012. We observed 42 classrooms of teachers with study children, achieving a response rate of 96 percent on the CLASS-T. We observed 20 home visits for a response rate of 87 percent on the HOVRS-A. In addition, we completed 44 teacher interviews and 29 home visitor interviews, for response rates of 100 percent for each. Teachers and home visitors were again receptive to completing the SCRs. We received a total of 82 SCRs from 50 teachers and 32 home visitors for a 98 percent completion rate.

We Attempted to Contact Families Who Left Early or Transitioned out of Early Head Start

In spring 2011, we attempted to contact 169 families that left (132) or transitioned (37) out of their Early Head Start programs after the spring 2010 data collection. Overall, we completed 95 (56 percent) interviews (see Table B.4). We were able to gather partial data from an additional 3 parents (2 percent). We could not locate a working telephone number for 47 parents (28 percent) and could not reach or set a suitable interview time with 18 parents (11 percent). In addition, 5 parents (3 percent) refused to participate in this exit interview. Of the 132 families who left early, we completed 68 interviews and 3 partial interviews (54 percent). Furthermore, of the 37 families who transitioned out of EHS, we completed 27 interviews (73 percent).

Table B.3. Baby FACES Annual Response Rates by Cohort

Instrument	2009 Number Completed (Percentage)			2010 Number Completed (Percentage)			2011 Number Completed (Percentage)			2012 Number Completed (Percentage)
	Newborn Cohort	1-year old Cohort	Both	Newborn Cohort	1-year old Cohort	Both	Newborn Cohort	1-year old Cohort	Both	Newborn Cohort
Staff-Child Report	185 (95.3)	748 (98.1)	933 (95.5)	128 (94.8)	575 (95.8)	703 (95.6)	93 (96.9)	445 (96.1)	538 (96.2)	82 (97.6)
Parent Interview (CATI)	175 (90.2)	719 (91.9)	894 (91.6)	108 (80.0)	475 (79.1)	583 (79.3)	84 (87.5)	361 (77.9)	445 (79.6)	61 (72.6)
Parent SAQ	--	--	--	--	537 (89.5)	--	87 (90.6)	394 (85.1)	481 (86.0)	70 (83.3)
Child Assessment	--	--	--	--	547 (91.2)	--	89 (92.7)	414 (89.4)	503 (90.0)	76 (90.5)
Caregiver Interview	--	--	229 (93.1)	--	--	267 (98.9)	--	--	232 (98.7)	44 (100)
Home Visitor Interview	--	--	323 (96.7)	--	--	225 (97.0)	--	--	174 (99.4)	29 (100)
ITERS-R	--	--	223 (94.9)	53 (98.1)	--	--	--	--	--	--
CLASS-T	--	--	--	--	220 (98.7)	--	--	--	231 (99.1)	42 (95.5)
HOVRS-A	--	--	242 (89.3)	--	--	193 (83.2)	--	--	139 (84.2)	20 (87.0)
Two Bags/PICCOLO	--	--	--	--	522 (87.0)	--	81 (84.3)	384 (82.9)	465 (83.2)	68 (81.0)
Early Communication Indicator	--	--	--	--	519 (86.5)	--	80 (83.3)	393 (84.8)	473 (84.6)	73 (86.9)

Source: Sample Management System.

Notes: Cohort-specific response rates are not meaningful for caregiver/home visitor interviews because caregivers and home visitors were interviewed once each year, and each could have children from both cohorts in their classrooms or on their caseloads. In addition, cohort-specific response rates are not always meaningful for observations (ITERS-R, CLASS-T, HOVRS-A) for the same reason. However, in 2010 the ITERS-R was only conducted with caregivers of the Newborn cohort, while the CLASS-T was only conducted with caregivers of the 1-year-old cohort.

ITERS-R = Infant Toddler Environment Rating Scale-Revised; HOVRS-A = Home Visitor Rating Scale-Adapted; CLASS-T = Classroom Assessment Scoring System-Toddler version.

Table B.4. Exit/Matriculation Interview Response Rates

	Round One Exit Interview (Oct – Dec 2009)		Round Two Exit Interview (March – June 2010)		Round Three Exit/Matriculation Interview (February – June 2011)		Round Four Matriculation Interview 1-year old Cohort only (August–October 2011)	
	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
Complete	62	54.9	76	38.8	98	57.9	337	72.8
Not complete	51	45.1	120	61.2	71	42.0	126	27.2
Total	113	100.0	196	100.0	169	100.0	463	100.0

Source: Sample management system.

Note: The total number of cases released for round two includes the 51 incomplete responses from round one. Overall, between rounds one and two, we released a total of 258 unique cases. The combined response rate of rounds one and two is 54 percent.

In the fall of 2011, we attempted to re-contact all the 1-year-old Cohort parents (whose children were still enrolled in Early Head Start or eligible during the spring 2011 data collection) and complete a matriculation interview (see Table B.4). In total, we attempted to reach 463 parents and were able to complete matriculation interviews with 337 (73 percent). We were unable to locate a working telephone number for 38 parents (8 percent) and could not reach or complete an interview with 76 parents (16 percent). Furthermore, 12 parents (3 percent) refused to participate in this final interview. No exit or matriculation interviews were planned for the Newborn Cohort parents.

Interviews Were Completed with All Program Directors for a Third Straight Year

We conducted telephone interviews with a program director or designee in all 89 programs (100 percent) in spring 2011. In most cases (70 out of 89), the interview was conducted with the program director. In the other 19 cases, the OSC or another person designated by the program director completed the telephone interview. In 16 cases, more than one respondent participated in the telephone interview. We did not conduct program director interviews in 2012.

Data Processing

Receipt Control Involved Several Steps and Validation Procedures

Receipt processes were the same as in spring 2011 and 2012 as in past years. After field materials returned to the SOC, SOC field staff reviewed the materials for each site, looking to ensure that all materials had arrived and that they matched the data collection plan.[12] If a child's teacher/home visitor had changed or was different than stated in the data collection plan, SOC field staff verified this change with the team leader and BFC and updated the SMS. After this review, SOC field staff transferred all materials to the receipting department at the SOC with a transmittal form to be scanned into the SMS. After being receipted, documents were placed into batches according to instrument type and transferred to quality control for review.

Quality Control and Data Entry use Two-Person Process

Quality control and editing of documents. We trained staff to review and edit documents for quality control (QC) and retrained one QC supervisors to oversee the entire process. In January 2011, we retrained three QC staff to review and edit the observation instruments (HOVRS-A, and CLASS-T). Given the small sample size in 2012, survey staff that had previously trained QC staff opted to complete the review and editing themselves. QC staff reviewed the instruments for completeness and checked that the skip logic had been followed correctly. Where possible, QC staff made appropriate edits to the instruments based on pre-established specifications. The specifications dictated how QC staff should review scores, when to mark data as missing, and when to set data aside and flag it for review. Following review, project staff would instruct QC staff how to proceed.

Quality control of videos. Video recordings of the Two Bags task and ECI were captured on mini-DVDs that were transferred to the QC department after receipt. QC staff imported the videos onto PCs, added time stamps, and saved the video files to a secure drive for coding. If QC staff received a blank DVD or a DVD with no usable video (less than 3 percent of videos received), the video instruments were marked in the SMS as "Received Blank." At this stage, the QC supervisor

[12] The field materials returned included teacher and home visitor interviews, HOVRS-A, , CLASS-T, SCRs, 24-month or 36 month Child Assessment Record Forms, and parent SAQs.

created coding sheets from the SMS and assigned the videos to coders. The coders reviewed the videos and completed the appropriate coding sheet. If necessary, coders would place videos in supervisor review for reasons including poor lighting or audio, parents or children not being in the frame, and failure to meet the required length. However, videos were most commonly placed in supervisor review when an English coder was assigned a video including Spanish speakers. In each instance, the supervisor, in collaboration with project survey staff, reviewed the situation and provided the coder with directions to proceed. Completed coding sheets were placed into batches and transferred to data entry.

Data entry and coding. SOC staff entered all data into the data entry program and coded responses. In 2011 and 2012, we again used a two-person data entry process to ensure 100 percent verification. Project staff reviewed all verbatim and "other specify" responses in the coding database. Staff back-coded responses into pre-existing answer options, built new codes if enough responses expressed the same concept, or left responses as verbatim text in the data file. After data were entered, the statuses of the instruments were updated in the SMS as "complete."

Data Cleaning Consisted of Frequency Review and Data Editing

Frequency review. After data from all instruments were entered and considered complete, we ran frequencies for each data file. Survey staff responsible for each instrument reviewed the frequencies to verify that (1) the number of completed cases in the data file was correct; (2) the number of completed cases by cohort in the data file was correct; (3) the skip logic was followed correctly and each variable had the appropriate number of responses; (4) the frequency for each variable was feasible; (5) there were no missing data, additional data, or outliers; and (6) labels and variable names were correct.

Data editing. The staff responsible for each instrument made edits to the data when necessary after reviewing frequencies. The SAS programmer produced a spreadsheet for editing in which survey staff selected a variable to edit, entered the current value, entered the new value, and entered the reason why the value was being edited. A programmer read the specifications from these documents and updated the data file. All data edits were documented and saved in a designated file. Most data edits corrected minor data entry errors or interviewer/assessor coding errors identified during frequency review (for example, filled in missing data with "M" or cleared out "other: specify" verbatim data when the response had been back-coded). Each time a data file was updated, a new set of frequencies was run and reviewed. This process continued until all of the data files were clean and ready for analysis.

This page is left blank for double-sided printing.

APPENDIX C. MEASURES

This appendix describes the measures used during spring 2011 and 2012 data collection efforts to assess children at age 3. Given the longitudinal nature of Baby FACES, many of these measures were also used in previous rounds of data collection.

We considered several factors when selecting among available child and family outcome measures for inclusion in the study. Among our requirements were adequate reliability and validity of the measures, appropriateness for use with children and families from diverse backgrounds, comparability with other large-scale research projects, burden on children and families, ease of administration and scoring, and appropriateness for use by Early Head Start programs. We also considered the need to complement well-established measures with those that are new to large-scale research efforts and fill existing measurement gaps. We attempted to select measures that related to the cognitive, language, and social-emotional development outcome domains identified by the National Education Goals Panel.

In addition to our review of the literature, we worked closely with experts from our technical work group (TWG), other authorities in the field, and the test developers themselves to select and modify measures for Baby FACES. The final list of measures presented here reflects the feedback of dozens of experts in the early childhood development field.

Measure Assessment and Scoring

We assessed the constructs arising from Baby FACES measures based on the user's guide for the measures or using a scoring approach consistent with the current literature. In addition, we used the following criteria in variable constructions:

- ***Sufficient Item-Level Data.*** If an individual was missing data from more than 25 percent of the items that made up a constructed variable, we did not compute a score for that individual. If the individual was missing up to 25 percent of the items, we imputed values based on the means of the items that were present. We used the specifications described in the user's guide to impute item values for the Ages and Stages Questionnaires, Third Edition (ASQ-3), and the Brief Infant Toddler Social Emotional Assessment (BITSEA). For methodological reasons, we did not impute missing data at the item level for the MacArthur-Bates Communicative Development Inventories (CDI)—Infant Short Form; the Preschool Language Scale, Fourth Edition (PLS-4); the Early Communication Indicator (ECI); the Peabody Picture Vocabulary Test-4th Edition (PPVT-4), the Home Visit Rating Scale-Adapted (HOVRS-A); the Classroom Assessment Scoring System-Toddler (CLASS-T); and the Parent-Caregiver Relationship Scale (PCRS).[13]
- ***Adequate Internal Consistency Reliability.*** Methods of estimating reliability that require only a single test administration are referred to as measures of internal consistency or homogeneity. They are based on estimates of how well items within a

[13] Because the HOVRS-A scales are composed of mean scores across a number of component items, we did not impute means for missing values. Similar considerations precluded imputing values for the CLASS-T and PCRS. The ECI score is based on occurrences of four communication elements, all of which need to be present to compute a total score. The raw scores for the CDI and PLS-4 are counts of correct answers, and the missing items are not counted for cases with 25 percent or fewer items missing. The scores are set to missing for cases with more than 25 percent of the items missing.

scale or instrument measure the same cognitive domain or construct. We chose Cronbach's coefficient alpha, which captures the correlation among items on an assessment. The greater the covariance among items, the higher the reliability (and thus the higher the value of Cronbach's coefficient alpha). Values of the alpha can range from -1.0 to 1.0, with greater values indicating stronger internal consistency. Cronbach's coefficient alpha is an extension of Kuder-Richardson Formula 20, a measure of internal consistency that is used when the items are dichotomous (Cronbach 1951). We consider an alpha of 0.65 or higher as adequate for the constructed measures.

Psychometric Properties of Constructs

This section provides information on the selection criteria, normative samples, and psychometric properties reported by the developers of the measures for nine child outcome measures: the (1) ASQ-3 (Squires 2009), (2) MacArthur-Bates CDI (Fenson 2000), (3) PLS-4 (Zimmerman et al. 2002), (4) ECI (Luze et al. 2001; Carta et al. 2010), (5) Peabody Picture Vocabulary Test-4th Edition (PPVT-4; Dunn and Dunn 2007), (6) BITSEA (Briggs-Gowan and Carter 2006), (7) Bayley Behavioral Rating Scale (BRS; Bayley 1993), (8) Behavior Problems Index (BPI; Zill and Peterson 1986), and (9) child constructs derived from the Parent-Child Interaction Rating Scales for the Two-Bag Assessment (Mathematica Policy Research 2010). We also provide a brief background for other measures gathered during this wave of data collection, such as measures of parenting and the home environment, parent mental health, and home visit and classroom quality.

For the measures listed above, we analyzed the psychometric properties from the Baby FACES spring 2011 and 2012 data collection and their difference sources. Tables C.1 through C.11 present the psychometric data for the constructed variables derived from the parent self-administered questionnaire (SAQ), staff-child report (SCR), direct child assessment, classroom and home visiting observations, and parent-child play assessment for the Baby FACES sample at age 3. The tables are organized by measurement domain. We include the sample size, the possible range of values for each variable, the reported range in the Baby FACES sample, the unweighted sample mean, standard deviation, and the internal consistency reliability (coefficient alpha). Most of the constructed measures have internal consistency reliability of 0.65 or higher.

Measure of General Child Development

Ages and Stages Questionnaires, Third Edition. The ASQ-3 is a parent-report tool for screening infants and young children for developmental delays (Squires et al. 2009). The 21 questionnaires included in the ASQ-3 are appropriate for children ages 1 month to 5-1/2 years and focus on assessment of five key developmental areas: (1) Communication, (2) Gross Motor, (3) Fine Motor, (4) Personal-Social, and (5) Problem Solving. Parents are asked to rate questions such as "Does your child make sentences that are three or four words long?" on a scale of "not yet," "sometimes," or "most of the time." There are six items in each of the five developmental areas. Possible raw scores in each developmental area range from 0 to 60, and the ASQ-3 total area score could range from 0 to 300.

Due to the ASQ's widespread use by Early Head Start programs, we included it as a measure of a child's general development. Among the ASQ's advantages are its short administration time, psychometric soundness, relatively low cost, and availability in Spanish. The ASQ has demonstrated reliability, validity, and accuracy in distinguishing between children with and without developmental delays. Early Head Start programs often used this instrument to identify children with (or at risk for) development delays.

The normative sample includes 15,138 children between 1 month and 66 months of age throughout the United States. The sample includes more boys (53 percent) than girls (47 percent). Approximately two-thirds of children are white, 12 percent are African American, and 15 percent are Hispanic; other races make up the remaining 5 percent. More than half (54 percent) of mothers had at least four years of college, and only 3.5 percent had not completed high school. Most (57 percent) of the families have annual incomes greater than $40,000.

The psychometric studies on the ASQ-3 demonstrate adequate reliability and concurrent validity of the questionnaires. Intraclass correlations ranged from 0.75 to 0.82, indicating strong test-retest reliability across developmental domains. Estimates of inter-rater reliability are less robust; intraclass correlations by area range from 0.43 to 0.69. Cronbach's alphas range from 0.51 to 0.87. The ASQ-3 classifications have moderate to high agreement with the Battelle Developmental Inventory (BDI) classifications (Newborg et al. 1984, Newborg 2004), with an aggregated sensitivity or specificity of 86 percent across all age intervals.

Cutoff points, which vary by age and indicate the need for further assessment, were derived by subtracting two standard deviations from the mean for each area of development (children scoring two standard deviations below the mean or lower are in the at-risk range). For example, the cutoff point in Communication is 25.36 for the 33-month form and 30.99 for the 36-month form. The cutoff point of two standard deviations has a sensitivity and specificity of 0.86. In other words, children whose scores are two standard deviations below the mean or lower have an 86 percent chance of being identified for further assessment. Children whose scores fall in the monitoring zone defined by the ASQ-3 authors (between one and two standard deviations below the mean) might benefit from practicing skills in a specific area of development. As expected, the cutoff point of one standard deviation has a high sensitivity (0.98) but a low specificity (0.59). Therefore, some children who are developing normally will be classified as needing further assessment (Squires et al. 2009).

Table C.1 illustrates the average, standard deviation, range, and internal consistency of the ASQ-3 scores among 3-year-olds in the Baby FACES study. Cronbach's alphas for the study's sample are similar to previous studies.

Measures of Child Language Development

MacArthur-Bates Communicative Development Inventory. The CDI is designed to assess children's early receptive and expressive language and communication skills through parent report (Fenson et al. 2000). At age 3 Baby FACES data collection, Early Head Start staff (teachers and home visitors) completed the CDI-III vocabulary checklist. The 100-item CDI-III vocabulary checklist is a short measure of expressive vocabulary for children 30–37 months of age. We also asked Early Head Start staff to report on 3-year-olds' receptive vocabulary using the checklist. Two measures were derived from this form:

- ***Vocabulary Comprehension*** measures the number of words the child understands. Teachers/home visitors are asked whether the child "understands" or both "understands and says" each of 100 specific words.
- ***Vocabulary Production*** measures the number of words in the child's spoken vocabulary. Early Head Start teachers and home visitors report whether the child "understands and says" each of 100 specific words.

The raw scores for both Vocabulary Comprehension and Vocabulary Production range from 0 to 100. Teachers and home visitors who reported they spoke Spanish completed the English CDI-

III and the Spanish CDI-III (Vagh, Mançilla-Martinez, and Pan, unpublished manuscript) for children identified as understanding Spanish.

In addition to staff reports, parents also report on children's English or Spanish Vocabulary Production using the CDI-III in the SAQ. For age 3 children in Baby FACES, parent- and staff-reported vocabulary production scores are moderately correlated for English (r = 0.35) and uncorrelated for Spanish (r = 0.08).

The CDI was used successfully in the Early Head Start Research and Evaluation Project (EHSREP) despite concerns about the norming sample's appropriateness. EHSREP researchers found that Early Head Start had a significant positive impact on 24-month-old children's language production. All versions of the CDI also show concurrent validity with other measures such as the Bayley language subscales. The ability to have both parents and teachers/home visitors provide data on this instrument made this measure of language development in Baby FACES a valuable tool.

The initial norming sample for the English CDI-III includes 356 children between 30 to 37 months of age from families obtained through a university subject pool database and an ongoing study in San Diego, California. More than two-thirds (69.8 percent) of parents hold a college diploma, 20.1 percent have some college education, and only 0.6 percent have not completed high school. The upwardly skewed socioeconomic status (SES) distribution of the normative sample may limit the applicability of the norms to children from low-SES families. The racial/ethnic composition of the sample and home language of the children are not reported.

In one study that includes 19 children of 36 to 37 months of age, the CDI-III vocabulary production was correlated at 0.63 with PLS-3 total scale score, 0.58 with the Auditory Comprehension score, and 0.47 with the Expressive Communication score. Other studies demonstrate moderate correlations of the CDI-III vocabulary production scores with the PPVT-R (r = 0.41 – 0.53) and the McCarthy Scales of Children's Abilities (r = 0.44 – 0.62) (Fenson et al. 2007).

Table C.1 illustrates the average, standard deviation, range, and internal consistency of the CDI scores among 3-year-olds in the Baby FACES study.

Preschool Language Scale, Fourth Edition (PLS-4). The PLS-4 is a direct child assessment used to evaluate children's receptive and expressive language skills, as well as information about children's understanding and use of grammatical rules from birth through 7 years of age (Zimmerman et al. 2002). It comprises two subscales: (1) Auditory Comprehension (AC) and (2) Expressive Communication (EC). We used the AC subscale of the English and Spanish editions of the PLS-4. This subscale assesses comprehension of basic vocabulary concepts and grammatical markers in preschoolers. Children who were exposed to Spanish at home received the Spanish version; children with no exposure to Spanish received the English version.

In discussion with the test publisher, we developed a procedure to derive a "conceptual score" for dual language learners, giving children credit for their knowledge of both English and Spanish. To capture emerging language in children exposed to both languages, we first administered the PLS-4 in Spanish. Once the ceiling was established in Spanish, the assessor transitioned to the English PLS-4, administering the direct English translation of all items for which the child did not receive credit in Spanish. Testing continued until a ceiling was established in English. We calculated the

conceptual scores by giving children credit for items that they answered correctly in Spanish or English, and derived the bilingual standard scores using the norms for the Spanish Edition.

The standardization sample for the PLS-4 English included 1,564 children ages 2 days to 6 years, 11 months (about 100 children for each age group from 18 months to age 3). A stratified representative sample was selected based on parent education level, geographic area, and race. About one-third of the sample was from the South; children from the West and North Central each made up a quarter of the sample; and 18 percent of the sample were from the Northeast. Within each age level, the proportion of boys and girls was split evenly. About 17 percent of the sampled children's primary caregivers were without a high school diploma; 32 percent of primary caregivers had a high school diploma or GED; 28 percent had earned some college credit; and 23 percent had completed four or more years of college. Most (97 percent) of the children spoke English only, and 3 percent spoke languages other than English.

The standardization sample for the PLS-4 Spanish included 1,188 children ages 2 days to 6 years, 11 months. A stratified representative sample was selected based on parent education level and geographic area. About half of the sampled children were from the South; 43 percent from the West; and the remaining 7 percent from the Northeast and North Central. The proportion of boys and girls was split evenly within each age level in the sample. About 62 percent of the sample was white; 15 percent African American; 17 percent Hispanic; and 5 percent comprising other races. Slightly less than half of the sample included children whose primary caregivers were without a high school diploma; 22 percent of primary caregivers had a high school diploma or GED; 14 percent had spent one to three years in college or a technical school; and 17 percent had completed four or more years of college. Almost all children in the sample were Hispanic, and most of them (81 percent) spoke a Spanish dialect used in Mexico.

The psychometric evidence indicates that the PLS-4 English and Spanish editions can provide reliable and valid inferences about a child's language ability (Zimmerman et al. 2002). The internal consistency reliability coefficients for the AC subscale of PLS-4 English edition range from 0.91 to 0.94 for children ages 18 months to 41 months. The test-retest reliability estimates range from 0.87 to 0.95 for children ages 24 months to 41 months. The PLS-4 correctly identified 79 percent of 3-year-old children with a previously diagnosed language disorder and 92 percent of typically developing 3-year-old children.

