Learning As We Go: A First Snapshot of Early Head Start Programs, Staff, Families, and Children

February, 2011

OPRE 2011-7

Baby FACES 2009 First Report: Technical Appendices

Learning As We Go: A First Snapshot of Early Head Start Programs, Staff, Families, and Children

Volume II: Technical Appendices

Submitted to:

Rachel Chazan Cohen
Office of Planning, Research, and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Cheri A. Vogel
Kimberly Boller
Yange Xue
Randall Blair
Nikki Aikens
Andrew Burwick
Yevgeny Shrago
Barbara Lepidus Carlson
Laura Kalb
Linda Mendenko
Judith Cannon
Sean Harrington
Jillian Stein
Mathematica Policy Research

Project Director:

Cheri A. Vogel, Mathematica Policy Research

Contract Number: HHSP23320072914YC
Mathematica Reference Number: 6432.150

Suggested citation:

Administration for Children and Families. Learning As We Go: A First Snapshot of Early Head Start Programs, Staff, Families, and Children, Technical Appendices. C.A. Vogel, K. Boller, Y. Xue, R. Blair, N. Aikens, A. Burwick, Y. Shrago, B.L. Carlton, L. Kalb, L. Mendenko, J. Cannon, S. Harrington, J. Stein, Washington, DC. U.S. Department of Health and Human Services. 2011.

OPRE
Office of Planning, Research and Evaluation
Administration for Children and Families

MATHEMATICA
Policy Research, Inc.

CONTENTS

APPENDIX A

SAMPLE

We Selected a Nationally Representative Sample of Programs

We designed the Baby FACES sample to be representative of the population of Early Head Start programs at the national level. Within programs, the families being served by Early Head Start represent the population of parents of newborn and 1-year-old children enrolled in Early Head Start in spring 2009. To achieve the goal of an efficient, representative national sample of sufficient size to detect developmental or programmatically meaningful differences over time and within key subgroups, we use a stratified clustered sample design. We selected all children receiving center- and/or home-based services from a probability sample of Early Head Start programs (including those receiving services through partnership arrangements). Children whose date of birth (or due date, for expectant mothers) fell within the study-defined windows were selected for the Baby FACES sample.[1] We oversampled larger programs to yield more children and families.

The sampling frame we used to select programs was the most recent Head Start Program Information Report (PIR) data, which at the time of sampling covered program year 2006-2007. All Head Start and Early Head Start programs must submit PIR data annually. For this reason, the PIR is a reliable source of information about the programs and the type of families they serve (for example, PIR data include the number of children served and the demographic characteristics of families in each program). The sampling unit was the PIR reporting level—that is, the grantee or delegate agency (or "program"). According to the 2006-2007 PIR, there were 734 Early Head Start programs nationwide; 640 of these programs met the study eligibility criteria (Table A.1). We excluded from the study's sampling frame those Early Head Start programs that:

- Were administered by Region XI (American Indian/Alaska Native programs) and Region XII (Migrant/Seasonal Worker programs)[2]
- Were under the management of the national interim grantee contractor
- Provided family child care services only
- Did not directly provide services to children[3]
- Provided services to fewer than 25 enrolled families
- Were located in Alaska, Hawaii, Puerto Rico, or U.S. territories

We excluded programs with fewer than 25 enrolled families because by the time we would have narrowed the families down by date of birth (or due date) to determine study eligibility, we expected the program to have too few study families to be logistically sustainable for the duration of the study. We excluded programs in Alaska and Hawaii (which have been included in various national Head Start studies) because there are only approximately four Early Head Start programs in Hawaii and Alaska (that are in Regions IX and X), and the costs to conduct the study there are prohibitive.

[1] Note that centers, classes, teachers, and home visitors became part of the sample only if they provided services to a study-eligible child.

[2] Any Early Head Start programs serving those populations but not under the administration of those two regional offices were eligible for this study.

[3] Some programs provide only administrative services to their delegate agencies and are excluded here.

While there are about 16 Early Head Start programs in Puerto Rico, the programs and populations in Puerto Rico differ greatly from those in the states, and may need to be omitted from some analyses. Programs serving pregnant women, infants, and toddlers funded by the Migrant and Seasonal Head Start branch and programs funded by the American Indian/Alaska Native branch were not included because these programs often operate on a different schedule than other programs, which increases the data collection costs. In addition, the Office of Head Start and the Office of Planning, Research and Evaluation are funding other research activities to address the unique needs of these programs. There are few programs that offer family child care as their only service approach and the number of families served is small enough to preclude comparisons to other service options. We did not exclude any programs that were involved in other studies of Head Start or Early Head Start.

In fall 2008, we selected programs for the study. From experience conducting similar studies, we anticipated that some programs would refuse to participate or be found ineligible after sampling. To address this, we initially selected twice the number of programs needed per stratum. We then formed sequential pairs of selected programs, as sorted by explicit and implicit stratification variables (described in next section), so that adjacent programs were likely to be similar. One program within each pair was randomly selected to be the main release, and the other was available as a replacement for the released program, should one have been needed. In the case of using replacement programs, both programs would be treated as "released" into the sample for purposes of calculating weights and response rates. This method provides an uncomplicated way of replacing a nonparticipating program with a similar program. It also virtually ensured meeting the target number of 90 participating programs while enabling us to quantify the probability of selection.

After the programs were sampled, we verified that programs were in good standing (a condition of eligibility) with the Office of Head Start and the ACF regional offices, explained the study to programs, and recruited program personnel for participation in the study. We also requested that programs provide a list of Early Head Start centers, home visitors, classrooms, and children served by each home visitor and in each classroom (children receiving both home visits and classroom services were randomly assigned to either home visits or classroom services for the observation, teacher/home visitor interviews, and Staff Child Rating [SCR] report). When a sampled program was projected to have too few eligible families during our study period to make a data collection visit worth the costs,[4] we excluded the program from the study as ineligible and released its sampled pair. We further excluded any programs that served a transient population or programs that served those who were unlikely to be enrolled in the program for more than 18 months (for example, those that served teenage mothers and were based in a high school). Our consent rate for programs was 93.7 percent. Our sample closely resembles the sample frame on the characteristics included in the design (see Table A.1).

[4] The eligibility threshold we set was that a program had to have a minimum of five children in the specified Newborn or 1-year-old Cohort age range in at least one of the data collection weeks to be included in the sample. We projected the number of children who would be eligible by computing how old each of the children on the eligibility roster would be during each of the 14 data collection weeks. These eligibility criteria required us to substitute 22 of the originally selected programs (16 due to eligibility; 6 refused to participate in the study). If a program and its match were both ineligible, we selected a "wildcard" program from among unused program matches (this occurred six times).

About two weeks before visiting each program in spring 2009, we asked the programs for an updated list of all enrolled families that included children's birth dates and the name of their main service provider. At this time, we also requested the address of centers for those in center-based service options. The lists also included pregnant women and their expected due dates.

From these dates of birth, we identified the children and pregnant women whose birth dates or due dates qualified them to be in one of the two cohorts for this study. We used the date of the program visit to calculate age (including gestational age for children who were not yet born). Because we took all children in selected programs whose birth dates fell within the specified window, the resulting census of children in each cohort represents families and children of these ages enrolled in Early Head Start in spring 2009.

We also asked the center directors to identify any siblings (twins and otherwise) and soon-to-be siblings (in which a child was in the Early Head Start program and his/her pregnant mother was also receiving Early Head Start services). To minimize burden on these families, we randomly selected only one eligible child from each family to participate in Baby FACES.

Stratification Maximized the Number of Children Eligible for the Study and Ensured Representation of Important Subgroups

One challenge we faced during the sampling process was the need to balance statistical issues related to sampling a large number of programs (and including a large number of children) with the per-program cost of sending data collection staff to many programs.[5] In addition, ACF sought to draw a sample that included a large enough number of children with important characteristics to support subgroup analyses. To address these needs and maximize the number of children per program included in the sample, we oversampled larger programs by stratifying the frame based on annual enrollment as reported by programs in the PIR. We also used an optimal allocation across strata, which balanced variance and unit cost.[6] We obtained information on the stratification variables—size, service approach, and location of each Early Head Start program—from the most recent PIR. Table A.2 shows the optimal sample allocation after we divided all Early Head Start programs into four strata according to the estimated number of study-eligible families (size), with stratum 1 containing the smallest programs and stratum 4 containing the largest programs. In this design, we sampled programs with equal probability within each stratum. Table A.3 illustrates the expected child-level sample at baseline using this design, based on the age distribution of children within program strata from the PIR.

Children with limited English proficiency are a high-interest population among policymakers; services for these children are emphasized in the recent Head Start Reauthorization Act. We wanted to ensure that we did not select a sample with too few dual-language learners (DLLs) to analyze

[5] Attrition from the program was another concern. If a program had only a small number of eligible children, the program would be more likely to drop out of the study early if those children left the program before age 3.

[6] To do an optimal sample allocation across strata, one must make assumptions about the cost and variance. For these calculations, we assumed that the cost per program (including recruiting the program and traveling to the program) was 50 times the marginal cost per child (to do each child's assessments and interviews), and used this to calculate the overall cost per child as the program cost divided by the number of children per program plus the marginal cost per child. We also assumed an intraclass correction factor of .05.

them as a separate subgroup.[7] This examination of the expected size of the DLL sample from the PIR concluded that, by substratifying on whether the majority of families in programs were Spanish-speaking, we would have a sufficient sample size for a subgroup analysis. As a result, we did not need to oversample majority-DLL programs. This was a positive finding, as oversampling would have adversely affected the precision of estimates for the full sample. Table A.4 shows the expected program- and child-level sample sizes under the DLL-stratified design.

[7] For this design, DLL children do not include children whose primary language is anything other than Spanish.

Table A1. Comparison of Study-Eligible Programs to Those in Final Sample

		Eligible Programs on PIR n = 640	Participating Sampled Programs n = 89
Number of States (plus DC) Represented		49	38
State (Percent)	California	10.16	12.36
	Texas	6.25	5.62
	New York	6.09	5.62
	Other	77.5	76.4
Census Region (Percent)	Northeast	18.28	14.61
	Midwest	27.19	25.84
	South	33.44	32.58
	West	21.09	26.97
Metro (Percent)[a]		67.66	66.29
ACF Regional Office (Percent)[a]	Region 04	17.03	15.73
	Region 05	17.34	14.61
	Other Eligible Regions	65.63	69.66[c]
Service Type (Percent)[a]	Center	23.75	17.98
	Home	15.16	12.36
	Mixed	61.09	69.66
Grantee/Delegate (Percent)	Grantee with no delegates	86.56	89.89
	Grantee with delegates	1.88	2.25
	Delegate agency	11.56	7.87
Percent with Majority Spanish Speaking[b]		11.72	11.24
Mean Proportion of Enrollees Spanish-Speaking		0.18	0.17
Mean Proportion of Enrollees Hispanic		0.28	0.28
Mean Proportion of Enrollees Black		0.28	0.22
Mean Proportion of Enrollees White		0.46	0.50
Mean Proportion of Enrollees Pregnant		0.11	0.11
Size Variables – Oversampled Larger Programs			
Percent by Size Stratum[b] (1 = smallest, 4 = largest)	1	24.7	10.1
	2	25.3	19.1
	3	25.0	24.7
	4	25.0	46.1
Mean Actual Enrollment		138.29	187.99
Mean Enrolled Children		123.3	167.1
Mean Center-Based Enrollment		43.8	58.4
Mean Home-Based Enrollment		36.8	53.9
Mean Estimated Number of Newborn Cohort		10.5	14.5
Mean Estimated Number of Age 1 Cohort		8.2	11.0
Mean Estimated Number of Study Eligible		18.7	25.5

[a] Sample control variable (implicit stratification).

[b] Explicit stratification variable.

[c] All 10 regional offices represented in sample.

Table A.2. Optimal Allocation of Program - and Child - Level Samples

Program Stratum	Number of Programs	Mean Annual Enrollment[a]	Enrollment Adjusted for Eligibility Window[b]	Optimal Allocation of Children[c]	Number of Programs to Sample
1 – Smallest	174	50	7.2	93.1	13
2	168	90	13.4	240.6	18
3	169	135	20.1	461.8	23
4 – Largest	169	261	39.3	1,413.0	36
Combined	680	133	19.9	2,208.5	90
Effective Sample Size				**936**	

[a] To ensure that programs of different sizes would be represented in the sample, we looked at the distribution of enrollment, and used the quartile values to divide the programs into four strata with about the same number of programs in each.

[b] We assumed a deflation factor of .66 to get from annual to point-in-time enrollment, and a deflation factor of .33 (four-month birthday window) to get from all children to those falling within the four-month eligibility windows.

[c] Optimal sample allocation balances variance and unit cost when allocating the sample across strata.

Table A.3. Expected Sample Sizes in Spring 2009

		Spring 2009—90 Programs	
Cohort	Data Collection Respondent	Within Age Range (Selected)	Eligible/With Consent/Responding (90 Percent)
Newborn	Parent	1,262	1,136
	Child[a]	-	-
1-year-old	Parent	946	851
	Child	946	851
Both cohorts combined	Parent	2,208	1,987
	Child	946	851

Note: Ultimately, 89 programs participated in the study.

[a] No child assessments will be conducted in the Newborn Cohort at baseline.

Table A.4. Sample Design Stratifying by Program Size and DLL Status

Stratum	Program Size (1 = Smallest, 4 = Largest)	Majority DLL	Sampled Programs	Study-Eligible Children in Sampled Programs	Study-Eligible DLL Children in Sampled Programs
1	1	No	11	81	6
2	1	Yes	2	11	9
3	2	No	15	200	19
4	2	Yes	3	40	28
5	3	No	21	421	44
6	3	Yes	2	41	27
7	4	No	32	1,223	180
8	4	Yes	4	192	127
Total	**n.a.**	**n.a.**	**90**	**2,209**	**441**
Effective Sample Size				**933**	**274**

Note: The expected sample size, based on PIR data, was 2,209 children.

n.a. = not applicable.

In addition to using an explicit stratification approach (based on the size of the study-eligible population and majority/minority DLL), we used implicit stratification when selecting programs. Implicit stratification is achieved by sorting the sampling frame by specified characteristics within explicit strata before sampling. This helps make the sample resemble the population in terms of these characteristics. The first sorting variable has the most influence and acts similarly to an explicit stratification variable. When selecting the program sample, we implicitly stratified by program service approach (center-based, home-based, or mixed)[8], then by urbanicity (metropolitan statistical area [MSA] versus non-MSA) and ACF regional office. We used a sequential sampling technique based on a procedure developed by Chromy (1979); this technique is available as a sampling option within SAS.[9] This procedure offers all the advantages of a systematic sampling approach but eliminates the risk of bias associated with that procedure. The Chromy procedure allows for explicit and implicit stratification, as described above.

We Defined Cohorts by Children's Age at Time of the Data Collection Visit

Within each selected program, we included all children and pregnant women who fell within our study eligibility windows in the study sample. We used the Monday of the site visit week identified for each program (which we will refer to as the "focus date" in this document) to determine the eligibility window for the Newborn Cohort and the 1-year-old Cohort. For the Newborn Cohort, we included pregnant women whose due dates were two months or less beyond

[8] The "mixed" category (home- and center-based) also includes programs with "combination" enrollment and those with "locally designed options."

[9] The procedure makes independent selections within each of the sampling intervals while controlling the selection opportunities for units crossing interval boundaries.

the focus date, and babies whose birth dates were within two months of the focus date.[10] The Newborn Cohort therefore included babies up to age 2 months at the time of the spring 2009 site visit, plus any pregnancies likely to result in a baby who will be between 10 and 14 months at the time of the spring 2010 site visit.[11]

The 1-year-old Cohort included children who were between 10 and 15 months of age at the time of the spring 2009 site visit—that is, the site visit was 2 months (or less) before the child's first birthday, or the child's first birthday occurred less than 3 months before the site visit. We allowed an extra month in this window to increase the sample size without adversely affecting appropriateness of the data collection instruments for 1-year-old children in terms of the assumed age range.

The Sample Design Accounts for Attrition and Inability to Locate Families

Both sample cohorts (Newborn and 1-year-old) will be followed each spring until the children are within the defined window of their third birthday. We projected that 15 percent of the children (and their parents) will leave the Early Head Start program each year before they reach age 3. Included in this projection are pregnant women whose pregnancies do not result in live births and those who give their babies up for adoption. We will make an additional telephone contact with the parents around the time the children in the 1-year-old Cohort turn 3½ to learn about transitions into other programs after Early Head Start. For those who have left the program since our last contact with them, we will conduct a brief telephone survey at the next scheduled data collection period. We will not collect other data from those who have left the program and will not continue to follow them in our sample. We estimate that, despite our best locating efforts, we will be unable to contact about 10 percent of the sample still in the program at each one-year interval (and 5 percent between ages 3 and 3½). Tables A.5 and A.6 show revised figures based on actual sample sizes, consent rates, and completion rates from the spring 2009 data collection.

[10] The final newborn sample comprised 174 children: 100 children born before the site visit, and 74 children born after the site visit.

[11] Note the following sample anomalies that were possible for children who were not born as of the time of the site visit (illustrated in chart below). Those whose due date was within the Newborn window were included in the sample, even if it turned out that they were born outside the window. Those whose due date was outside the Newborn window were excluded from the sample, even if it turned out that they were born within the window.

Due Date	Child Born Within Window After Site Visit	Child Born After Window
Outside of window	*Not in sample but within window*	Not in sample
Within window	In sample	*In sample but outside of window*

Table A.5. Original Expected Sample Sizes Throughout Data Collection

		Responding				
Cohort	Data Collection Respondent	Spring 2009	Spring 2010	Spring 2011	Spring 2012	Age 3½ (Fall 2011)
Newborn	Parent	1,136	1,023	782	598	-
	Child	-	869	665	509	-
1-year-old	Parent	851	766	586	-	473
	Child	851	651	498	-	-
All cohorts	Parent	1,987	1,789	1,368	598	473
	Child	851	1,520	1,163	509	-

Note: These estimates assume sequential sampling with optimal allocation. We assume that 15 percent of the eligible participating baseline families will leave the program each year. We also assume that we will be able to locate, contact, and obtain consent for 90 percent of families per wave that remain in the program. When combining all age cohorts, and after accounting for the impact of the sample design on the variance, the *effective* sample size at baseline (for the 1,987 parent interviews) was expected to be about 891.

Table A.6. Actual and Expected Sample Sizes Throughout Data Collection, Revised

		Responding				
Cohort	Data Collection Respondent	Actual Spring 2009	Spring 2010	Spring 2011	Spring 2012	Age 3½ (Fall 2011)
Newborn	Parent	175	158	121	92	-
	Child	-	148	114	87	-
1-year-old	Parent	719	647	495	-	435
	Child	-	599	458	-	-
All cohorts	Parent	894	805	616	92	435
	Child	-	747	572	87	-

Note: We assume that 15 percent of the eligible participating baseline families will leave the program each year. We also assume that we will be able to locate, contact, and obtain consent for 90 percent of families per wave who remain in the program. When combining all age cohorts, and after accounting for the impact of the sample design on the variance, the effective sample size at age 1 (for the 894 estimated parent interviews) is expected to be about 570.

Our estimates of the eligible sample in each age cohort at the design stage were larger than those we found in reality. Although the size of the sample we enrolled in the study is sufficient for meaningful subgroup comparisons (described in the next section), if higher-than-expected attrition or locatability issues occur in subsequent years, it will affect power to detect small or moderate differences (described later in this appendix). Because the actual sample is smaller than we expected, we used PIR data and program rosters to investigate and reached the following conclusions:

- Comparing 2007-2008 and 2006-2007 PIR data (the latter being the one used for sampling), the programs in our sample seem to be reporting fewer enrolled pregnant women and children age 1 and younger, a higher number of enrolled children ages 2 and older, and lower enrollments overall. For the population overall, we also saw a decrease in the number of pregnant women and an increase in the number of children ages 2 and older, suggesting a population-wide trend. We also overestimated the size of the Newborn Cohort because, in our sample design, we summed two PIR variables that include (1) all pregnant women and (2) infants who enroll over the course of the year. This will overestimate newborn enrollment if their mothers were also served by the programs during pregnancy.