For the AC subscale of the PLS-4 Spanish edition, the internal consistency coefficients range from 0.84 to 0.89 for children ages 18 to 42 months. The test-retest reliability estimates for children ages 24 months to 47 months range from 0.73 to 0.84. PLS-4 correctly identified 87 percent of children with a previously diagnosed language disorder and 57 percent of typically-developing 3-year-old children.

Table C.1 illustrates the average, standard deviation, range, and internal consistency of the PLS-4 scores among children in the Baby FACES study.

Early Communication Indicator. As part of the assessment activities conducted with 3-year-olds, interviewers administered the ECI—a semi-structured, play-based communication task designed to measure the expressive communication of infants and toddlers between the ages of 6 months and 36 months (Luze et al. 2001; Carta et al. 2002, 2010). The ECI comprises four key skill elements, or communicative behaviors:

1. ***Gestures*** are nonverbal, intentional actions that convey communicative intent (such as pointing to direct another's attention to an object).
2. ***Vocalizations*** consist of nonword, verbal utterances (such as cooing or babbling), or verbalizations voiced by the child that are otherwise unintelligible.
3. ***Single-word utterances*** are defined as single words voiced by the child that are recognizable and readily understood (such as "pig" and "bye-bye").
4. ***Multiple-word utterances*** consist of two or more voiced words that fit together in a meaningful way to approximate a statement or sentence (such as "Piggy sleeps" and "Cow eats food").

Interactions were video-recorded for subsequent coding by staff at Mathematica. Coders record the frequency of occurrences of each observed skill element over the six-minute assessment; observed instances are later combined to yield a total communication score. Specifically, the total communication score reflects the weighted combination of the child's gestures, vocalizations, and single- and multiple-word utterances; the latter two are given weights of two and three, respectively, to account for the greater complexity of skill associated with their use.[14] The total weighted score is then converted to a rate score that reflects the number of communicative acts per minute over the course of the six-minute play assessment. An age-based, standardized score with a mean of 100 (SD = 15) is also calculated. Two cutoff scores identify children with (or at risk for) expressive language delays. Children scoring between one and one and a half standard deviations below the mean are in the at-risk range; those with scores one and a half standard deviations below the mean or lower are identified as demonstrating delays in expressive language (Greenwood et al. 2006, 2010).

Beyond providing a measure of children's communication proficiency at discrete periods during early development, the ECI can also be administered on an ongoing basis to monitor the short-term growth and development of children's expressive communication over time (Carta et al. 2002; Luze et al. 2001). Other notable advantages include its short administration time and psychometric soundness.

The feasibility and psychometric properties of the ECI have been documented across a number of studies, including a longitudinal study of 50 children in center-based care (Luze et al. 2001); a cross-sectional study of 1,486 infants and toddlers served by Early Head Start, community-based childcare, and early intervention programs (Greenwood et al. 2006); and, most recently, a large-scale study of 5,883 children enrolled in Early Head Start programs across two states. The broader goal of the large-scale study was to develop a normative sample based on children served by Early Head Start (Greenwood et al. 2010).

Normative growth estimates are based on a composite sample of more than 1,400 children that combined data from three smaller study samples (Greenwood et al. 2006). This aggregate sample represents children drawn from 22 center- and 14 home-based programs from 1999 to 2004. Most (90 percent) of these children were participating in Early Head Start. Collectively, children were racially diverse, with 49 percent from African American, Hispanic, or racially mixed minority backgrounds. Slightly more than half of all children (55 percent) were male, and 12 percent were receiving Part C early intervention services with Individualized Family Service Plans. Ten percent of families reported languages other than English were spoken in the home. Although the combined sample was

[14] Weighting is used in the calculation of the total communication score to create a growth-based metric that reflects growth in communication proficiency by offsetting declines in prelinguistic communication (gestures and vocalizations) that occur as children acquire greater proficiency in spoken language skills. The weighted calculation also approximates an absolute estimate of total words produced by the child.

predominantly low-income, it did include some children from middle- to high-SES families. The total weighted communication score for 3-year-old children in the sample averaged 20.2. Normative estimates derived from this sample were used in the creation of ECI total communication standard scores for children in the Baby FACES study.

A more recent, population-specific normative sample is based on data obtained between 2002 and 2007 from 5,883 children served by Early Head Start in two midwestern states (Greenwood et al. 2010). However, normative estimates by child age (in months) are currently unavailable. Assessments were obtained at six-month intervals between 6 months and 36 months, with 18.4 months as the average age at the first assessment. Overall, the median number of observations per child was three, and the mean number of assessments obtained at each month of age was 428 (SD = 198; range = 23 to 727). Children were regionally and racially diverse, with nearly all children residing in homes in which English (90 percent) or Spanish (9 percent) were spoken. Representation by gender was roughly equivalent (48 percent male). In accordance with Early Head Start mandates, 8 percent of children were receiving Part C services. Among 3-year-old children in the sample, the total weighted communication score averaged 21.6.

Studies of the ECI demonstrate that it can be reliably administered, with reported split-half reliabilities of 0.89 and interobserver agreement of 90 percent (Luze et al. 2001; Greenwood et al. 2005). The ECI has also been shown to demonstrate significant, positive associations with known measures of early communication, including the Preschool Language Scale, Third Edition (PLS-3; Zimmerman et al. 1992) and the Caregiver Communication Measure (CCM; Walker et al. 1998; r = 0.62 and 0.51, respectively). Cross-sectional and longitudinal studies have documented the ECI's sensitivity to individual differences in communication proficiency within age and across early development, sensitivity to differences in the performance of children with disabilities, and sensitivity to short-term early interventions with infants and toddlers in Early Head Start (Carta et al. 2004; Greenwood et al. 2002, 2006; Luze et al. 2001).

In Baby FACES, inter-rater reliabilities between the team leaders and coders were established to a criterion of 80 percent agreement. Thereafter, the team conducted weekly inter-rater reliability checks on two to five randomly selected videos. A total of 57 videos (10.6 percent of the 539 codable videos) served as reliability videos.[15] Agreement averaged 85.2 percent across all coders, with a range of 77 to 86 percent.[16]

Table C.1 illustrates the average, standard deviation, and range of the ECI scores among children in the Baby FACES study.

Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4). The PPVT-4 is a norm-referenced standardized test designed as a measure of receptive vocabulary and screening test for

[15] As part of the 2011 data collection, we completed a total of 473 videos of ECI administrations with 2- and 3-year-old children. We were unable to code two of these videos. In one video, the child spoke a language other than English or Spanish; in a second case, the interaction was not recorded due to field interviewer error. An additional five videos were shorter than the required duration and were excluded from the analyses. Therefore, we computed scores from a total of 466 videos (387 of which were observations conducted with 3-year-olds). For the 2012 round of data collection, we received an additional 73 videos of ECI administrations from the field, all of which were coded.

[16] Reported inter-reliabilities are for the combined 2011 and 2012 data collection waves. In 2011, we simultaneously coded observations of children in the 1-year-old Cohort at age 3 and in the newborn Cohort at age 2. Reliability videos were selected at random from among all of the codable videos; thus, the estimates reported here include a small subset of observations conducted with 2-year-old children.

verbal ability and is suitable for a wide range of ages, from 2 1/2 through adulthood (Dunn and Dunn 2007). We administered the PPVT-4 to all children regardless of their primary language at Baby FACES age 3 data collection. Children are asked to say, or indicate by pointing, which of four pictures best shows the meaning of a word that is said aloud by the assessor. The standardization sample included 3,540 individuals aged 2 1/2 years through 90 years and older at 320 sites throughout the U.S. At each age, the sample matched the U.S. population in terms of sex, race/ethnicity, SES, geographic region, and special education status. The samples were divided into six-month age intervals at ages 2 1/2 years through 6 years, with 100 cases in each age group.

The PPVT-4 established good reliability evidence, with split-half and alpha reliability ranging from 0.95 to 0.97 for age 3 groups. The test-retest reliability was 0.91 for children of 2 to 4 years of age. Studies carried out during the PPVT-4 standardization also provided evidence of convergent validity. The PPVT-4 standard scores were correlated with the Expressive Vocabulary Test, Second Edition (EVT-2) at 0.81 for children of 2 to 4 years of age, and with the Comprehensive Assessment of Spoken Language (CASL) at 0.46 for children of 3 to 5 years of age.

Table C.1 presents the mean, standard deviation, range, and internal consistency of the PPVT-4 scores among children in the Baby FACES study.

Correlations Among Language Measures Used in Baby FACES at Age 3. To compare relationships and test the validity of the language measures and sources of report, we ran correlations of all language measures for children at age 3 (Table C.2).

The correlations between the CDI parent-reported English and Spanish vocabulary production scores and the ASQ-3 Communication IRT score are around 0.60 (0.55 for English and 0.64 for Spanish). The correlations of the CDI staff-reported vocabulary production scores with the ASQ-3 IRT score are lower than parent reports (0.30 for English and 0.23 for Spanish). The ASQ-3 Communication score is correlated with the CDI staff-reported vocabulary comprehension score in the low range (r = 0.23 for English and 0.13 for Spanish).

The correlations of the CDI parent- and staff-reported English vocabulary production with the PLS-4 English scores and the PPVT-4 are in the moderate range (r = 0.34 – 0.48). Staff-reported CDI English production scores are negatively correlated the PLS-4 Spanish scores (r = -0.18). The parent-reported CDI Spanish production score is correlated with the PLS-4 Spanish score and bilingual score at 0.30 and 0.23, respectively. However, the correlations of the staff-reported CDI Spanish production or comprehension score with the PLS-4 Spanish score or bilingual score are low to none. Parent- or staff-reported CDI Spanish scores are negatively correlated with the PPVT-4 (r = -0.14 to -0.30).

The correlations between parent- and staff-reported CDI English scores and the ECI total communication score are in the low range (r = 0.12 – 0.30). Parent- and staff-reported CDI Spanish scores are not correlated with the ECI total communication score. The correlations of the ASQ-3 Communication with the PLS-4, PPVT-4, and the ECI total communication scores are in the low to moderate range (r = 0.24 – 0.44). The ECI total communication scores are correlated with the PLS-4 English scores (r = 0.21), and uncorrelated with the PLS-4 Spanish or bilingual scores. The two direct assessments of children's language, the PLS-4 and PPVT-4, are moderately correlated (r = 0.64).

Measures of Child Social-Emotional Development

Brief Infant Toddler Social Emotional Assessment. The BITSEA is the screener version of the longer Infant Toddler Social Emotional Assessment (ITSEA), which is designed to detect delays in the acquisition of social-emotional competencies as well as social-emotional and behavior problems in children 12 months to 36 months old (Briggs-Gowan and Carter 2006). The 42-item parent and staff report focuses on the development of competencies (for example, hugs or feeds dolls or stuffed animals), as well as problem behaviors (for example, avoids physical contact).

We selected the BITSEA as our measure of social-emotional development due to its dual focus on social competencies and behavior problems, Spanish language availability, and the possibility of administering it to both parents and staff.

The 31-item BITSEA Problem scale assesses social-emotional/behavioral problems such as aggression, defiance, overactivity, negative emotionality, anxiety, and withdrawal. Higher scores indicate more problems. The 11-item BITSEA Competence scale assesses social-emotional abilities such as empathy, prosocial behaviors, and compliance. Lower scores indicate less competence. Respondents are asked to rate each item as "not true/rarely," "somewhat true/sometimes," or "very true/often." The BITSEA is available in both English and Spanish, and we administered it to both parents and teachers/home visitors in Baby FACES. The raw score ranges from 0 to 22 for the competence domain and 0 to 62 for the problem domain.

We created cutoff scores to indicate a high degree of problems or low competence. We calculated cutoff points in six-month age bands according to child gender by using cutoff points established with the national standardization sample. For the BITSEA Problem scale, the cutoff point indicates scores at the 75th percentile or higher. For the BITSEA Competence scale, the cutoff point indicates scores at the 15th percentile or lower. A score in this range suggests that delays in social-emotional competence may be present. Scoring in the cutoff range in either or both domains (that is, many problems and/or low competence) indicates "screening positive" on the BITSEA. At the time of Baby FACES age 3 data collection, approximately 60 percent of children were 36 months or older, but we still created the cutoff scores using the oldest age band (30–36 months).

The nationally normative sample includes 600 children between 12 months and 35 months, 30 days of age, with 150 children (75 boys and 75 girls) in each age band: 12 to 17 months, 18 to 23 months, 24 to 29 months, and 30 to 35 months. Each age band was stratified to match the 2002 U.S. Census on race/ethnicity, parent education level, and region.

The BITSEA has adequate test-retest reliability (r = 0.82 – 0.92), inter-rater reliability (r = 0.67 – 0.74) (Briggs-Gowan and Carter 2006), and internal consistency. On the Parent Form, Cronbach's alpha for the Problem and Competence scales are 0.79 and 0.65, respectively. On the Childcare Provider Form, Cronbach's alpha for the Problem and Competence scales are 0.80 and 0.66, respectively (Briggs-Gowan 2004).

The BITSEA has demonstrated construct validity through expected associations with other measures of the same construct (Briggs-Gowan and Carter 2006). The BITSEA Parent Form Problem and Competence scores were both moderately correlated with the ASQ: Social-Emotional (ASQ: SE) (Squires et al. 2002) (r = 0.55 and -0.55, respectively). The correlations between the BITSEA Problem score and the Child Behavior Checklist 1.5–5 (CBCL 1.5–5) (Achenbach and Rescorla 2000) Internalizing, Externalizing, and total scores range from 0.46 to 0.60, and the

correlations between the BITSEA Competence score and the CBCL scores range from -0.30 to -0.42. The BITSEA scores were moderately correlated with the Adaptive Behavior Assessment System—Second Edition (ABAS-II) (Harrison and Oakland 2003) domain-specific skill scores (Conceptual, Social, and Practical), with the correlations ranging from 0.39 to 0.56 for Competence and -0.31 to -0.36 for Problem. The BITSEA scores also demonstrated small to modest correlations with the Bayley Scales of Infant and Toddler Development—Third Edition (Bayley-III) (Bayley 2006) Cognitive Assessment and Language Scale, ranging from 0.25 to 0.32 for the Competence score and -0.19 to -0.28 for the Problem score. The correlation between the BITSEA Problem score and the Bayley-III Social Emotional score was -0.27, and the correlation between the BITSEA Competence score and the Bayley Social-Emotional score was 0.51.

The BITSEA has also demonstrated validity in discriminating children with clinically significant problems from matched control subjects (Briggs-Gowan and Carter 2006). The BITSEA Competence scale demonstrates excellent sensitivity (100 percent) and good specificity (91 percent) in detecting autistic disorder. The Problem scale provides excellent specificity (97 percent) and some sensitivity (64 percent).

The BITSEA validation study (Briggs-Gowan 2004) reported that the parent and child care provider correlation was higher than expected for Competence ($r = 0.59$) and typical for Problems ($r = 0.28$), because children may behave differently in the two contexts. For Baby FACES age 3 data collection, the correlations between parent ratings and Early Head Start staff ratings are lower than those found in the BITSEA validation study ($r = 0.25$ for Competence and 0.22 for Problem). Although still low, for the Problem scale, home visitor ratings are more highly correlated with parent ratings ($r = 0.42$) than are teacher ratings with the parent ratings ($r = 0.13$) on the Problem scale; for the Competence scale, teacher ratings are more highly correlated with parent ratings ($r = 0.28$) than are home visitor ratings and parent ratings ($r = 0.22$).

Table C.3 illustrates the average, standard deviation, range, and internal consistency of the BITSEA scores among 3-year-old children in the Baby FACES study.

Bayley Behavioral Rating Scale (BRS). The BRS measures a child's behavior during child assessment. The BRS is one of the three component scales of the Bayley Scales of Infant Development—Second Edition (Bayley 1993). Baby FACES uses the following two subscales:

- ***Orientation/Engagement*** measures the child's cooperation with the interviewer during the assessment, positive affect, and interest in the test materials.
- ***Emotional Regulation*** measures the child's ability to change tasks and test materials, negative affect, and frustration with tasks during the assessment.

The assessor rates the child's behavior by scoring items on a five-point scale, with 5 indicating more positive behavior (for example, more cooperation and less frustration). Scores are the total of the items in the subscale. Possible scores range from 9 to 45 for Orientation/Engagement and 10 to 50 for Emotional Regulation. The test-retest reliability coefficients range from 0.60 to 0.71 for ages 24 months and 36 months. The correlations between the two subscales and the Bayley Scales of Infant Development Mental Development Index (MDI) and Bayley Scales of Infant Development Psychomotor Development Index (PDI) are in the low range ($r = 0.22 – 0.34$), suggesting that the BRS taps a unique source of variance from the Mental and Motor Scales. The BRS "nonoptimal" cutoff scores indicate raw scores at or below the 10th percentile, and "questionable" cutoff scores indicate raw scores between the 11th and 25th percentiles. The BRS cutoff scores differentiate

children with significant impairment (identified by the MDI and PDI or diagnosed with medical conditions that are associated with severe impairment) from normally developed children.

Table C.3 illustrates the average, standard deviation, range, and internal consistency of BRS scores among 3-year-old children in the Baby FACES study.

Behavior Problems Index (BPI). The Behavior Problems Index (BPI) was developed by Zill and Peterson (1986) to measure child externalizing behavior problems (such as aggression and hyperactivity) and internalizing behavior problems (such as anxiety and depression). Most of the items are from the Achenbach Child Behavior Checklist (Achenbach and Edelbrock 1981). The National Longitudinal Survey of Youth (NLSY; Center for Human Resource Research 2009) used the BPI with children 4 years of age or older and the Panel Study of Income Dynamics (PSID) Child Development Supplement used the BPI with children 3 years of age or older. Baby FACES used the BPI, but included slightly fewer items than the previous two studies. We followed the scoring method used in the PSID as much as possible in order to compare our results to national data. The questions in the BPI ask about specific behaviors that the child may have exhibited in the past 3 months. Three response categories were used in the questionnaire: (1) "often true," (2) "sometimes true," and (3) "not true." The responses were reverse coded so that higher scores indicate more behavior problems. In addition to the BPI total score, two subscale scores were created: externalizing and internalizing behaviors. At age 3, both parents and Early Head Start staff rated children's behavior problems on the BPI.

In the PSID, the Total Behavior Problems Scale and the Externalizing and Internalizing subscales of the BPI each showed strong internal reliabilities, 0.90, 0.86, and, 0.81, respectively for children ages 3 to 12. No validity information is available.

Table C.3 illustrates the average, standard deviation, range, and internal consistency of BPI scores among children in the Baby FACES study.

Correlations Among Social-Emotional Measures Used in Baby FACES at Age 3. To compare relationships and test the validity of the social-emotional measures and sources of report, we ran correlations of all social-emotional measures for children at age 3 (Table C.4).

The correlations within data sources are in the moderate to high range. The absolute values of the correlations range from 0.26 to 0.66 between parent-reported measures and from 0.37 to 0.70 between staff-reported measures. The assessor ratings are highly correlated (r = 0.75). The observational measures are moderately to highly correlated, with the absolute values of the correlations ranging from 0.33 to 0.72.

The correlations between different data sources are in the low range. The absolute values of the correlations between parent and staff reports range from 0.07 to 0.25. The absolute values of the correlations range from 0.10 to 0.22 between assessor ratings and parent reports and from 0.10 to 0.23 between assessor ratings and staff reports. The absolute values of the correlations range from 0.05 to 0.29 between observational measures and parent reports and from 0.00 to 0.22 between observational measures and staff reports. The absolute values of the correlations between assessor ratings and observational measures range from 0.12 to 0.29.

Parent-Child Interaction Rating Scales for the Two-Bag Assessment–Child Scales. As part of the assessment activities conducted with 3-year-olds, we administered an eight-minute semi-

structured play-based task to parents and children (Two-Bag Task). Child behaviors were assessed using the Parent-Child Interaction (PCI) Rating Scales for the Two-Bag Assessment (Mathematica Policy Research 2010). Collectively, the 12 rating scales assess a range of child and parent behaviors. Nine of the component scales are derived from the 24- and 36-Month Child-Parent Interaction Rating Scales for the Three-Bag Assessment that was used as part of the EHSREP (ACF 2002; Brady-Smith et al. 1999, 2000). Three additional scales that originated in the NICHD Study of Early Child Care Three Box Task (NICHD Early Child Care Research Network 1997, 1999) were adapted from Cox (1997). The coding scheme includes four scales that assess child behaviors; seven scales that focus on parent behaviors; and one scale that addresses the quality of the dyadic interaction as a whole. Each area is assessed on a seven-point scale, ranging from a very low incidence of the behavior to a very high incidence of the behavior. Ratings along the scale are anchored by a description of the behaviors (and associated exemplars) that warrant a specific score. Overall, the scales measure both the prevalence and intensity of the observed behaviors.

Four scales assess the child's (1) engagement of parent (extent to which the child initiates and/or maintains interaction with the parent); (2) sustained attention with objects (degree of involvement with and focused exploration of the play materials); (3) enthusiasm (degree of vigor and confidence during the task); and (4) negativity toward parent (displays of anger, hostility, or disdain). (See Parent-Child Interaction Rating Scales for the Two-Bag Assessment–Parent Scales for information on the parenting-related scales, including psychometric properties). Based on findings from preliminary analyses of the Baby FACES data, we retained the child scales as individual scales. The scales were significantly associated, with associations that were moderate to high in magnitude (r = -0.33 – 0.72).

The psychometric properties of the scales have been documented in other large-scale studies, including the EHSREP and ECLS-B.[17] Interobserver agreement (exact or within one point) on the national coding scales in the EHSREP averaged 90 percent at 14 months, 93 percent at 24 months, and 94 percent at 36 months (with a range of 83 percent to 100 percent across the three ages; Brady-Smith et al. 2005). In ECLS-B, interobserver agreement on the child scales at age 2 was 94.7 percent (Andreassen et al. 2007). The scales have also been shown to relate significantly to widely used instruments that tap similar child constructs (Ispa et al. 2004; Tamis-LeMonda et al. 2004).

In Baby FACES, a total of 60 videos (11.4 percent of the 524 codable videos) served as reliability videos.[18] Agreement (exact or within one point) averaged 92 percent across all coders, with

[17] The task on which coding of the scales was based varied slightly across these studies. In the EHSREP, parents and children engaged in play with materials (age-appropriate book and toys) provided in three numbered bags (Three-Bag Task), while in ECLS-B, parents and their 2-year-old children were asked to play with materials provided in two numbered bags (Two-Bag Task). In the EHSREP, the Three-Bag Task was administered to parent-child dyads when children were approximately 14, 24, and 36 months of age. In ECLS-B, a comparable task was administered to parents and their 2-year-old children.

[18] During the 2011 data collection, we received a total of 465 Two-Bag Task administrations from the field. We were unable to code 9 using the Parent-Child Interaction Rating Scales for a number of reasons, including a parent speaking a language other than English or Spanish (n = 5) and poor audio quality (n = 1). An additional three videos were shorter than the required duration and were excluded from the analyses. Therefore, we computed scores from a total of 456 videos (376 of which were observations conducted with 3-year-olds). For the 2012 round of data collection, we received an additional 68 videos from the field, all of which were coded.

a range of 85 to 97 percent.[19] Cohen's kappa[20] for scales assessing child behaviors ranged from 0.55 to 0.67.

Table C.3 illustrates the average, standard deviation, range, and internal consistency of scores for the child constructs derived from the Parent-Child Interaction Rating Scales for the Two-Bag Assessment in Baby FACES.

Measures of Positive and Negative Parenting Behaviors During Play

Ratings of parent behaviors were based on coding of the Two-Bag Task according to two coding schemes: the Parent-Child Interaction Rating Scales for the Two-Bag Assessment (Mathematica Policy Research 2010) and an adaptation of the Parenting Interactions with Children: Checklist of Observations Linked to Outcomes (PICCOLO; Roggman et al. 2009).

Parent-Child Interaction Rating Scales for the Two-Bag Assessment–Parent Scales. As described above, the Parent-Child Interaction Rating Scales (Mathematica Policy Research 2010) comprises 12 scales that assess a range of child and parent behaviors, seven of which are related to aspects of parenting and one that addresses the quality of the dyadic interaction as a whole. (See Parent-Child Interaction Rating Scales for the Two-Bag Assessment–Child Scales for more information on the development of the measure, and for psychometric information on the Child Scales.)

Collectively, the parenting and dyadic scales address: (1) sensitivity (the extent to which the parent acknowledges the child's perspective, accurately perceives the child's signals, and promptly and appropriately responds to these signals); (2) positive regard (displays of love, respect, and/or admiration); (3) stimulation of cognitive development (effortful teaching aimed at expanding the child's abilities); (4) intrusiveness (over-involvement and over-control); (5) detachment (under-involvement with, lack of awareness of, attention to, and engagement of the child); (6) negative regard (expressions of discontent with, anger toward, and rejection of the child); (7) relationship quality (degree of relatedness and mutual engagement); and (8) boundary dissolution (extent to which the parent fails to maintain an appropriate parental role in his or her interaction with the child).

The psychometric properties of the scales have been documented in other large-scale studies, including the EHSREP and ECLS-B (see Parent-Child Interaction Rating Scales for the Two-Bag Assessment–Child Scales for interobserver agreement estimates in the EHSREP study sample). In ECLS-B, interobserver agreement at age 2 on the parent scales was 96.5 percent (Andreassen et al. 2007). Cronbach's alpha for the composite measure of supportive parenting created in the EHSREP—derived from average scores on sensitivity, cognitive stimulation, and positive regard, all of which were strongly intercorrelated (rs = 0.50 to 0.71)—ranged from 0.82 to 0.83 at the three

[19] Reported inter-reliabilities are for the combined 2011 and 2012 data collection waves. In 2011, we simultaneously coded observations of children in the 1-year-old Cohort at age 3 and in the Newborn Cohort at age 2. Reliability videos were selected at random from among all of the codable videos; thus, the estimates reported here include a small subset of observations conducted with 2-year-old children.

[20] Coefficients were weighted to reflect the degree of disagreement among coder ratings. Whereas unweighted kappa treats all disagreements equally, weighted kappa attaches greater emphasis to large differences between ratings along the ordinal scale. For example, disagreement by one scale point is seen as less serious than disagreement by two scale points, and so on.

ages (Brady-Smith et al. 2005). A composite derived from these three scales was also created for the ECLS-B age 2 data collection, although detailed psychometric information is not reported.

The parent rating scales have also been shown to relate significantly to widely used instruments that tap similar parent constructs (Ispa et al. 2004; Fuligni and Brooks-Gunn 2013). Positive associations were demonstrated between intrusiveness ratings and scores on the Traditional subscale (r = 0.22) of the Parental Modernity Scale (Schaefer and Edgerton 1985), and ratings of parents' positive regard and the Emotional Responsivity subscale (r = 0.30) of the Infant/Toddler Home Observation for Measurement of the Environment (HOME; Caldwell and Bradley 1984). Negative associations between dyadic mutuality ratings and mothers' concurrent scores on the Parent-Child Dysfunctional Interaction subscale (r = -0.17) of the Parenting Stress Index (PSI; Abidin 1995) have also been reported.

In Baby FACES, Cohen's kappa[21] for scales assessing parent behaviors ranged from 0.50 to 0.69. We conducted preliminary analyses to examine patterns of association among the scales, possible underlying factors, and internal consistency reliability. Based on our analyses, we created a composite parenting score, "synchronicity" (Cronbach's alpha = 0.85), by computing a mean score derived from scores on parental sensitivity, positive regard, and relationship quality—all of which were moderately to highly correlated (ranged from 0.57 to 0.78).[22] We retained the scales assessing negative parenting behaviors (intrusiveness, detachment, negative regard, and boundary dissolution) as individual scales. The correlations among the four negative parenting scales were small to moderate and statistically significant (0.21 to 0.48), with the exception of associations of negative regard with intrusiveness and boundary dissolution (0.60 and 0.52, respectively).

Parenting Interactions with Children: Checklist of Observations Linked to Outcomes. The PICCOLO is an observational instrument designed to measure "developmental parenting" along four key domains known to support children's development across a number of areas (Cook and Roggman 2009; Roggman et al. 2009). The measure was developed for use by practitioners working with parents of young children and has applications to other settings, including research and intervention efforts. Specifically, the PICCOLO rates 29 positive parenting behaviors along four domains: (1) Affection (displays of warmth, physical closeness, and positive expressions toward the child); (2) Responsiveness (to the child's cues, emotions, vocalizations, interests, and behaviors); (3) Encouragement (attempts to support the child's exploration, effort, skills, initiative, curiosity, creativity, and play); and (4) Teaching (the degree to which the parent engages in shared conversation and play, provides cognitive stimulation, and extends the child's verbalizations). All items are rated on a three-point scale, ranging from 0 (absent) to 2 (clearly evident and frequent in their occurrence and/or intensity). A score of 1 indicates emerging behaviors that are briefly

[21] Coefficients were weighted to reflect the degree of disagreement among coder ratings. Whereas unweighted kappa treats all disagreements equally, weighted kappa attaches greater emphasis to large differences between ratings along the ordinal scale. For example, disagreement by one scale point is seen as less serious than disagreement by two scale points, and so on.

[22] To allow for comparisons to other large-scale studies such as EHSREP, a second composite score of positive parenting, "supportiveness," (Cronbach's alpha = 0.70) was derived from scores on parental sensitivity, positive regard, and cognitive stimulation (correlations ranged from 0.32 to 0.69). Notably, associations between parental cognitive stimulation and parental sensitivity (r = 0.35), positive regard (r = 0.41), and relationship quality (r = 0.38) were only moderate in magnitude, and inclusion of cognitive stimulation in the composite score reduced the overall alpha. Given the overall lower internal consistency reliability and intercorrelation of its components, we reserve discussion of this construct for anchoring our findings to those reported in other national studies. Of note, concurrent associations between key child development outcomes and synchronicity were consistently more robust than were associations to supportiveness, lending further support to the validity of this composite measure.

observed. The domains of Affection, Responsiveness, and Encouragement each comprise seven items; the Teaching domain consists of eight items.

Evidence of the PICCOLO's validity is based on data derived from two research samples—the EHSREP and the Bilingual Early Language and Literacy Supports (BELLS) project. Collectively, the PICCOLO was validated on more than 4,500 videotaped interactions of 2,199 parents and their 10- to 40-month-old children obtained across repeated assessments. Although not intended to be nationally representative, the study samples represent ethnically diverse, low-income families, with 41 percent of parents identified as European American; 37 percent as African American, and 22 percent as Latino. Approximately 60 percent of mothers were teens at the time of the focus child's birth, and up to two-thirds of mothers had at least a high school diploma or equivalent (52 percent for African Americans, 66 percent for European Americans, 28 percent for Latinos).

Psychometric evidence supports the reliability and validity of the measure (Roggman et al. 2009). Average interobserver agreement is 80 percent for Affection, 76 percent for Responsiveness, 83 percent for Encouragement, and 69 percent for Teaching. Within each domain, factor loadings are in the moderate to high range, and internal consistency reliability coefficients range from 0.75 to 0.80. Among children who average 3 years of age, intercorrelations among the domains are moderate to strong in magnitude (r = 0.43 – 0.69); cross-age associations within domains demonstrate moderate stability over time (0.36 to 0.52). There is also evidence of the PICCOLO's convergent and predictive validity. Specifically, the domains of parenting measured by the PICCOLO related significantly to measures of supportive parenting derived from the Early Head Start Parent-Child Interaction Rating Scales for the Three-Bag Assessment (ACF 2002), including parental sensitivity (r = 0.31 – 0.50), cognitive stimulation (r = 0.26 – 0.56), and positive regard (r = 0.31 – 0.57). Each domain also related to children's outcomes both concurrently and over time, including measures of children's cognitive development, emotion regulation, vocabulary production, receptive language, emergent literacy, and problem solving (Roggman et al. 2009).

Inter-rater reliability between the team leaders and members of the coding team was established on the 29-item binary scale to a criterion of 80 percent exact agreement. A total of 59 videos (11.3 percent of the 524 codable videos) served as reliability videos.[23] Across all coders, agreement averaged 83 percent overall, with a range of 79 to 91 percent. Agreement averaged 91 percent for Affection, 81 percent for Encouragement, 81 percent for Responsiveness, and 79 percent for Teaching.[24]

An overall score of positive parenting was derived by calculating a mean score across each of the four domain scores (Cronbach's alpha = 0.77). To allow for comparisons to the EHSREP and other comparison samples, average and total sum scores were also computed for each domain. The authors provide age-based scoring rubrics which specify cut-points for low, moderate, and high

[23] During the 2011 data collection, we received a total of 465 Two-Bag Task administrations from the field. We were unable to code 9 using the PICCOLO for a number of reasons, including a parent speaking a language other than English or Spanish (n = 5) and poor audio quality (n = 1). An additional three videos were shorter than the required duration and were excluded from the analyses. Therefore, we computed scores from a total of 456 videos (376 of which were observations conducted with 3-year-olds). For the 2012 round of data collection, we received an additional 68 videos from the field, all of which were coded.

[24] Reported inter-reliabilities are for the combined 2011 and 2012 data collection waves. In 2011, we simultaneously coded observations of children in the 1-year-old Cohort at age 3 and in the Newborn Cohort at age 2. Reliability videos were selected at random from among all of the codable videos; thus, the estimates reported here include a small subset of observations conducted with 2-year-old children.

levels of support for each of the four PICCOLO domain scores (Roggman et al. 2009). Among children who average 3 years of age, the high range is represented by total scores greater than or equal to 10 (out of 14) on Affection and Encouragement; scores greater than or equal to 11 (out of 14) on Responsiveness; and scores greater than or equal to 8 (out of 16) on Teaching.

Measures of Parent Mental Health

Center for Epidemiologic Studies Depression Scale (CES-D). The CES-D is a self-administered screening tool used to identify symptoms of depression or psychological distress (Radloff 1977). The full version of the CES-D consists of 20 items, and the short form (CESD-SF) (Ross et al. 1983) consists of 12 items. Respondents are asked to rate how often each of the items applied to them in the past week on a four-point scale from "rarely or never" (score of 0) to "most or all of the time" (score of 3). Symptoms include poor appetite, restless sleep, loneliness, sadness, and lack of energy. Raw scores range from 0 to 36 for the short form, with higher scores indicating more depressive symptoms.

The CESD-SF has been used as a measure of parent well-being in large-scale studies such as the EHSREP and the Head Start Family and Child Experiences Survey (FACES). We chose the CESD-SF because of its use in previous Early Head Start studies, well-established psychometric properties, and short administration time.

Parents with scores on the CESD-SF of 15 or higher are considered as having severe depressive symptoms; those with scores of 10 to 15 are considered as having moderate depressive symptoms; and those who score between 5 and 10 are considered as having mild depressive symptoms.

Table C.5 illustrates the average, standard deviation, range, and internal consistency of CESD-SF scores among parents of 3-year-olds in the Baby FACES study.

The Parenting Stress Index—Short Form (PSI-SF). The PSI-SF measures the degree of stress in the parent-child relationship stemming from three sources: (1) the child's challenging temperament, (2) parental depression, and (3) negative reinforcement of parent-child interactions (Abidin 1995). We employed the PSI-SF due to its previous use in the EHSREP and ease of administration. We included the Parental Distress and Parent-Child Dysfunctional Interaction subscales in Baby FACES.

The Parental Distress subscale (five items) measures the level of distress the mother or father is feeling in his or her role as a parent, including a low sense of competence and a high level of stress due to perceived restrictions stemming from parenting. The parent answers whether or not he or she agrees with statements such as "You have been unable to do new and different things," and "You feel trapped by your responsibilities as a parent." Parents rate each item on a five-point scale from "strongly disagree" to "strongly agree." Scores can range from 5 to 25. Higher scores indicate higher levels of parental distress.

The Parent-Child Dysfunctional Interaction subscale (six items) measures a parent's perception that his or her child does not meet expectations and that interactions with the child are not reinforcing to the parent. The parent answers whether he or she agrees with statements such as "Most times, you feel that your child does not like you and does not want to be close to you" and "When you do things for your child, you get the feeling that your efforts are not appreciated very

much." Parents rate each item on a five-point scale from "strongly disagree" to "strongly agree." Scores can range from 6 to 30. Higher scores indicate a more dysfunctional parent-child interaction.

Table C.5 illustrates the average, standard deviation, range, and internal consistency of PSI scores among parents of 3-year-olds in the Baby FACES study.

Measure of Functioning of Families

The Family Environment Scale, Family Conflict Subscale (FES). The FES measures the extent to which the open expression of anger and aggression and conflict-filled interactions are characteristic of the family (Moos 2002). Parents rated five items on a four-point scale, where a 4 indicates higher levels of agreement with statements such as "We fight a lot" and "We sometimes hit each other." Scores can range from 1 to 4. The subscale score is then the mean of the five individual item scores. We included the FES because it had been previously included in the EHSREP. For the Baby FACES sample, however, we removed one item: "We hardly ever lose our tempers." It had a low correlation with the rest of the items in the scale and therefore reduced the overall alpha of the measure.

Table C.6 illustrates the average, standard deviation, range, and internal consistency of FES scores among parents in the Baby FACES study.

Measures of Home and Neighborhood Environment

Home Observation for Measurement of the Environment. The HOME measures the quality of stimulation and support available to a child in the home environment (Caldwell and Bradley 1984). It has separate inventories for infants and toddlers (birth to 3 years old), early childhood (ages 3 to 6), and middle childhood (ages 6 to 10). Information needed to score the inventory is obtained through a combination of interview and observation conducted in the home with the child's parent while the child is present. We used selected items from the infant version of the HOME inventory, the internal environment items from the Early Childhood version of the HOME, and neighborhood rating items from the Project on Human Development in Chicago Neighborhoods (PHDCN). We derived five subscales from this assessment, as well as the total score:

- ***Emotional Responsivity*** measures responsive and supportive parenting behavior observed by the interviewer during the home visit. Interviewer observations of the parent and child during the interview inform the items in this subscale, and explore such questions as whether the mother praised the child, whether she expressed warmth and affection toward the child, and whether she responded verbally to the child's verbalizations during the interview.
- ***Maternal Verbal-Social Skills*** measures the parent's ability to speak freely and clearly to the interviewer. This subscale comprises interviewer observations of the parent during the interview.
- ***Support of Cognitive, Language, and Literacy Environment*** measures the provision of a variety of developmentally stimulating toys and furnishings, as well as whether the parent provides toys for the child during the visit, reads to the child several times per week, and talks to the child while doing household chores. Items are obtained by a combination of parent report and interviewer observation. We also created another

measure of ***Enhanced*** **Cognitive, Language, and Literacy Environment** by crediting parents only when more toys are provided to the child.

- ***Absence of Punitive Interactions*** measures harsh or punitive parenting behavior observed during the home interview. Interviewer observations of the parent and child during the interview inform the items in this subscale, and include such events as shouting at, expressing annoyance or hostility toward, hitting, scolding, or restricting the child. Items received a score of 1 if the parent did not engage in harsh or punitive behaviors during the home visit.
- ***Internal Physical Environment*** measures the cleanliness, organization, and warmth of the home environment. This subscale comprises interviewer observations during the interview.
- ***Total Score*** measures the cognitive stimulation and emotional support provided by the parent in the home environment. It includes all 30 items used in the five previous subscales.

We also derived an **External Environment score** (not included in the HOME total), which measures the physical and social environment of the face block[25] where the family lives based on some neighborhood rating items from the PHDCN. This subscale comprises interviewer observations of the neighborhood. Examples include the general condition of most of the housing units, presence of garbage in the street or on the sidewalk, volume of traffic, and people arguing or fighting in the street. The items are recoded as 1 (yes) or 0 (no), and then summed.

The internal consistency reliability coefficients (Cronbach's alphas) were 0.84 for the original infants and toddlers HOME inventory and ranged from 0.49 to 0.78 for the subscales. Kuder-Richardson coefficients were 0.89 for the inventory and ranged from 0.44 to 0.89 for the subscales. The test-retest reliability estimates were 0.77 for the inventory and ranged from 0.30 to 0.77 when administered at ages 12 and 24 months. The intraclass correlation, which measures stability by comparing the similarity of paired scores relative to the total variation of all scores, resulted in slightly lower values. The intraclass correlation coefficients were 0.76 for the inventory and 0.30 to 0.76 at ages 12 and 24 months. The inter-rater reliability estimates ranged from 0.76 to 1.0 for the HOME.

Families' HOME inventory scores administered when the child was 6, 12, and 24 months old were compared with the child's scores on the Bayley Scales of Infant Development Mental Development Index (MDI) at 6 months and 12 months, the Stanford-Binet at 36 months and 54 months, and the Illinois Test of Psycholinguistic Abilities (ITPA) at 37 months. The HOME was found to be a better predictor of intelligence than socio-economic measures and was a stronger predictor for females and whites. The HOME was also compared with the Supplement to the HOME for Impoverished Families (SHIF), the Nursing Child Assessment Feeding Scale (NCAFS) and the Nursing Child Assessment Teaching Scale (NCATS). Below are the comparisons between the measures:

- Comparison with the Bayley MDI: The correlations between the HOME inventory score at 6 months and the Bayley MDI at 6 months and 12 months were 0.14 and 0.16, respectively (subscale correlations ranged from 0.01 to 0.27). The correlation between

[25] A face block is defined as the two sides of one street between intersecting streets.

the HOME at 12 months and the Bayley MDI score at 12 months was 0.30 (subscales ranged from 0.01 to 0.28).

- Comparison with the Stanford-Binet: The correlations between the HOME inventory score at 6 months and the Stanford-Binet at 36 months and 54 months were 0.50 (subscales ranged from 0.24 to 0.41) and 0.44 (subscales ranged from 0.10 to 0.44), respectively. The correlation between the HOME at 12 months and the Stanford-Binet at 36 months was 0.58 (subscales ranged from 0.24 to 0.56), respectively. The correlations between the HOME at 24 months and the Stanford-Binet at 36 months and 54 months were 0.71 (subscales ranged from 0.41 to 0.64) and 0.57 (subscales ranged from 0.28 to 0.56), respectively.
- Comparison with the ITPA: The correlations between the HOME inventory scores at 6 months and 24 months and the total ITPA score at 37 months were 0.39 and 0.61, respectively.
- Comparison with SHIF: The correlation between the HOME and the SHIF was 0.69 for families with children aged 3 or younger.
- Comparison with the NCAFS and the NCATS: In a nonrepresentative sample of impoverished urban families, the correlations were 0.55 and 0.42 between the HOME and the NCAFS and NCATS, respectively for families with children aged 3 or younger.

Neighborhood Disorder measures the physical and social environment of the face block where the family lives. Items in this subscale are based entirely on interviewer observations of the neighborhood, and include such variables as the general condition of most of the housing units, garbage in the street or on the sidewalk, traffic volume, and people arguing or fighting in the street. The scale score is the mean of the item z-scores. Higher scores indicate more disorder.

Table C.7 illustrates the average, standard deviation, range, and internal consistency for the HOME and neighborhood disorder scores of families with 3-year-olds in the Baby FACES study.

Exposure to Violence. The Exposure to Violence scale measures how many violent incidents (out of four) a child has observed in his or her lifetime. Items come from the ITSEA (Carter and Briggs-Gowan 2000), which asks parents to respond yes or no to questions that explore, for example, whether a child has "seen violence in their neighborhood" or "seen someone hit, push or kick a family member."

Measures of Home Visit and Classroom Quality

Home Visit Rating Scale-Adapted. The HOVRS-A[26] assesses a variety of dimensions of home-visiting quality and content, including home visitor responsiveness, nonintrusiveness, support

[26] Four main modifications were made in creating the HOVRS-A from the HOVRS. First, to make the measure easier to score, the number of scale rating points was reduced from seven to five. This step helped to establish inter-rater reliability, because there are fewer subtle distinctions to make between one rating point and another. Second, the indicators were aligned across each of the three anchors (1, 3, and 5) to ensure that they are consistent and that the same types of behaviors are assessed at each level. Third, the Home Visitor Relationship with Family item was adapted so that it taps both the home visitor's engagement and relationship with the family and the family's relationship with the home visitor. Finally, we created two versions of the last item, Child Engagement During Home Visit, one for visits with a focus child up to 12 months old (Infant Engagement During Home Visit) and another for visits with toddlers 12 to 24

of parent-child interaction, and parent and child engagement in the visit (Roggman et al. 2009). The HOVRS-A consists of seven items that are rated from 1 to 5, with anchors of 1 (minimal), 3 (moderate), and 5 (good). The scores can be combined to form a total score and two subscale scores:

- ***Visitor Strategies Quality*** focuses on the home visitor's responsiveness to the parent and child and consists of four items: (1) home visitor facilitation of parent-child interaction; (2) home visitor-family relationship; (3) home visitor responsiveness to family; and (4) home visitor nonintrusiveness.
- ***Visitor Effectiveness Quality*** assesses the parent's and child's engagement with each other and with the home visitor and consists of three items: (1) parent-child interaction during home visit; (2) parent engagement during home visit; and (3) child (infant or toddler) engagement during home visit.

The overall quality score, as well as the Visitor Strategies subscale, have high internal consistency (0.80 and 0.84, respectively). However, the Effectiveness Quality subscale has somewhat lower internal consistency (0.53). This score is slightly lower than the 0.70 standard in the field but higher than estimates reported by Peterson and Roggman (2006).

Table C.8 illustrates the average, standard deviation, range, and internal consistency for the HOVRS-A scores for home visits observed during the spring 2011 and 2012 Baby FACES data collection.