- Further, because we collected rosters from programs over the fall 2008 recruiting period to determine program eligibility for the study, we were able to compare the number of children who would have been in our age windows had we conducted a fall rather than a spring data collection.[12] We then obtained updated rosters just before the spring 2009 visits (and field staff confirmed these while on site to ensure any newly enrolled eligible children could be included in the study). We found evidence of seasonality in births, with a greater number of children eligible for Baby FACES given a fall data collection, compared to one in the spring. That is, fewer children in our sample programs have spring birthdates compared to those with fall birthdates.

As described above, the study is designed to support statistical comparisons of child-level outcomes across subgroups of children/families and child-level characteristics based on subgroups of programs. We calculated the study's statistical power to detect differences at: (1) the child level, and (2) the program level. Next, we present the statistical power to detect these differences.

Power for Child-Level Analysis. The sample sizes described in Tables A.5 and A.6 should be large enough to detect developmentally meaningful differences, given various assumptions about the sample design and its impact on the variance of estimates. Table A.7 shows the half-confidence intervals, or margins of error,[13] for the child assessments. Tables A.8–A.10 show the minimum detectable differences (MDDs) and effect sizes (ES) with 80 percent power and various sample and subgroup sizes, and with different assumptions about the impact of weighting and clustering on the variance of estimates from the child assessments. We assume an intracluster correlation of .05 and, for the change over time estimates, an average correlation between measures at baseline and age 3 of 0.5. To yield the effective sample sizes in the table, we also adjust the nominal sample size for design effects due to clustering and unequal weighting according to the oversampling design described above.

[12] The spring data collection was necessitated by the receipt of OMB clearance in September 2008.

[13] The half-confidence interval (sometimes referred to as the margin of error) is the amount of variation above or below an estimate within which we are fairly certain the true value lies. In this case (where our level of certainty is 95 percent), it is 1.96 times the standard error, where the standard error has been adjusted for the design effect.

Table A.7. Half-Confidence Intervals (95 Percent)—Child-Level Data

				Half-Confidence Intervals	
Cohort	Time Period	Nominal Sample Size	Effective Sample Size (Accounting for Sample Design)	Proportional p = 0.50 Std. Dev. = 0.50	Normalized Variable Mean = 100 Std. Dev. = 15
Newborn	Age 1	134	84	.108	3.206
	Age 3	78	57	.131	3.904
1-year-old	Age 1	719	304	.056	1.685
	Age 3	421	339	.053	1.597

Note: Two-sided α = .05.

These values would be used for estimating confidence intervals around descriptive statistics.

[a] We show the most conservative situation here—an estimated proportion of 0.5 has the largest variance among all proportions. Proportions higher or lower than 0.5 will have a smaller variance and, therefore, a smaller margin of error than shown here. The same holds for Table A.10. For Tables A.7–A.9 and A.11–A.13, the smaller variance for other proportions will allow for the detection of smaller differences between subgroups. For Tables A.9 and A.14, the smaller variance for other proportions will allow for the detection of smaller changes over time.

Table A.8. Child-Level Data MDDs and ES Comparing Two Program-Defined Subgroups at a Point in Time

			Effective Sample Sizes		MDDs Between Subgroups		
Cohort			Subgroup 1	Subgroup 2	Proportion p = .50 Std. Dev. = 0.50	Normalized Variable Mean = 100 Std. Dev. = 15	ES
Newborn	Age 1	1/2, 1/2	42.1	42.1	.309	9.160	.61
		1/3, 2/3	28.0	56.1	.329	9.715	.65
	Age 3	1/2, 1/2	28.4	28.4	.379	11.155	.74
		1/3, 2/3	18.9	37.8	.403	11.832	.79
1-year-old	Age 1	1/2, 1/2	152.2	152.2	.161	4.815	.32
		1/3, 2/3	101.5	202.9	.171	5.107	.34
	Age 3	1/2, 1/2	169.4	169.4	.153	4.564	.30
		1/3, 2/3	112.9	225.8	.162	4.841	.32
Combined	Age 1	1/2, 1/2	284.7	284.7	.118	3.521	.24
		1/3, 2/3	189.8	379.5	.125	3.734	. 25
	Age 3	1/2, 1/2	192.3	192.3	.143	4.284	.29
		1/3, 2/3	128.2	256.3	.152	4.544	.30

Note: Two-sided α = .05. Power = .80.

An example would be comparing average child cognitive outcomes for children in center-based versus other program options (most closely represented by the 1/3, 2/3 rows).

Table A.9. Child-Level Data MDDs and ES Comparing Two Child-Defined Subgroups at a Point in Time

			Effective Sample Sizes		MDDs Between Subgroups		
Cohort			Subgroup 1	Subgroup 2	Proportion p = .50 Std. Dev. = 0.50	Normalized Variable Mean = 100 Std. Dev. = 15	ES
Newborn	Age 1	1/2, 1/2	43.1	43.1	.305	9.049	.60
		1/3, 2/3	29.0	57.0	.324	9.585	.64
	Age 3	1/2, 1/2	28.8	28.8	.375	11.065	.74
		1/3, 2/3	19.3	38.2	.400	11.726	.78
1-year-old	Age 1	1/2, 1/2	166.6	166.6	.154	4.601	.31
		1/3, 2/3	114.7	215.4	.162	4.855	.32
	Age 3	1/2, 1/2	187.4	187.4	.145	4.339	.29
		1/3, 2/3	129.5	241.3	.153	4.575	.30

Note: Two-sided α = .05. Power = .80.

An example would be comparing average child cognitive outcomes for children receiving higher-intensity services to those receiving lower-intensity services.

Table A.10. Child-Level MDDs and ES for Comparisons Over Time (Age 1 to Age 3)

	Effective Sample Sizes		MDDs Over Time		
Cohort	Time 1 (Age 1)	Time 2 (Age 3)	Proportion p = .50 Std. Dev. = 0.50	Normalized Variable Mean = 100 Std. Dev. = 15	ES
Newborn	84	57	.197	5.876	.39
1-year-old	304	339	.074	2.209	.15

Note: Two-sided α = .05. Power = .80. Assume correlation over time = 0.5.

Point-in-Time Comparisons by Cohort. As depicted in Table A.8, if we compared standardized assessment scores (mean of 100, standard deviation of 15) of 1-year-old Cohort children at age 3 for two approximately equal-sized program-defined subgroups (that is, each having about half the programs, 45 out of 89, and about half the total sample, or about 169 children), this design will allow us to detect a minimum difference of 5 points with 80 percent power (or an ES of .30). Table A.9 shows comparable minimum differences for subgroups defined at the child level, where all 89 programs are included. One would use Table A.8 to get a sense of what size differences in program-level variables (for example, home- versus center-based or average teacher education level) would need to be observed to be significant predictors of child-level assessment outcomes in a regression model. Table A.9 gives a sense of what size differences in child-level variables (for example, attendance rate) would need to be observed to be significant predictors of child-level assessment outcomes. Classroom-level predictors (for example, classroom quality or teacher qualifications) fall somewhere in between.[14]

Point-in-Time Comparisons Combining Cohorts. We can detect smaller differences in analyses that combine cohorts and examine outcomes at a given age (as small as 0.24 of a standard deviation; see the 50/50 subgroup at age 1 in Table A.8). Differences of one-quarter to one-third of a standard deviation were found on some outcomes in some subgroups in the Early Head Start Research and Evaluation Project (EHSREP), but these are considered to be large differences.

Change Over Time Comparisons. If we compare a child outcome measure such proportion of children screening positive on the Brief Infant Toddler Social Emotional Assessment (BITSEA) over time (age 1 to age 3) for children in the 1-year-old Cohort (that is, a sample of about 339 children at age 3), we would be able to detect a minimum difference of .074 points (an ES of 0.15; see Table A.10). This is more than enough power to detect developmentally meaningful change.

Table A.11. Quality Measures Half-Confidence Intervals (95 Percent)

				Half-Confidence Intervals	
Cohort	Time Period	Nominal Sample Size	Effective Sample Size (Accounting for Sample Design)	Proportional p = 0.50 Std. Dev. = 0.50	Normalized Variable Mean = 100 Std. Dev. = 15
Newborn	Age 1	67	43	.152	4.504
	Age 3	39	32	.176	5.205
1-year-old	Age 1	359	191	.071	2.125
	Age 3	211	183	.073	2.172

Note: Two-sided α = .05.

These values would be used for estimating confidence intervals around descriptive statistics.

[14] The MDDs and ES shown are for subgroups defined by categorical variables, and can be thought of as categorical predictive variables in a regression context. Continuous predictive variables, while not represented in these tables, are likely to have somewhat smaller detectable differences than categorical variables, assuming they have a linear relationship with the outcome variable.

Table A.12. Quality Measures ES Comparing Two Program-Defined Subgroups at a Point in Time

			Effective Sample Size		
Cohort	Time Period	Subgroups	Subgroup 1	Subgroup 2	ES
Newborn	Age 1	1/2, 1/2	21.3	21.3	.858
		1/3, 2/3	14.2	28.4	.910
	Age 3	1/2, 1/2	16.0	16.0	.991
		1/3, 2/3	10.6	21.3	1.052
1-year-old	Age 1	1/2, 1/2	95.7	95.7	.405
		1/3, 2/3	63.8	127.6	.429
	Age 3	1/2, 1/2	91.6	91.6	.414
		1/3, 2/3	61.1	122.1	.439
Combined	Age 1	1/2, 1/2	162.1	162.1	.311
		1/3, 2/3	108.0	216.1	.330
	Age 3	1/2, 1/2	105.2	105.2	.386
		1/3, 2/3	70.1	140.2	.410

Note: Two-sided α = .05. Power = .80.

An example would be comparing average program quality for children in programs with higher average staff education to those in programs with lower average staff education.

Table A.13. Quality Measures ES Comparing Two Child-Defined Subgroups at a Point in Time

			Effective Sample Size		
Cohort	Time Period	Subgroups	Subgroup 1	Subgroup 2	ES
Newborn	Age 1	1/2, 1/2	21.6	21.6	.852
		1/3, 2/3	14.5	28.7	.902
	Age 3	1/2, 1/2	16.1	16.1	.986
		1/3, 2/3	10.8	21.4	1.045
1-year-old	Age 1	1/2, 1/2	102.6	102.6	.391
		1/3, 2/3	70.1	133.6	.413
	Age 3	1/2, 1/2	97.9	97.9	.400
		1/3, 2/3	66.8	127.6	.423

Note: Two-sided α = .05. Power = .80.

An example would be comparing average program quality for children receiving higher-intensity services to those receiving lower-intensity services.

Table A.14. Quality Measures ES for Comparisons Over Time (Age 1 to Age 3)

	Effective Sample Size		
Cohort	Time 1 (Age 1)	Time 2 (Age 3)	ES
Newborn	43	32	.519
1-year-old	191	183	.209

Note: Two-sided α = .05. Power = .80.
Assume correlation over time = 0.5.

Quality Measures. For estimates and analyses based on classroom and home visit quality measures (Tables A.11–A.14), we assume that about one-half of the children would be receiving each type of service, and about 80 percent of the programs (72 of 89) would be providing each type of service.[15] In Table A.12, we see that, for the 1-year-old Cohort, we can detect a quality measure effect size at the child level between two equal-sized program-defined subgroups at age 3 of .41.[16] An example of this type of analysis is comparing the quality of home visits of children in urban versus rural programs. Table A.13 shows similar effect sizes when subgroups are defined at the child level rather than at the program level, where all programs providing a certain type of service would be included. Table A.12 shows what size differences in program-level variables need to be observed to be significant predictors of child-level quality outcomes in a regression model. Table A.13 shows what size differences in child-level variables need to be observed to be significant predictors of child-level quality outcomes. Classroom-level predictors of child-level quality outcomes fall somewhere in between. Table A.14 shows that, for quality over time, we can detect an effect size of .21 for 1-year-old Cohort children between ages 1 and 3 with 80 percent power.

[15] Because the home visit observations will be more difficult to schedule and coordinate during the week of the site visit than the classroom observations, we expected that the home visit quality measures would have a lower response rate than the classroom quality measures. Because the nominal sample size for the home visit quality measures will likely be smaller than shown in Tables A.9–A.12, their MDDs will be commensurately larger than the ones shown here.

[16] Because the standard deviation for quality measures is 1, the ES is identical to the MDD.

Table A.15. Program Measures MDDs and ES Comparing Two Program Defined Subgroups at a Point in Time

Subgroups	Subgroup 1 (Nominal Sample Size)	Subgroup 2 (Nominal Sample Size)	MDDs Between Subgroups for a Proportion p = .50	Minimum Detectable Effects (Presented As a Proportion of a Standard Deviation)
1/2, 1/2	45.0	45.0	.321	.63
1/3, 1/3	29.7	29.7	.397	.78
1/3, 2/3	29.7	60.3	.342	.68
1/4, 3/4	22.5	67.5	.373	.73
1/5, 4/5	18.0	72.0	.406	.79
4/10, 6/10	36.0	54.0	.328	.65

Note: Two-sided α = .05. Power = .80.

An example would be comparing average staff education levels between center-based and other programs.

Power for Program-Level Analysis. Estimates of program-level outcomes across program subgroups are shown in Table A.15 for the 90-program design. Because the analysis would be at the program level, there are no clustering effects—just a small unequal weighting effect (1.155), which increases the variance.

If we want to compare two equally sized subgroups of programs—for example, staff education or turnover in center-based versus home-based programs—we are able to detect a difference of about two-thirds of a standard deviation with a sample size of 89 programs. The row labeled "1/5, 4/5" is the most appropriate for comparisons of center-based programs to the other two program groups. One would use Table A.15 to get a sense of what size differences in program-level variables would need to be observed to be significant predictors of program-level outcomes in a regression model. These detectable differences and effect sizes are fairly large, which is to be expected for a nominal sample size of 89 programs. Our ability to detect much smaller differences in child outcomes between program subgroups is substantially greater.

We Constructed Weights at the Program and Child Levels

We constructed analysis weights at the program and child levels. These weights make the sample representative of the target population by adjusting for differential probabilities of selection and response patterns.

Program Weights. The program-level weight can be used for analysis of the 89 participating programs (for example, data from the program director interview). The program-level weight is also a building block for the child-level weights. This weight accounts for the initial probability of selection of each program within stratum, whether it was released into the sample, its eligibility status, and its participation status.

As described above, we initially selected twice the number of programs needed per stratum. We then formed sequential pairs of selected programs, sorted by the explicit and implicit stratification variables, so that adjacent programs were likely to be similar. We randomly selected one program within each pair to be the main release, and the other was available as a replacement for the released program if the primary release was a refusal or was ineligible.[17] In all, 111 programs were released into the sample, and 89 programs participated (Table A.16).

Table A.16. Program Sample Release and Final Status

Primary Release	Backup Release	Number of Pairs	Number of Participating Programs	Number of Ineligible Programs	Number of Refusing Programs
Participating	Not Released	68	68	0	0
	Participating*	6	12	0	0
Ineligible	Participating	8	8	8	0
	Ineligible	2	0	4	0
	Refusal	3	0	3	3
Refusal	Participating	1	1	0	1
	Ineligible	1	0	1	1
	Not Released*	1	0	0	1
Total		**90**	**89**	**16**	**6**

When only one member of a selected pair of programs was released into the sample (69 pairs), the initial probability of selection (based on 180 selections) was multiplied by 0.5. For the other 21 pairs, this factor was not applied. We then adjusted the probability weights for the participating programs to account for the six refusing programs either within the pair (one case) or within the stratum (five cases), then excluded the ineligible programs to construct the final program-level weight. After excluding the ineligible programs, the program weights sum up to 570 programs. This represents our best estimate of the total number of Early Head Start programs that met our study's eligibility criteria.

Child Weights. There are 1,194 children or pregnant women who we believed to be enrolled in one of the selected programs and whose birth date or due date falls within one of our defined windows, based on enrollment rosters provided by the programs. Among these, there were 33 pairs of siblings from which we randomly sampled one child per household.[18] This left us with 1,161 in the sample. There were an additional 55 children who met the study criteria and were added at the time of the site visit. Of these 1,217 children, we confirmed 1,108 to be enrolled and born within the defined windows at the time of the data collection visit, while 108 had dropped out or had originally had an incorrect birth date assigned (or were pregnant women with a due date inside the window who actually gave birth before the window).

[17] Because six of these backup programs also turned out to be ineligible or refusals, we released six backup programs from other randomly chosen pairs in which the primary release was a participant. In addition, one program that was willing to cooperate had institutional review board (IRB) delays that made it impossible for them to participate, and at that point it was too late to release its backup.

[18] We defined "siblings" as any set of children who shared a primary caregiver (parent or guardian). We subsampled among these children (should more than one have met our study criteria based on birthday windows) to minimize the burden of the study on the parent or guardian.

The child weight begins with the final program weight assigned to each child in the program, with an adjustment for the sibling subsampling for 33 cases. The sum of the sibling-adjusted weights for these 1,108 eligible children is 6,229. This is our best estimate of the number of study-eligible children being served by study-eligible Early Head Start programs.

Among these 1,108 eligible children, we obtained parental consent to participate in the study for 976. The next step in the weighting process was to adjust the consented children to account for those whose parents did not consent to participate in the study, or more precisely, to adjust the sibling-adjusted weights for the consented children to account for those of the nonconsented children. We ran forward and backward stepwise logistic regressions to predict consent, with the pool of independent variables comprised of the child's age cohort, and the program's size stratum, service type (home, center, mixed), urbanicity (metro versus nonmetro), and census region. Both the forward and backward procedures yielded the same significant predictors of consent: size stratum and service type. We used the inverse of the propensity score as the nonconsent adjustment, weighting up the consented children and assigning weights of zero to the nonconsented children. The sum of the consent-adjusted weights for the 976 consented children is 6,219.

Then we created two child-level weights for the consented children based on responses to the parent interview, SCR, teacher or home visitor interview, and completion of the classroom (Infant Toddler Environmental Rating Scale-Revised [ITERS-R]) or home visitor (Home Visitor Rating Scales-Adapted [HOVRS-A]) observation. One weight is for analysis involving child-specific information obtained in the parent interview or SCR, and is positive if either of these was completed (Table A.17). Of the 976 consented children, 972 had completed one of these instruments at baseline (spring 2009).

Table A.17. Child-Level Instrument Completion Rates

Parent Interview Completed	SCR Completed	Consented Children
No	No	4
No	Yes	78
Yes	No	40
Yes	Yes	854
Total		**976**

Because only four children were missing both data components, we simply weighted up the consent-adjusted weight to account for them within weighting cells formed by crossing the size stratum with the service type. The sum of this weight for the 972 children with at least one of these instruments completed is 6,219.

Data on the quality of the services the child receives can be obtained from either the teacher or home visitor interview, or the classroom or home visitor observation (Table A.18).

Table A.18. Teacher- /Home Visitor-Level Instrument Completion Rates

Teacher or Home Visitor Interview Completed	Classroom or Home Visitor Observation Completed	Consented Children
No	No	22
No	Yes	11
Yes	No	159
Yes	Yes	784
Total		**976**

This next weight is for analysis involving both child-specific information (as found in the parent interview or SCR) and this quality information. This weight is positive if we have (1) either the parent interview or SCR, and (2) either the teacher/home visitor interview or the classroom/home visitor observation (Table A.19).

Table A.19. Observation Weight Constituents and Completion Rates

Parent Interview or SCR Completed	Teacher/Home Visitor Interview or Observation Completed	Consented Children
No	No	0
No	Yes	4
Yes	No	22
Yes	Yes	950
Total		**976**

We ran forward and backward stepwise logistic regressions to predict this completion pattern with the same pool of independent variables. Both the forward and backward procedures yielded the same significant predictors of consent: age cohort, size stratum, and service type. We used the inverse of the propensity score as the nonresponse adjustment, weighting up the responding children and assigning weights of zero to the nonresponding children. After applying the score (and doing a minor poststratification adjustment within the same classes used for the weighting adjustment described above), the sum of this nonresponse-adjusted weight for the 950 children is 6,219.

APPENDIX B

DATA COLLECTION

APPENDIX B. DATA COLLECTION

This chapter provides details about the process we followed to recruit and enroll programs and families into Baby FACES, conduct data collection and staff training, and prepare the data for analysis.