Classroom Assessment Scoring System-Toddler. The CLASS-T is an adaptation of the Pre-K CLASS (Pianta et al. 2008) that focuses on teacher-child interaction quality in toddler child care classrooms (La Paro et al. 2012; Pianta et al. 2010). Compared with other established quality measures of global or structural quality, the CLASS-T measures process quality along eight dimensions: (1) Positive Climate, (2) Negative Climate, (3) Teacher Sensitivity, (4) Regard for Child Perspectives, (5) Behavior Guidance, (6) Facilitation of Learning and Development, (7) Quality of Feedback, and (8) Language Modeling. These dimensions are represented in two overarching domains that describe teachers' interactions with children: Emotional and Behavioral Support and Engaged Support for Learning. Dimensions are defined by observable indicators along a seven-point scale, with ratings reflecting scores in the low (1-2), middle (3-5), and high (6-7) range.

We used the CLASS-T for observations of classrooms serving 2- and 3-year-old children during the spring 2011 and 2012 data collections. Observations also included counts of infants and toddlers and the adults caring for them that we used to compute child-adult ratios and group sizes. We computed dimensions scores by averaging ratings obtained across four independent observation cycles, and calculated domain scores for each classroom by averaging the scores for the component dimensions on which the domain scores were based. We reverse-coded scores on Negative Climate prior to calculating the domain mean score.

To ensure the reliability of our data, we assessed the inter-rater reliability of the field staff who observed classrooms serving 2- and 3-year-old children using the CLASS-T. Overall, reliability estimates were in accord with standards established by the measure developers—80 percent agreement (the same score or within one point) with the scores of the developer-certified gold

months (Toddler Engagement During Home Visit), and ensured that the indicator wording on all items is appropriate for infants and toddlers.

standard group leaders. As Table C.9 shows, trained observers scored within 1 rating point of gold standard observers' scores, who rated the same classrooms on each of the seven-point CLASS-T dimensions, 96 to 99 percent of the time.[27] Observers were thus adequately trained and maintained an adequate degree of reliability in CLASS-T scoring throughout Baby FACES data collection.

As Table C.10 shows, the internal consistency reliability of the CLASS-T Emotional and Behavioral Support and Engaged Support for Learning domain scores in our field observation was 0.91 and 0.95, respectively. These estimates are higher than those reported for the domain of Emotional Climate (0.88) in a pilot study of the measure conducted in 30 toddler classrooms (Thomason and LaParo 2009).[28] As reported by the study authors, evidence of the validity of the adapted measure was supported by associations between a number of CLASS-T dimension scores and other indicators of quality. Specifically, observed scores on Positive Climate and Teacher Sensitivity were most consistently associated with characteristics of the classroom, including teacher education level, group size, and child-teacher ratios (correlations range from 0.33 to -0.61).

Table C.10 illustrates the average, standard deviation, range, and internal consistency for CLASS-T scores for classrooms observed during the spring 2011 and 2012 Baby FACES data collections.

Measure of Quality of the Parent-Caregiver Relationship

Parent-Caregiver Relationship Scale. The PCRS assesses the perceived quality of the relationship between parents and the child's home visitor or teacher (Elicker et al. 1997). Parents reported on the quality of their relationship with the home visitor or teacher; staff, in turn, provided similar reports on their relationship with the parent. The PCRS is intended to provide focused information on multiple dimensions and specific perceptions of the dyadic relationship.

Items on PCRS focus on important dimensions of the parent-caregiver relationship, including trust and confidence, communication, respect/acceptance, caring, competence/knowledge, partnership/collaboration, and shared values. Respondents complete the questionnaire in reference to a specific caregiver or parent, indicating on a five-point scale their level of agreement or disagreement with a statement. Example statements include, "If there is a problem, my child's teacher or home visitor and I always talk about it soon," and "If there is a problem, this child's parent and I always talk about it soon." Scale scores in Baby FACES reports represent the average across a subset of these items. Separate scores are calculated for staff (using six items) and parents (using seven items).

The full PCRS scale includes 35 items, which were narrowed down to the 13 items used for the spring 2011 and 2012 data collections. We shortened the measure in response to a call after spring 2009 data collection expressing concern about the appropriateness of some items for staff and to

[27] Includes observations of classrooms serving Newborn and 1-year-old Cohort in spring 2011 (ages 2 and 3, respectively) and Newborn Cohort only in spring 2012 (age 3).

[28] The measure used by Thomason and LaParo (2009) was an age appropriate revision of the pre-K CLASS (Pianta et al. 2008) that included only six of the eight component dimensions. Consequently, the study authors did not report findings for the dimensions of Facilitation of Learning and Development and Quality of Feedback, or the resulting composite domain score derived from these dimensions. The Emotional Climate domain score reported by the authors is similar to the composite measure of Emotional and Behavioral Support derived in Baby FACES; a key dissimilarity between the measures is the inclusion of Behavior Guidance in the Emotional and Behavioral Support composite score (see Appendix D for findings of principal components factor analysis using the Baby FACES study sample).

reduce the burden on respondents. We selected a subset of items for use in the spring 2011 and 2012 data collection rounds that focused on areas of importance and demonstrated acceptable internal consistency and reliability estimates.

In a study of 217 parents and caregivers (Elicker et al. 1997), the PCRS was correlated with aspects of the infant care environment, including the amount of time the infant received care from the caregiver and caregiver work satisfaction. The authors reported internal consistency reliabilities of 0.93 for parents and 0.94 for caregivers on the measure. Correlations among the parent and caregiver scales were not significant, however, suggesting that parent-caregiver reports were incongruent. Typically, on such measures of perceived relationships, caregivers rate parents lower than parents rate caregivers, with caregivers' responses varying according to demographic characteristics of parents (such as, age, education, income, and marital status). The internal consistency reliability of the PCRS total score in our age 3 field observations for parents in home-based and center-based settings was 0.93 and 0.92, respectively. It ranged between 0.88 and 0.92 for caregivers in centers and home-based settings.

Table C.11 presents the mean, standard deviation, range, and internal consistency for PCRS scores of parents and staff serving 3-year-old children.

Table C.1 Child General and Language Development at Age 3

Outcome	Possible Range		Reported Range		Mean/ Percentage	Standard Deviation	Cronbach Alpha
	Min.	Max.	Min.	Max.			
ASQ-3 Area Score[a]							
Communication	0	60	5	60	50.75	10.14	0.58-0.82
Gross Motor	0	60	10	60	52.91	8.92	0.31-0.70
Fine Motor	0	60	0	60	44.22	14.85	0.77-0.83
Problem Solving	0	60	10	60	50.98	10.15	0.64-0.70
Personal-Social	0	60	10	60	53.71	8.13	0.64-0.76
Total Score	0	300	70	300	252.37	39.68	0.70-0.88
ASQ-3 Total Scale Score[b]							
Communication	0	100	10	100	84.61	16.88	0.83
Gross Motor	0	90	25	90	79.19	12.35	0.74
Fine Motor	0	90	0	90	63.85	20.17	0.81
Problem Solving	0	100	0	100	82.95	17.17	0.77
Personal-Social	0	100	25	100	88.73	13.32	0.76
ASQ-3 IRT Score							
Communication	.	.	30	67	56.61	7.90	.
Gross Motor	.	.	32	65	55.35	8.03	.
Fine Motor	.	.	24	73	56.15	9.02	.
Problem Solving	.	.	32	69	57.07	8.51	.
Personal-Social	.	.	30	66	56.53	8.03	.
ASQ Cut-Off Score (2 SDs below the mean or lower)							
Communication	0	1	0	1	0.94	9.66	.
Gross Motor	0	1	0	1	7.19	25.86	.
Fine Motor	0	1	0	1	6.11	23.98	.
Problem Solving	0	1	0	1	5.71	23.24	.
Personal-Social	0	1	0	1	3.90	19.39	.
ASQ in the Monitoring Zone (1 to 2 SDs below the mean)							
Communication	0	1	0	1	8.99	28.64	.
Gross Motor	0	1	0	1	13.07	33.75	.
Fine Motor	0	1	0	1	13.76	34.48	.
Problem Solving	0	1	0	1	12.53	33.14	.
Personal-Social	0	1	0	1	6.07	23.91	.
Staff-Reported CDI (English) Raw Score							
Vocabulary Comprehension	0	100	0	100	64.85	25.35	0.98
Vocabulary Production	0	100	0	100	41.88	27.60	0.98
Staff-Reported CDI (English) IRT Score	.	.	37	76	57.64	5.42	.
Staff-Reported CDI (Spanish) Raw Score							

Outcome	Possible Range Min.	Possible Range Max.	Reported Range Min.	Reported Range Max.	Mean/ Percentage	Standard Deviation	Cronbach Alpha
Vocabulary Comprehension	0	100	1	100	69.24	26.66	0.99
Vocabulary Production	0	100	0	100	38.62	29.88	0.99
Staff-Reported CDI (Spanish) IRT Score	.	.	31	81	58.78	7.97	.
Parent-Reported CDI English							
Vocabulary Production	0	100	0	100	56.60	28.67	0.98
Parent-Reported CDI Spanish							
Vocabulary Production	0	100	0	100	55.40	27.18	0.98
PLS-4 Standard Score							
English	50	150	50	141	97.36	15.22	0.89
Spanish	50	150	50	141	95.95	18.11	0.94
Bilingual	50	150	57	147	104.12	16.19	0.92
PPVT-4 English Standard Score	50	150	60	136	90.69	13.80	.
ECI – Expressive Language							
Standard Score	50	150	63	148	93.04	15.80	.
Language delay (percentage 1.5 SDs below the mean or lower)	0	1	0	1	17.03	37.63	.
At-risk for language delay (percentage 1 to 1.5 SDs below the mean)	0	1	0	1	16.59	37.24	.
Sample Size							
Parent SAQ ASQ-3			**379-461**				
SCR English CDI			**440-497**				
SCR Spanish CDI			**95-86**				
Parent SAQ English CDI			**351**				
Parent SAQ Spanish CDI			**103**				
PLS-4 English			**338**				
PLS-4 Spanish			**129**				
PLS-4 Bilingual			**135**				
PPVT-4 English			**391**				
ECI			**458**				

Source: Spring 2011/2012 Parent Self-Administered Questionnaire (SAQ), Staff-Child Report (SCR), Direct Child Assessment, and Videotaped Parent-Child Interaction.

Note: Sample includes both the 1-year-old Cohort and the Newborn Cohort at age 3.

[a] Includes only items on the age-specific forms.
[b] Includes all items across age forms.

ASQ-3 = Ages & Stages Questionnaires (Third Edition). Depending on the age of the child on the day of the parent interview, the age range of children at the spring 2011/2012 wave required administration of the ASQ-3 33-, 36-, or 42-month questionnaire.

CDI = MacArthur Communicative Development Inventories; PLS-4 = Preschool Language Scale (Fourth Edition); PPVT-4 = Peabody Picture Vocabulary Test (Fourth Edition); ECI = Early Communication Indicator.

Table C.2. Correlations Among Language Measures at Age 3

	ASQ-3 Communication	Parent-Reported CDI English Production	Parent-Reported CDI Spanish Production	Staff-Reported CDI English Production	Staff-Reported CDI Spanish Production	Staff-Reported CDI English Comprehension	Staff-Reported CDI Spanish Comprehension	PLS-4 English	PLS-4 Spanish	PLS-4 Bilingual	PPVT-4
ASQ-3 Communication IRT score	--										
Parent-Reported CDI English Vocabulary Production	0.55***	--									
Parent-Reported CDI Spanish Vocabulary Production	0.64***	--	--								
Staff-Reported CDI English Vocabulary Production	0.30***	0.35***	-0.05	--							
Staff-Reported CDI Spanish Vocabulary Production	0.23*	--	0.08	0.22*	--						
Staff-Reported CDI English Vocabulary Comprehension	0.23***	0.31***	0.06	0.74***	0.17	--					
Staff-Reported CDI Spanish Vocabulary Comprehension	0.13	--	0.05	-0.06	0.75***	0.05	--				
PLS-4 English	0.44***	0.48***	--	0.48***	--	0.43***	--	--			
PLS-4 Spanish	0.35***	-0.16	0.30**	-0.18+	-0.03	-0.12	0.14	--	--		
PLS-4 Bilingual	0.28***	0.13	0.23*	0.04	-0.13	0.08	-0.02	--	0.76***	--	
PPVT-4 English	0.27***	0.34***	-0.18	0.40***	-0.14	0.31***	-0.30*	0.64***	-0.08	0.40***	--
ECI Total Communication	0.24***	0.30***	0.04	0.19***	-0.01	0.12*	-0.04	0.21***	-0.00	0.03	0.13*

Source: Spring 2011/2012 Parent Self-Administered Questionnaire (SAQ), Staff-Child Report (SCR), Direct Child Assessment, and Videotaped Parent-Child Interaction.

Note: Sample includes both the 1-year-old Cohort and the Newborn Cohort at age 3.

ASQ-3 = Ages & Stages Questionnaires (Third Edition); CDI = MacArthur Communicative Development Inventories; PLS-4 = Preschool Language Scale (Fourth Edition); PPVT-4 = Peabody Picture Vocabulary Test (Fourth Edition); ECI = Early Communication Indicator.

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

Table C.3 Child Social-Emotional Development at Age 3

Outcome	Possible Range		Reported Range		Mean/ Percentage	Standard Deviation	Cronbach Alpha
	Min.	Max.	Min.	Max.			
Parent-Reported BITSEA Raw Score							
Problem Domain	0	62	0	47	11.07	7.53	0.88
Competence Domain	0	22	1	22	17.81	3.22	0.74
Staff-Reported BITSEA Raw Score							
Problem Domain	0	62	0	34	6.95	5.97	0.84
Competence Domain	0	22	4	22	16.63	3.50	0.77
Parent-Reported BITSEA Cut-Off Score							
Problem Domain	0	1	0	1	31.36	46.45	.
Competence Domain	0	1	0	1	16.16	36.85	.
Staff-Reported BITSEA Cut-Off Score							
Problem Domain	0	1	0	1	21.28	40.97	.
Competence Domain	0	1	0	1	13.40	34.10	.
Parent-Reported BITSEA Screening Positive	0	1	0	1	39.25	48.89	.
Staff-Reported BITSEA Screening Positive	0	1	0	1	28.43	45.15	.
Parent-Reported BPI							
Externalizing Behaviors	0	15	0	15	3.84	3.68	0.86
Internalizing Behaviors	0	10	0	10	0.91	1.56	0.77
Total Score	0	24	0	24	4.62	4.72	0.89
Staff-Reported BPI							
Externalizing Behaviors	0	15	0	15	4.58	4.56	0.92
Internalizing Behaviors	0	10	0	10	1.27	2.12	0.86
Total Score	0	24	0	24	5.60	5.91	0.93
Assessor-Reported BRS Total Scale Score							
Orientation/Engagement	9	45	13	45	36.54	7.00	0.93
Emotional Regulation	10	50	12	50	41.81	8.05	0.94
Assessor-Reported BRS in Questionable Range							
Orientation/Engagement	0	1	0	1	12.22	32.78	.
Emotional Regulation	0	1	0	1	8.77	28.31	.
Assessor-Reported BRS in Non-Optimal Range							
Orientation/Engagement	0	1	0	1	13.36	34.06	.
Emotional Regulation	0	1	0	1	10.44	30.61	.
PCI Rating Scales							
Engagement of Parent	1	7	1	7	4.77	0.94	.
Sustained Attention	1	7	1	7	4.99	0.87	.
Negativity Toward Parent	1	7	1	7	2.21	1.25	.
Enthusiasm	1	7	1	7	4.90	0.95	.

Outcome	Possible Range Min.	Possible Range Max.	Reported Range Min.	Reported Range Max.	Mean/ Percentage	Standard Deviation	Cronbach Alpha
Sample Size							
Parent SAQ			**451-458**				
SCR			**515-522**				
Assessor Rating			**479-483**				
Videotaping			**444**				

Source: Spring 2011/2012 Parent Self-Administered Questionnaire (SAQ), Staff-Child Report (SCR), Direct Child Assessment, and Parent-Child Two-Bag Task.

Note: Sample includes both the 1-year-old Cohort and the Newborn Cohort at age 3.

BITSEA = Brief Infant-Toddler Social-Emotional Assessment; BPI = Behavior Problems Index; BRS = Bayley II Behavior Rating Scale; PCI = Parent-Child Interaction.

Table C.4. Correlations Among Social-Emotional Measures at Age 3

	Parent-Rprtd BITSEA Comp.	Parent-Rprtd BITSEA Problem	Staff-Rprtd BITSEA Comp.	Staff-Rprtd BITSEA Problem	BRS Orient /Engagmt	BRS Emo Reg	Eng of Parent	Sustained Attention	Neg Toward Parent	Enthusm	Parent-Rprtd BPI Ext.	Parent-Rprtd BPI Int.	Parent-Rprtd BPI Total	Staff-Rprtd BPI Ext.	Staff Rprtd BPI Int.	Staff Rprtd BPI Total
Parent-Reported BITSEA Raw Score																
Competence Score	--															
Problem Score	-0.37***	--														
Staff-Reported BITSEA Raw Score																
Competence Score	0.25***	-0.22***	--													
Problem Score	-0.17***	0.22***	-0.47***	--												
BRS Scale Score																
Orientation/ Engagement	0.17***	-0.17***	0.18***	-0.17***	--											
Emotional Regulation	0.21***	-0.22***	0.25***	-0.23***	0.75***	--										
PCI Rating Scales																
Engagement of Parent	0.29***	-0.15**	0.22***	-0.08+	0.26***	0.26***	--									
Sustained Attention	0.21***	-0.15**	0.17***	-0.10*	0.26***	0.29***	0.66***	--								
Negativity Toward Parent	-0.13**	0.16**	-0.11*	0.07	-0.12**	-0.29***	-0.45***	-0.37***	--							
Enthusiasm	0.20***	-0.12*	0.14**	-0.04	0.31***	0.26***	0.65***	0.72***	-0.33***	--						
Parent-Reported BPI																
Externalizing	-0.34***	0.62***	-0.23***	0.19***	-0.12*	-0.22***	-0.13**	-0.13**	0.14**	-0.10*	--					
Internalizing	-0.39***	0.60***	-0.21***	0.09+	-0.10*	-0.10*	-0.11*	-0.15**	0.05	-0.14**	0.66***	--				
Total	-0.37***	0.66***	-0.24***	0.16***	-0.13**	-0.21***	-0.13**	-0.16***	0.13**	-0.13**	0.97***	0.82***	--			
Staff-Reported BPI																
Externalizing	-0.18***	0.14**	-0.41***	0.66***	-0.10*	-0.20***	-0.05	-0.04	0.15**	-0.00	0.23***	0.08	0.19***	--		
Internalizing	-0.15**	0.07	-0.37***	0.47***	-0.12*	-0.12*	-0.00	-0.04	0.07	-0.01	0.14**	0.06	0.12*	0.70***	--	
Total	-0.19***	0.13**	-0.42***	0.64***	-0.12*	-0.19***	-0.04	-0.04	0.13**	-0.01	0.21***	0.07	0.18***	0.97***	0.85***	--
ASQ-3 Personal Social	0.32***	-0.26***	0.21***	-0.18***	0.20***	0.22***	0.19***	0.17***	-0.13**	0.12*	-0.18***	-0.14**	-0.17***	-0.11*	-0.12*	-0.12*

Source: Spring 2011 /2012 Parent Self-Administered Questionnaire (SAQ), Staff Child Report (SCR), Direct Child Assessment, and Parent-Child Two-Bag Task.

Note: Sample includes both the 1-year-old Cohort and the Newborn Cohort at age 3.

BITSEA = Brief Infant-Toddler Social & Emotional Assessment; BRS = Bayley II Behavior Rating Scale; BPI = Behavior Problems Index; ASQ-3 = Ages & Stages Questionnaires (Third Edition); PCI = Parent-Child Interaction.

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

Table C.5 Parent Mental Health When Focus Child Is 3 Years Old

Outcome	Possible Range		Reported Range		Mean/ Percentage	Standard Deviation	Cronbach Alpha
	Min.	Max.	Min.	Max.			
CESD-SF							
Raw Score	0	36	0	36	3.99	5.95	0.92
Severe Depressive Symptoms	0	1	0	1	5.42	22.66	.
Moderate Depressive Symptoms	0	1	0	1	8.35	27.70	.
Mild Depressive Symptoms	0	1	0	1	15.35	36.09	.
No Depressive Symptoms	0	1	0	1	70.88	45.48	.
PSI							
Parental Distress	5	25	5	25	9.62	4.70	0.83
Parent-Child Dysfunctional Interaction	6	30	6	30	8.67	4.71	0.85
Sample Size							
Parent Interview				**444**			

Source: Spring 2011/2012 Parent Interview.

Note: Sample includes parents of both the 1-year-old Cohort and the Newborn Cohort at age 3.

PSI = Parenting Stress Index; CESD-SF = Center for Epidemiologic Studies Depression Scale Short Form. Severe depressive symptoms = scores of 15 or higher; moderate depressive symptoms = scores of 10 or higher but lower than 15; mild depressive symptoms = scores of 5 or higher but lower than 10; no depressive symptoms = scores lower than 5.

Table C.6 Functioning of Families at Age 3

Outcome	Possible Range		Reported Range		Mean/ Percentage	Standard Deviation	Cronbach Alpha
	Min.	Max.	Min.	Max.			
FES-Family Conflict[a]	1	4	1	3	1.42	0.42	0.64
Sample Size							
Parent Interview			**337**				

Source: Spring 2011/2012 Parent Interview.

Note: Sample includes parents of both the 1-year-old Cohort and the Newborn Cohort at age 3.

FES = Family Environment Scale.

Table C.7 Home and Neighborhood Environment at Age 3

Outcome	Possible Range Min.	Possible Range Max.	Reported Range Min.	Reported Range Max.	Mean/ Percentage	Standard Deviation	Cronbach Alpha
HOME							
Parental Warmth	0	7	0	3	2.49	0.83	0.62
Parental Lack of Hostility	0	5	0	3	2.50	1.03	0.92
Support of Cognitive, Language, and Literacy Environment	0	12	5	13	11.40	1.70	0.78
Internal Physical Environment	0	3	0	3	2.32	0.89	0.52
HOME Total Score	0	30	9	22	18.81	2.82	0.79
Neighborhood Disorder	--	--	-1	3	0.00	0.59	0.73
Sample Size							
Parent Interview			**356-458**				

Source: Spring 2011/2012 Parent Interview.

Note: Sample includes parents of both the 1-year-old Cohort and the Newborn Cohort at age 3.

Table C.8. Summary Statistics for Baby FACES Child Care Quality Data: HOVRS-A

HOVRS-A Scales	Possible Range Min.	Possible Range Max.	Reported Range Min.	Reported Range Max	Mean	Standard Deviation	Cronbach's Alpha
HOVRS-A Overall Quality	1	5	2	5	3.45	0.76	0.80
Visitor Strategies Quality	1	5	1	5	3.20	0.90	0.84
Effectiveness Quality	1	5	2	5	3.78	0.83	0.53
Sample Size			**181**				

Source: Spring 2011 and 2012 Home Visit Observations.