Baby FACES Coordinators Worked to Recruit Programs

Given the complexity and longitudinal nature of the study, we worked to establish personal relationships between researchers and Early Head Start program staff to ensure strong ongoing communication between the Mathematica team and the programs. We formed and sustained these relationships with a staffing structure that includes a cadre of 10 trained full-time Mathematica staff who act as Baby FACES coordinators (BFCs). We assigned each BFC a set of programs to recruit and work with closely throughout the study.[1]

The recruiting process began in fall 2008 when we selected the sample of potential programs and confirmed that they were in good standing with the Office of Head Start. First, we mailed program directors a packet of information about the study. Shortly afterward, BFCs contacted program directors to recruit them into the study. As discussed in the sampling section, at this point BFCs asked program directors to provide preliminary rosters of children and pregnant women receiving Early Head Start services (along with their birth or due dates). We used this roster to determine each program's eligibility for the study. When an eligible program agreed to participate, BFCs asked directors to confirm a date for a potential data collection site visit, and to designate an on-site coordinator (OSC) with whom the BFC could work throughout the duration of the study. BFCs recruited programs from November 2008 through February 2009.

On-Site Coordinators and Baby FACES Coordinators Collaborated to Secure Family Participation

After BFCs successfully recruited each program and determined eligibility (see Section C), they worked with the identified OSC to find the best week to conduct the spring data collection. Scheduling took into account center closings, abbreviated weeks, and program regional monitoring visits. We also made it a priority to avoid data collection during weeks in which the sampled program did not have enough eligible children (five or more within each cohort). The BFCs and OSCs conducted telephone planning meetings in the months leading up to data collection in which they finalized the logistics associated with the visit, including when the team would be arriving; procedures for scheduling observations and teacher/home visitor interviews; and steps to obtain informed consent from potential study participants. For seven programs, part of that work included helping programs meet local IRB requirements in addition to the study IRB approval obtained by Mathematica at the study's inception.[2] The result of this process was a data collection plan for the Mathematica field team to follow. The data collection plan detailed the location and time of each

[1] Each BFC is responsible for 9 programs, on average (ranging from 1 to 15).

[2] All but one of the seven programs obtained local IRB approval and participated in the study. One program was unable to meet the requirements of the IRB (program staff did not complete human subject certification) and was not included in baseline data collection.

classroom that was scheduled to be observed, and the name and contact information of each home visitor to be observed.

In the spring, approximately two weeks before the site visit, BFCs requested an updated roster from each program's OSC. We used this roster to determine which children and pregnant women would be eligible for Baby FACES so that we could begin gathering consent from parents. Approximately one week before the data collection visit, the BFC confirmed the roster of eligible children and mailed consent forms to OSCs. In each program, the OSC distributed consent forms, along with information about the study, and gathered parents' consent. When consent forms had been obtained, they were faxed back to the BFC at Mathematica and were entered into the sample management system (SMS). When the OSC was unable to obtain consent before a scheduled visit, it became the responsibility of the data collection team leader to do so before gathering any data on the child. The consent forms themselves included information about the study—including the longitudinal nature of data collection—and requested information about the parent or guardian, including contact information and preferred language, as well as information on the selected child, such as name, gender, and date of birth or due date. In four programs, IRBs required consent from teachers and home visitors of the children selected to participate in the study. This process of confirming "adds and drops" was repeated by the leader of the site visit team during the actual week of data collection. We considered any child or pregnant woman enrolled in the Early Head Start program at the time of the site visit (and who met the birthday window parameters) as eligible for Baby FACES.

Training and Quality Assurance

To prepare for the spring 2009 interviewer training, we trained a group of lead trainers for each measure and data collection approach, certified a set of "gold standard" service quality observers, and hired and trained field interviewers. Mathematica's experience collecting similar data informed these steps, and they conformed to, or exceeded, data collection quality standards in the field. First, we present the field staff training approach. Next, we describe parent survey training and program director telephone interviewer training.

Pretraining Preparation for Field Staff Trainers and Quality Assurance Staff Was Extensive

We took a number of steps to prepare for training field staff. In mid-January 2009, the authors of the ITERS-R came to Mathematica's Princeton, New Jersey, office to train our group leaders for the field staff training and gold standard quality assurance observers. They trained 10 people; 4 of the 10 were reliable after the three-day training.[3] Nine of 10 were reliable by the final gold standard observation in mid-February. Only those who met the reliability standard led trainee groups and conducted quality assurance observations in the field.[4]

To prepare the HOVRS-A for research data collection, we made the following adaptations to the original scale:

[3] Reliability for a gold standard observer as defined by the ITERS-R authors is 95 percent reliability within one point on each item.

[4] Of the nine gold standard quality assurance observers on the HOVRS-A and ITERS-R, three are bilingual.

- ***Arranged Indicators Across Anchors to Be Parallel.*** To streamline the use of the scale by observers, the indicators on each were reorganized so that (1) the indicators assessed the same types of interactions within an anchor, and (2) the indicators would be parallel between anchors.
- ***Revised Difficult-to-Operationalize Indicators.*** Several indicators on the original scale were judged to be subjective and, therefore, possibly difficult to operationalize and reliably train observers. In these cases, the wording of the indicator was revised. In some cases, indicators that captured the same interaction as another indicator were removed to avoid redundancy.
- ***Added User-Friendly Features.*** To ensure that observers could easily use the HOVRS-A, each indicator was numbered, and "yes" and "no" check boxes were added.
- ***Developed a Manual.*** The manual included instructions for observers on how to prepare for a home visit observation, general guidelines for conducting an observation, and descriptions of how to score each scale.

Preparation for training included the development of the Home Visit Observation Content and Characteristics Form. This form collects information about the characteristics of the participants in the visit, the content of the visit, and whether the home visitor believed that the objectives of the visit had been accomplished.

The Mathematica staff member responsible for developing the HOVRS-A training served as the gold standard for all observers and developed a video-based training sequence for the field staff training.

Field Staff Training, Certification, and Quality Assurance Procedures Ensured High-Quality Data

In January 2009, we recruited 10 two-person teams and seven additional "floaters" (staff who worked outside of defined teams), for a total of 27 people, to conduct the first round of data collection. To minimize long-distance travel, team members were drawn from several states around the country. Each team had a designated team leader responsible for managing on-site activities, including scheduling home visit and classroom quality observations. Team leaders were also the main point of contact with the OSC during the site visit week. Twelve of the team members were bilingual in Spanish and English. All were experienced data collectors, many of whom (17) had worked on similar early education data collection projects, either for Mathematica or for other research organizations.

The field training lasted seven days and took place in Princeton, New Jersey, between February 15 and 21, 2009, two weeks before the start of baseline data collection. Field staff received a total of 40 hours of training during the week. Training included both lectures and practice sessions in the field.

Most of the training focused on observation (discussed later). Interviewers were also instructed in conducting the teacher/home visitor interview and resolving logistical issues. Training on the teacher/home visitor interview included a section-by-section review, paired practice, and a review of the instruments. Field staff was also familiarized with Staff Child Rating (SCR) forms and procedures for distributing, collecting, and answering questions about the forms. All team members learned how to complete the hard-copy forms and how to conduct a final review of all materials before sending them back to Mathematica.

Before the in-person training, field staff received the Baby FACES Observer Manual and a DVD of an Early Head Start home visit. We asked staff to watch the DVD before training to familiarize themselves with the aspects of a home visit. All observers received 18 hours of training on conducting home visit observations and administering the HOVRS-A. Training on the home observation began with an overview of home visits, the observer's role during the home visit, guidelines for conducting a home visit observation, and instructions on how to complete the Home Visit Observation Content and Characteristics Form (See Table B.1). The second and third day of observation training focused on the HOVRS-A. Training on the HOVRS-A began with an overview of the rating scales, a discussion of the key terms used throughout the scale, and a description of the seven individual scales. The trainer then described how the rating scale points were used as "anchors" and how indicators were used as descriptions of behaviors observed during the visits. After the review of each scale and its associated indicators, field staff viewed short video vignettes and practiced coding each of the seven items separately. The trainer then presented the "gold standard" ratings and provided justification for each rating. After each of the seven scales had been discussed individually, trainees viewed a longer home visit video clip and rated all seven scales for practice. Field staff viewed and rated three more home visits during the training session to establish reliability. Observers were required to match the gold standard score for each HOVRS scale (for example, Home Visitor Family Relationship, Parent Engagement) for at least two of the three ratings of video clips. We offered help sessions in the evenings to answer questions on an individual basis. All observers became reliable and received certification to conduct the HOVRS-A.

Table B.1. Training Agenda—Spring 2009

Day 1	Introductions and Overview of the Baby FACES Study Discussion of and Practice with the Home Visit and Teacher Interviews Introduction to the Home Visit and Practice with the Home Visit Observation Content and Characteristics Form
Day 2	HOVRS-A
Day 3	HOVRS-A training continued and reliability testing
Day 4	ITERS-R classroom training
Day 5	ITERS-R field practice – A.M. Small group discussion and debriefing – P.M.
Day 6	ITERS-R field reliability – A.M. Observation wrap-up and data entry – P.M.
Day 7	Wrap-up and administrative details (time sheets and expense reports)

Training on the ITERS-R involved one day of classroom presentation, a video segment review and quizzes, one day of practice in actual classrooms followed by group discussion, and one day for certification. Training included lectures on the components of each item in the observation measures, a discussion of how to score more difficult items and not easily observed items, a review of how to conduct an observation, and coding practice to achieve reliability across observers on all measures. The observations were recorded on hard-copy instruments. Gold standard trainers and groups of three trainees visited two local child care centers—the first day for practice, the second to establish reliability. Trainees who did not meet reliability standards established by the developers of the ITERS-R (80 percent agreement within one rating point with the author-certified gold-standard group leaders) conducted additional practice observations until certified. By the end of training, all trainees except one were certified to conduct the ITERS-R.[5]

Quality Assurance Field Visits Ensured Reliability on HOVRS-A and ITERS-R

Nine people (six from Mathematica and three from Branch Associates, our subcontractor) were certified to conduct quality assurance (QA) visits on the HOVRS-A and ITERS-R. Three of these people were bilingual. QA visits occurred in weeks 7 through 12, with a total of 15 sites visited. QA staff were on-site an average of two days and observed all team members conducting observations. In total, we observed 23 field staff for quality assurance. We were able to rate 13 staff on both the HOVRS-A and ITERS-R. For three, we observed only the HOVRS-A, and for seven, we observed only the ITERS-R. Eighty-one percent of the field staff who were monitored on the HOVRS-A and 95 percent of the field staff who were monitored on the ITERS-R were reliable on 80 percent of the items or better.

Telephone Interviewers Were Trained to Administer the Parent Survey

Two different groups of telephone interviewers (daytime and evening interviewers) received four hours of training for the parent survey in early April 2009. We trained 31 telephone interviewers on the project (10 were bilingual in English and Spanish). In addition, five interview monitors (three of whom were bilingual) and three 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 instruments. At the conclusion of the formal training, interviewers were paired up to conduct mock interviews with one another under the supervision of trainers and supervisors. The practice interviews were conducted using the Computer-Assisted Telephone Interview (CATI) instrument. 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.

Two Researchers Administered the Program Director Survey

In late March 2009, the Baby FACES project and survey directors trained two researchers for four hours to conduct the program director survey. The training involved detailed discussions about the structure of programs and the intent of the survey items so they could gather enough

[5] That observer conducted only HOVRS-A observations during the field period.

information from the program director to accurately record the information on the questionnaire form. Extensive spreadsheets were created to capture additional information outside the questionnaire form to fully understand programs' structure and activities.

The project director conducted and audiotaped the first interview and then reviewed it with the two researchers. The first interviews conducted by the two researchers were conducted as a team effort, with one person conducting the interview while the second person listened. Both recorded responses and compared their coding after the interview. Each researcher reviewed their own interviews, entered verbatim comments into the spreadsheet, reviewed the SAQ for completeness, and determined if a call-back was needed. Because the program director interview resembled a semistructured executive interview, the interviewers recorded extensive additional information on spreadsheets to capture information that went outside the questionnaire form and would help to fully understand program activities. The project director also listened to tapes of these first interviews by each of the researchers and debriefed them.

Interviews and Observations

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

After they were certified as reliable to collect the data, teams of field interviewers visited 89 sites over a 15-week period during spring 2009. Teams consisted of a team leader and one or more field interviewers. Forty-one programs were visited by a single field interviewer, and 32 programs were visited by teams of two field interviewers. Larger teams of 3, 4, and 5 members visited 15 programs, and one very large program required a team of 13 field interviewers. If children received any services in Spanish, at least one bilingual member was assigned to the team visiting the program. Upon arrival at the site, the team leader met with the OSC to schedule classroom and home visit observations, as well as in-person interviews with teachers and home visitors. Only teachers and home visitors serving 1-year-olds in the study were observed in spring 2009. Furthermore, only one observation per teacher/home visitor was conducted. If a teacher or home visitor provided services to more than one study child, only one observation was needed.

Classroom Observations. We conducted observations of teachers providing services to 1-year-olds for a two- to three-hour period during the study week. Whenever possible, we scheduled the classroom observation in the morning. Only two observations were conducted in the afternoon hours. During each observation, the field interviewer conducted the ITERS-R, which included recording information on space and furnishings, personal care routines, listening and talking, activities, and program structure. The observers also completed two Counts of Children and Adults (spaced at least an hour apart) and completed three Post Visit Ratings. During these observations, there was no interaction with the children or the teachers. At the end of the observation, the observers asked the teachers questions about things that were not observed (for example, "Since I was not here to naptime, can you please describe how nap is handled?"). After gathering these last pieces of information, the observers assigned their final scores. We gave a gift bag of classroom supplies worth $25 to the teacher in each observed classroom.

Home Visitor Observations. Using the HOVRS-A, we also conducted an observation of one home visit by home visitors providing services to 1-year-old children in the study. This observation tool focused 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. As with the classroom observation, only one observation per home visitor was conducted, regardless of the number of study children each home visitor may see. On average, each home visit, and therefore each HOVRS observation, lasted an average of one and a half hours. We gave home visitors we observed a gift bag of classroom supplies that was identical to those we gave to teachers.

Teacher/Home Visitor Interview and SCR. The in-person teacher and home visitor interviews lasted approximately 30 minutes and focused on their background, training, services they provided to families, and expectations the program placed on them. The field interviewer 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 providing services in the home versus a classroom.

At the start of the week of the site visit, we distributed Staff-Child Report forms (SCR) to each teacher and home visitor of study children; teachers and visitors were instructed to complete this self-administered questionnaire about each of the study children with parental consent. 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. On average, each teacher/home visitor completed 1.7 SCR forms, with 295 teachers/home visitors completing only one SCR, and 256 completing multiple forms.

There were three different versions of the SCR: one for 1-year-olds, one for newborns, and a different version for women who were pregnant with their child. Those providing services to 1-year-old children were asked to report on the child's social skills, language development, and parent-staff relationships. Those providing services to newborns and pregnant women were asked to report on prenatal and pregnancy services provided and parent-staff relationships.

After the visit, the team leader collected and reviewed all completed instruments and sent the documents to Mathematica for receipt and review. In a few cases, field staff could not schedule or conduct home visits during the initial visit week. In these 10 cases, interviewers returned to the program for a second visit one week later.

Mathematica Conducted Parent Interviews and Program Director Interviews by Telephone

Parent Interviews. To obtain important information on children's home environment, we asked all parents to complete a telephone interview. Mathematica telephone interviewers conducted the interview at the Survey Operations Center (SOC) in Princeton, New Jersey. The interview was programmed and administered using Computer-Assisted Telephone Interviewing (CATI), thereby allowing the individual path of each interview to be determined based on the responses given to previous questions or preloaded information (such as cohort). The interview was conducted in Spanish when necessary. The parent interview had 21 sections, although not all parents were asked questions in every section. Table B.2 shows the section titles of the parent interview. The parent interview instrument will be available on the ACF website, along with all other study instruments.

We sent parents an advance letter describing the interview on the Thursday before the week they were scheduled to be called. The letter explained the importance of the interview and provided a toll-free call-in number. We released sample weekly throughout the three-month field period, and did not begin interviewing parents until at least a week after we visited the program in person. We conducted parent interviews between April 13 and July 27, 2009. Parents completing the interview received a $35 check in exchange for their participation.

Table B.2. Parent Interview Sections

Section Prefix	Section Title
SCREENER	
HH	ABOUT HOUSEHOLD
B	HOUSEHOLD LANGUAGES
C	PROGRAM SERVICES
D	STAFF-PARENT RELATIONSHIPS
E	CHILD BEHAVIOR
F	CHILD HEALTH
G	CHILD CARE
H	ABOUT CHILD'S MOTHER
I	ABOUT CHILD'S FATHER
J	ABOUT FATHER FIGURE
K	ABOUT RESPONDENT
L	HEALTH CARE SERVICES
M	PARENT HEALTH
N	RAISING A CHILD
O	FAMILY ROUTINES
Q	FAMILY GOALS
R	SOCIAL SUPPORT
S	NEEDS AND RESOURCES
U	INCOME AND HOUSING
V	TRACKING INFORMATION
W	INTERVIEW RATINGS

Program Director Interviews. To learn about program practices, policies, and overall enrollment, we conducted interviews with program directors. The program director interview was broad in scope and asked directors about the entire program, including (if applicable) all of their Early Head Start centers (not just those selected for the study). We gathered program-level information in two ways: (1) an hour-long telephone interview, and (2) a self-administered questionnaire (SAQ). The interview focused on program structure, involvement with community partners, approaches to serving DLLs, and program goals. In the SAQ, we included questions that might have required review of records or consultation with others. These included items that asked program directors to quantify the number and education level of staff in various positions. In addition, the SAQ asked program directors to rate their program's level of implementation in five cornerstone areas: (1) child development, (2) family development, (3) staff development, (4) community building, and (5) management systems and procedures. This implementation rating form was developed from indicators used in the Early Head Start Research and Evaluation Project (EHSREP) and pilot tested in the Survey of Early Head Start Programs (SEHSP). We continue to explore the best ways to learn about program implementation.

We conducted program director telephone interviews between April and July 2009. Program directors received a letter via Federal Express inviting them to participate in the interview. The letter informed them that researchers from Mathematica would be calling to schedule the telephone interview. The letter included a short list of topics so directors could decide if they wanted others from their program to participate in the call with them. The packet also included a copy of the SAQ and a prepaid return envelope.

At the end of the telephone interview, if Mathematica had not yet received the SAQ, the researcher asked directors when we could expect it and if they had any questions or issues. If Mathematica had already received the questionnaire, we thanked them for receipt of the self-administered survey. Programs that participate in the study receive a $500 honorarium annually. We sent the checks to programs after they completed the program director interview. (The payment was made upon receipt of the SAQ, generally after the telephone interview was completed. For the three sites that failed to complete the SAQ, we sent payment in August 2009.)

Mathematica Launched the Family Services Tracking System During 2009

Family Services Tracking System. Mathematica developed an instrument to track the services each study family receives over time: the Family Services Tracking (FST) system. The FST asks the staff member with primary responsibility for study children (either a teacher or home visitor) to complete a brief weekly report of the services provided to each study child, including:

- Whether there was a change in their service type or teacher/home visitor
- The child's expected and actual attendance in the program on a week-to-week basis
- Reasons for absence
- Any referrals made or special events or activities attended

We introduced the FST to programs on a rolling basis; the first seven programs received their materials and instructions in late April 2009 and the last program in July 2009.[6] Before sending information to programs, BFCs introduced the topic during regular telephone calls with the OSC. Then BFCs sent a packet of information, including a cover letter that described how to complete the forms, information on web log-in IDs and passwords for each teacher and home visitor serving study children, and a packet of labeled paper forms to distribute among staff. Based on feedback from programs, we devised a user manual and added a set of frequently asked questions and answers to the web system. We asked programs to begin FST reporting after all on-site data collection was completed.