Note: Includes observations of home visits to the Newborn and 1-year-old Cohort in spring 2011 and Newborn Cohort only in spring 2012.

HOVRS-A = Home Visit Rating Scale-Adapted.

Table C.9. Gold Standard Reviewers and Field Staff Show Close Agreement on CLASS-T Dimension Scores (Percentages)

CLASS-T Dimension Scores	Agreement Within 1 Point
Positive Climate	0.97
Negative Climate	0.98
Teacher Sensitivity	0.97
Regard for Child Perspectives	0.99
Behavior Guidance	0.98
Facilitation of Learning and Development	0.96
Quality of Feedback	0.97
Language Modeling	0.99
Sample Size	**23**

Source: Spring 2011 and 2012 classroom observations.

Note: Includes observations of classrooms serving Newborn and 1-year-old Cohort in spring 2011 (ages 2 and 3, respectively) and Newborn Cohort only in spring 2012 (age 3). Reliability estimates are based on 14 observations in 2011 and 9 observations in 2012.

Table C.10. Summary Statistics for Baby FACES Child Care Quality Data: CLASS-T at Age 3

	Possible Range		Reported Range				
CLASS-T Scales	Min.	Max.	Min.	Max	Mean	Standard Deviation	Cronbach's Alpha
Emotional and Behavioral Support	1	7	1	7	5.28	0.89	0.91
Positive Climate	1	7	1	7	5.57	1.15	0.93
Negative Climate	1	7	1	6	1.34	0.71	0.85
Teacher Sensitivity	1	7	2	7	4.83	1.04	0.87
Regard for Child Perspectives	1	7	1	7	4.71	1.00	0.87
Behavior Guidance	1	7	2	7	4.67	1.16	0.89
Engaged Support for Learning	1	7	1	7	3.23	1.16	0.95
Facilitation of Learning and Development	1	7	1	7	3.73	1.13	0.86
Quality of Feedback	1	7	1	7	3.08	1.19	0.89
Language Modeling	1	7	1	7	2.87	1.32	0.91
Sample Size			**310-314**				

Source: Spring 2011 and 2012 Classroom Observations.

CLASS-T = Classroom Assessment Scoring System-Toddler.

Table C.11. Staff-Parent Relationship Quality at Age 3

PCRS Scales	Possible Range		Reported Range		Mean	Standard Deviation	Cronbach's Alpha
	Min.	Max.	Min.	Max			
PCRS Scores for Children Served in Centers							
Parent Report	1	5	1	5	4.56	0.62	0.92
Teacher Report	1	5	1	5	4.15	0.83	0.92
PCRS Scores for Children Receiving Services by Home Visits							
Parent Report	1	5	1	5	4.66	0.53	0.93
Home Visitor Report	1	5	1	5	4.36	0.65	0.88
Sample Size			**169-305**				

Sources: Spring 2011 and 2012 Parent, Teacher, and Home Visitor Interviews.

PCRS = Parent-Caregiver Relationship Scale.

APPENDIX D. ANALYTICAL ISSUES

To fully support our findings and inform our analytical decisions, we further explored various analytical issues, many of which arose as a result of our data collection administration and analytic modeling needs. This appendix focuses on seven main analytic issues: (1) comparison of the children who leave the program early compared with children who graduate, (2) use of the Classroom Assessment Scoring System-Toddler (CLASS-T) scores and factor analysis of the CLASS-T, (3) approaches to growth curve analysis, (4) procedure of creating the MacArthur-Bates Communicative Development Inventories (CDI) item response theory scores, (5) statistical methods and sensitivity analyses for modeling experience and outcomes, (6) imputing missing data and (7) construction of an overall implementation score.

Comparisons of Early Exiters and Graduates

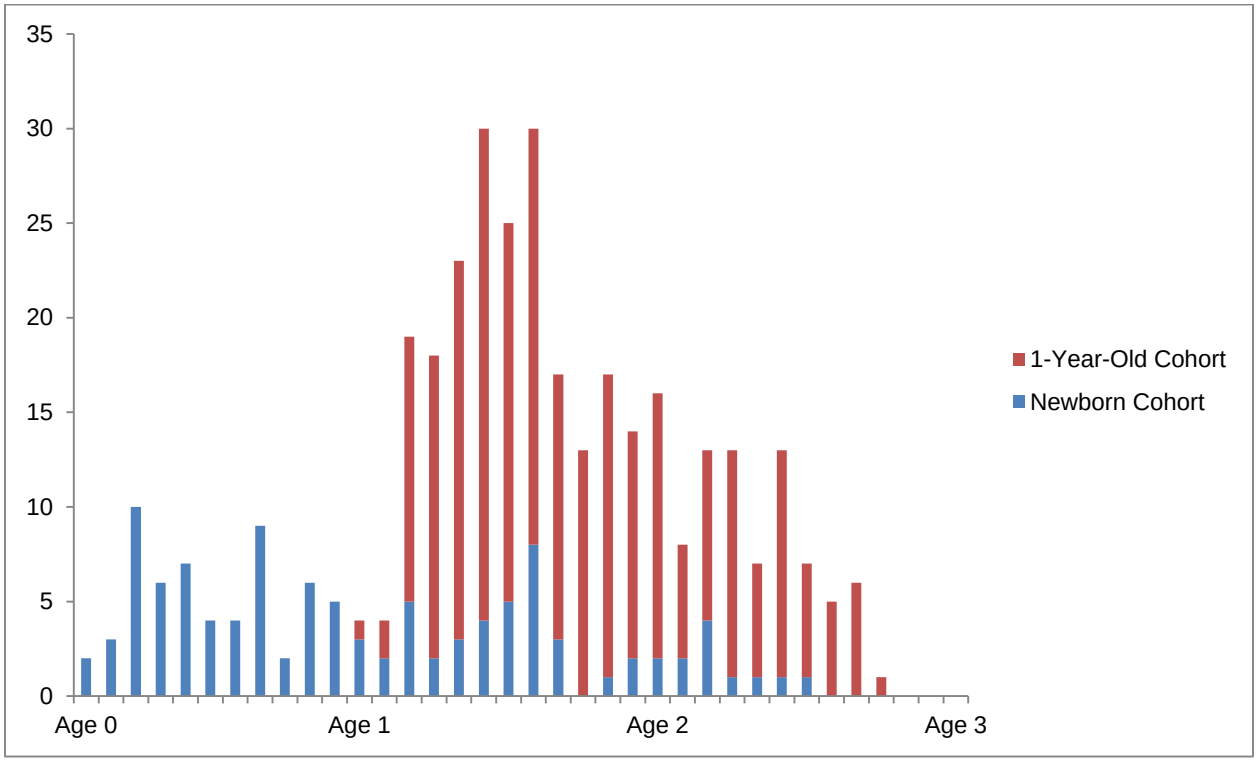
In Chapter II, we described attrition from Early Head Start and its potential effects on data analysis. Children who left their programs early ceased to be eligible for the study and were not part of subsequent data collection efforts. As such, we had less data to draw on in later years. In this section, we provide additional detail about when early exits occurred and explore the extent to which missing data due to early exiting may introduce bias in our estimates.

If early exiters are systematically more disadvantaged than graduates, then our estimates of growth may be upwardly biased—that is, we may estimate higher levels of growth than we would have if we were also able to follow the development of early exiters and include them in our analysis. For other analyses, the effect of sample attrition may be more subtle and it is more challenging to determine how our estimates would change if we had data on early exiters. For example, it is unclear whether we would expect estimates of average center attendance or home visit completion to be different if early exiters continued to receive services. We provide additional information about characteristics of early exiters in this section so that readers can take these into account in interpreting the findings presented in this report. The information included here complements the discussion in Chapter II by providing a detailed look at potential differences between the early exiters and graduates that may influence various analyses conducted.

In total, 361 children (37 percent of the sample) left their programs early (before their eligibility ended), 108 from the Newborn Cohort (56 percent of the newborn sample) and 253 from the 1-year-old Cohort (33 percent of the 1-year-old sample).[29] As illustrated in Figure D.1, most of the exits took place between ages 1 to 2, when 214 children (20 percent of the Newborn Cohort and 23 percent of the 1-year-old Cohort) left their programs early. An additional 89 children (6 percent of the Newborn Cohort and 10 percent of the 1-year-old Cohort) left their programs between ages 2 to 3. A large number of exits also occurred for the Newborn Cohort between ages 0 to 1, during which 58 children (30 percent of the Newborn Cohort) left their programs.

[29] The difference in exit rates across cohorts is significant at the < 0.001 level (p-value = 0.00).

Figure D.1. Number of Early Exits Throughout the Duration of the Study



Source: Sample Management System

In Table D.1, we compare baseline characteristics of children who left their program early with children who eventually graduated. Early exiters have similar characteristics to children who stay in the program. Most of the differences we observe between the two groups are not statistically significant. The few exceptions are that early exiters (42 percent) were more likely than graduates (28 percent) to have moved in the year prior to the baseline interview. More early exiters (59 percent) than graduates (50 percent) were born to mothers who had their first birth as a teenager.[30]

[30] The pattern of differences was similar when we conducted the analysis separately by cohort. However, differences in the proportion of mothers who had their first birth as a teenager was not statistically significant for the newborn cohort.

Table D.1. Baseline Child and Family Characteristics of Continuing Participants and Early Exiters

	Graduates Weighted Mean/Percentages (Standard Error)	Early Exiters Weighted Mean/Percentages (Standard Error)
Male	52.6 (2.41)	55.2 (2.92)
Race[a]		
White	35.7 (3.93)	36.9 (3.93)
Black	19.7 (3.33)	15.1 (2.87)
Hispanic	34.1 (4.01)	36.8 (3.95)
Other	10.4 (1.77)	11.3 (2.21)
Dual language learner status	41.2 (3.60)	36.3 (3.69)
Moved in past year	27.7 (1.91)	41.9 (3.10)***
Maternal risk factors (total number)		
Receiving public assistance	68.4 (2.42)	71.2 (3.49)
Not currently employed, in classes, or training	38.5 (2.66)	42.9 (3.86)
Less than high school education	36.7 (2.83)	41.0 (3.71)
Teenage mother at first birth	50.2 (2.85)	58.6 (2.93)*
Single mother	44.7 (2.56)	48.6 (3.46)
Income to needs ratio	1.2 (0.15)	0.9 (0.15)+
Sample size	**503-604**	**245-355**

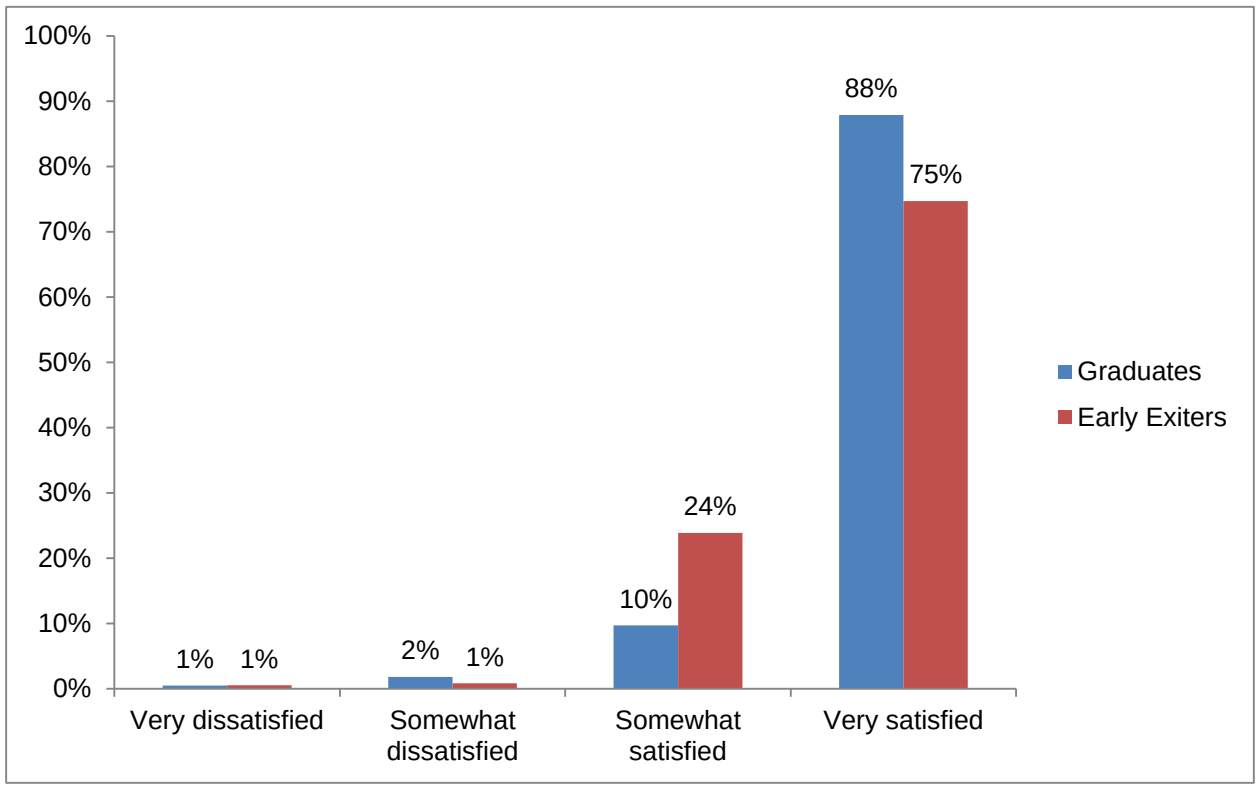
Source: Spring 2009 Parent Interview.

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

We also tested whether early exiters differed from graduates in terms of their overall satisfaction with their Early Head Start program.[31] In interviews conducted upon program exit or when graduates were 3.5 years old, parents were asked to rate their satisfaction with the program on a 4-point scale (with 1 = very dissatisfied and 4 = very satisfied). We found a statistically-significant difference between the mean satisfaction rating of early exiters (3.72) compared to the mean rating of graduates (3.85). Figure D.2 illustrates that this difference is due to a larger number of early exiters who were only "somewhat satisfied" (24 percent) compared to graduates (10 percent) and a smaller number of early exiters who were "very satisfied" (75 percent) compared to graduates (88 percent).

[31] This analysis was restricted to the 1-year-old Cohort only since we did not conduct "matriculation" interviews for the Newborn Cohort (data collection for the study ended when these children were 3).

Figure D.2. Overall Program Satisfaction of Graduates and Early Exiters



Source: Spring 2009, 2010, and 2011 Exit Interviews.

Note: Sample restricted to 1-year-old Cohort only. Differences in the "somewhat satisfied" and "very satisfied" categories are statistically significant ($p < .05$).

CLASS-T

We addressed a number of conceptual and analytic questions related to scoring the Classroom Assessment Scoring System-Toddler (CLASS-T; Pianta et al. 2010). As noted in Appendix C, the CLASS-T is a downward extension of the CLASS (Pianta et al. 2008), which addresses teacher-child interaction quality in toddler child care classrooms. In Baby FACES, we used the CLASS-T in spring 2010 to observe interactions in classrooms serving 2-year-old children; and again in spring 2011 and 2012 for observations of classrooms serving 2- and 3-year-old children. As elaborated in Appendix C, there is preliminary evidence for the validity of the adapted, pilot measure, albeit in a relatively small sample of classrooms ($N = 30$; Thomason and LaParo 2009). Baby FACES represents the first large-scale effort to establish the reliability and validity of the measure.

Use of the CLASS-T in Baby FACES

The CLASS-T measure (La Paro et al. 2012) was under development throughout the course of the Baby FACES study. However, a pilot version of the instrument was available (Pianta et al. 2010). Despite ongoing changes to the instrument by the developers during the course of the project, the pilot version of the instrument was used for the duration of the study. The measure developers trained the field staff according to the guidelines, indicators, and exemplars in the pilot manual.

Since the spring 2010 Baby FACES data collection, the CLASS-T has undergone a number of modifications and refinements, resulting in an updated version of the instrument (La Paro et al.

2012). Among the most significant changes to the instrument is the reconceptualization of the domains that characterize teacher-child interaction quality. Specifically, the CLASS-T pilot manual used in Baby FACES (Pianta et al. 2010) defines classroom interactions along eight dimensions grouped into three overarching domains: Emotional Support (Positive and Negative Climate, Teacher Sensitivity, Regard for Child Perspectives), Classroom Organization (Behavior Guidance, Facilitation of Learning and Development), and Instructional Support (Quality of Feedback, Language Modeling). The authors provide support for the validation of this "organization structure" using data from more than 3,000 classrooms ranging from preschool to fifth grade (Hamre et al. 2006); there is less available evidence to support the validity of its use with classrooms serving toddlers.

In ongoing refinements of the instrument, the authors proposed a revised organizational structure in which the dimension of Facilitation of Learning and Development—previously part of the Classroom Organization domain—became a component of the Instructional Support domain. The dimension of Behavior Guidance became subsumed under the Emotional Support domain. Other noteworthy changes included the provision of additional details, exemplars, and guidance to observers for the indicators within dimension ratings. Thus, while assessing teacher-child interactions along the same eight quality features, the updated version the instrument (La Paro et al. 2012) offers two overarching domains of toddlers' classroom experience: Emotional and Behavioral Support and Engaged Support for Learning.

Factor Analysis of CLASS-T Data

The modifications of the CLASS-T described above represent noteworthy implications for the calculation of the domain scores. We thus conducted analyses to cross-validate the underlying factor structure of the CLASS-T in our sample and guide our approach to scoring at the domain level. In spring 2010, we conducted a principal components factor analysis with varimax rotation using the eight component dimensions of the CLASS-T using data from 217 Early Head Start classrooms serving 2-year-old children (See Vogel et al. 2015 for results).We repeated this analysis with 315 classroom observations of three-year-olds collected during spring 2011 and 2012. Findings paralleled those obtained in classrooms serving 2-year-old children. As shown in Table D.2, item loadings for the factors ranged from 0.73 to 0.95 and aligned closely with the domains identified by the CLASS-T developers in the updated manual (La Paro et al. 2012).[32] The two-factor solution demonstrated high internal consistency and explained a substantial portion of common variance. Thus, two composite scores, Emotional and Behavioral Support and Engaged Support for Learning, were created by averaging the component items corresponding to each of the derived factors.

[32] The positive dimensions corresponding to the first factor were strongly and significantly correlated (0.77 to 0.86); associations between Negative Climate and the positive dimensions comprising this factor were moderate in magnitude (-0.46 to -0.50). Notably, associations between Behavior Guidance and the positive dimensions comprising the first factor ranged from 0.77 to 0.79. Associations among the three component dimensions corresponding to the second factor were strongly associated and statistically significant (0.86 to 0.90).

Table D.2. CLASS-T Dimensions Load into a Two-Factor Solution, Spring 2011 and Spring 2012

Dimension	Factor	
	Factor 1: Emotional and Behavioral Support	Factor 2: Engaged Support For Learning
Positive Climate	0.88	0.27
Negative Climate	0.73	-0.05
Teacher Sensitivity	0.83	0.43
Regard for Child Perspectives	0.84	0.37
Behavior Guidance	0.81	0.37
Facilitation of Learning and Development	0.40	0.87
Quality of Feedback	0.18	0.94
Language Modeling	0.16	0.95
Mean (Standard Deviation)	5.30 (0.89)	3.20 (1.16)
Standardized Alpha	0.91	0.95
Percentage of Total Variance Explained	64.48	18.31
Sample Size	**315**	

Source: Spring 2011 and 2012 Classroom Observations.

Note: Includes observations of classrooms serving three-year-olds (the 1-year-old Cohort in spring 2011 (and the Newborn Cohort in spring 2012).

[a] Standardized alpha calculated among items with loadings of 0.45 or higher.

CLASS-T = Classroom Assessment Scoring System-Toddler.

Analytic Strategies for Growth Curves

The longitudinal data on the child/family outcomes as well as the nested design resulted in the following hierarchical data structure in this study: child/family outcomes varying across different time points within individuals (level 1), between individuals within programs (level 2), and between programs (level 3). Using multilevel modeling will improve the precision of estimates of the associations of change/growth with child/family and program characteristics by taking the nested data structure into account. Therefore, three-level growth curve models are formulated in this study using Hierarchical Linear modeling (HLM) software (Raudenbush and Bryk 2002).

The level-1 model estimates change over time within individuals and demonstrates the change trajectory over time for each of the outcome—whether it is linear or nonlinear. We first model a nonlinear trajectory for each of the outcome measures, which include linear and quadratic (non-linear) terms, and then drop the quadratic term from the model if it is not significant. We specify the following model of change within individuals across waves:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}(d_{tij}) + \pi_{2ij}(d_{tij})^2 + e_{tii}$$

where, d_{tij} is the time measure. We handle the time measure in two ways in the analyses. First, for child outcomes, d_{tij} is the age of child i in program j at time t minus that child's mean age across all waves (that is, the age variable is centered around mean age). Thus, the intercept, π_{0ij}, represents the expected mean level of a particular outcome of child i in program j across all waves. Second, for family outcomes, d_{tij} is the wave of data collection for parent i in program j. We center the time point at baseline so that the intercept, π_{0ij}, represents the outcome of parent i in program j at baseline. The linear component, π_{1ij}, is the growth rate for child or parent i in program j, and π_{2ij} captures the acceleration (quadratic component) in each growth trajectory. e_{tii} is the random effect.

Level-2 models variation between individuals within programs. By including the child/family characteristics in the model, we examine whether the change trajectories of the outcomes differ by these variables. The intercept and growth rate parameters from the level-1 model become outcomes in level-2 models:

$$\pi_{0ij} = \beta_{00j} + \beta_{0pj}(X_{pij}) + r_{0ij}$$

$$\pi_{1ij} = \beta_{10j} + \beta_{1pj}(X_{pij}) + r_{1ij}$$

$$\pi_{2ij} = \beta_{20j}$$

where, X_{pij} is the individual level child/family characteristic variables we explore, such as gender, race/ethnicity, DLL status, family income/needs ratio, and maternal characteristics. β_{0p} represent the associations between individual level variables and the mean level of a particular outcome measure across all waves, and β_{1p} represents the association between individual level variables and the growth rate of the outcome measure. r_{0ij} and r_{1ij} represent the random effects on the intercept and linear change, respectively. Our data contained three time points, which does not allow us to estimate random effects for all growth parameters simultaneously, therefore, we fixed the quadratic parameter in the model.

Level-3 models variation between programs. The purpose of the program-level model is to examine how program characteristics are related to individual change in the outcome measures:

$$\beta_{00j} = \gamma_{000} + \gamma_{00q}(W_{qj}) + u_{00j}$$

$$\beta_{10j} = \gamma_{100} + \gamma_{10q}(W_{qj}) + u_{10j}$$

$$\beta_{0pj} = \gamma_{0p0}$$

$$\beta_{1pj} = \gamma_{1p0}$$

where, W_{qj} is the program-level variables, such as program approach. γ_{00q} represents the association between program variables and the mean level of the outcome measure; and γ_{10q} represents the association between program variables and the growth rate of the outcome measure. u_{00j}, u_{01j}, and u_{10j} represent the random effects.

Procedure for Creating the CDI Item Response Theory (IRT) Scores

The three age forms for the MacArthur-Bates Communicative Development Inventories (CDI) are different, with some items that overlap across the adjacent age forms. In order to examine growth over time in children's language abilities as measured by the CDI, we created the item response theory (IRT) scores across the three different age forms (Infant, Toddler, and CDI-III) by scaling the forms together and anchoring on the overlapped items. We did this separately for English and Spanish forms. To ensure we had a sufficient number of items that overlapped across forms at the adjacent ages and make it possible for us to scale the forms together, we selected some words from the Toddler form to include in the English CDI-III and the Spanish CDI-III. These items were selected with attention given to item difficulty and representation of conceptual categories. To minimize respondent burden we added the minimum number of items needed to scale the forms together.

We used IRT analysis, specifically a one-parameter Rasch rating scale model,[33] to scale the three CDI forms together and create the IRT scores. In the Rasch model, the probability of a specified rating is modeled as a function of a child's ability and the item difficulty.[34] Item difficulty and child ability are placed on the same scale and expressed as log odds, providing an interval measurement scale. The higher a child's ability level is, the greater the probability that the Early Head Start staff will report that the child understands and says a particular word.

Statistical Methods and Sensitivity Analyses

In Chapter IX, we report results from models relating Early Head Start experiences including family participation, service quality, and program implementation to child, parenting, and family outcomes at age 3, taking into account family and child characteristics. Box IX.3 discusses the statistical methods used in these analyses. As a first step in modeling child and family outcomes, we assessed the bivariate relationship between each experience and outcome variable. We used each outcome that had a significant bivariate association (*p*-value of 0.05 or less) with an Early Head Start experience measure as the dependent variable in a multivariate linear regression. For the dichotomous experience variables, we used a technique called doubly robust estimation in addition to multivariate regression, as a check that our multivariate linear regression results were not sensitive to the models' assumptions about the nature of the relationship between control variables and outcomes (i.e., functional form assumptions).

As discussed in Chapter II, there are limitations in our analysis and conclusions because Baby FACES is not an experimental study, and therefore causal interpretation of relationships between variables is only warranted under very specific assumptions. The most important assumption that must hold for multiple linear as well as doubly robust regression to estimate a causal effect is that all variables that predict participation, quality, or implementation and that are relevant for the outcome are included in the model. Even with the wealth of data Baby FACES provides, however, we cannot say with certainty that we have controlled for all relevant variables or that the results we report warrant a causal interpretation.

To determine how much of a problem omitted variables might be for the models relating Early Head Start experiences to child and family outcomes, we conducted sensitivity analyses. These sensitivity analyses allow us to assess how robust are findings are to omitted variable bias. In each instance that we observed a significant bivariate relationship, proceeded with multivariate analysis, and found that the relationship remained significant with the full set of controls, we re-estimated the relationship using three additional regression models. Model 1 contains basic child characteristics. Model 2 contains these characteristics plus baseline child health and cognitive and social-emotional development measures. Model 3 adds baseline parenting and family characteristics. Finally, Model 4 adds program characteristics. These models are described in Table D.3. We chose these variables, in consultation with our Technical Workgroup and internal quality assurance reviewer, in order to control for the most obvious sources of bias—that is, in order to account for differences we might observe in the outcomes that are not due to different Early Head Start experiences. If the estimates of the coefficient of interest (i.e., that on the Early Head Start experience variable) are stable across

[33] Rasch models assume unidimensionality.

[34] In our analysis of the CDI scores, item difficulty refers to the difficulties of the words. The easiest words are most likely to endorse while more difficult words are less likely to endorse.

specifications, we have evidence that our findings are robust.[35] On the other hand, if the coefficient changes substantially as different covariates are added, omitted variable bias may be a problem—we are left to wonder how much the estimates would continue to change if we were able to add other, currently unobservable, covariates.

We examined changes in the coefficient for high family involvement as a predictor of child and parent outcomes as covariates are added across the four models (Table D.4). The dependent variables are BRS Emotional Regulation, BITSEA Problem and Competence domains, and age 3 maternal demographic risk. For each outcome, the pattern is similar. The coefficient on high family involvement becomes smaller in magnitude, shrinking in most cases most noticeably between Model 2 and Model 3. The change in the coefficients between these two models provides some evidence that there may be omitted parenting and family characteristics that could shrink the coefficients even further if we were to include them in the regression models. The relationships between high involvement and the BITSEA Problem and Competence domain measures, however, remain significant at the five percent level or lower across all four models, suggestive of a strong relationship that is robust to different model specifications and less likely to be due to omitted variable bias.

[35] The sequential adding of covariates is a common method to assess the robustness of estimates in non-experimental studies, but is not without its detractors. See Gelbach (2009).

Table D.3. Four Regression Models for Sensitivity Analysis

Variable	Model 1	Model 2	Model 3	Model 4
Basic child characteristics				
Cohort	X	X	X	X
Black	X	X	X	X
Hispanic	X	X	X	X
Other	X	X	X	X
Male	X	X	X	X
Age in months, Spring 2009	X	X	X	X
DLL	X	X	X	X
HV only	X	X	X	X
Other service type	X	X	X	X
Baseline child health and cognitive/social-emotional development				
Low birth weight		X	X	X
Child health level		X	X	X
ASQ-Communication		X	X	X
ASQ-Problem Solving		X	X	X
ASQ-Personal Social		X	X	X
Baseline parenting and family characteristics				
Parenting Stress Index			X	X
High involvement at age 1			X	X
Maternal depressive symptoms			X	X
Maternal risk at age 1			X	X
Program characteristics				
Multiple-approach program				X
Urban/rural status				X
Total enrollment				X
Percentage of families speaking Spanish				X
Majority of families served by program have more than 3 demographic risks				X
Majority of families served by program have mental health or substance abuse problems				X
Majority of families served by program reside in unsafe neighborhoods or experience family violence				X
Head Start region indicators				X

ASQ = Ages and Stages Questionnaire; DLL = dual language learner; HV = home visit.

Table D.4. Results from Four Multiple Linear Regression Models for Sensitivity Analysis, Consistently High Involvement

	Model 1		Model 2		Model 3		Model 4	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
BRS Emotional Regulation	2.31**	0.71	2.03**	0.69	1.61*	0.68	1.77*	0.71
BITSEA Problem Domain (Staff Report)	-2.29***	0.57	-2.11***	0.51	-1.73**	0.54	-1.77**	-0.50
BITSEA Competence Domain (Staff Report)	1.46***	0.29	1.38***	0.28	1.32***	0.29	1.39***	0.30
Age 3 Maternal Risk	-0.44**	0.14	-0.44**	0.15	-0.25**	0.92	-0.25*	0.10
Variables Included								
Basic child characteristics	Yes		Yes		Yes		Yes	
Baseline child health and cognitive/social-emotional development	No		Yes		Yes		Yes	
Baseline parenting and family characteristics	No		No		Yes		Yes	
Program characteristics	No		No		No		Yes	

Source: Parent Interview, Staff-Child Report, Direct Child Assessment, Home Visit Observation, Survey Management System.

Note: Results calculated using multiply imputed data. Weighted results reported.

BITSEA = Brief Infant Toddler Social Emotional Assessment; BRS = Behavioral Rating Scale.

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

We observed some coefficient instability in the next relationship we considered. Table D.5 contains the coefficients from the four models on the indicator for the HOVRS-A Visitor Strategies subscale meeting or exceeding the threshold of 3. The dependent variable is the Spanish PLS-4. The coefficient shrinks moving from Model 1 to Model 3 and becomes only a trend (significant at the 10 percent level) across these models. Once program characteristics are added in Model 4, however, the coefficient increases in magnitude and becomes significant at the five percent level. This instability suggests that omitted variable bias might be a problem. That coupled with the marginal significance levels indicates that we should not place too much confidence in the Model 4 coefficient representing the true relationship between the HOVRS-A threshold and the Spanish PLS-4.

Table D.5. Results from Four Multiple Linear Regression Models for Sensitivity Analysis, Home Visitor Strategies Quality Threshold of 3

	Model 1		Model 2		Model 3		Model 4	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
PLS-4 (Spanish)	11.23+	5.94	11.30+	5.80	10.76	5.37+	14.18*	6.44
Variables Included								
Basic child characteristics	Yes		Yes		Yes		Yes	
Baseline child health and cognitive/social-emotional development	No		Yes		Yes		Yes	
Baseline parenting and family characteristics	No		No		Yes		Yes	
Program characteristics	No		No		No		Yes	

Source: Parent Interview, Staff-Child Report, Direct Child Assessment, Home Visit Observation, Survey Management System.

Note: Results calculated using multiply imputed data. Weighted results reported. Sample limited to children in the home visit option who did not change service type.

PLS = Preschool Language Scale.

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

The relationship between center quality and language development appears to be more stable. Table D.6 shows how the coefficient on the CLASS-T Engaged Support for Learning subscale average at ages 2 and 3 changes under the different models in regressions with the PPVT-4 as the dependent variable. Overall, the coefficient becomes smaller as more covariates are added though it increases slightly between Model 2 and Model 3. The coefficient is significant at the five percent level in all models. Because this coefficient is relatively stable it is suggestive of a relationship between this classroom quality measure and language development as measured by the PPVT-4.

Finally, Table D.7 contains the coefficients on the indicator for 50 percent or greater center attendance across the four models. The dependent variable is the English PLS-4. The coefficient shrinks between Model 1 and Model 2, increases between Model 2 and Model 3, then decreases between Model 3 and Model 4. It is significant at the five percent level in all models. This coefficient is unstable in magnitude but stable in terms of statistical significance across the models, indicating that further exploration of this relationship may be warranted.

Table D.6. Results from Four Multiple Linear Regression Models for Sensitivity Analysis, Average Center Quality, Ages 2-3 (CLASS-T Engaged Support for Learning)

	Model 1		Model 2		Model 3		Model 4	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
PPVT-4	3.00*	1.17	2.84*	1.16	2.92*	1.31	2.76*	1.19
Variables Included								
Basic child characteristics	Yes		Yes		Yes		Yes	
Baseline child health and cognitive/social-emotional development	No		Yes		Yes		Yes	
Baseline parenting and family characteristics	No		No		Yes		Yes	
Program characteristics	No		No		No		Yes	

Source: Parent Interview, Staff-Child Report, Direct Child Assessment, Home Visit Observation, Classroom Observation, Survey Management System.

Note: Results calculated using multiply imputed data. Weighted results reported. Sample limited to children in the center-based option who did not change service type.

PPVT = Peabody Picture Vocabulary Test.

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

Table D.7. Results from Four Multiple Linear Regression Models for Sensitivity Analysis, Center Attendance Threshold (Attended at Least 50 Percent of Recommended Days)

	Model 1		Model 2		Model 3		Model 4	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
PLS-4 (English)	7.81*	3.85	7.65*	3.61	9.80*	3.73	9.07*	3.54
Variables Included								
Basic child characteristics	Yes		Yes		Yes		Yes	
Baseline child health and cognitive/social-emotional development	No		Yes		Yes		Yes	
Baseline parenting and family characteristics	No		No		Yes		Yes	
Program characteristics	No		No		No		Yes	

Source: Parent Interview, Staff-Child Report, Direct Child Assessment, Home Visit Observation, Survey Management System, Family Services Tracking System.

Note: Results calculated using multiply imputed data. Weighted results reported. Sample limited to children in the center-based option who did not change service type.

PLS = Preschool Language Scale.

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

Multiple Imputation of Missing Data for Analytic Models

We used multiple imputation to address potential biases from data that could be systematically missing in the analytic models in chapters IV, VI, VIII, and IX. In this section, we describe the patterns of missing data in the variables used in our analyses, the correlations between data completeness and child, family, and staff characteristics, and the steps we took to impute missing data at the program, staff, and child or family levels and reduce bias.

Patterns of Missing Data

Our data on program, staff, family child, home visit, and classroom characteristics come from interviews with program directors, teachers, home visitors, and parents, direct assessments of children in their homes, and observations of home visit and classroom quality. These data (henceforth referred to as "spring data collection instruments") were collected in spring 2009, 2010, 2011, and 2012. (See Appendix B for details on data collection.) In addition to data collected each spring, we also obtained weekly data on the services families received through the FST system. For each of our data collection instruments and at each wave, we have missing data due to unit and item nonresponse.[36]

Spring Data Collection Instruments

We assessed the extent of missing data on the covariates used in our analytic models by calculating the percentage of programs, staff, home visits, classrooms, and children for whom we had non-missing data for each variable. We observed varying levels of completeness across the different covariates, especially for child-level variables. Across program-level covariates, we had complete data for 98 to 100 percent of the programs in the study. Across staff-level covariates, we had complete data for 83 to 100 percent (mean = 94 percent) of staff. In terms of classroom and home visit characteristics, we had complete data for 66 to 85 percent of home visits (mean = 75 percent), and 81 to 100 percent of classrooms (mean = 96 percent). Finally, for child-level covariates we had complete data for 62 to 100 percent of children (mean = 84 percent).

Family Services Tracking

For FST data on weekly service receipt, we measured the extent of missing data by calculating the FST reporting rate. The FST reporting rate is equal to the number of weeks in which we received an FST report for a child divided by the number of weeks the child was enrolled in (i.e., had not exited) his or her program. As discussed in Chapter II, we received at least one report for 82 percent of the children in our sample. Table D.8 shows that reporting rates vary by cohort and age. They are higher between age 1 and 2 than between age 2 and 3, and higher for the Newborn Cohort than for the 1-year-old Cohort. Figure D.3 shows that reporting rates vary throughout the year, with a notable drop for both cohorts in the summer of their age 3 year; that is, near the end of their time in Early Head Start.

[36] We also have missing data due to attrition from the program (i.e., early exiters). Early exiters were ineligible to participate in Baby FACES so we did not collect follow-up data on them after their exits. The extent of early exit and its associations with key family and child characteristic are described elsewhere (Chapter 2 and "Early Exiters" section of Appendix D) in this report.

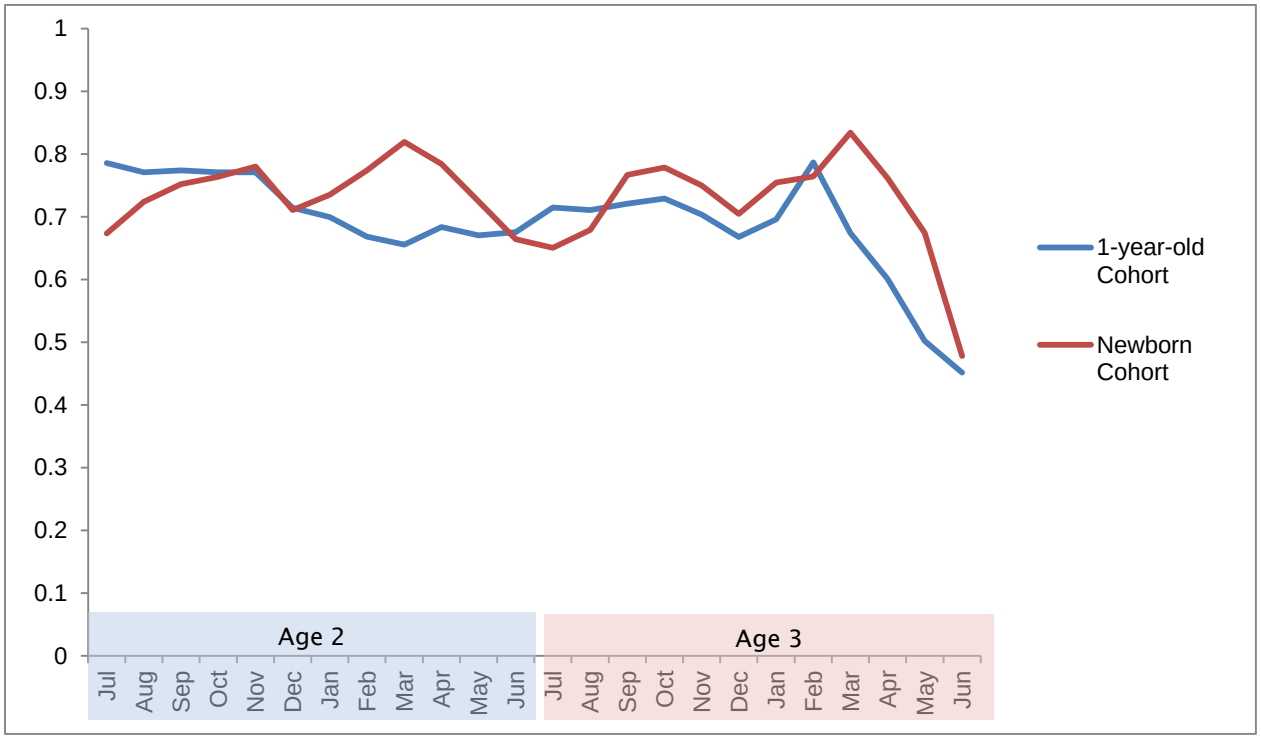
Table D.8. FST Reporting Rates by Cohort and Age

Cohort	Age	Average FST Reporting Rate, Percent (SE)	Sample Size
1-year old	1-2	72.3 (1.17)	676
1-year old	2-3	68.1 (1.45)	450
Newborn	1-2	75.2 (2.84)	96
Newborn	2-3	72.5 (3.72)	67

Source: Family Services Tracking System.

Note: Data are not weighted and include only those children for whom we received at least one FST report during a given one-year period. The FST reporting rate is equal to the number of weeks in which we received an FST report for a child divided by the number of weeks the child was enrolled in his or her program.

Figure D.3. Family Services Tracking System Reporting Rates by Cohort and Month



Reporting rates varied by program. Out of 89 programs, 88 submitted at least one FST report. FST reporting rates within programs varied from 2 percent to 99 percent. Six programs had FST reporting rates of less than 25 percent. A plurality of programs (39) had FST reporting rates of at least 75 percent.

We also observed gaps between FST reports received. We expected to receive one report per week as long as a child was enrolled in his or her program, even if the child did not receive services in a particular week (staff could note that on the FST form). A gap is defined as having two or more weeks between consecutive FST reports. Table D.9 shows data on gaps by cohort and age (for example the year between age 1 and 2 or between age 2 and 3). Approximately 60 percent of children had at least one gap. For children with at least one gap, the average number of gaps was less than three. Children had up to 14 gaps in one year, with a maximum gap length (i.e., the number of weeks between two consecutive FST reports) of 47 weeks. The average gap length was less than five weeks for all cohorts and ages, with the Newborn Cohort having shorter gaps overall.

Table D.9. Gaps in Family Services Tracking Reports Received

Cohort	Age	Percentage of Children with at Least One Gap	Average Number of Gaps[a] (SE)	Maximum Number of Gaps[a]	Average Gap Length in Weeks[a] (SE)	Maximum Gap Length in Weeks[a]
1-year-old	1-2	62.7	2.8 (0.10)	11	4.8 (0.31)	47
1-year-old	2-3	57.6	2.8 (0.14)	14	3.7 (0.27)	39
Newborn	1-2	58.3	2.8 (0.31)	9	3.5 (0.38)	29
Newborn	2-3	58.2	2.9 (0.40)	12	2.6 (0.15)	13

Source: Family Services Tracking System.

Note: Data are not weighted and include only those children for whom we received at least one FST report during a given one-year period.

[a]For children with one or more gaps. A gap is defined as having two or more weeks between consecutive FST reports.

Child, Family and Staff Characteristics Are Associated with Variable Completeness

Exploratory analyses suggest that data are not missing completely at random (MCAR). For illustration, we examined correlations between child, family, and staff characteristics and measures of data completeness on several key instruments. Children for whom we have more complete data on covariates from the spring data collection instruments are more likely to be from the 1-year old Cohort and dual language learners (Table D.10). We also examined correlations between characteristics and the completeness of the FST. Children in the home-based service option and children who are white tend to have a higher percentage of FST reports completed. Dual language learners and Hispanic children have lower FST reporting rates, though all correlations are in the low range (Table D.11).

Observing significant associations between child, family, and staff characteristics and the extent of missing data across several instruments means that estimates obtained from these data would likely be biased if we ignored the missingness; that is, if we used casewise deletion. Thus, we chose to impute missing data, taking advantage of a comprehensive imputation model made possible by Baby FACES' rich data sources. The imputation method is described below.

Table D.10. Correlations Between Completeness of Child-level Covariates and Baseline Child and Family Characteristics

	Percentage of Non-missing Child-level Covariates
In home-based service option (vs. center or combination)	0.02
Family enrolled in pregnancy	0.01
1-year old Cohort (vs. Newborn Cohort)	0.12***
Male child	-0.02
Dual language learner	0.09**
Race/Ethnicity (vs. all other races)	
White	-0.03
African American	0.00
Hispanic	0.04
Other	-0.02
Maternal demographic risks	-0.06+
Sample Size	**810-971**

Sources: Survey Management System, Parent Interview.

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

Table D.11. Correlations Between FST Reporting Rate and Baseline Child and Family Characteristics

	FST Reporting Rate
In home-based service option (vs. center or combination)	0.12***
Family enrolled in pregnancy	-0.02
1-year old Cohort (vs. Newborn Cohort)	-0.05
Male child	0.03
Dual language learner	-0.09*
Race/Ethnicity (vs. all other races)	
White	0.12***
African American	-0.06+
Hispanic	-0.10**
Other	0.06
Maternal demographic risks	-0.05
Sample Size	**712-786**

Sources: Survey Management System, Family Services Tracking System, Parent Interview.

FST = Family Services Tracking system.

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

Imputation Method

We imputed missing data through multiple imputation by chained equations (MICE) using the "mi" suite of commands in Stata. Baby FACES data have some notable characteristics. First, the data are nested (children within staff within program)[37] at each time point. Second, the data are longitudinal. We have up to 3 or 4 waves of interview, staff-child report, home and classroom observation, and assessment data, and up to 104 weekly observations per child from the FST. In this section, we describe our method for imputing data from the spring data collection instruments and the FST. It is also important to describe when we do not impute data. For example, we used available data but did not impute missing data if the reason for missingness is that a teacher or home visitor is no longer affiliated with a study child, the child exits the program early, or the family withdrew consent to participate in the study.

Imputing Data from Spring Instruments

We conducted separate imputations within each level and merged the imputed datasets to create a complete analytic dataset. We used "wide" datasets in each stage so that we could impute variables in later time points using variables from earlier waves.[38] Box D.1 lists the variables in our imputation models. We included measures with missing values that were to be dependent variables in final analyses in the imputation model but used the non-imputed versions in final analyses.

We imputed the few missing observations on program characteristics by creating a program-level dataset containing variables from three waves of program director interviews and questionnaires. Our imputation model included variables used to stratify the sample of programs

[37] Each of the 89 programs enrolls as few as 2 and up to 26 study children (median = 7) and employs as few as 1 and up to 28 different staff members (median = 3). Each staff member serves as primary caregiver for as many as 6 children (median = 1).