We provided many options for staff to complete the forms, including a paper- or web-based form. We expected that staff would find it easiest to complete the paper form during the day and enter it into the web-based form later. Some programs collect forms from staff over a period of time and have a designated staff member enter data from all forms into the web-based data system; other programs collect their forms and mail them to Mathematica for entry. We continue to work with programs to complete the FST regularly for all study children and to address questions as they occur. We will present data from the FST in the next report.

[6] Two sites requested that they begin the FST in September rather than during summer months; these sites were mailed materials in September 2009.

Response Rates

Field Data Collection Achieved High Consent and Completion Rates

Baby FACES recruiting and data collection were successful. As noted earlier, we were able to recruit 89 programs into the study, with a consent rate of 93.7 percent. Mathematica staff visited these 89 sites over a 15-week field period from the week of March 1 to the week of June 7, 2009. Within these programs, we approached families of children in our age cohort windows and obtained consent from 88.1 percent of families (976 of 1,108, Table B.3). The sample consists of 80 percent 1-year-old Cohort members and 20 percent Newborn Cohort members. The consent rate within cohorts is similar: 88.8 percent for the 1-year-old Cohort and 87.4 percent for the Newborn Cohort. This difference is not statistically significant.

Table B.3. Baby FACES Consent and Completion Rates

Week	Date	Num of Selected Children	Num of Consents Received	% Consent Rate	Num of SCRs Received	% SCRs Received	Num of HOVRS received	% HOVRS received	Num of ITERS received	% of ITERS received	Num of TI Received	% of TI Received	Num of HV Ints Received	Num of HV Ints Received
1	3/1	95	79	83.2	75	94.9	30	90.9	12	100.0	12	100.0	39	95.1
2	3/8	96	89	92.7	87	98.9	17	94.4	32	100.0	31	96.9	22	100.0
3	3/15	95	91	95.8	91	100.0	11	84.6	32	100.0	32	100.0	24	100.0
4	3/22	93	78	83.9	77	98.7	26	89.7	17	100.0	17	100.0	32	100.0
5	3/29	73	64	87.7	64	100.0	22	95.7	11	100.0	11	100.0	26	100.0
6	4/5	24	23	95.8	23	100.0	9	100.0	2	100.0	2	100.0	11	100.0
7	4/12	96	90	93.8	88	98.9	13	86.7	26	100.0	28	100.0	25	100.0
8	4/19	77	63	81.8	59	93.7	8	100.0	15	88.2	19	90.5	11	100.0
9	4/26	110	92	83.6	86	92.5	25	92.6	14	66.7	16	61.5	27	90.0
10	5/3	106	88	83.0	75	85.2	24	82.8	17	94.4	17	94.4	32	94.1
11	5/10	69	65	94.2	63	96.9	16	80.0	14	93.3	15	100.0	21	91.3
12	5/17	125	112	89.6	105	92.9	29	85.3	22	95.7	20	87.0	38	95.0
13	5/24	4	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
14	5/31	45	42	93.3	41	95.3	12	92.3	9	100.0	9	100.0	15	100.0
Total		**1,108**	**976**	**88.1**	**933**	**95.5**	**242**	**89.3**	**223**	**94.9**	**229**	**93.1**	**323**	**96.7**

Most Consenting Families Completed Parent Interviews

Among families who gave consent, we conducted parent interviews from April 13 through July 27, 2009. On April 13, we released the initial sample load of sites visited in the first seven weeks. After the initial sample release, we released new sample weekly through the end of May. Until July 16, we released sample periodically as we received late consents.

We sent advance letters to parents the Thursday before the Monday start of telephone interviewing. We sent reminder postcards to those who had not completed the telephone interview by June 12 for English interviews and June 15 for Spanish interviews.

We completed parent interviews with 825 parents (84.6 percent). An additional 37 parents (3.8 percent) completed through the household composition section of the interviews, and another 32 (3.3 percent) completed through Section E (the child development section). Among the group who gave consent, 88 percent of Newborn Cohort parents completed the parent interview, while 84 percent of 1-year-old Cohort parents completed the interview. We completed a total of 130 interviews in Spanish (22 from the Newborn Cohort and 108 from the 1-year-old Cohort) and the rest in English. The average length of the parent interview was 109 minutes, with 1-year-old Cohort interviews running 120 minutes, on average, and Newborn Cohort interviews lasting 65 minutes. Spanish interviews took longer than English (121 minutes overall) with the 1-year-old Cohort taking 134 minutes and the Newborn Cohort taking 77 minutes, on average.

Very few parents (1 percent) refused the interview, and we were not able to conduct an interview in 6.5 percent of the cases. In half of these cases (2.8 percent), we could not locate a working telephone number. As interviewers identified incorrect or nonworking telephone numbers throughout the telephone field period, the BFC asked the program to confirm the number or if any other numbers were known to them. If that was unsuccessful, staff at the SOC attempted to find telephone numbers through directory assistance and online sources.

All parents who completed the interview (partial or complete) were mailed a $35 thank-you check, usually within two weeks of completing the interview. In a few cases, we learned that parents had incurred costs from the telephone interview, as they had used prepaid minutes on their cell phones. In the 18 cases where this was made known to Mathematica, the parents received an addition $12 payment to reimburse them for the minutes used ($0.10 per minute for 120 minutes). Moreover, since the telephone interview lasted substantially longer than anticipated, we decided to mail all parents a picture magnet with the Baby FACES logo and toll-free telephone number as an additional thank-you.

We observed a total of 223 (out of 235) classrooms for a 94.9 percent completion rate, and 242 (out of 271) home visitors for an 89.3 percent completion rate. In addition, a total of 552 (out of 580) teachers/home visitors completed the teacher/home visitor interviews (95.2 percent completion rate). We received 934 completed SCRs from 553 teachers for a 95.6 percent completion rate.

Program Director Interviews Achieved a Perfect Response Rate

We mailed survey packets to program directors on a rolling basis, beginning on March 26, of those programs visited in the first four weeks of the field period. Mailings to later sites occurred on April 23 (weeks 5 to 8), May 8 (weeks 9 to 12), and on May 27 and June 17 for sites visited in weeks

14 and 15, respectively. We sent multiple email reminders to programs that did not return the SAQ within a month of the mailing. In addition, 14 programs received second packets via express mail at their request.

We conducted telephone interviews with a program director or designee in all 89 programs (100 percent). The first interview was completed on March 31, and the final interview was completed on July 22. The telephone interview averaged one hour in length. In most cases (79 out of 89), the interview was conducted with the program director. In the other 10 cases, the OSC or another person designated by the program director completed the telephone interview. In four cases, more than one respondent participated in the telephone interview.

We received a total of 86 SAQs from program directors (97 percent). The field period for this portion of the program director interview began on April 7, with the receipt of the first SAQ, and extended until July 27, when we received the final SAQs. We also included one additional SAQ received after the close of the field on September 3, 2009.

Data Processing

Receipt Control Involved Several Steps and Validation Procedures

The team leader reviewed all field materials before sending the package to the SOC. The field materials included consent forms, teacher/home visitor interviews, HOVRS-A, ITERS-R, and some SCRs. The remaining SCRs were mailed in directly by the teachers/home visitors after they completed them.

Field materials generally arrived at the SOC the Wednesday following each site visit. SOC field staff reviewed the materials for each site, looking specifically to make sure all materials had arrived and that they matched the data collection plan. In some cases, a child's teacher/home visitor had changed or was different than stated in the data collection plan. The SOC field staff would verify this change with the team leader and BFC. The BFC would make the change(s) in the SMS or add a new teacher and assign that teacher to the child. SOC field staff would then produce a new data collection plan consistent with the field materials.

In two instances, teachers—each with a unique sample child—shared the same classroom. Instead of completing two ITERS-R observations of the same classroom, staff completed one ITERS-R observation on-site for these cases. A duplicate observation booklet was then created so that an ITERS-R could be receipted and data entered for both teachers. Staff administered separate teacher interviews to both teachers.

Receipting Documents. After review, SOC field staff transferred all materials to the receipt department with a transmittal form with which they were scanned into a tracking system. The receipting staff entered consent forms into the system first, enabling the associated instruments to be processed successfully. The instruments that scanned successfully were placed into batches for quality control, editing, and data entry. However, if staff scanned any instruments into the system that corresponded to study participants for whom we had not yet obtained consent, those instruments were flagged, given a special status code, and set aside. When consents were obtained, the flags were dropped, and the instruments were allowed to be sorted for quality control, editing, and data entry.

Quality Control Procedures Ensured Accurate Coding

Quality Control and Editing. In early April, we trained three quality control staff to review and edit the observation instruments (HOVRS-A, ITERS-R), and trained three additional quality control staff to review and edit the teacher/home visitor interview and SCRs. In addition, we trained one supervisor to oversee all survey instruments and another supervisor to specifically oversee the teacher/home visitor interview and SCR quality control process. Two members of the project survey staff monitored the process. In general, quality control staff reviewed the instruments for completeness and checked that the skip logic had been followed correctly. Where possible, quality control staff made appropriate edits to the instruments based on preestablished specifications. Specifically, the ITERS-R/HOVRS-A specifications dictated how quality control staff should review scores, when to mark data as missing, and when to set aside for project survey staff to review.

Data Entry. After being reviewed and/or edited, instrument data were keyed into the data entry program. One staff member entered the data initially, and a second staff member entered the data a second time to ensure accuracy; this ensured a 100 percent verification of all data. After data were entered, the statuses of the instruments were updated in the system as "complete."

Coding. After all instruments had been data entered, all verbatim or "other specify" responses were loaded into a coding database. Data entry staff reviewed each response in the coding database. The responses could either be back-coded into preexisting answer options, built into new codes if enough responses expressed the same concept, or left as verbatim text in the data file. If data entry staff had questions about certain responses, the database allowed them to flag the responses for supervisor review. Project survey staff monitored the coding process.

We Completed Reconciliations for SMS Data, Birth Dates, Consent, and Gender

Survey staff gave a list of special requests to the information systems team to reconcile all of the information and data from all sources, including rosters, data collection plans, consent forms, and survey instruments. The purpose of reconciliation was to identify mismatches, duplicates, and possible discrepancies. The information systems team ran these requests through the SMS and produced reports for survey staff to review.

SMS and Data File Reconciliation. The first step in the reconciliation process was an extensive check to verify that completed interviews in each data file matched the SMS database. If completed interviews in a data file did not show up as complete in the SMS, staff investigated the issue to determine the source of the discrepancy. This process also confirmed that the correct instrument type was completed for each case. In a few instances, the wrong SCR was handed out in the field (for example, a home visitor SCR to a teacher). Many of the questions are identical in the instruments. When these cases were identified during the reconciliation process, the data were transferred from the incorrect instrument to the correct instrument, and the data were entered in the appropriate file.

Consent Reconciliation. We did a final check during reconciliation to ensure that all instruments considered final were associated with families who had consented to be in the study. If we determined that consent had not been granted, any instruments associated with that family were removed from the data file.

Date of Birth Reconciliation. The dates of birth first entered into the SMS originated from rosters provided by the programs. After the original entry, dates of birth could be updated from BFC discussions with OSCs, consent forms, or information provided by parents in the telephone parent interview. Some discrepancies resulted from dates of birth coming from multiple sources. In some cases, the original or updated date of birth did not fall in the program sample window, and we had to reconcile the date of birth to make certain that accurate dates of birth were in the system and that they fell in the sample window. When there were date of birth discrepancies and one of the dates of births fell in the sample window, the BFCs discussed this with OSCs to determine the correct date of birth. This reconciliation ensured that each year of birth matched the child's cohort. A date of birth for a child in the newborn cohort could only have a year of 2009, and a date of birth for a 1-year-old could have a year of 2007 or 2008. BFCs investigated any discrepancies and updated dates of birth in the SMS. In addition, the parent interview data file was edited to include the correct date of birth when the date of birth originally collected was incorrect.

Dates of Birth from Pregnant Mothers. By the end of the field period, all of the pregnant mothers in the sample had given birth. These dates of birth were generally collected during the parent interview and automatically updated in SMS. However, in several cases the mother was still pregnant during the parent interview or an interview was not completed. Therefore, information systems staff created a reconciliation report listing all cases with a missing date of birth in SMS. The dates of birth were obtained from the programs by the BFCs and entered into SMS.

Gender. Similar to date of birth, we collected children's gender from multiple sources, including rosters, consent forms, and the parent interview. In several instances, the gender was not originally entered into the SMS because the child had not been born yet. Information systems staff created reconciliation reports that listed each of the children who had gender missing in the SMS or each case where the child had a gender discrepancy between the SMS and the parent interview. The BFCs obtained the missing or correct genders from the OSCs and entered the information into the SMS. The parent interview data file was also edited to include the correct gender.

Data Cleaning Consisted of Frequency Review and Data Editing

Frequency Review. After data from all instruments had been entered and considered complete, we ran frequencies for each data set. Survey staff responsible for each instrument reviewed the frequencies to check 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 that each variable had the correct 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 survey team developed a document 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 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). In a few instances during reconciliation and cleaning, we discovered that some parent interview cases had been loaded with the wrong cohort or that the wrong SCR had been completed by a teacher. In these cases, large amounts of valid data existed that had to either be moved to different variables or moved to different instrument types altogether. If the data needed to be moved to different variables, staff made these changes using the edit document. If data had to be moved to a different instrument, the SAS programmer transferred the data manually to the appropriate instrument. Each time a data file was updated, a new set of frequencies was run and reviewed. This process continued until all of the data were clean for each instrument.

APPENDIX C

MEASURES

APPENDIX C. MEASURES

We selected child and family outcome measures according to many different considerations. Among our requirements were the reliability and validity of the measures, appropriateness for use with children and families from diverse backgrounds, comparability with other large research projects, burden on children and families, ease of administration and scoring, appropriateness for use by Early Head Start programs, and the need to complement well-established measures with those that are new to large-scale research and fill existing measurement gaps. Given the longitudinal nature of Baby FACES, we also attempted to select measures that could be used at all or at least multiple assessment points. We attempted to select measures that related to the cognitive, language, and social-emotional development outcome domains of the National Education Goals Panel. A description of all measures employed in this wave of data collection can be found in this chapter.

In addition to our review of the literature, we worked closely with experts from our technical work group (TWG), other experts in the field, and the test developers themselves to select and modify measures for Baby FACES. Over the course of nearly two years of planning, we hosted dozens of conference calls to discuss our approaches to assessing child development and gathering the most useful information on the language development of dual language learners (DLLs). The final list of measures presented here reflects the feedback of many 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 had missing data on more than 25 percent of the items that comprised a constructed variable, we did not compute a score for that individual. If the individual had 25 percent or fewer of the items missing, we imputed values based on the means of the nonmissing items. We used the specifications described in the user's guide to impute item values for the Ages & Stages Questionnaires, Third Edition (ASQ-3) and the Brief Infant Toddler Social Emotional Assessment (BITSEA). For methodological reasons, we did not impute missing data for the MacArthur-Bates Communicative Development Inventories—Infant Short Form (CDI), the Home Visit Rating Scale-Adapted (HOVRS-A), the Infant Toddler Environment Rating Scale-Revised (ITERS-R), and the Parent-Caregiver Relationship Scale (PCRS).[1]

- ***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 scale or instrument measure the same cognitive domain or construct. We chose Cronbach's coefficient alpha, which captures the correlation among items on an

[1] Because the HOVRS-A and ITERS-R scales are comprised of means, we did not impute means for missing values. Similar considerations precluded imputing values for the CDI and PCRS.

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.

- ***Consistent Reliability Across Key Subgroups.*** We examined internal consistency reliability across key subgroups, such as race/ethnicity and DLL status, to determine whether the constructed measures had similar levels of internal consistency across the subgroups.

Measures of Home Visit and Classroom Quality

To assess key aspects of the quality of both home- and center-based services, field staff conducted structured observational assessments of home visits of children in the 1-year-old Cohort. We observed center-based classrooms for the 1-year-old Cohort children receiving child development services primarily through center-based care, and home visits for those receiving child development services primarily through home visits. For classroom observations we used ITERS-R (Harms, Cryer, and Clifford 2003); for home visits we used HOVRS-A (Roggman et al. 2009), an adaptation of HOVRS (Roggman et al. 2008).

We Adapted HOVRS to Increase the Ease of Measuring Home Visit Quality

For children in home-based settings, home visiting is intended to provide supports for children's development, parenting outcomes, and the parent-child relationship (Roggman et al. 2008; Sweet and Applebaum 2004). Home visiting typically involves a trained home visitor working with the parent, child, and other family members. Although typically taking place in the home, visits may occur in a number of settings and typically focus on activities to facilitate parent-child interactions and support the parent-child relationship. Home visiting strategies and evaluations often focus on the quality of the home visitor-family relationship. Visitors typically adopt strengths-based approaches or those that develop rapport and trust and empower parents. Although evidence for the efficacy of home visiting strategies is mixed, stronger effectiveness is likely when the quality of the home visit is high and when the relationship quality between the home visitor and the family is strong.

Few tools exist for assessing home visit content and quality. HOVRS (Roggman et al. 2006) assesses a variety of dimensions, including home visitor responsiveness, nonintrusiveness, support of parent-child interaction, and parent and child engagement in the visit. Ratings on HOVRS have been associated with the quality of the home environment and with children's vocabulary via the home environment (Roggman et al. 2006). Peterson and Roggman (2006) reported internal consistency reliabilities for visitor quality and effectiveness quality subscales at about 0.65, lower than the 0.70 standard in the field. The overall quality score had good internal consistency (alpha = 0.78). The instrument's primary strengths are its focus on key dimensions of quality, the manner in which services are delivered, and the level of family engagement. Understanding the quality of the parent-home visitor relationship and the quality and content of the home visits provides important insight about service delivery improvements that might be needed.

To assess key aspects of the quality of home-based services, field staff conducted structured observational assessments of home visits of children in the 1-year-old Cohort. Dr. Roggman and her research team designed HOVRS collaboratively with home visitors from programs receiving professional development and training on home visiting from the research team. It includes seven scales, four that focus on the home visitor's responsiveness to the parent and child (Home Visitor Responsiveness to Family, Home Visitor-Family Relationship, Home Visitor Facilitation of Parent-Child Interaction, and Home Visitor Non-Intrusiveness), and three that assess the parent and child engagement with each other during the home visit (Parent-Child Interaction During Home Visit, Parent Engagement During Home Visit, and Child Engagement During Home Visit).

HOVRS-A[2] consists of seven items that can be combined to form a total score and two subscale scores: Visitor Strategies (4 items) and Visitor Effectiveness (3 items). Visitor Strategies focuses on the home visitor's responsiveness to the parent and child, and Visitor Effectiveness assesses the parent's and child's engagement with each other and with the home visitor. Items on HOVRS-A are rated from 1 to 5, with anchors of 1 (minimal), 3 (moderate), and 5 (good practice). During observations of home visits, field staff also completed the Home Visit Content and Characteristics Form that collected data on the topics, activities, and structure of the home visit. These observations provided information on the characteristics of participants in the visit, the content of the visit, and whether the home visitor believed the objectives of the visit were accomplished. Items included in the HOVRS-A subscales are noted below:

Visitor Strategies Quality

- Home Visitor Facilitation of Parent-Child Interaction
- Home Visitor-Family Relationship
- Home Visitor Responsiveness to Family
- Home Visitor Non-Intrusiveness

Effectiveness Quality (at involving and engaging family)

- Parent-Child Interaction During Home Visit
- Parent Engagement During Home Visit
- Child [Infant or Toddler] Engagement During Home Visit

[2] Four main modifications were made in creating HOVRS-A. First, to make the measure easier to score, the number of scale rating points was reduced from seven to five. This was intended to help 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 months (Toddler Engagement During Home Visit), and ensured that the indicator wording on all items is appropriate for infants and toddlers.

Table C.1 presents HOVRS-A scores for classrooms observed during Baby FACES data collection. As illustrated, the overall quality score has high internal consistency (0.84). However, the effectiveness quality subscale has somewhat lower internal consistency (0.69). This is slightly lower than the 0.70 standard in the field, but higher than estimates reported by Peterson and Roggman (2006).

Table C.1. Summary Statistics for Baby FACES Child Care Quality Data: HOVRS-A, Spring 2009

	Possible Range		Reported Range				
Measure	Min.	Max.	Min.	Max	Mean/ Percentage	Standard Deviation	Cronbach's Alpha
HOVRS-A Overall Quality	1	5	1.00	5.00	3.34	0.87	0.84
Visitor Strategies Quality	1	5	1.00	5.00	3.18	0.98	0.82
Effectiveness Quality	1	5	1.00	5.00	3.56	0.95	0.69
Sample Size	**366**						

Source: Spring 2009 Home Visit Observations.

ITERS-R Provided a Measure of Classroom Quality

The most common Early Head Start option is providing center-based care to children. Although widespread agreement exists that higher-quality care leads to better outcomes for children, defining what constitutes quality child care and how to measure it is challenging. Providing and measuring quality care for infants and toddlers cause particular challenges due to the high level of individual attention needed by young children.

ITERS emphasizes positive relationships, health and safety, and a stimulating learning environment (Harms, Cryer, and Clifford 1990). The scale is based on a format and scoring system previously employed for the Early Childhood Environment Rating Scale (ECERS) by Harms, Cryer, and Clifford 1980 at the University of North Carolina-Chapel Hill's Frank Porter Graham Child Development Institute. The authors specifically designed ITERS to rate care quality for children ages birth to 30 months. The ITERS-R edition provides a revision to the original ITERS, creating a more formal system of indicators to enable more exact scoring. The authors also combined, revised, added, or dropped questions from most subscales. The revisions are based on advances in research on early childhood development, feedback from ITERS users, and extensive testing of the original instrument (Development of the ITERS-R, FPG Child Development Institute 2002).

The full ITERS-R consists of 39 items organized under seven subscales: (1) Space and Furnishings (5 items), (2) Personal Care Routines (6 items), (3) Listening and Talking (3 items), (4) Activities (10 items), (5) Interaction (4 items), (6) Program Structure (4 items), and (7) Parents and Staff (7 items). Items on ITERS-R are rated from 1 to 7, with descriptors provided by the authors for ratings of 1 (inadequate), 3 (minimal), 5 (good), and 7 (excellent). The Baby FACES study used a modified 32-item ITERS-R scale that excluded all Parents and Staff subscale items. We excluded these items because they rely heavily on staff reports rather than observation. Classroom observations also included counts of children and the adults caring for them that we used to compute child-adult ratios and group sizes. We computed each classroom's full ITERS-R score by averaging the scores on all items collected for that classroom, and computed the six mean subscale scores for each classroom by averaging the classroom's scores on the items in each subscale.

The internal consistency reliability of the ITERS-R total score in our field observation was 0.88 (Table C.2). Author-defined subscale scores had lower internal consistency, with alphas ranging from 0.57 (Program Structure) to 0.76 (Activities). Four of the six subscales had alphas below 0.70 (Space and Furnishings at 0.65, Personal Care Routines at 0.67, Listening and Talking at 0.69, and Program Structure at 0.57). None was higher than 0.80.

Table C.2. Summary Statistics for Baby FACES Child Care Quality Data: ITERS-R, Spring 2009

	Possible Range		Reported Range				
Measure	Min.	Max.	Min.	Max	Mean/ Percentage	Standard Deviation	Cronbach's Alpha
ITERS-R total	1	7	1.86	5.84	3.83	0.80	0.88
Personal care	1	7	1.17	6.50	3.12	1.13	0.67
Furnishings	1	7	1.60	7.00	3.94	1.07	0.65
Listening and talking	1	7	1.33	7.00	4.36	1.20	0.69
Activities	1	7	1.57	6.11	3.51	0.96	0.76
Interaction/social	1	7	1.30	7.00	4.68	1.14	0.71
Program structure	1	7	1.33	7.00	4.19	1.32	0.57
Sample Size	**368**						

Source: Spring 2009 Classroom Observations.

PCRS Assessed the Quality of the Parent-Caregiver Relationship

Communication between parents and teachers/home visitors, as well as agreement between parents and teachers in their attitudes toward child care, has been related to child outcomes. Particularly with home visiting services, the relationship quality between the home visitor and the parent may influence the effectiveness of care and the extent and quality of parent engagement and involvement (Roggman et al. 2008). In fact, the home visitor-parent relationship is associated with parents' engagement and involvement in their child care program (Roggman et al. 2008). This relationship is increasingly emphasized as an important aspect of high-quality programs (Elicker et al. 1997). Emphasis on this relationship highlights the importance of the network of relationships experienced by infants/toddlers and their caregivers.

We included items from PCRS (Elicker et al. 1997) to assess the quality of the relationship between parents and the child's home visitor or teacher, a key aspect of overall service quality. Parents reported on the quality of their relationship with the home visitor or teacher; they in turn provided similar reports on their relationship with the parent. PCRS measures the perceived relationship between the parent and the caregiver (that is, provider, teacher, or home visitor) of infants and toddlers. It was intended to provide focused information on multiple dimensions and specific perceptions of the dyadic relationship. Typically, on such measures caregivers have less positive ratings of parents than parents have of caregivers, with caregivers' ratings varying with demographic characteristics of parents (for example, age, education, income, and marital status). In a study of 217 parents and caregivers (Elicker et al. 1997), 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 not congruent.

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. The spring 2009 Baby FACES instruments included items across these dimensions. The full scale includes 35 items, each comprising a statement about the relationship. 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. For example, they respond to statements such as, "If there is a problem, my child's teacher or home visitor and I always talk about it soon" and "I feel that my child's home visitor or teacher genuinely cares for him/her." Scale scores in Baby FACES represent the average across a subset of these items (six and seven items for staff and parents, respectively). We needed to reduce the burden on respondents and had received a call expressing concern about the appropriateness of some items for staff, and decided to shorten the measure by selecting a subset of items with acceptable internal consistency and reliability and that focused on areas of importance. Selected items will be retained in the parent interview and Staff Child Report (SCR) in future data collection rounds.

Psychometric Properties of Constructs

The following tables present the psychometric data for the constructed variables derived from the parent interview. 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. One exception is the Parental Modernity Scale, in which Cronbach's alphas are close to 0.60 (0.59 and 0.58 for Traditional and Progressive Attitudes, respectively; Table C.12). Additional tables in this chapter present the internal consistency reliability for the constructed measures by child race/ethnicity and DLL status. Generally, the levels of internal consistency are similar across the subgroups.

Next, we describe the measures in detail. We provide information on the selection criteria, normative samples, and psychometric properties reported by the developers of the measures for three child outcome measures: ASQ-3 (Squires 2009), CDI (Fenson 2000), and BITSEA (Briggs-Gowan and Carter 2006). We also provide a brief background for other measures gathered during this wave of data collection**.**

Ages & Stages Questionnaires (Third Edition) (ASQ-3). The ASQ-3 is a parent-report tool for screening infants and young children for developmental delays (ASQ-3) (Squires, Twombly, Bricker, and Potter 2009). The 21 questionnaires included in the ASQ-3, which are appropriate for children ages 1 month to 5-1/2 years, 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 walk along furniture while holding on with only one hand?" on a scale of "Not yet," "Sometimes," or "Most of the time." There are six items in each of the five developmental areas. The raw score in each developmental area could range from 0 to 60, and the total ASQ-3 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 cognitive 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 of) development delays.

The normative sample includes 15,138 children between one and 66 months of age throughout the United States. There are more boys (53 percent) than girls (47 percent) in the sample. Approximately two-thirds of children are white, 12 percent are African American, 15 percent are Hispanic, and the remaining 5 percent are of other races. 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. Inter-rater reliability is less strong; intraclass correlations by area range from 0.43 to 0.69. Cronbach's alphas range from 0.51 to 0.87. ASQ classifications have moderate to high agreement with the Battelle Developmental Inventory (BDI) (Newborg 1984; Newborg 2004) classifications, with an aggregated sensitivity or specificity of 86 percent across all age intervals.

The 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 22.87 for the 10-month form and 15.64 for the 12-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). This means that some children who are developing normally will be classified as needing further assessment (Squires et al. 2009).

Tables C.3, C.4, and C.5 illustrate the average, standard deviation, range, and internal consistency of ASQ-3 scores among children in the Baby FACES study. Cronbach's alphas for the study's sample are similar to previous studies.

Table C.3. Child Cognitive and Language Development

	Possible Range		Reported Range		Mean/ Percentage	Standard Deviation	Cronbach's Alpha
	Min.	Max	Min.	Max.			
ASQ-3[a] Raw Score							
Communication	0	60	0	60	40.32	14.00	0.65–0.73
Gross motor	0	60	0	60	50.67	13.71	0.79–0.85
Fine motor	0	60	5	60	43.40	13.03	0.69–0.73
Problem solving	0	60	0	60	40.23	14.24	0.68–0.77
Personal-social	0	60	0	60	42.96	12.90	0.61–0.70
Total score	0	300	15	300	216.23	50.54	0.78–0.84
ASQ Cutoff Score (2 SDs below the mean or lower)							
Communication	0	1	0	1	7.12	25.74	.
Gross motor	0	1	0	1	10.65	30.88	.
Fine motor	0	1	0	1	13.70	34.42	.
Problem solving	0	1	0	1	20.00	40.04	.
Personal-social	0	1	0	1	8.70	28.21	.
ASQ in the Monitoring Zone (1 to 2 SDs below the mean)							
Communication	0	1	0	1	22.55	41.82	.
Gross motor	0	1	0	1	8.26	27.56	.
Fine motor	0	1	0	1	17.39	37.94	.
Problem solving	0	1	0	1	21.30	40.99	.
Personal-social	0	1	0	1	23.70	42.57	.
CDI[b] (English) Raw Score							
Vocabulary comprehension	0	89	0	89	30.34	20.90	0.98
Vocabulary production	0	89	0	72	2.86	6.31	0.95
CDI[b] (Spanish) Raw Score							
Vocabulary comprehension	0	89	0	89	35.86	22.49	0.98
Vocabulary production	0	89	0	20	2.16	3.66	0.87
Sample Size							
Parent interview	**674**						
Parent interview[c]	**460**						
SCR English CDI	**692**						
SCR Spanish CDI	**113**						

Source: Spring 2009 Parent Interview and Staff Child Report (SCR).

Note: Sample restricted to 1-year-old Cohort. Depending on the age of the child on the day of the parent interview, the age range of children at the baseline required administration of the ASQ-3 10-, 12-, 14-, 16-, or 18-month questionnaire. In error we administered the wrong version of the ASQ to parents of children ages 11 and 12 months in all domains except Communication, and therefore report only Communication scores for this group of children.

[a] Parent report.

[b] Teacher/home visitor report.

[c] Pertains to ASQ Gross Motor, Fine Motor, Problem Solving, and Personal-Social. Excludes 12-month group.

ASQ-3 = Ages & Stages Questionnaires (Third Edition); CDI = MacArthur-Bates Communicative Development Inventories; SD = standard deviation.

Table C.4. Internal Consistency Reliability of Child Cognitive and Language Development Measures, by Race/Ethnicity

	White		African American		Hispanic		Other	
	N	Cronbach's Alpha	N	Cronbach's Alpha	N	Cronbach's Alpha	N	Cronbach's Alpha
ASQ-3[a] Raw Score								
Communication	226	0.64–0.78	121	0.59–0.64	260	0.67–0.68	57	0.65–0.74
Gross motor	156	0.82–0.89	85	0.88–0.91	175	0.69–0.80	35	0.63–0.81
Fine motor	156	0.63–0.76	85	0.77–0.77	175	0.65–0.65	35	0.75–0.86
Problem solving	150	0.68–0.77	83	0.65–0.75	170	0.71–0.78	35	0.62–0.80
Personal-social	156	0.68–0.68	85	0.63–0.74	171	0.54–0.69	35	0.63–0.74
Total score	156	0.79–0.80	85	0.78–0.81	175	0.77–0.82	35	0.72–0.78
CDI[b] (English) Raw Score								
Vocabulary comprehension	224	0.97	124	0.98	244	0.98	58	0.97
Vocabulary production	224	0.94	124	0.97	244	0.89	58	0.77

Source: Spring 2009 Parent Interview and Staff Child Report (SCR).

Note: Sample restricted to 1-year-old Cohort. Depending on the age of the child on the day of the parent interview, the age range of children at the baseline required administration of the ASQ-3 10-, 12-, 14-, 16-, or 18-month questionnaire. In error we administered the wrong version of the ASQ to parents of children ages 11 and 12 months in all domains except Communication, and therefore report only Communication scores for this group of children.

[a] Parent report.

[b] Teacher/home visitor report.

ASQ-3 = Ages & Stages Questionnaires (Third Edition); CDI = MacArthur-Bates Communicative Development Inventories.

Table C.5. Internal Consistency Reliability of Child Cognitive and Language Development Measures, by DLL status

	English		Spanish		Other[c]	
Measures	N	Cronbach's Alpha	N	Cronbach's Alpha	N	Cronbach's Alpha
ASQ-3[a] Raw Score						
Communication	407	0.58–0.77	230	0.64–0.71	27	.
Gross motor	276	0.85–0.88	158	0.68–0.80	17	.
Fine motor	276	0.64–0.75	158	0.65–0.69	17	.
Problem solving	270	0.66–0.85	153	0.75–0.79	15	.
Personal-social	276	0.65–0.71	154	0.56–0.70	17	.
Total score	276	0.79–0.86	158	0.77–0.80	17	.
CDI[b] (English) Raw Score						
Vocabulary comprehension	409	0.98	215	0.98	26	0.98
Vocabulary production	409	0.95	215	0.90	26	0.90

Source: Spring 2009 Parent Interview and Staff Child Report (SCR).

Note: Sample restricted to 1-year-old Cohort. Depending on the age of the child on the day of the parent interview, the age range of children at the baseline required administration of the ASQ-3 10-, 12-, 14-, 16-, or 18-month questionnaire. In error we administered the wrong version of the ASQ to parents of children ages 11 and 12 months in all domains except Communication. Cronbach's alphas are computed separately by age group for each developmental area. (Cronbach's alphas are not computed for the 12-month group for areas other than Communication.)

[a] Parent report.
[b] Teacher/home visitor report.
[c] Cronbach alphas are not computed for the ASQ-3 because of small sample sizes by age group.
ASQ-3 = Ages & Stages Questionnaires (Third Edition); CDI = MacArthur-Bates Communicative Development Inventories; N = number.

MacArthur-Bates Communicative Development Inventory (CDI). The CDI is designed to assess children's early receptive and expressive language and communication skills through parent report (Fenson et al. 2000). In the baseline wave of Baby FACES, teachers and home visitors completed the English Infant Short Form (an 89-word vocabulary checklist for 8- to 18-month-olds) for all children. Two measures were derived from this form:

1. *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 89 specific words.
2. *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 89 specific words. The raw scores for both Vocabulary Comprehension and Vocabulary Production range from 0 to 89.

The CDI was used successfully in the 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 home visitors provide data on this instrument made this an attractive measure of language development in Baby FACES.

The norming sample for the English Infant Short Form includes 481 infants between 8 and 18 months of age from three locations in the United States: New Haven, Connecticut; Seattle, Washington; and San Diego, California. The majority (89 percent) of children are white. Black and Asian children each comprise 3 percent of the sample. The remaining 5 percent are of other races. More than half (53 percent) of parents hold a college diploma, and only 2 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 normative sample was also limited to children whose primary language was English. Approximately 14 percent of the infants in the sample had exposure to more than one language.

Cronbach's alpha was 0.97 for the infant form in the normative sample. The correlations between the short and long infant forms are 0.88 for vocabulary comprehension and 0.90 for vocabulary production, suggesting that the short form provides an effective alternative to the long form.

Baby FACES used the Spanish infant form from the EHSREP, which was a direct translation of the English form and not the official Spanish version. Teachers and home visitors who reported they spoke Spanish also completed the Spanish form for children identified as understanding Spanish (staff completed Spanish CDIs for 137 children). There are 86 words in the Baby FACES Spanish CDI form administered at the baseline that overlap with the full-version official Spanish CDI. The developer of the full official version of the Spanish CDI helped us create the virtual norms for the Baby FACES Spanish CDI by using the 86 overlapping words in our version and the full version (Jackson-Maldonado 2003).

We also derived the CDI conceptual scores for Spanish-speaking children. For each word in the 89-word checklist, we coded the child as understanding or producing the word concept if the Early Head Start staff reported that the child understood or produced the word in English and/or Spanish. The conceptual scores range from 0 to 89.

The normative sample for the Spanish CDI full infant form includes 778 children between 8 and 18 months of age from eight cities in Mexico. About one-third of mothers have some college education and another one-third have not completed high school. The normative sample was limited to children for whom Mexican Spanish was the primary language. About 6 percent of children were exposed to a second language.

At Baby FACES baseline, the correlations of CDI scores with ASQ-3 Communication are small; this indicates little or no relationship between them. The ASQ-3 Communication scores were correlated with English Vocabulary Comprehension at 0.08 and with English Vocabulary Production at 0.14. The correlations with ASQ-3 Communication were 0.02 for Spanish Vocabulary Comprehension and 0.12 for Spanish Vocabulary Production.

Tables C.3, C.4, and C.5 illustrate the average, standard deviation, range, and internal consistency of CDI scores among children in the Baby FACES study.

Brief Infant Toddler Social-Emotional Assessment (BITSEA). The BITSEA (Briggs-Gowan and Carter 2006) is the screener version of the longer 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 to 36 months old. 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 both social competencies and behavior problems, such as internalizing and externalizing behavior. Spanish language availability and the possibility of administering it to both parents and staff provided further reason to employ the BITSEA.

The 31-item BITSEA Problem scale assesses social-emotional/behavioral problems such as aggression, defiance, over-activity, negative emotionality, anxiety, and withdrawal. Higher scores indicate more problems. The 11-item BITSEA Competence scale assesses social-emotional abilities such as empathy, pro-social behaviors, and compliance. Lower scores indicate lesser 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 the baseline wave of data collection. 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 lower than the 15th percentile. 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, high problems and/or low competence) indicates "screening positive" on the BITSEA.

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. The 12- to 17-month sample has a racial breakdown of 56 to 60 percent white, 16 percent African American, 20 to 21 percent Hispanic, 4 to 5 percent

Asian, and 0 to 1 percent of another race. This matches the 2002 U.S. Census. In addition, about 60 percent of children's parents have at least some college education, about one-quarter (25 to 27 percent) have completed high school, and approximately 13 to 15 percent of parents have less than a high school education.

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 (Cronbach's alpha of 0.79 for the Problem scale and 0.65 for the Competence scale on the Parent Form, as well as Cronbach's alpha of 0.80 for the Problem scale and 0.66 for the Competence scale on the Childcare Provider Form) (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 2002) (r = 0.55 and r = -0.55, respectively). The correlations between the BITSEA Problem score and the Child Behavior Checklist 1.5–5 (CBCL 1.5–5) (Achenbach 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 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 Competence and from -0.19 to -0.28 for Problem. 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 (0.59) and typical for Problems (0.28) because children may behave differently in the two contexts. In Baby FACES, the correlations between parent ratings and Early Head Start staff ratings are lower than those found in the BITSEA validation study. We find that the parent report is not correlated with the staff report on the Problem scale (*r* = 0.01; nonsignificant), and the correlations between ratings on the Competence scale from the two sources are low (*r* = 0.18; *p* < .001). Although still low, for the Problem scale, home visitor ratings are more highly correlated with parent ratings (*r* = 0.13), than are teacher ratings with the parent ratings (*r* = -0.09) on the Problem scale; for the Competence scale, teacher ratings are more highly correlated with parent ratings (*r* = 0.19) than are home visitor ratings and parent ratings (*r* = 0.17).

Tables C.6, C.7, and C.8 illustrate the average, standard deviation, range, and internal consistency of BITSEA scores among children in the Baby FACES study.