[38] A wide dataset is one in which measurements from each data collection wave are represented by individual variables. For example, maternal employment status was collected in each spring parent interview. In a wide dataset, there would be up to four variables for maternal employment status for each family. In contrast, in a long dataset, this would be one variable with up to four observations per family.

selected for the study including urbanicity, program size, and the percentage of Spanish speaking families enrolled. We created 20 imputed program-level datasets.

To impute staff data, we created staff-level datasets containing variables from all three waves. We created separate datasets for home visitors and teachers. Since staff members were only interviewed in years in which they were attached to a study child, we only imputed characteristics for staff who were eligible (that is, associated with a study child) in each year.[39] Our imputation model included staff characteristics as well as program-level variables with no missing data.[40] We created 20 imputed staff datasets.

To impute child and family characteristics, we created a child-level dataset containing variables across waves. The imputation model included child and family characteristics as well as program variables with no missing data. Because some children have missing data due to program attrition, our imputation model included a dichotomous variable to identify early exiters. We did not impute data for early exiters after they left their programs, nor did we impute data for the few children who had revoked consent in later periods. We imputed for all three time periods in each child-level dataset to create 20 imputed child-level datasets. To create our final analytic set of imputed datasets, we merged imputed program and staff data to the imputed child-level datasets, keeping only the staff data that pertained to the relevant periods for each child.

Box D.1. Variables in Imputation Models

- ***Program characteristics***
  - Urbanicity
  - Program size
  - Percentage of Spanish-speaking families enrolled
  - Yearly implementation ratings
  - Yearly staff turnover rates
  - Program approach
  - Whether more than 50 percent of families had mental health or substance abuse problems
  - Whether more than 50 percent of families lived in unsafe neighborhoods or were experiencing family violence
  - Whether more than 50 percent of families had more than three demographic risks
- ***Staff characteristics***
  - Has a bachelor's degree or higher
  - Race/ethnicity
  - Years of experience in Early Head Start
  - Has a degree in early childhood education
  - Has a Child Development Associate (CDA) credential
  - Speaks a language other than English
  - Depressive symptoms (Center for Epidemiologic Studies Depression Scale [CES-D])
  - Assigned a mentor or coach

[39] We interviewed a total of 757 unique staff members over three years but only 151 have three years of data.

[40] We included variables with no missing observations since using variables with missing values as predictors in our imputation models could result in missing imputed values.

- ***Home visit characteristics***
  - Percentage of time spent on family-focused activities
  - Percentage of time spent on parent-child activities
  - Percentage of time spent on staff-family relationship building
  - Percentage of time spent on crisis management activities
  - Alignment with visit plan
  - Interference from environmental distractions
  - Presence of other children
  - Presence of other adults
  - HOVRS-A Visitor Strategies rating
  - HOVRS-A Visitor Effectiveness rating
- ***Classroom characteristics***
  - Class size
  - Number of dual language learners in the classroom
  - Adult-child ratio
  - CLASS-T Engaged Support for Learning score
  - CLASS-T Emotional and Behavioral Support score
- ***Child characteristics***
  - Race/ethnicity
  - Gender
  - Age in months
  - Dual language learner status
  - Child born with low or very low birth weight
  - Overall health status
  - Premature birth
  - Social-Emotional development
    - Brief Infant Toddler Social Emotional Assessment (BITSEA) (staff- and parent-reported)
    - Bayley Behavioral Rating Scale (BRS)
    - Ages & Stages Questionnaires, Third Edition (ASQ-3)
  - Language and cognitive development
    - Ages & Stages Questionnaires, Third Edition (ASQ-3)
    - Preschool Language Scale-4th Edition, Auditory Comprehension subscale (PLS-4)
    - Peabody Picture Vocabulary Test-4th Edition (PPVT-4)
    - MacArthur-Bates Communicative Development Inventories (CDI)—Infant Short Form
    - Early Communication Indicator (ECI)
- ***Family characteristics***
  - Household income-to-needs ratio
  - Maternal demographic risk
  - Psychological risk
  - Early Head Start service option in spring
  - Quality of relationship with Early Head Start provider (staff- and parent-reported)
  - Staff rating of family involvement

- Whether the family was an early exiter
- Length of program enrollment
- Parental Depressive symptoms (CES-D)
- Parenting
  - Parenting Stress Index-Short Form (PSI-SF)
  - Parent-Child Interaction Rating Scales for the Two-Bag Assessment (PCI)
  - The Parenting Interactions with Children: Checklist of Observations Linked to Outcomes
  - Parent Support for Child Learning Index
  - Parent Provision of Learning Materials Index

- ***Weekly service variables***
  - Center days offered
  - Center days attended
  - Home visits offered
  - Home visits received

Imputing Weekly Service Data

We report on FST data collected from July 2009 to June 2010 for the 1-year-old Cohort (corresponding to ages 1 to 2), from July 2010 to June 2011 for both cohorts (ages 1 to 2 for the Newborn Cohort and ages 2 to 3 for the 1-year-old Cohort), and from July 2011 to June 2012 for the Newborn Cohort (corresponding to ages 2 to 3). The end of each of the FST periods roughly corresponds to the spring data collection period, so we classify data collected from ages 1 to 2 as age 2 data and data collected from ages 2 to 3 as age 3 data.

We imputed data for weeks in which a child was eligible for services but in which we did not receive an FST report. In our analyses, we used the weekly data to create aggregate yearly measures of services received (such as the number of home visits received or center days attended during the year). We imputed at the weekly level rather than imputing the aggregate measure so that we could account for attendance and reporting patterns that vary throughout the year. (See Vogel et al. 2015 for a discussion of the seasonal attendance patterns we observed for the 1-year-old Cohort at age 2.)

We imputed 20 datasets separately by cohort and age. Taking advantage of the long time-series dimension of the FST data, we used program, child, and month fixed effects instead of sets of program- or child-level variables. An advantage of using fixed effects instead of observed variables is that they control for observable as well as unobservable characteristics (as long as these characteristics are time invariant).

After imputing, we merged the imputed datasets to create 20 "wide" datasets so that we could calculate aggregate yearly measures (for example, total number of home visits received at age 2 and age 3) for each child at each age he or she was in the Baby FACES sample.

Estimation and Analysis

Our estimates are based on analyses conducted across 20 multiply imputed datasets. We obtained each estimate by calculating the statistic of interest (such as a mean or a regression coefficient) within each imputation and then averaging across imputations. The standard errors of

these statistics account for two sources of variance: (1) sampling variance, which is based on average standard errors within each imputed dataset; and (2) imputation variance, which is the variance in estimated averages across imputations.

As described above, we used the non-imputed versions of variables that we consider as outcomes in final analyses. The reported results are weighted to account for nonresponse on the outcome variables as well as to adjust for consent status and the probability of selection into the sample.

Construction of an Overall Implementation Score

We explored various measures of program implementation for analyses. We decided it would be most useful to creating a single cross-year indicator of program implementation that (1) capitalized on the multiple waves of data available but provided a stable indicator of implementation quality that was less prone to year-to-year shifts in program resources, staff turnover, as well as changes in data collection approaches, and (2) facilitated a meaningful comparison between programs with varying levels of implementation.

Description of Implementation Data

Data Collection Protocol and Response Rates. Implementation ratings in 2009 were based on program director's responses on a self-administered questionnaire. Directors rated various aspects of their program's implementation on a scale ranging from 1 (Low) to 5 (Enhanced). The instrument used provided comprehensive descriptors for each anchor point, which was heavily based on summary ratings used in the EHSREP. In 2010 and 2011 a different approach to measuring implementation was used. The team developed items with concrete response categories that tracked to each element in the rating form. This approach was less subjective, first because program directors could not see the ratings they were awarding themselves and second, because it allowed us to ensure that all requirements for a given anchor were met. Program directors responded to these survey items during the program director interview, and these were then coded by the analysis team into ratings for each cornerstone on a scale ranging from 1 (Low/Minimal), 2 (Moderate), 3 (Full), 4 (Enhanced). Response rates were slightly lower in 2009 compared to 2010 and 2011 (Table D.12). However, there was at least one complete set of ratings for each program in the study.

Table D.12. Means and Standard Deviations of Cornerstone and Overall Ratings, by Year

	Means (Standard Deviations)			
	2009	2010	2011	Cross-Year
Community Building	3.29 (0.40)	3.19 (0.16)	3.21 (0.14)	3.36 (0.18)
Child Development	3.30 (0.59)	3.16 (0.20)	3.12 (0.21)	3.19 (0.20)
Family Development	3.28 (0.49)	3.16 (0.31)	3.09 (0.31)	3.18 (0.30)
Management Systems	3.71 (0.46)	3.41 (0.17)	3.44 (0.12)	3.53 (0.18)
Staff Development	3.74 (0.40)	3.31 (0.29)	3.31 (0.27)	3.30 (0.26)
Overall Rating	3.46 (0.36)	3.24 (0.13)	3.23 (0.11)	3.31 (0.16)
Sample Size	**84-86**	**85-89**	**77-88**	**89**

Notes: 2009 ratings were recoded to a 1-4 scale (with the lowest 2 categories combined) for comparison purposes. The overall average in each year is calculated by taking the average of cornerstone ratings in that year. The cross-year cornerstone ratings are based on average scores across 3 years. The cross-year overall rating is based on the average of the cross-year cornerstone ratings.

Comparison of Ratings Over Time. The ratings for each cornerstone are similar across years, with the average program obtaining a rating above 3 (adjusting the 2009 ratings to be on the same 1-4 scale as the later years' ratings). Ratings in 2009 were slightly higher, on average, compared to the following years (Table 1). We also found more variability in 2009 ratings as evidenced by the larger standard deviations.

Proposed Approach for Creating a Cross-Year Rating

The approach for creating a cross-year rating was to calculate a cross-year average for each cornerstone based on all available data from 2009, 2010, and 2011. Then we calculated an overall rating based on the average of cross-year cornerstone ratings. The cross-year averages are presented in the last column of Table D.12. We considered excluding the 2009 data in creating a cross-year average, due to the difference in data collection protocol and mode from the protocol and mode of administration in later years. However, doing so would restrict the variation in scores since 2009 ratings were more variable. Although the more subjective nature of the instrument in 2009 may have caused some artificial inflation in 2009 ratings, it is unlikely that the inflation would have impacted the ratings of some programs more than others. The 2009 ratings also provide an important data point for programs since these ratings were obtained at baseline and shortly before some programs received ARRA expansion funds in 2010.

APPENDIX E. SUPPLEMENTAL TABLES

This appendix provides supplemental tables for Chapter IX and the implementation rating scale.

Chapter IX Supplemental Tables

Table E.IX.1. Full Multivariate Linear Regression Results, High Family Involvement at Age 3

	PLS-4		BRS-ER		BITSEA-PD		BITSEA-CD		Maternal Risk	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
High involvement at ages 2 and 3	2.67	2.69	1.77*	0.71	-1.77**	0.50	1.39***	0.30	-0.25*	0.10
Cohort	17.48+	10.07	-7.52*	2.96	-0.86	2.25	0.85	1.38	0.08	0.42
Black	-6.53*	2.91	-0.65	1.43	0.54	0.88	-0.64	0.58	0.01	0.15
Hispanic	-5.31	3.46	-1.31	1.18	0.32	0.87	-0.59	0.63	0.12	0.21
Other	2.54	3.09	0.85	1.13	-1.38+	0.77	0.85	0.47	-0.25+	0.13
Male	-1.64	1.88	-2.63***	0.72	1.17*	0.57	-0.88**	0.31	-0.09	0.09
Age in months, Spring 2009	-1.75*	0.78	0.60*	0.25	-0.04	0.18	-0.11	0.12	-0.01	0.04
DLL	-0.94	2.79	1.14	1.07	-0.94	0.63	0.39	0.40	-0.01	0.16
HV only	-7.23**	2.67	-3.96***	0.94	-3.15***	0.71	-0.67	0.49	0.15	0.12
Other service type	-7.29*	3.13+	-2.28	1.24	-3.02***	0.81	-0.3	0.53	0.13	0.13
Low birth weight	-3.11	3.00	-2.48	1.53	2.51	1.31	-1.63*	0.76	-0.24	0.16
Child health level	0.96	2.29	0.83	0.82	0.76	0.61	-0.46	0.41	0.00	0.11
ASQ-Communication	0.19+	0.10	0.02	0.04	-0.04+	0.02	0.03*	0.01	-0.01+	0.00
ASQ-Problem Solving	0.12	0.09	0.02	0.04	-0.02	0.03	0.00	0.02	0.01	0.00
ASQ-Personal Social	-0.02	0.09	-0.03	0.05	0.01	0.03	-0.01	0.02	0.00	0.01
Parenting Stress Index	-0.13	0.23	-0.12	0.09	0.01	0.05	-0.02	0.04	-0.01	0.01
High involvement at age 1	0.47	2.18	0.65	0.71	-0.5	0.56	0.39	0.33	-0.30**	0.10
Maternal depressive symptoms	-0.01	0.15	-0.11	0.08	0.07	0.05	-0.04	0.03	0.01	0.01
Maternal risk at age 1	-1.48*	0.71	-0.85**	0.30	0.04	0.21	0.05	0.14	0.72***	0.04
Multiple-approach program	-2.56	2.15	-1.01	1.03	2.21***	0.64	-0.35	0.38	0.07	0.11
Urban/rural status	1.66	2.15	1.42+	0.81	-0.05	0.59	0.06	0.33	-0.09	0.12
Total enrollment	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Percentage of families speaking Spanish	2.78	8.02	3.01	1.59	-1.93+	1.15	0.79	0.78	0.26	0.23
Majority of families served by program have more than 3 demographic risks	5.60*	2.56	1.49	1.03	0.21	0.74	-0.41	0.53	0.08	0.17
Majority of families served by program have mental health or substance abuse problems	3.05	3.50	1.09	1.34	-0.61	0.64	-0.23	0.46	-0.30	0.21
Majority of families served by program reside in unsafe neighbor-hoods or experience family violence	-3.84	2.65	-1.01	0.96	1.28*	0.52	-0.22	0.34	-0.05	0.12
Constant	99.76***	7.97	44.1***	4.08	8.53**	2.62	19.18***	1.34	1.43**	0.51
Sample size	**360**		**489**		**485**		**484**		**517**	

Source: Parent Interview, Staff-Child Report, Direct Child Assessment, Home Visit Observation, Survey Management System.

Note: Results calculated using multiply imputed data. Weighted results reported. Dummies for Head Start region are included in the regressions but their coefficients are not reported.

ASQ = Ages and Stages Questionnaire
BITSEA = Brief Infant Toddler Social Emotional Assessment (PD = Problem Domain, CD = Competence Domain)
BRS = Behavioral Rating Scale
DLL = Dual language learner
HV = Home visiting
PLS = Preschool Language Scale
PPVT = Peabody Picture Vocabulary Test

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

Table E.IX.2. Full Multivariate Linear Regression Results, Enrollment during Pregnancy

	PPVT-4		Maternal Risk	
	Coeff.	SE	Coeff.	SE
Enrollment during pregnancy	-2.37	1.59	-0.17+	0.10
Cohort	10.00	6.29	0.01	0.41
Black	-0.87	2.68	0.08	0.16
Hispanic	-2.13	2.98	0.18	0.20
Other	3.85	2.70	-0.21	0.14
Male	-0.96	1.47	-0.08	0.09
Age in months, Spring 2009	-0.88+	0.49	-0.01	0.04
DLL	-1.97	2.58	-0.02	0.17
HV only	-4.73*	2.10	0.14	0.13
Other service type	-5.00*	1.96	0.17	0.14
Low birth weight	3.53	4.11	-0.18	0.16
Child health level	2.69	2.15	-0.01	0.11
ASQ-Communication	-0.05	0.07	-0.01+	0.00
ASQ-Problem Solving	0.10	0.08	0.01	0.00
ASQ-Personal Social	0.16+	0.09	0.00	0.01
Parenting Stress Index	-0.37*	0.14	-0.01	0.01
High involvement at age 1	-2.32	1.80	-0.33***	0.10
Maternal depressive symptoms	0.16	0.14	0.01	0.01
Maternal risk at age 1	-2.06**	0.70	0.72***	0.04
Multiple-approach program	-3.20	2.09	0.03	0.12
Urban/rural status	-0.44	1.86	-0.07	0.12
Total enrollment	0.00	0.01	0.00	0.00
Percentage of families speaking Spanish	-8.24+	4.59	0.24	0.23
Majority of families served by program have more than 3 demographic risks	4.68*	1.92	0.05	0.18
Majority of families served by program have mental health or substance abuse problems	2.95	2.23	-0.27	0.22
Majority of families served by program reside in unsafe neighborhoods or experience family violence	-2.66	1.95	-0.06	0.12
Constant	95.51***	7.32	1.56**	0.54
Sample size	**411**		**517**	

Source: Parent Interview, Staff-Child Report, Direct Child Assessment, Home Visit Observation, Survey Management System.

Note: Results calculated using multiply imputed data. Weighted results reported. Dummies for Head Start region are included in the regressions but their coefficients are not reported.

ASQ = Ages and Stages Questionnaire
DLL = Dual language learner
HV = Home visiting
PPVT = Peabody Picture Vocabulary Test

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

Table E.IX.3. Full Multivariate Linear Regression Results, Enrollment Length 30-month Threshold

	PLS-4 (Bilingual)[a]	
	Coeff.	SE
Enrolled for longer than 30 months	-0.79	3.28
Cohort	43.35***	12.91
Black	N/A	N/A
Hispanic	N/A	N/A
Other	N/A	N/A
Male	-7.30*	2.80
Age in months, Spring 2009	-3.63***	1.10
DLL	N/A	N/A
HV only	-1.06	4.38
Other service type	6.41	3.71
Low birth weight	-14.58+	8.49
Child health level	-0.30	2.85
ASQ-Communication	0.00	0.12
ASQ-Problem Solving	0.01	0.18
ASQ-Personal Social	0.01	0.21
Parenting Stress Index	0.27	0.38
High involvement at age 1	4.10	3.44
Maternal depressive symptoms	0.08	0.42
Maternal risk at age 1	0.69	1.09
Multiple-approach program	-10.51	6.79
Urban/rural status	2.09	4.88
Total enrollment	-0.02*	0.01
Percentage of families speaking Spanish	-6.99	6.54
Majority of families served by program have more than 3 demographic risks	6.09	4.27
Majority of families served by program have mental health or substance abuse problems	-1.41	4.30
Majority of families served by program reside in unsafe neighborhoods or experience family violence	-0.64	4.50
Constant	131.91***	14.06
Sample size	**189**	

Source: Parent Interview, Staff-Child Report, Direct Child Assessment, Home Visit Observation, Survey Management System.

Note: Results calculated using multiply imputed data. Weighted results reported. Dummies for Head Start region are included in the regressions but their coefficients are not reported.

[a]Models with Spanish or bilingual assessments as the dependent variable do not contain race/ethnicity or DLL indicators because of low variability in these indicators when these assessments are considered.

ASQ = Ages and Stages Questionnaire
DLL = Dual language learner
HV = Home visiting
PLS = Preschool Language Scale

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

Table E.IX.4. Full Multivariate Linear Regression Results, HOVRS-A Visitor Strategies Threshold of 3

	PLS-4 (Spanish)[a]	
	Coeff.	SE
HOVRS-A, average of 3 or more at ages 2-3	14.18*	6.44
Cohort	52.82*	20.93
Black	N/A	N/A
Hispanic	N/A	N/A
Other	N/A	N/A
Male	-12.26+	7.11
Age in months, Spring 2009	-4.86**	1.64
DLL	N/A	N/A
HV only	N/A	N/A
Other service type	N/A	N/A
Low birth weight	-9.82	5.93
Child health level	-7.88	6.46
ASQ-Communication	0.01	0.18
ASQ-Problem Solving	-0.10	0.30
ASQ-Personal Social	-0.33	0.29
Parenting Stress Index	-0.48	0.91
High involvement at age 1	-1.76	6.11
Maternal depressive symptoms	1.32**	0.47
Maternal risk at age 1	0.48	2.44
Multiple-approach program	-34.28***	9.44
Urban/rural status	15.91+	7.95
Total enrollment	-0.05**	0.01
Percentage of families speaking Spanish	8.88	22.26
Majority of families served by program have more than 3 demographic risks	14.64*	6.44
Majority of families served by program have mental health or substance abuse problems	14.56	11.24
Majority of families served by program reside in unsafe neighborhoods or experience family violence	-4.45	5.71
Constant	149.43***	32.73
Sample size	**114**	

Source: Parent Interview, Staff-Child Report, Direct Child Assessment, Home Visit Observation, Survey Management System.

Note: Results calculated using multiply imputed data. Weighted results reported. Sample limited to children in the home visit option who did not change service type. Dummies for Head Start region are included in the regressions but their coefficients are not reported.

[a]Models with Spanish or bilingual assessments as the dependent variable do not contain race/ethnicity or DLL indicators because of low variability in these indicators when these assessments are considered.

ASQ = Ages and Stages Questionnaire
DLL = Dual language learner
HOVRS-A = Home Visit Rating Scales-Adapted
HV = Home visiting
PLS = Preschool Language Scale

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

Table E.IX.5. Full Multivariate Linear Regression Results, Average Center Quality (CLASS-T Emotional and Behavioral Support), Ages 2-3

	PPVT-4		BITSEA-PD	
	Coeff.	SE	Coeff.	SE
CLASS-T Emotional and Behavioral Support, average at ages 2-3	2.99	1.88	-1.43	1.05
Cohort	18.15+	9.25	-5.65	4.60
Black	0.20	2.72	-0.87	1.44
Hispanic	-3.27	3.75	0.33	1.49
Other	11.47**	3.85	-2.46	1.95
Male	-3.26	2.13	1.25	1.05
Age in months, Spring 2009	-1.30+	0.70	0.37	0.34
DLL	0.92	3.03	-1.65	1.93
HV only	N/A	N/A	N/A	N/A
Other service type	N/A	N/A	N/A	N/A
Low birth weight	1.68	4.87	-0.84	1.95
Child health level	2.67	3.07	1.42	1.50
ASQ-Communication	-0.16+	0.10	-0.06	0.04
ASQ-Problem Solving	0.07	0.10	-0.01	0.05
ASQ-Personal Social	0.05	0.14	0.03	0.07
Parenting Stress Index	-0.17	0.23	-0.05	0.10
High involvement at age 1	0.15	2.35	-2.88*	1.27
Maternal depressive symptoms	0.03	0.22	0.25+	0.13
Maternal risk at age 1	-1.17	0.79	0.18	0.41
Multiple-approach program	3.54	3.00	2.68*	1.19
Urban/rural status	-0.26	2.66	2.17	1.46
Total enrollment	-0.01	0.02	0.00	0.01
Percentage of families speaking Spanish	-13.86+	7.96	-6.44*	3.01
Majority of families served by program have more than 3 demographic risks	1.06	3.01	0.48	1.37
Majority of families served by program have mental health or substance abuse problems	8.97*	3.54	3.51*	1.40
Majority of families served by program reside in unsafe neighborhoods or experience family violence	-0.79	2.58	0.16	1.20
Constant	85.63***	15.02	15.35+	7.88
Sample size	**244**		**230**	

Source: Parent Interview, Staff-Child Report, Direct Child Assessment, Home Visit Observation, Classroom Observation, Survey Management System.