Table C.6. Child Social Emotional Development

	Possible Range		Reported Range				
Outcome	Min.	Max.	Min.	Max.	Mean/ Percentage	Standard Deviation	Cronbach's Alpha
Parent Interview BITSEA Raw Score							
Problem domain	0	62	0	40	10.57	6.31	0.79
Competence domain	0	22	5	22	16.16	3.38	0.66
SCR BITSEA Raw Score							
Problem domain	0	62	0	27	6.29	4.70	0.78
Competence domain	0	22	0	22	12.82	3.53	0.73
Parent Interview BITSEA Cutoff Score							
Problem domain	0	1	0	1	26.80	44.33	.
Competence domain	0	1	0	1	9.81	29.76	.
SCR BITSEA Cutoff Score							
Problem domain	0	1	0	1	13.62	34.33	.
Competence domain	0	1	0	1	14.65	35.39	.
Parent Interview BITSEA Screen Positive	0	1	0	1	32.99	47.05	.
SCR BITSEA Screen Positive	0	1	0	1	24.77	43.20	.
Sample Size							
Parent Interview	**679**						
SCR	**740**						

Source: Spring 2009 Parent Interview, Staff Child Report (SCR).

Note: Sample restricted to 1-year-old Cohort.

BITSEA = Brief Infant-Toddler Social and Emotional Assessment; SCR = Staff Child Report.

Table C.7. Internal Consistency Reliability of Child Social Emotional Development Measures, by Race/Ethnicity

	White		African American		Hispanic		Other	
Measures	N	Cronbach's Alpha	N	Cronbach's Alpha	N	Cronbach's Alpha	N	Cronbach's Alpha
Parent Interview BITSEA Raw Score								
Problem domain	228	0.77	125	0.78	268	0.77	60	0.85
Competence domain	228	0.64	125	0.69	268	0.65	60	0.71
SCR BITSEA Raw Score								
Problem domain	224	0.78	125	0.78	245	0.69	59	0.79
Competence domain	224	0.75	125	0.76	245	0.74	59	0.61

Source: Spring 2009 Parent Interview, Staff Child Report (SCR).

Note: Sample restricted to 1-year-old Cohort.

BITSEA = Brief Infant-Toddler Social and Emotional Assessment; N = number; SCR = Staff Child Report.

Table C.8. Internal Consistency Reliability of Child Social Emotional Development Measures, by DLL Status

	English		Spanish		Other	
Measures	N	Cronbach's Alpha	N	Cronbach's Alpha	N	Cronbach's Alpha
Parent Interview BITSEA raw score						
Problem domain	416	0.78	237	0.77	28	0.86
Competence domain	416	0.65	237	0.64	28	0.83
SCR BITSEA Raw Score						
Problem domain	411	0.77	216	0.69	26	0.87
Competence domain	411	0.74	216	0.75	26	0.69

Source: Spring 2009 Parent Interview, Staff Child Report (SCR).

Note: Sample restricted to 1-year-old Cohort.

BITSEA = Brief Infant-Toddler Social and Emotional Assessment; N = number; SCR = Staff Child Report.

Center for Epidemiologic Studies Depression Scale (CES-D). The CES-D (Radloff 1977) is a self-administered screening tool used to identify symptoms of depression or psychological distress. 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 4-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 or higher but lower than 15 are considered as having moderate depressive symptoms; and those who score between 5 and 10 are considered as having mild depressive symptoms.

Tables C.9, C.10, and C.11 illustrate the average, standard deviation, range, and internal consistency of CESD-SF scores among parents in the Baby FACES study.

Table C.9. Parent Mental Health

	Possible Range		Reported Range				
Outcomes	Min.	Max.	Min.	Max.	Mean/ Percentage	Standard Deviation	Cronbach's Alpha
CESD-SF raw score	0	36	0	35	5.46	5.64	0.84
CESD-SF: severe depressive symptoms	0	1	0	1	7.89	26.97	.
CESD-SF: mild depressive symptoms	0	1	0	1	24.76	43.19	.
CESD-SF: no depressive symptoms	0	1	0	1	57.65	49.44	.
PSI: parental distress	5	25	5	25	10.86	4.64	0.73
PSI: parent-child dysfunctional interaction	6	30	6	30	8.79	4.15	0.78
FES-family conflict	1	4	1	4	1.58	0.52	0.70
Social support	13	39	13	39	30.91	7.40	0.93
Parenting alliance measure	10	50	10	50	45.96	5.81	0.94
Sample Size							
Parent Interview	**825**						
FES-family Conflict[a]	**155**						

Source: Spring 2009 Parent Interview.

Note: 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.

[a]Asked only of the Newborn Cohort in spring 2009.

CESD-SF = Center for Epidemiologic Studies Depression Scale Short Form; PSI = Parenting Stress Index.

Table C.10. Internal Consistency Reliability of Parent Mental Health and Family Functioning Measures, by Race/Ethnicity

	White		African American		Hispanic		Other	
Outcome	N	Cronbach's Alpha	N	Cronbach's Alpha	N	Cronbach's Alpha	N	Cronbach's Alpha
Parent's Mental Health								
PSI: parental distress	219	0.68	121	0.77	261	0.74	59	0.68
PSI: parent-child dysfunctional interaction	219	0.61	121	0.66	260	0.82	59	0.81
CESD-SF raw score	251	0.85	150	0.82	310	0.85	79	0.80
Family Functioning								
FES-family conflict[a]	35	0.76	31	0.68	54	0.68	19	0.79
Social support	253	0.92	151	0.93	306	0.94	78	0.92
Parenting alliance measure	138	0.95	31	0.91	178	0.93	27	0.91

Source: Spring 2009 Parent Interview.

[a] Asked only of 1-year-old Cohort in spring 2009.

CESD-SF = Center for Epidemiologic Studies Depression Scale Short Form; FES = Family Environment Scale; PSI = Parenting Stress Index.

Table C.11. Internal Consistency Reliability of Parent Mental Health and Family Functioning Measures, by DLL Status

	English		Spanish		Other	
Outcome	N	Cronbach's Alpha	N	Cronbach's Alpha	N	Cronbach's Alpha
Parent's Mental Health						
PSI: Parental Distress	402	0.72	231	0.73	27	0.74
PSI: Parent-Child Dysfunctional Interaction	402	0.67	230	0.83	27	0.75
CESD-SF Raw Score	508	0.84	280	0.84	28	0.63
Family Functioning						
FES-Family Conflict[a]	102	0.68	52	0.72	2	.
Social Support	507	0.93	279	0.94	28	0.92
Parenting Alliance Measure	185	0.94	168	0.94	21	0.91

Source: Spring 2009 Parent Interview.

[a] Asked only of 1-year-old Cohort in spring 2009.

CESD-SF = Center for Epidemiologic Studies Depression Scale Short Form; FES = Family Environment Scale; N = number; PSI = Parenting Stress Index.

The Parenting Stress Index—Short Form (PSI-SF). The PSI-SF measures the degree of stress in parent-child relationships stemming from three possible sources: (1) the child's challenging temperament, (2) parental depression, and (3) negatively reinforcing 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 parent is feeling in his or her role as a parent, including a low sense of competence and a high level of stress because of 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 high 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 interactions with the child do not reinforce the parent. The parent answers whether he or she agrees or disagrees 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 higher levels of parent-child dysfunctional interaction.

Tables C.9, C.10, and C.11 illustrate the average, standard deviation, and range of PSI scores among parents in the Baby FACES study.

The Family Environment Scale, Family Conflict Subscale (FES) (Moos 2002) was designed to measure the extent to which the open expression of anger and aggression and conflict-filled interactions are characteristic of the family. Parents rated 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. We included the FES because it had been previously included in the EHSREP. Short administration time, acceptable reliability, and the inclusion of distressed families within the normative sample all contributed to our decision to employ this measure.

The Parenting Alliance Measure (PAM). The PAM (Abidin 1999) is a 20-item self-report instrument that measures a parent's perspective on how cooperative, communicative, and mutually respectful he or she is with his or her partner in regard to caring for their children. We included 10 items of the PAM in Baby FACES. Parents responded to items such as "(The father/mother) and I are a good team" and "(The father/mother) makes my job of being a parent easier" on a five-point rating scale ranging from "strongly agree" to "strongly disagree." The items are reverse coded and raw scores range from 10 to 50, with higher scores indicating stronger and more positive parenting alliance.

Tables C.9, C.10, and C.11 illustrate the average, standard deviation, range, and internal consistency of FES and PAM scores among parents in the Baby FACES study.

In addition to the measures presented earlier, we developed the following three measures of parent mental health specifically for Baby FACES data collection and analysis:

Social Support. We measured social support by asking parents questions about whether there is someone they can count on for physical and emotional help. Parents rated the 13 items on a three-point scale ranging from "not at all," to "sometimes," to "all or most of the time." Raw scores range from 13 to 39, with higher scores indicating higher levels of social support.

Problems with People. Parents reported whether they are having problems with a range of different people (including their neighbors, landlord, current or past spouse or partner, others living in the home, bill collectors, or coworkers). We present the proportion of parents who reported they are having no problems with any of these people.

Community Participation. Parents are asked about their participation in community organizations spanning many different areas (church/religion; a community group, such as tenant association; a school group, such as PTA, Early Head Start, or another early childhood parent group; or a political advocacy group). We present the proportion of parents who reported that they participated in any of these organizations.

The Parental Modernity Scale (PMS). The PMS (Schaefer 1985) is a 30-item measure of parents' attitudes toward children and childrearing practices (traditional, authoritarian parental beliefs and progressive, democratic beliefs). Parents responded to items on a five-point scale ranging from "strongly disagree" to "strongly agree." We included 10 of the 30 items in Baby FACES, yielding two subscales: (1) Traditional Beliefs, and (2) Progressive Beliefs. Raw scores range from 5 to 25 for each scale, with higher scores indicating more traditional beliefs and more progressive beliefs, respectively.

Tables C.12, C.13, and C.14 illustrate the average, standard deviation, range, and internal consistency of PMS scores among parents in the Baby FACES study.

Table C.12. Parenting Outcomes

Outcome	Min.	Max.	Min.	Max.	Mean/ Percentage	Standard Deviation	Cronbach's Alpha
Parental Modernity Scale							
Traditional Beliefs	5	25	5	25	19.78	3.55	0.59
Progressive Beliefs	5	25	5	25	20.07	3.45	0.58
Sample Size							
Parent Interview	**654**						

Source: Spring 2009 Parent Interview.

Table C.13. Internal Consistency Reliability of Parenting Outcomes Measures, by Race/Ethnicity

	White		African American		Hispanic		Other	
Outcome	N	Cronbach's Alpha	N	Cronbach's Alpha	N	Cronbach's Alpha	N	Cronbach's Alpha
Parental Modernity Scale								
Traditional Beliefs	219	0.58	121	0.54	261	0.62	58	0.64
Progressive Beliefs	220	0.57	121	0.58	261	0.53	59	0.58

Source: Spring 2009 Parent Interview.

N = number.

Table C.14. Internal Consistency Reliability of Parenting Outcomes Measures, by DLL Status

	English		Spanish		Other	
Outcome	N	Cronbach's Alpha	N	Cronbach's Alpha	N	Cronbach's Alpha
Parental Modernity Scale						
Traditional Beliefs	402	0.58	231	0.59	26	0.73
Progressive Beliefs	403	0.61	231	0.51	27	0.64

Source: Spring 2009 Parent Interview.

N = number.

Spanked Child in Past Week measures a parent's report that he or she used physical punishment in the past week by spanking the child.

Implementation

The main source of data on program implementation is the Program Director Self-Administered Questionnaire (SAQ). In the SAQ, we asked directors to evaluate their programs by rating them on individual items that make up five cornerstones of program implementation These cornerstones are Child Development, Family Development, Staff Development, Community Building, and Management Systems and Procedures.[3] Within each cornerstone, the director ranked

[3] The Child Development cornerstone elements were frequency of child development services and developmental assessments, availability of health services, child care and group socializations, the level of parent involvement in child development services, and the degree of individualization of services. The Family Development cornerstone focused on the presence of individualized family partnership agreements, the availability and frequency of family development services, and the level of parent involvement in the program. The Staff Development cornerstone consists of quantity of supervision, training, and turnover. The Community Building cornerstone involves the quantity and quality of collaborative relationships between the program and other service providers and the existence of transition plans for children approaching their third birthday. The Management Systems and Procedures cornerstone elements were existence and quality of a communication system, goals and objectives, a self-assessment, and a community needs assessment.

the program's implementation of individual elements on a scale of 1 to 5; a score of 4 signifies that the program is fully implemented and meets the performance standards for that element; a score of 5 signifies that the program exceeded the standards ("enhanced" implementation). Each cornerstone score was the average score for all its program elements. In the Survey of Early Head Start Programs (SEHSP), these cornerstones were the basis for classifying programs' implementation. The implementation ratings originated from the initial implementation study in the EHSREP and were used by researchers to summarize the data they collected through interviews and record reviews over several days. These ratings have been very useful as a program subgroup in the impact study and we adapted these sheets as an SAQ that program directors completed on their own.

We administered the implementation ratings quite differently than in past studies. In the EHSREP research, staff completed the forms based on a great deal of in-depth information gathered in site visits. In a small pilot test in the SEHSP, 17 program directors completed the form in its current version, but did so in the company of a researcher who discussed each item with them to help them choose the proper rating. In the current study, directors completed the form without any assistance or guidance.

Missing Values and Imputation

We followed rules to guide our approach to missing data that balanced the objective of retaining as many sample members' data as possible without unduly compromising the accuracy of the data. Missing values for scale elements could have arisen if a parent or staff did not know an answer or if a respondent refused to answer or otherwise skipped the question. Noncompletion of the interview also produced missing values for measures.

For example, in the BITSEA, "don't know" responses arose primarily for questions that asked about whether a child did things such as "point[ed] to show something far away" or "won't touch objects because of how they feel." The interview format prevented parents and teachers from checking if a child acted in a certain way when presented with an unfamiliar situation. Refusal tended to arise from more personal questions. On the BITSEA, the sole "refused" response came on a question about whether the child hits, kicks, or bites the parent. "Not applicable" responses came for questions that hinged on the child's interactions with other nonsibling children. Because children in home-based care may not have had the opportunity to interact with other children, parents would have no basis for answering the question. We asked BITSEA questions only of parents and teachers/home visitors of 1-year-old Cohort children. As a result, no Newborn Cohort children had a BITSEA score.

Because most measures involve summing scores for a number of individual elements, simply ignoring elements with missing values would bias estimates downward for individuals with one or more missing values within their scale. We have chosen to impute missing values with the mean of the remaining measure elements for that individual. The high alphas for most of the scales provide a theoretical justification for applying this rule. In keeping with the 25 percent rule mentioned earlier, imputation occurred only if fewer than 25 percent of the items that make up the scale were missing. Responses missing because the questions were not applicable did not count as truly missing, and we did not impute values for them. Table C.15 shows a list of measures in which imputation occurred, the number of observations with nonmissing values for that measure, and the percentage of observations in which we imputed one or more scale elements.

Table C.15. Imputation for Key Baby FACES Measures

Measure	Number of Nonmissing Values for Scale
Parent Interview	
BITSEA Problem domain	679
BITSEA Competence domain	673
CES-D long form	837
CESD-SF (short form)	837
PMS-Traditional	659
PMS-Progressive	661
PSI-Parent-Child Dysfunctional Interaction	659
PSI-Parental Distress	660
Family conflict	158
Sources of social support	828
Parenting alliance measure	374
Family financial difficulties	825
Family food insecurities	824
Staff-parent relationship short form	694
ASQ-3 Communication	675
ASQ-3 Gross Motor	675
ASQ-3 Fine Motor	674
ASQ-3 Problem Solving	672
ASQ-3 Personal-Social	674
Maternal risk	666
Staff Child Report	
Parent-Caregiver Relationship short form	737
Teacher/Home Visitor Interview	
CESD-SF	550

ASQ-3 = Ages & Stages Questionnaires (Third Edition); BITSEA = Brief Infant Toddler Social Emotional Assessment; CES-D = Center for Epidemiologic Studies Depression Scale; PMS = Parental Modernity Scale; PSI = Parenting Stress Index.

APPENDIX D

ANALYTICAL ISSUES

APPENDIX D. ANALYTICAL ISSUES

As we processed and analyzed Baby FACES survey data, we addressed a few conceptual and administration issues that arose. We devoted significant effort to exploring these issues fully to ensure that we understand our data and can stand behind our findings. These analyses also informed our revisions of survey instruments for future rounds of data collection. Next, we briefly discuss five analytical issues that arose. These issues are (1) categorizing families by service approach, (2) examining variability in self-reported program implementation ratings, (3) comparing ITERS-R scores with earlier ITERS scores in the Early Head Start Research and Evaluation Project (EHSREP), (4) conducting a factor analysis of ITERS-R data, and (5) addressing errors in administration of the MacArthur-Bates Communicative Development Inventories (CDI) and the Ages & Stages Questionnaires, Third Edition (ASQ-3).

Family-Level Early Head Start Service Approach

There are two main data sources regarding Early Head Start program service approach: the program director screener and the parent interview:

Program Director Screener. The screener administered as part of the program director interview captured information on program service approach at the *program level.* Directors were asked to indicate whether they offer only center-based services (center-based programs), only home-based services (home-based programs), or both center-based and home-based services to families ("multiple-approach" programs)[1]. Directors of multiple-approach programs were asked whether all families get either center- or home-based services, or whether some families simultaneously get both center- *and* home-based services ("combination" services). Finally, directors of programs offering more than one type of service were asked to indicate the percentage of families that received each type of service.

Parent Interview. The parent interview captured information on program service receipt at the *family level.* Parents of 1-year-old children indicated whether they received center-based services, home-based services, family child care services, or another type of service.[2] Parents also indicated the frequency of their receipt of services, both in terms of center attendance and home visits. Parents were not given the option of reporting simultaneous center- and home-based services (combination services). However, the frequencies of service receipt can be used to define whether each family receives strictly home-based, strictly center-based, or combination services.

[1] In the Survey of Early Head Start Programs (SEHSP), programs that offered both center-based and home-based services to families could be "multiple" or "combination." Programs that offered regular center-based services and monthly or more frequent home visits to *all* families were termed "combination" programs. Those that offered both home- and center-based options, but did not provide both simultaneously to *all* enrolled families were termed "multiple." In this study, there are no combination programs meeting the SEHSP definition. As such, our basic classification of service approach at the program level is restricted to center-based, home-based, and multiple programs. However, families can report "combination" services, in which their children receive center- and home-based services simultaneously. The two locally designed programs included in our sample were classified as multiple-approach programs using our definition.

[2] Pregnant mothers and parents of newborns were not asked questions about program services in the parent interview.

Table D.1 shows the number of programs and families in each service approach based on data from the program director screener. Of 89 programs in the sample, 17 percent are center-based, 16 percent are home-based, and 67 percent are multiple-approach programs. Of the 978 consenting families in the study, center-based and home-based programs account for about one-third of families; multiple-approach programs account for two-thirds of families. Distinguishing family-level service approach from program-level approach is problematic for most families in our sample, as we cannot easily classify families in multiple-approach programs as center-based, home-based, or combination. Instead, we relied on reported parent interview data for those families enrolled in multiple-approach programs.

Table D.1. Most Programs Provide Multiple Service Options

Approach	Number of Programs	Percentage of Programs	Number of Families	Percentage of Families
Center-based	15	17	193	20
Home-based	14	16	126	13
Multiple-approach	60	67	659	67
Families receive either center-based or home-based services	52	58	n.a.	n.a.
Some families receive combination services	8	9	n.a.	n.a.
Total	**89**	**100**	**978**	**100**

Source: Spring 2009 Program Director Interview.

n.a. = not available.

We Used Information from Parents to Complement Information from the Program Director Screener

Later in this section, we outline our method of classifying receipt of Early Head Start services at the family level. This method involved combining data from the director and parent interviews in the following ways for families with 1-year-olds:

For all families served by programs that reported offering exclusively center-based services or exclusively home-based services, we assigned families the classification reported by the director—either center- or home-based services. By definition, this is the only type of service these families can receive.[3]

For all families served by multiple-approach programs, we examined parents' self-reported frequency of service receipt. If a family reports at least one home visit a month and does not report receiving center-based care, it is defined as primarily home-based. If a family reports any weekly center visits and home visits less than monthly, it is defined as primarily center-based. If a family

[3] There is one minor exception to this rule: families that reported being serviced by a family child care provider kept this designation regardless of the directors' reported program approach.

reports both weekly center-based services and at least one home visit a month, it is defined as combination.[4]

If we could not determine the service type of families in multiple-approach programs from their reported service usage, we assigned them the program approach that they initially reported in the parent interview—either center-based, home-based, or family child care.[5]

Combining the data this way produced a higher degree of confidence that we had classified each family correctly. Because directors have the best knowledge of their programs' services, we preferred their answers in situations in which they allowed us to assign child service type definitively. In situations in which the programs offer multiple approaches, parent reports on the frequency of service receipt provided additional information to make a more accurate family-level classification. Relying solely on parent reports is problematic because the questions asked in the parent interview did not provide sufficient detail for parents to differentiate between child care and visits to the center for services other than child care or for group socializations. The questions also did not distinguish between a home visit and other types of visits that staff might make to the home, such as for child screening and assessments. For these reasons, we consulted parents' reported service receipt only in cases in which information provided by directors was insufficient to determine a family's service option. Finally, our interview instrument did not define the levels of service required to be considered center-based or home-based with sufficient precision to use the responses as our first choice for constructing the family service receipt variable.

Table D.2 provides a comparison of director-reported program services and our family-level service receipt classification scheme. According to the criteria outlined earlier, families in multiple-approach programs have been sorted into center-based, home-based, and combination services. A total of 28 families in seven multiple-approach programs reported at least one home visit a month as well as weekly center visits, and were thus classified as receiving combination services. The majority of families, however, were classified either as home-based (364 families) or center-based (332 families) using the criteria.

Figure D.1 illustrates our proposed service receipt classification at the family level.

[4] Families could be classified as combination only if they were in a program in which the director reported that combination services were provided to some families.

[5] In item C2.1 of the parent interview, the parents were asked to choose which of these three types of care best describe what the child receives from the program. Combination care was not an option, although some parents did select "other program option" and verbally specified a combination program. Parents specifying combination services were classified as such.

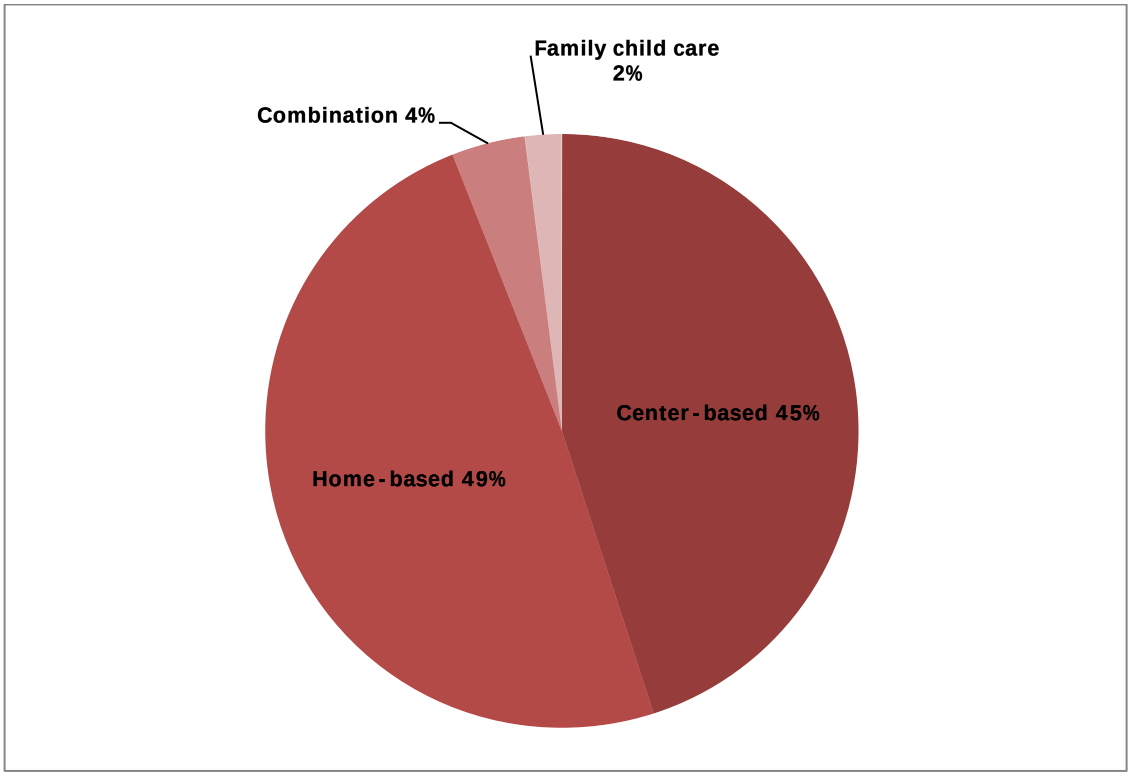
Table D.2. Families Are Nearly Evenly Split Between Center-Based and Home-Based Care Options

Approach According to Directors	Number of Families in Each Approach				
	Center-Based	Home-Based	Combination	Family Child Care	**Total**
Center-Based	141	0	0	0	141
	100%	0%	0%	0%	100%
Home-Based	0	107	0	0	107
	0%	100%	0%	2%	100%
Multiple-Approach	194	260	29	5	487
	40%	54%	6%	1%	100%
Total	334	367	29	5	735
Total Percentage	**44%**	**52%**	**4%**	**1%**	**100%**

Source: Spring 2009 Program Director Interview, Spring 2009 Parent Interview.

Note: There are 46 missing values; these values reflect all families in multiple-approach programs that did not complete service receipt questions in the parent interview.

Figure D.1. Families Split Nearly Evenly Between Home- and Center-Based Care



Sources: Spring 2009 Program Director Interview, Spring 2009 Parent Interview.

Program-Level Implementation

We have not systematically collected information on program implementation since the EHSREP and explored ways to do so in the nationally representative Baby FACES sample. In the EHSREP, researchers developed elements of program implementation derived from the performance standards and from the conceptual framework. The cornerstones include child development, family development, staff development, community building, and management systems and procedures.[6] After intensive site visits that included interviews and record reviews, researchers used a consensus process to rate implementation on each cornerstone element, and then create cornerstone ratings that average elements within each cornerstone. Final classification of programs' implementation in the ESHREP averaged cornerstone ratings with the child development cornerstone weighted twice as much as the others. These ratings have been very useful as a program subgroup in the impact study; we hoped we could use the rating forms as a self-administered questionnaire (SAQ) that program directors completed on their own.

In Baby FACES we used the same overall rating form with labeled columns for rating each element on a scale of 1 to 5—from low to enhanced implementation—as part of an SAQ.[7] We asked directors to evaluate their programs by rating them on these items and noted in the instructions that a rating of 5 went *beyond* the requirements in the performance standards.[8]

Results of the implementation ratings indicated that most program directors report programs are generally very well implemented in every cornerstone (Table D.3), although variability is limited. To further investigate the ratings, we examined overall and cornerstone scores relative to other, related items in the director survey. We found limited correspondence between either individual element ratings, cornerstone scores, or overall implementation and related items (for example, rate of staff turnover, quality outcomes, and parental participation). The restricted range of the ratings may also have reduced correlations with other variables.

[6] The child development cornerstone elements were frequency of child development services and developmental assessments, availability of health services, child care and group socializations, the level of parent involvement in child development services, and the degree of individualization of services. The family development cornerstone focused on the presence of individualized family partnership agreements, the availability and frequency of family development services, and the level of parent involvement in the program. The staff development cornerstone consists of quantity of supervision, training, and turnover. The community building cornerstone involves the quantity and quality of collaborative relationships between the program and other service providers and the existence of transition plans for children approaching their third birthday. The management systems and procedures cornerstone elements were existence and quality of a communication system, goals and objectives, a self-assessment, and a community needs assessment.

[7] In the earlier SEHSP study we explored whether we could obtain reasonable program implementation information without the intensive process used in the EHSREP. Program directors completed summary rating forms that EHSREP researchers used to collate data on all the individual elements, by cornerstone. At the end of the site visit a researcher discussed the implementation rating form with the program directors to help them determine the proper rating.

[8] Our observations that some program directors who were at full implementation for a given element disliked that they were not at the highest rating even though they were meeting all the requirements in the performance standards. Because the format of the administration was as part of an interview, the researcher was able to describe the rationale for the ratings with directors. We attempted to address the possibility of inflated ratings in the instruction indicating that "enhanced" implementation included things that went beyond the standards.

In part because of the limited range in ratings and low correspondence between ratings and related interview items, we speculate that the format and presentation of the implementation items may have created a positive response demand. That is, because directors completed the form without any assistance or guidance and because responses to all cornerstone items were ordered and labeled from "Low" to "High" implementation, directors might have been very conscious of how their responses to cornerstone items would affect their overall implementation scores. Perhaps partly due to the presentation of the cornerstone elements, director scores are highly concentrated in the 4 to 5 range for all cornerstones (between full and enhanced implementation), as shown in Table D.3.

Further, the rating form was developed as a research tool, not as a survey instrument, and directors might have been confused by the complexity of the descriptors that accompanied ratings for each element. Most anchor descriptors included in the individual elements actually consisted of several dimensions rolled into a single rating. For instance, for a program to achieve a score of 3 on the "frequency of child development services" element required "most" families to receive both "two instances of child development services" and "monthly parent education." Moving up to a score of 4 requires "almost all" families to "receive child development services three times a month." Directors with programs that do not fit neatly within any single score description might have had difficulty rating some elements.

The relatively low variability in ratings coupled with a lack of strong correlations between ratings of cornerstone items and survey information suggests that the implementation self-ratings from this round of data collection not be treated as a definitive measure of program implementation. Instead we recommend continuing to explore ways to collect this important information. For the next wave of data collection in spring 2010, we will adapt the director interview and implementation rating forms to more concretely and explicitly capture the individual cornerstone elements. For instance, each item will include a rating on only one dimension, and we will administer the items as part of the telephone interview rather than as an SAQ, to allow for questions, and eliminate the visual presentation in terms of low to high levels of implementation. We do not recommend using program implementation as a subgroup for this baseline round of analysis.

Table D.3. Most Directors Rate Their Programs as Fully Implemented at Each Cornerstone

	Cornerstone									
	Child Development		Family Development		Staff Development		Community Building		Management Systems	
Score Range	N	Percentage	N	Percentage	N	Percentage	N	Percentage	N	Percentage
2–2.99	0	0	4	4	1	1	0	0	1	1
3–3.99	15	17	11	13	12	13	4	4	1	1
4–4.99	69	78	62	70	59	67	27	31	39	44
5	1	1	8	9	12	13	54	61	44	50
Missing	4	4	4	4	5	6	4	4	4	4
Total	**89**	**100**	**89**	**100**	**89**	**100**	**89**	**100**	**89**	**100**

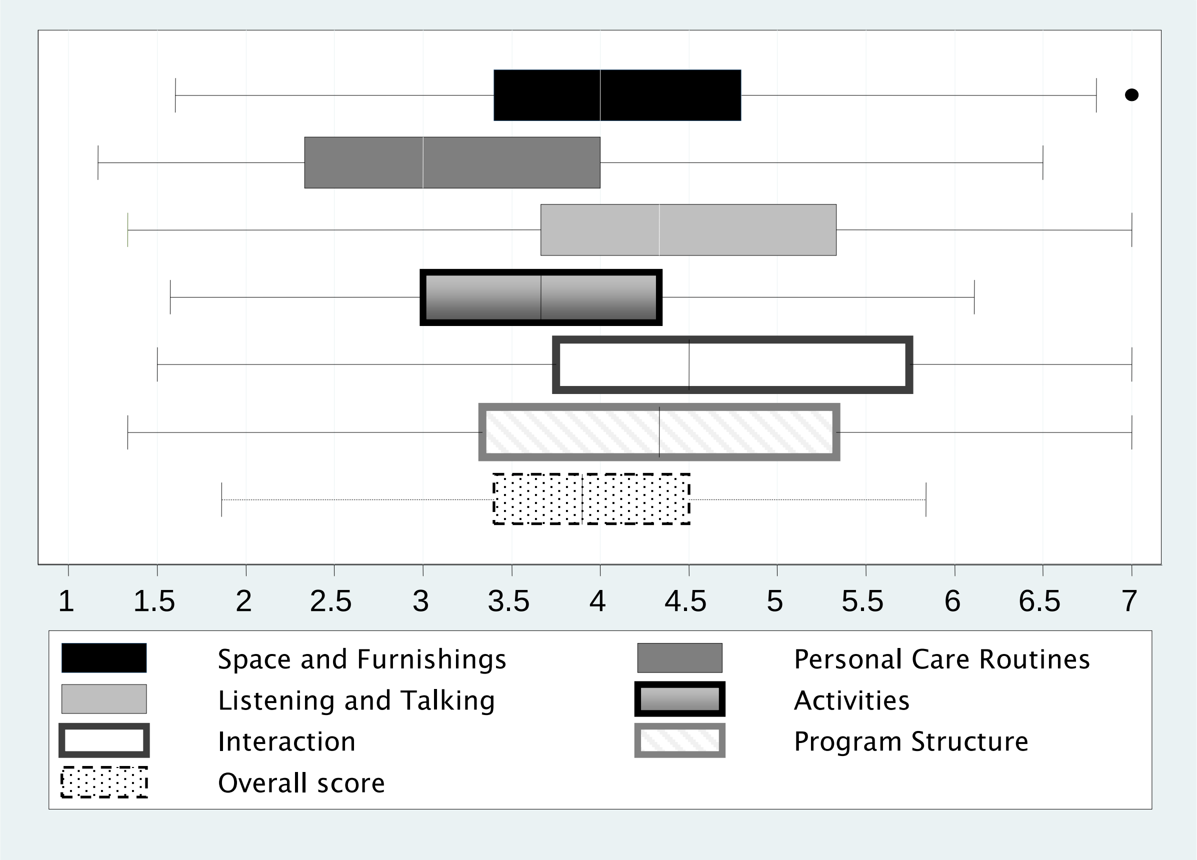
ITERS and ITERS-R Scores

In this section, we present a brief analysis of Early Head Start programs' scores on ITERS-R (Harms, Cryer, and Clifford 2006). First, we compare Early Head Start programs' ITERS-R scores from Baby FACES data collection in 2009 to Early Head Start programs' scores on the original ITERS instrument (ITERS; Harms, Clifford, and Cryer 1990) from the EHSREP. Next, we present an item-level comparison of ITERS-R scores to ITERS scores.

Median ITERS-R Scores from Baby FACES Are a Point Lower than Median ITERS Scores from the EHSREP

Figure D.2 displays the distribution of subscale and overall ITERS-R scores for the 223 Early Head Start classrooms that were observed and rated during spring 2009. Half of classroom scores fall within the shaded boxes (the line in the shaded region represents the median score); the horizontal lines provide the range of ITERS scores in each subscale (dots outside of the range indicate outlier scores). As illustrated, median scores for most subscales fall between 3.5 and 4.5 (between minimum and good quality levels), with the exception of Personal Care Routines (with a median score around 3, which is the minimum quality level). Early Head Start classrooms had a median overall ITERS score of 3.9, between minimum and good quality levels.

Figure D.2. Distribution of ITERS-R Subscale and Overall Scores for EHS Classrooms Serving 1-Year-Olds, Baby FACES, Spring 2009

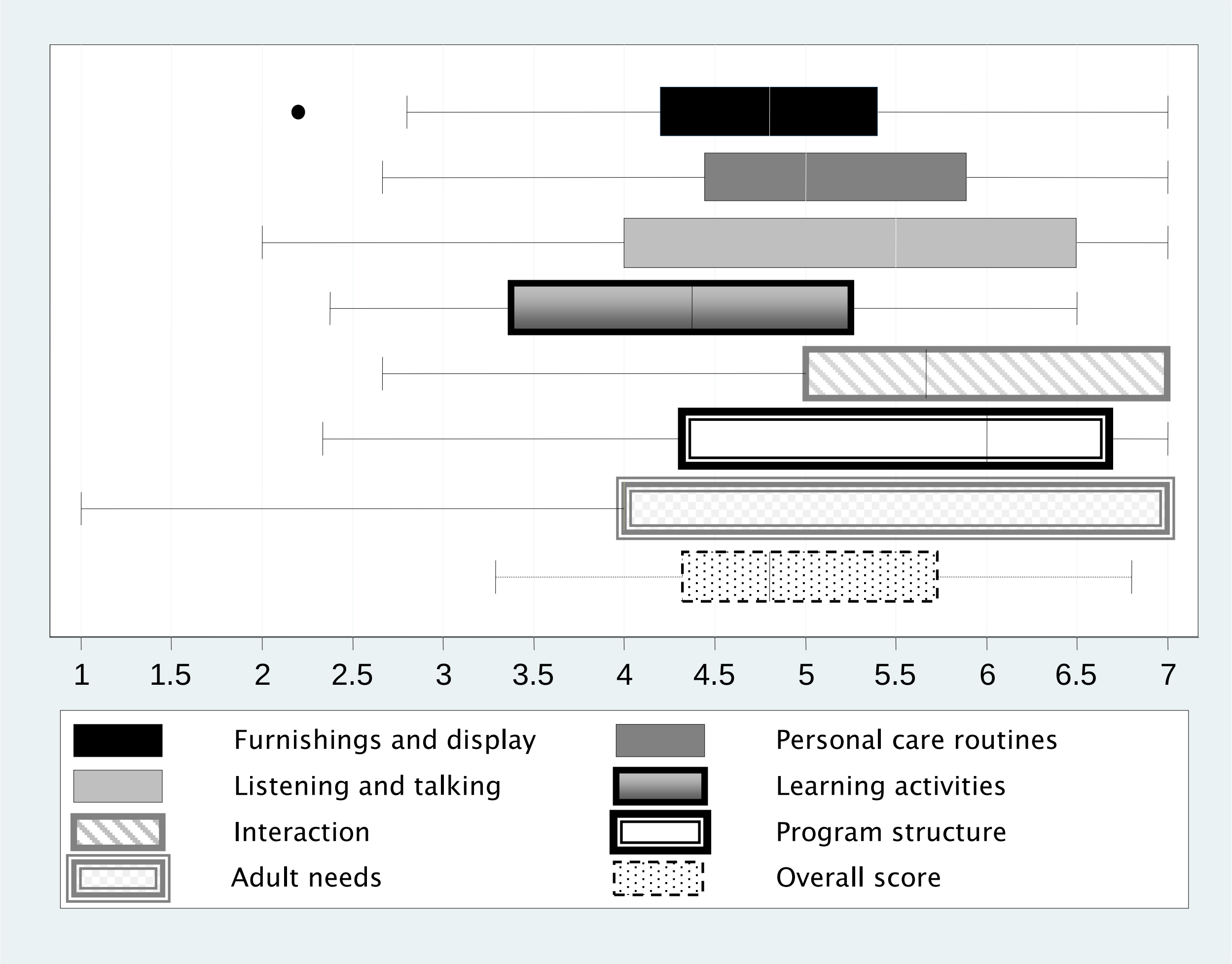


Source: Spring 2009 Baby FACES classroom observations.

Sample Size = 223.

Figure D.3 is arranged in the same way as Figure D.3 and displays the distribution of subscale and overall ITERS scores for the 214 Early Head Start classrooms serving 14-month-old children in the EHSREP, collected about five years earlier.[9] Here, median scores for most subscales fall between 5 and 6 (above the good quality level). The lowest subscale score is Personal Care Routines (median score is slightly below 4.5). Classrooms had a median overall ITERS score of 4.8, which is just below the good quality level.

Figure D.3. Distribution of ITERS-R Subscale and Overall Scores for Early Head Start Classrooms Serving 14-Month-Olds, Child Care Policy Report, 2004



Source: EHSREP classroom observations.

Sample Size = 214.

[9] ITERS scores for this sample of 214 Early Head Start classrooms are also reported on page 70 of the 2004 Administration for Children and Families report, "The Role of Early Head Start Programs in Addressing the Childcare needs of Low-income Families with Infants and Toddlers." The report can be accessed at http://www.acf.hhs.gov/programs/opre/ehs/ehs_resrch/reports/role_ehs_cc/role_ehs_cc.pdf

Comparing Figures D.2 and D.3, we find that median Baby FACES ITERS-R subscale and overall scores are about one point lower than EHSREP ITERS scores. The largest difference between ITERS-R and ITERS median subscale scores is found in Personal Care Routines (3 for ITERS-R versus above 5 for ITERS—the difference between good and minimum quality). In addition, subscale and overall scores for ITERS data are more positively skewed than ITERS-R data (illustrated by a lack of symmetry in box-plots), particularly for the Listening and Talking and Program Structure subscales. In the next section, we consider some possible reasons for these pronounced differences between EHSREP ITERS scores and Baby FACES ITERS-R scores.