Note: Results calculated using multiply imputed data. Weighted results reported. Sample limited to children in the center-based option who did not change service type. Dummies for Head Start region are included in the regressions but their coefficients are not reported.

ASQ = Ages and Stages Questionnaire
BITSEA = Brief Infant Toddler Social Emotional Assessment (PD = Problem Domain, CD = Competence Domain)
CLASS-T = Classroom Assessment Scoring System-Toddler Version
DLL = Dual language learner
HV = Home visiting
PPVT = Peabody Picture Vocabulary Test

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

Table E.IX.6. Full Multivariate Linear Regression Results, Average Center Quality (CLASS-T Engaged Support for Learning), Ages 2-3

	PPVT-4		PLS-4 (Bilingual)[a]	
	Coeff.	SE	Coeff.	SE
CLASS-T Engaged Support for Learning, average at ages 2-3	2.76*	1.19	4.05	3.21
Cohort	17.60+	9.06	70.44***	14.98
Black	-0.16	2.64	N/A	N/A
Hispanic	-2.57	3.66	N/A	N/A
Other	10.51*	4.09	N/A	N/A
Male	-3.27	2.15	-3.87	4.67
Age in months, Spring 2009	-1.29+	0.69	-6.12***	1.11
DLL	0.47	2.94	N/A	N/A
HV only	N/A	N/A	N/A	N/A
Other service type	N/A	N/A	N/A	N/A
Low birth weight	2.65	4.83	21.15	12.50
Child health level	2.96	2.94	1.90	4.64
ASQ-Communication	-0.15	0.10	-0.22+	0.12
ASQ-Problem Solving	0.06	0.11	-0.18	0.23
ASQ-Personal Social	0.06	0.13	0.14	0.30
Parenting Stress Index	-0.14	0.24	0.47	0.85
High involvement at age 1	-0.02	2.38	-6.06	5.24
Maternal depressive symptoms	0.03	0.22	-1.08	0.85
Maternal risk at age 1	-1.14	0.80	-6.05*	2.38
Multiple-approach program	3.81	2.98	-2.74	9.16
Urban/rural status	0.24	2.71	-15.48	16.30
Total enrollment	-0.01	0.02	0.02	0.04
Percentage of families speaking Spanish	-13.33	7.74	-0.70	12.60
Majority of families served by program have more than 3 demographic risks	1.40	2.86	-0.21	8.28
Majority of families served by program have mental health or substance abuse problems	9.78**	3.70	-15.85	16.73
Majority of families served by program reside in unsafe neighborhoods or experience family violence	-1.69	2.51	-12.85	9.98
Constant	88.03***	11.4	156.16***	39.67
Sample size	**244**		**99**	

Source: Parent Interview, Staff-Child Report, Direct Child Assessment, Home Visit Observation, Classroom Observation, Survey Management System.

Note: Results calculated using multiply imputed data. Weighted results reported. Sample limited to children in the center-based option who did not change service type. Dummies for Head Start region are included in the regressions but their coefficients are not reported.

[a]Models with Spanish or bilingual assessments as the dependent variable do not contain race/ethnicity or DLL indicators because of low variability in these indicators when these assessments are considered.

ASQ = Ages and Stages Questionnaire
CLASS-T = Classroom Assessment Scoring System-Toddler Version
DLL = Dual language learner
HV = Home visiting
PLS = Preschool Language Scale
PPVT = Peabody Picture Vocabulary Test

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

Table E.IX.7. Full Multivariate Linear Regression Results, CLASS-T Engaged Support for Learning Threshold of 3

	PPVT-4		PLS-4 (Bilingual)[a]	
	Coeff.	SE	Coeff.	SE
CLASS-T Engaged Support for Learning, average of 3 or more at ages 2-3	4.00	2.43	7.28	8.35
Cohort	17.09+	9.12	68.56***	16.11
Black	-0.66	2.58	N/A	N/A
Hispanic	-3.24	3.65	N/A	N/A
Other	11.55**	4.04	N/A	N/A
Male	-3.13	2.17	-5.03	4.31
Age in months, Spring 2009	-1.23+	0.69	-6.07***	1.15
DLL	0.02	3.02	N/A	N/A
HV only	N/A	N/A	N/A	N/A
Other service type	N/A	N/A	N/A	N/A
Low birth weight	2.87	4.96	27.29+	16.15
Child health level	3.25	2.95	1.57	5.12
ASQ-Communication	-0.17+	0.10	-0.26	0.17
ASQ-Problem Solving	0.07	0.11	-0.15	0.23
ASQ-Personal Social	0.05	0.14	0.13	0.28
Parenting Stress Index	-0.17	0.24	0.24	0.89
High involvement at age 1	0.24	2.41	-6.64	5.87
Maternal depressive symptoms	0.04	0.22	-0.88	0.95
Maternal risk at age 1	-1.24	0.93	-6.62*	2.66
Multiple-approach program	3.22	3.10	-5.00	9.27
Urban/rural status	0.77	2.84	-18.68	17.09
Total enrollment	-0.01	0.02	0.02	0.05
Percentage of families speaking Spanish	-13.82+	7.63	-1.04	14.01
Majority of families served by program have more than 3 demographic risks	2.03	3.00	1.67	8.51
Majority of families served by program have mental health or substance abuse problems	9.51*	3.77	-13.36	21.91
Majority of families served by program reside in unsafe neighborhoods or experience family violence	-1.73	2.62	-12.96	11.35
Constant	96.95***	10.57	166.77***	37.43
Sample size	**244**		**99**	

Source: Parent Interview, Staff-Child Report, Direct Child Assessment, Home Visit Observation, Classroom Observation, Survey Management System.

Note: Results calculated using multiply imputed data. Weighted results reported. Sample limited to children in the center-based option who did not change service type. Dummies for Head Start region are included in the regressions but their coefficients are not reported.

[a]Models with Spanish or bilingual assessments as the dependent variable do not contain race/ethnicity or DLL indicators because of low variability in these indicators when these assessments are considered.

ASQ = Ages and Stages Questionnaire
CLASS-T = Classroom Assessment Scoring System-Toddler Version
DLL = Dual language learner
HV = Home visiting
PLS = Preschool Language Scale
PPVT = Peabody Picture Vocabulary Test

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

Table E.IX.8. Full Multivariate Linear Regression Results, Total Center Days Attended

	PPVT-4		PLS-4		Parent Support for Child Learning Index	
	Coeff.	SE	Coeff.	SE	Coeff.	SE
Total center days attended	0.01	0.01	0.04*	0.01	0.00+	0.00
Cohort	16.70+	9.56	26.77+	14.08	-0.18	0.30
Black	-1.43	3.18	-3.35	3.22	-0.01	0.12
Hispanic	-3.70	4.12	0.26	5.59	-0.14	0.14
Other	13.26**	4.46	16.59**	5.56	0.08	0.12
Male	-2.60	2.05	-2.71	2.85	0.12	0.08
Age in months, Spring 2009	-0.77	0.73	-1.08	0.94	0.03	0.02
DLL	0.24	3.31	-5.15	3.50	0.11	0.08
HV only	N/A	N/A	N/A	N/A	N/A	N/A
Other service type	N/A	N/A	N/A	N/A	N/A	N/A
Low birth weight	5.25	4.89	-4.79	4.68	0.23**	0.07
Child health level	3.25	3.00	1.88	3.41	0.00	0.07
ASQ-Communication	-0.17+	0.10	0.03	0.13	0.01+	0.00
ASQ-Problem Solving	0.08	0.11	0.19+	0.10	0.00	0.00
ASQ-Personal Social	0.06	0.14	0.03	0.13	0.01+	0.00
Parenting Stress Index	-0.16	0.23	0.17	0.34	0.00	0.01
High involvement at age 1	-0.22	2.42	-1.09	2.95	0.04	0.07
Maternal depressive symptoms	-0.02	0.25	0.08	0.27	-0.01	0.01
Maternal risk at age 1	-1.51+	0.80	-2.57*	1.07	-0.06*	0.02
Multiple-approach program	1.33	3.30	-0.76	2.96	-0.23**	0.08
Urban/rural status	2.80	3.22	1.72	3.65	0.26***	0.07
Total enrollment	-0.02	0.02	0.03	0.02	0.00+	0.00
Percentage of families speaking Spanish	-16.31+	8.53	-4.33	10.26	0.21	0.20
Majority of families served by program have more than 3 demographic risks	0.71	3.54	1.11	4.39	-0.06	0.12
Majority of families served by program have mental health or substance abuse problems	12.43**	4.72	8.96	5.75	0.22+	0.12
Majority of families served by program reside in unsafe neighborhoods or experience family violence	-1.97	2.88	2.65	3.36	0.03	0.07
Constant	88.58***	13.47	65.24***	13.34	-0.97*	0.40
Sample size	**219**		**190**		**297**	

Source: Parent Interview, Staff-Child Report, Direct Child Assessment, Home Visit Observation, Survey Management System, Family Services Tracking System.

Note: Results calculated using multiply imputed data. Weighted results reported. Sample limited to children in the center-based option who did not change service type.

ASQ = Ages and Stages Questionnaire
DLL = Dual language learner
HV = Home visiting
PLS = Preschool Language Scale
PPVT = Peabody Picture Vocabulary Test

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

Table E.IX.9. Full Multivariate Linear Regression Results, Center Attendance Threshold (Attended at Least 50 Percent of Recommended Days)

	PPVT-4		PLS-4		Parent Support for Child Learning Index	
	Coeff.	SE	Coeff.	SE	Coeff.	SE
Attended at least 50% of recommended center days	5.85+	3.27	9.07*	3.54	0.18+	0.10
Cohort	16.32+	9.71	26.80+	14.11	-0.16	0.31
Black	-0.79	3.24	-4.06	3.20	-0.01	0.11
Hispanic	-3.31	4.03	-1.52	5.37	-0.17	0.13
Other	13.34**	4.58	16.73**	5.50	0.08	0.12
Male	-2.68	2.05	-2.55	2.89	0.12	0.08
Age in months, Spring 2009	-0.79	0.72	-1.16	0.92	0.03	0.02
DLL	0.09	3.17	-5.85	3.56	0.11	0.08
HV only	N/A	N/A	N/A	N/A	N/A	N/A
Other service type	N/A	N/A	N/A	N/A	N/A	N/A
Low birth weight	4.48	4.87	-6.00	4.70	0.20**	0.07
Child health level	2.66	2.93	0.68	3.59	-0.01	0.07
ASQ-Communication	-0.17+	0.10	0.05	0.13	0.01	0.00
ASQ-Problem Solving	0.08	0.10	0.20+	0.10	0.00	0.00
ASQ-Personal Social	0.06	0.14	0.03	0.14	0.01	0.00
Parenting Stress Index	-0.21	0.23	0.07	0.34	0.00	0.01
High involvement at age 1	-0.58	2.33	-0.80	2.88	0.05	0.07
Maternal depressive symptoms	-0.03+	0.24	0.03	0.26	-0.01	0.01
Maternal risk at age 1	-1.45	0.79	-2.30*	1.06	-0.06*	0.02
Multiple-approach program	1.64	3.32	-0.89	3.12	-0.23**	0.08
Urban/rural status	2.94	3.06	1.08	3.70	0.23**	0.08
Total enrollment	-0.02	0.02	0.03	0.02	0.00+	0.00
Percentage of families speaking Spanish	-14.72+	8.35	-1.55	10.62	0.28	0.22
Majority of families served by program have more than 3 demographic risks	0.24	3.49	1.39	4.40	-0.06	0.12
Majority of families served by program have mental health or substance abuse problems	11.72*	4.75	6.85	5.88	0.20+	0.11
Majority of families served by program reside in unsafe neighborhoods or experience family violence	-1.59	2.84	2.75	3.33	0.03	0.07
Constant	89.98***	12.31	75.61***	12.79	-0.78*	0.36
Sample size	**219**		**190**		**297**	

Source: Parent Interview, Staff-Child Report, Direct Child Assessment, Home Visit Observation, Survey Management System, Family Services Tracking System.

Note: Results calculated using multiply imputed data. Weighted results reported. Sample limited to children in the center-based option who did not change service type.

ASQ = Ages and Stages Questionnaire
DLL = Dual language learner
HV = Home visiting
PLS = Preschool Language Scale
PPVT = Peabody Picture Vocabulary Test

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

Implementation Rating Scales

I. Child Development Cornerstone

Item	Rating			
	Low-Minimal (1)	**Moderate (2)**	**Full (3)**	**Enhanced (4)**
1. Frequency of child development services[a]				
S3c. Number of days children in center-based option typically attend the center	Less than 3x/week	3x/week	4x/week	More than 4x/week
S3d. Number of home visits offered to children in the home-based option	Less than biweekly or no information	Biweekly	Weekly	--
IC1. Percentage of home-based families that receives home visits 4 times/month or more	0-50%	51-74% or don't know	75-89%	90-100%
HV-A3. On average, how often home visitors report that families typically attend home visits as scheduled.	Rarely or never	Sometimes	Often	--
IC4. Percentage of center-based families for whom program has attendance concerns	30% or more	15-29% or don't know	5-14%	Less than 5%
IC5. Activities to encourage attendance for cases that have attendance issues	None or don't know	Send letter or call	Home visit	Multiple strategies
2. Developmental assessments				
IC6. Percentage of children that receives developmental screenings	0-50%	51-74%	75-89%	90-100%
IC7. Percentage of children that receives periodic developmental assessments	0-50%	51-74%	75-89%	90-100%
IC8. Results of periodic developmental assessments used to plan services for each child	--	No or missing	Yes	--
IC8A. Who uses results of periodic developmental assessments	Only primary caregiver/home visitor	Teacher/home visitor and frontline supervisors/Part C staff or child dev coordinator	Teacher/home visitor, frontline supervisors/ Part C staff, and child development coordinator	AND other program managers
3. Health services				
IC9. Percentage of families with a medical home	Less than 90% or don't know	90-99%	100%	--
IC10. Program has a formal system for following up on health-related referrals	--	No	Yes	--
SAQ8. Program helps families access health services	1 or 2 services	Many services	All services	All services
G4d. Program tracks information in children's health/immunization status	--	No	Yes	--
Q8c. Program provides prenatal care or OB/GYN services directly or by referral	--	No	Yes	--

Item	Rating			
	Low-Minimal (1)	**Moderate (2)**	**Full (3)**	**Enhanced (4)**
4. Child care				
IC11. Program provides child care services to all families who need them	No	For some families, provides care directly or through community partners (a, b) and/or provides referrals (c) and/or helps families apply for subsidies (d)	For all families, provides care directly or through community partners (a, b) and/or provides referrals (c) and/or helps families apply for subsidies (d)	--
IC11. If program has child care partners, ensures the quality of child care provided directly or through community partners	None	Assesses quality for some or all (e)	AND conducts ongoing monitoring for some or all (f)	AND offers training and support to improve quality for some or all (g)
5. Parent participation in child development services planning				
IC12. Percentage of families with at least one parent involved in planning child development services	Less than 50%	50-74% or don't know	75-89%	90-100%
T-A2. Average teacher-reported percentage of children with parents who participate in EHS as classroom volunteers	0-25%	26-75%	76-100%	--
IC13. Of families with father or father-figure, percentage of fathers that participate in child development services	Less than 50%	50-74% or don't know	75-89%	90-100%
6. Individualization				
IC14. Factors taken into account when placing children in classrooms or assigning h.v.	None or don't know	Family factors (a-e)	Child factors (f-j)	Child and family factors
IC15. Factors taken into account when planning curriculum and services	None or don't know	Family factors (a-e)	Child factors (f-j)	Child and family factors
IC16. Percentage of children with suspected disability referred to Part C	Less than 50%	50-74% or don't know or missing	75-89%	90-100%
7. Group socializations				
IC17. Frequency of group socializations for home-based families	Less than 1x/month	1x/month	2x/month	More than 2x/month
IC18. Percentage of families that regularly participates in group socializations	Less than 25%	25-39% or don't know	40-69%	70-100%
IC19. Frequency of parent meetings or ed. activities not including group socializations	Less than 1x/month	1x/month	2x/month	More than 2x/month
T-A2. Average teacher-reported percentage of children with parents who participate by attending parent education or group activities	0-25%	26-75%	76-100%	--
HV-A2. Average home visitor-reported percentage of children with parents who participate by attending parent education or	0-25%	26-75%	76-100%	--

Item	Rating			
	Low-Minimal (1)	Moderate (2)	Full (3)	Enhanced (4)
group activities				
IC20. Percentage of families that regularly participates in meetings or education activities	Less than 25%	25-39% or don't know	40-69%	70-100%

II. Family Development Cornerstone

Item	Rating			
	Low-Minimal (1-2)	Moderate (3)	Full (4)	Enhanced (5)
1. Individualized Family Partnership Agreements (IFPAs)				
D2. Percentage of families with an IFPA	Less than 75%	76-94% or don't know	95-99%	100%
IF2. Whether staff collaborate with other agencies/programs in developing IFPAs	--	--	No or don't know	Yes
D3a/b. Number of times/year IFPAs reviewed for families	Less than 1x/year or not applicable (no policy)	1x/year or don't know or missing	2x/year	More than 2x/year
2. Availability of services				
IF3. Activities to follow-up ensure service receipt and monitor quality	__	None	Discuss services with families	AND discuss with service providers
3. Frequency of family development services				
IF1. Percentage of families who have met with a family service worker in past month	Less than 50% or not applicable	50-74% or don't know or refused	75-89%	90-100%
4 Parent involvement in program activities				
IF4. Activities to encourage father involvement	No activities or not applicable	Accommodations to facilitate involvement (a-c) or don't know or missing	AND encourage father participation (g-h)	AND have male staff or father involvement coordinator (d-e)
IF5. Whether any staff members are current/former EHS parents	__	No or don't know or missing	Yes	__

III. Staff Development Cornerstone

Item	Rating			
	Low-Minimal (1-2)	**Moderate (3)**	**Full (4)**	**Enhanced (5)**
1. Supervision				
IS1. Types of supervision activities	None listed or don't know	Regular ongoing individual & group supervision with performance feedback	AND Supervisors observe staff regularly	AND Training on reflective supervision to all supervisors
IS2. Frequency of supervision meetings with individual staff members	Annually or never	2-4 times a year or don't know	5-11 times a year	Once a month or more
T-B4. Average teacher-reported frequency of supervision meetings	Annually or never	Once every 4-6 months	Once every 1-3 months	Once a month or more
HV-B5. Average home visitor-reported frequency of supervision meetings	Annually or never	Once every 4-6 months	Once every 1-3 months	Once a month or more
2. Training				
IS3. Program develops a training plan each year	--	No	Yes	--
IS4. Program solicits info on staff development needs from staff and supervisors	--	No	Yes	--
T-B1. Percentage of teachers who have an individual career or professional development plan	Less than 50%	50-74%	75-89%	90-100%
HV-B2. Percentage of home visitors who have an individual career or professional development plan	Less than 50%	50-74%	75-89%	90-100%
IS5. Percentage of front line staff who have attended at least 3 trainings in past year	Less than 50%	50-74%	75-89%	90-100%
IS6. Program offers trainings specifically for new staff members	--	No or don't know or refused or missing	Yes	--
3. Turnover				
Staff turnover rate (based on E7/E8/E9)	30 percent or more	20-29% or information not available	10-19%	Less than 10%
4. Compensation and morale				
F4. Staff salary levels	Below average	Below average	Average or don't know	Above average
T-B13. Percentage of teachers reporting receipt of paid health insurance, retirement/pension plan, or paid vacation, holidays or sick leave	Less than 50%	50-74%	75-89%	90-100%
HV-B13. Percentage of home visitors reporting receipt of paid health insurance, retirement/pension plan, or paid vacation, holidays or sick leave	Less than 50%	50-74%	75-89%	90-100%
F5a. Whether program has high morale	Strongly disagree	Disagree	Agree	Strongly agree

IV. Community Building Cornerstone

Item	Rating			
	Low-Minimal (1-2)	Moderate (3)	Full (4)	Enhanced (5)
1. Collaborative Relationships[b]				
ICB1. Program participates in a coordinating group of service providers	--	No	Yes	--
ICB2. Whether program has formal written partnerships with Part C provider	--	No or don't know	Yes	--
C3-C3a. Number of formal written partnerships with child care providers	--	0	1	More than 1
C4. Whether program has regular contacts will all child care partners	--	No	Yes	--
C4c. Frequency of contact with child care partners	Annually	Every few months	Monthly	More than once a month
ICB4. Program has formal written partnership with child protective services agency	--	No or don't know	Yes	--
2. Transition Plans				
ICB5. Whether program has established procedure for facilitating transitions	--	No	Yes	--
ICB6. Percentage of 2.5-year-old children with transition plan in place	Less than 50%	50-74% or don't know	75-94%	95-100%
ICB7. Percentage of parents involved in creating transition plan	Less than 50%	50-74% or refused	75-89%	90-100%
3. Advisory Committees				
ICB8. Whether program has health advisory committee that meets regularly	--	No or missing	Yes	--
ICB9. Frequency of health advisory committee meetings	--	None	Once a year	More than once a year
ICB10. Whether program has other advisory committees that meet 2x/year or more	--	No	Yes	--

V. Management Systems

Item	Rating			
	Low-Minimal (1-2)	**Moderate (3)**	**Full (4)**	**Enhanced (5)**
Policy Council				
IM1. Program has policy council that includes parents	__	No	Yes	--
IM2. Frequency of policy council meetings	1x/year or less	2-3x/year	4-6x/year	More than 6x/year
Communication Systems				
IM3/IM4. Program has regular system of communications	Among program staff	AND between staff and parents	AND with grantee agency AND policy council	AND system of communication is two-way (email or meetings) (IM4)
Goals, Objectives, and Plans				
IM5. Program has set of written goals, objectives, and plans	--	No	Yes	--
IM6. Frequency that goals/objectives/plans updated	Never	Less than 1x/year	1x/year	More than 1x/year
IM7. Who is involved in developing goals/objectives/plans	Director only or others only	Director/Managers/Staff	AND Policy Council or parents	AND advisory committees or community members
Self-Assessment				
IM9. Program has conducted written self assessment in past year	--	No	Yes	--
IM10. Who participated in the self-assessment	None	Director/Managers/Staff	AND Board of Directors or Policy Council or parents	AND advisory committees or community members
IM11. Program has made changes as result of self assessment	--	No	Yes	--
Community Needs Assessment				
IM13. Program has conducted and documented community needs assessment	--	No	Yes	--
IM14. When most recent community needs assessment conducted	Never	More than two years ago	1-2 years ago	In past year
IM15. Who participated in community needs assessment	None	Director/Managers/Staff	AND Board of Directors or Policy Council or parents	AND advisory committees or community members

This page is left blank for double-sided printing.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ WASHINGTON, DC

MATHEMATICA
Policy Research