ITERS-R Items and Subscales Differ Substantially from ITERS Items and Subscales

Table D.4 compares ITERS and ITERS-R items and subscales, as well as average classroom scores on these items and subscales. For items and subscales that appear in both ITERS and ITERS-R, the table shows the difference in mean scores as well as descriptions of discrepancies between the two instruments that could be responsible for these differences.

It appears that the main differences between the two forms is the greater amount of detail required in ITERS-R than in ITERS. For most items, ITERS-R provides more detail, examples, and guidance to observers than ITERS. Further, in several items, ITERS-R introduces higher benchmarks and additional conditions for classrooms to reach minimum, good, and excellent levels than ITERS.[10] The combination of greater detail and more stringent benchmarks and conditions could account for at least a portion of the apparent drop in Early Head Start classroom scores on ITERS-R compared with ITERS.

[10] A full account of the item-level differences between ITERS and ITERS-R can be found at http://www.fpg.unc.edu/~ECERS/iterscomparison_frame.html.

Table D.4. ITERS-R and ITERS Item and Subscale Scores Differ Due to Substantial Changes in the Scoring Criteria and Subscale Composition, Baby FACES, Spring 2009 and Child Care Policy Report, 2004

Baby FACES 2009 Observations			EHSREP Observations				
ITERS-R Item	Mean Score	Standard Deviation	ITERS Item	Mean Score	Standard Deviation	Difference	Possible Explanations for Difference
Indoor space	3.97	2.04	No equivalent	n.a.	n.a.		n.a.
Furniture for routine care and play	4.45	1.92	Use of furnishings for learning activities	6.50	1.17	-2.05	ITERS-R combines indicators from two furnishings items in ITERS
Furniture for routine care and play (repeated)	4.45	1.92	Furnishings for routine care	4.64	2.11	-0.19	ITERS-R combines indicators from two furnishings items in ITERS
Provision for relaxation and comfort	4.12	1.53	Furnishings for relaxation and comfort	4.76	1.69	-0.64	Three or more soft toys required for the minimum in ITERS-R versus two or more in ITERS; a cozy area must be "accessible for most of the day" in ITERS-R versus "available" in ITERS
Room arrangement	3.85	1.74	Room arrangement	3.95	2.32	-0.10	New indicator in ITERS-R requires the indoor space be accessible regardless of whether or not individuals with disabilities are currently part of the group
Display for children	3.84	1.21	Display for children	4.13	1.36	-0.29	At the excellent level, display must be changed monthly in ITERS-R versus periodically in ITERS
Space and Furnishings Subscale	*4.04*	*1.10*	*Furnishings and Display for Children Subscale*	*4.80*	*0.95*	*-0.76*	
Greeting/departing	5.13	1.76	Greeting/departing	6.51	1.10	-1.38	The minimum score in ITERS-R requires parents to bring children into child care areas, compared with the good score in ITERS
Meals/snacks	2.94	2.00	Meals/snacks	4.43	2.38	-1.49	To score above the minimum in ITERS-R, proper hand washing must occur 50 percent of the time versus only two missed hand washings in ITERS
Nap	3.55	2.47	Nap	5.79	2.03	-2.24	In ITERS-R, supervision must be pleasant, responsive, and warm to score in the 5s; ITERS addresses only the adequacy of supervision
Diapering/toileting	2.05	1.67	Diapering/toileting	3.81	2.39	-1.76	In order to get credit at the minimal level (3s) on ITERS-R, staff must consistently wash hands 75 percent of the time; in order to get credit at the minimal level (3s) on ITERS, staff must consistently wash hands; only missing or improperly washing hands twice is allowed
Health practices	2.42	1.55	Health practice	4.88	2.26	-2.46	ITERS-R addresses washing of staff's hands in the 1s, 3s, and 5s; specific examples of when hand washing is necessary are also listed in item's notes; ITERS addresses washing of staff's hands in the 1s and 3s

Table D.4 (continued)

Safety practices	3.13	1.95	Safety practice	3.86	2.72	-0.73	ITERS-R addresses safety hazards at the 1s, 3s, and 5s; ITERS addresses safety hazards in the 1s and 3s
No equivalent	n.a.	n.a.	Personal grooming	4.33	2.06		n.a.
No equivalent	n.a.	n.a.	Health policy	6.73	0.71		n.a.
No equivalent	n.a.	n.a.	Safety policy	5.88	1.60		n.a.
Personal Care Routines Subscale	*3.20*	*1.16*	*Personal Care Routines Subscale*	5.12	1.03	-1.92	
Help children understand language	4.84	1.49	Informal use of language	5.65	1.46	-0.81	ITERS-R splits this item into two new items
Help children use language	4.57	1.49	Informal use of language (repeated)	5.65	1.46	-1.08	ITERS-R splits this item into two new items
Using books	3.75	1.85	Books and pictures	4.53	2.08	-0.78	ITERS-R addresses only books
Listening and Talking Subscale	*4.39*	*1.26*	*Listening and Talking Subscale*	*5.10*	*1.39*	*-0.71*	
Fine motor	4.40	1.45	Eye-hand coordination	5.58	1.54	-1.18	The 5 level in ITERS-R addresses whether “many and varied” toys are accessible “for much of the day”; the 5 level in ITERS addresses whether a variety of toys are accessible for independent use daily
Active physical play	3.82	1.67	Active physical play	4.80	1.76	-0.98	At the excellent level (7s), two or more types of surfaces for play are required in ITERS-R; required types of play surfaces are not specifically addressed in ITERS
Art	3.61	2.00	Art	4.40	1.89	-0.79	When art is used with infants, the item is scored in ITERS-R regardless of whether or not problems occur; when art is used with infants in ITERS, the item is scored only if problems associated with art activities are observed
Music and movement	3.72	1.43	Music and movement	4.56	1.78	-0.84	At the minimal level (3s), a music activity is required daily in ITERS-R; at the minimal level (3s), a music activity is required three times a week in ITERS
Blocks	3.18	1.61	Blocks	4.43	1.68	-1.25	At a minimal level (3s) in ITERS-R, one set of six or more blocks must be accessible for much of the day; at the excellent level (7s), three sets of 10 or more blocks are required in order to receive credit; at a minimal level (3s) in ITERS, some blocks are needed
Dramatic play	4.09	1.51	Pretend play	4.32	2.04	-0.23	At the minimal level (3s) in ITERS-R, materials must be accessible daily for much of the day; at the minimal level (3s) in ITERS, materials must be accessible daily

Table D.4 (continued)

Sand and water play	3.92	1.74	Sand and water play	3.77	2.14	0.15	Similar scores across ITERS-R and ITERS
Promoting acceptance of diversity	3.41	1.34	Cultural awareness	3.39	1.44	0.02	Similar scores across ITERS-R and ITERS
Nature/science	3.22	1.45	No equivalent	n.a.	n.a.		n.a.
Use of TV, video, and/or computers	3.39	1.31	No equivalent	n.a.	n.a.		n.a.
Activities Subscale	*3.67*	*0.94*	*Learning Activities Subscale*	*4.42*	*1.12*	*-0.75*	
Supervision of play and learning	4.68	1.81	Supervision of daily activities	5.16	2.21	-0.48	At the good level (5s) in ITERS-R, staff give children help and encouragement when needed; help and encouragement from staff are not specifically addressed by an indicator in ITERS
Peer Interaction	4.92	1.52	Peer interaction	5.99	1.17	-1.07	The good level (5s) in ITERS-R addresses whether the staff facilitate positive peer interaction, among all children; the good level (5s) in ITERS addresses whether the interaction between peers is usually positive
Staff-child interaction	4.79	1.92	Caregiver-child interaction	5.54	1.63	-0.75	At the excellent level (7s) in ITERS, responsibility for a small group of children needs to be assigned to one caregiver; this issue is now addressed in the Staff Continuity item in ITERS-R
Discipline	4.47	1.30	Discipline	5.69	1.47	-1.22	Expectations for the children need to be realistic for their ages and abilities at the minimal level of quality (3s) for ITERS-R; expectations for the children need to be realistic for their ages and abilities at the good level of quality (5s) for ITERS
Interaction Subscale	*4.71*	*1.20*	*Interaction Subscale*	*5.75*	*1.16*	*-1.04*	
Schedule	4.03	1.77	Schedule of daily activities	5.04	2.05	-1.01	At the inadequate level (1s) in ITERS-R, caregivers do not have time to supervise children at play; at the inadequate level (1s) in ITERS, caregivers do not have time to talk and play with children
Free Play	4.29	1.62	Staff cooperation	6.05	1.32	-1.76	
Group play activities	4.67	1.87	No equivalent	n.a.	n.a.		
Provisions for children with disabilities	5.41	1.73	Not included in analysis	n.a.	n.a.		
Program Structure Subscale	*4.34*	*1.31*	*Program Structure Subscale*	*5.38*	*1.38*	*-1.04*	
Not included in analysis	n.a.	n.a.	Adult personal needs	4.55	1.81		n.a.
Total ITERS-R Score (32 items)	*3.93*	*0.81*	*Total ITERS Score (31 items)*	*4.97*	*0.89*	*-1.04*	

n.a. = not available.

More Stringent Scoring Guidelines May Explain Lower ITERS-R Scores

The apparent decline in classroom quality based on our observations was unexpected and we took several steps to better understand the data. Based on the analyses described earlier, our main conclusion is that changes in the ITERS instrument are likely major contributors to the observed decrease in scores. Most of the changes in ITERS-R add requirements and clarifications to the items to obtain higher scores that were not present in the original instrument. For example, the ITERS-R requirement for the nap item is that supervision must be pleasant, responsive, and warm, rather than simply "adequate" for the same rating in ITERS; this probably contributed to the lower score in that item. Similarly, in the discipline and peer interaction elements, standards that constituted a good score in ITERS now comprise the requirements for a minimal score in ITERS-R. These differences between the two instruments make it difficult to attribute lower ITERS-R scores in 2009 entirely to actual changes in Early Head Start program quality.

That said, we also cannot rule out the possibility that changes in the program may account for at least some of the differences in scores. We can illustrate this point by looking at the subscale with the largest decline (Personal Care Routines). Hand washing figures prominently and is counted in three separate items in this subscale. Interestingly, under certain circumstances the hand-washing requirements in ITERS-R may be *less* stringent than in ITERS. For example, given 10 hand-washing opportunities in the diapering item, the ITERS requirement to miss no more than two hand washings (80 percent) is more stringent than ITERS-R, which requires hand washing after diapering 75 percent of the time. The requirement of 50 percent hand washing for the meals and snacks item is also likely to be easier to meet in ITERS-R relative to ITERS (again, only two missed hand washings allowed). Observing the scores on these items over time illustrates a marked decline that in turn contributes to the decreased mean score on the Personal Care Routines subscale from good quality (5s) to minimal quality (3s). At least with regard to hand washing, the current lower scores on these items might reflect a decline over time in program staff's attention to this area.

Field Staff Demonstrated Strong Inter-Rater Reliability on ITERS-R Scores

We assessed the inter-rater reliability of data collectors who observed and scored Early Head Start classrooms using ITERS-R to ensure data were collected reliably. As shown in Table D.5, trained observers nearly always scored within one point (on each 7-point subscale as well as on the overall score) of gold standard observers who observed and scored the same classrooms—the standard that the ITERS-R developers require. Considering a higher standard of reliability, we also examined the percentage of time observers agreed with the gold standard within 0.5 of a point. Observers scored within 0.5 points of gold standard observers on subscale and total ITERS-R scores 62 to 81 percent of the time. We can conclude that reviewers were adequately trained and demonstrated a high degree of reliability in ITERS-R scoring throughout Baby FACES data collection.

Table D.5. Gold Standard Reviewers and Field Staff Show Close Agreement on ITERS-R Subscales and Total Scores, Baby FACES, Spring 2009 (Percentages)

ITERS-R Subscale	Agreement Within 1 Point	Agreement Within 0.5 Points	Agreement Within 0.25 Points
Space and Furnishings	86	62	52
Personal Care Routines	95	81	76
Listening and Talking	81	62	48
Activities	95	71	57
Interaction	95	62	38
Program Structure	95	67	48
Overall ITERS-R Score	**100**	**81**	**57**

Source: Spring 2009 Baby FACES classroom observations.

Sample Size = 35 observed classrooms.

Factor Analysis of ITERS-R Data

We calculated ITERS-R scores according to instructions in the manual using data from 223 Early Head Start classrooms observed in spring 2009. We also examined the psychometric properties of ITERS-R in our data by documenting the psychometric properties of the author-defined subscales and by conducting a principal components factor analysis.

Our initial findings are similar to Bisceglia et al. (2009), who found one global aspect of quality through an exploratory analysis of ITERS-R items. Our one-factor solution demonstrated high internal consistency (alpha = 0.88) and explained a substantial portion of common variance (25 percent). We tested the psychometric properties of subsets of ITERS-R items by selecting three 12-item subsets from ITERS-R. The alphas for these subsets of items were 0.76, 0.72, and 0.76, somewhat lower than, but comparable to, those found by Bisceglia and colleagues in their three subsets (0.80, 0.81, and 0.80). These findings suggest that ITERS-R could capture a global measure of classroom quality with a subset of its original 39 items.

We next examined other factor solutions and found that this global measure of quality can be divided into four distinct dimensions, which we termed (1) Language/Interaction, (2) Activities, (3) Routines, and (4) Space/Furnishings. The item loadings for these four factors are shown in Table D.6. Several previous studies (Helburn 1995; Hestenes et al. 2007; Tietze and Cryer 2004; Whitebook, Howes, and Phillips 1989) also identify one or several of these factors in their analysis of ITERS-R and ITERS.

Table D.6. ITERS-R Elements Load into a Four-Factor Solution, Baby FACES, Spring 2009

	Factor			
Item	Language/Interaction	Activities	Routines	Space/Furnishings
Staff-child interaction	0.72	-0.12	0.02	0.16
Discipline	0.72	0.06	0.05	0.23
Help children use language	0.72	0.19	0.19	-0.18
Help children understand language	0.70	0.16	0.09	-0.03
Peer Interaction	0.60	0.25	-0.04	0.03
Supervision of play and learning	0.56	0.15	0.07	0.21
Nature/science	0.18	0.71	0.00	0.01
Promoting acceptance of diversity	-0.04	0.68	-0.07	-0.08
Dramatic play	0.18	0.66	0.07	0.29
Blocks	0.03	0.63	0.22	0.15
Free Play	0.37	0.57	0.10	0.12
Fine motor	0.21	0.53	0.20	0.32
Health practices	0.12	0.04	0.83	0.13
Meal/snacks	0.17	0.05	0.80	0.08
Diapering/toileting	-0.03	0.13	0.77	0.03
Indoor space	0.02	0.05	0.03	0.76
Room arrangement	0.22	0.03	0.04	0.73
Furniture for routine care and play	-0.03	0.25	0.28	0.58
Provision for relaxation & comfort	0.13	0.42	-0.01	0.47
Mean (SD)	*4.73 (1.09)*	*3.76 (1.03)*	*2.48 (1.44)*	*4.11 (1.27)*
Standardized Alpha[a]	*0.78*	*0.77*	*0.76*	*0.65*
Percentage of Total Variance Explained	*25.58*	*10.48*	*9.52*	*7.62*
Sample Size		**220**		

Source: Spring 2009 Baby FACES classroom observations.

Note: Three of the 223 classrooms in the study were missing data for at least one item and were thus excluded from the factor analysis. As a result, the sample size was 220 classrooms.

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

Although we have empirical evidence that collecting data for 12-item subsets could provide a concise and accurate measure of global classroom quality, our exploratory analysis suggested that collecting data for all ITERS-R items provides richer information on classroom quality in the distinct domains of Language/Interaction, Activities, Space/Furnishings, and Routines. Therefore, future Baby FACES classroom observations will continue collecting information on all 32 items in the revised ITERS-R scale.

It is also important to consider how these data and any modified approach to scoring ITERS-R data might be useful to program staff. For our sample of programs, the author-defined subscales do not robustly assess the constructs they purport to measure. Using an empirically derived scoring structure might provide more explicit and meaningful guidance to program improvement efforts. Because our four-factor solution has higher internal consistency (alphas of higher than 0.65 for each factor) than the author-defined subscales, we believe classrooms' scores for these factors might be a better measure of their performance on the underlying dimensions of quality captured by ITERS-R.

Difficulties Administering the CDI and ASQ

As part of the spring 2009 Baby FACES data collection for the 1-year-old Cohort, teachers and home visitors completed the Macarthur-Bates Communicative Development Inventories (CDI) using the Infant Form word list in the Staff-Child Report (SCR). At 12 months, we asked teachers and home visitors to complete the 89-item English CDI for all children, including dual language learners (DLLs). We asked teachers and home visitors who understood Spanish and were reporting on children who understood Spanish to complete the CDI Spanish short form. There is a current, official 104-item version of the Spanish short form (Jackson-Maldonado et al. 2003) and the norms are currently being finalized and summarized in a forthcoming publication (Donna Jackson-Maldonado, personal communication, October 2009).[11] In error, we administered the EHSREP version of the Spanish form, a direct translation of the 89-item English short form (when the EHSREP was under way there was no official Spanish version of the CDI short form). The CDI authors report that the Spanish and English forms were developed separately to reflect the vocabulary and grammatical structure of each language. The English version did not serve as the basis for the Spanish version.

The Baby FACES and EHSREP Infant CDI forms differ by three items that were changed slightly, without changing either their meaning or difficulty. Between the 2003 official form and the Baby FACES SCR instrument, 30 items are identical and 4 additional items are the same except for the form of the word (gender for two, reflexive for one, and plural for one), which does not change the difficulty of the item or the meaning (Table D.7). Thus, there are 34 items in common across the two versions. This allowed for analyses of the items and their properties and yielded a useful way to bridge the two versions. In addition, 25 of the 34 items that overlap are in the toddler form as well (not shown), which might make it possible for us to scale the forms together and put both forms on the same scale for longitudinal analysis. We will not be able to explore this until after the spring 2010 CDI data collection.

[11] We are working with the authors to obtain the norming data for the Spanish short form.

Table D.7. The Baby FACES Spanish CDI Differs Substantially from the Published Spanish CDI

Published 2003 104-Word Spanish CDI Short Form	Baby FACES 89-Word Spanish CDI Teacher/Home Visitor Version	EHSREP 89-Word Spanish CDI Parent Version
Matching Words		
abuela	abuela	All match Baby FACES except:
afuera	afuera	
agua	agua	
am	am	
aqui	aqui	Cobija instead of cobiha
ay	Ay	
bano	bano	
bonita	bonita	
botella/mamilla	botella/mamilla	Buenos noches instead of buenas noches
caliente	caliente	
carro/coche	carro/coche	
cocina	cocina	
como	como	
flor	flor	
gato	gato	Mio/mia instead of mio
grande	grande	
hoy	hoy	
libro	libro	
mama/mami	mama/mami	
miau	miau	
ojos	ojos	
pan	pan	
pantalon	pantalon	
pelota	pelota	
perro	perro	
por favor	por favor	
quien	quien	
roto	roto	
television	television	
zapato	zapato	
Similar Words		
comer	comerse	
mio	mio/mia	
nino	nina	
un	unos	
Different Words		
adios/byebye	abajo	
ahi	acabarse	
ahorita/ahora	ayudar	
bebe	besitos	
boca	brincar	
brazos	buenas noches	
buenos dias	cabeza	
caerse	calcetines	
calle	cantar	
cama	casa	
camion/troca	cereal	
cansado	cobija	
carne	correr	
cielo	cuchara	
collar	cuna	
dinero	dientes	
donde esta	dulce	
dormirse	empujar	
ellas	esperar	
en	galleta	

Published 2003 104-Word Spanish CDI Short Form	Baby FACES 89-Word Spanish CDI Teacher/Home Visitor Version	EHSREP 89-Word Spanish CDI Parent Version
encima	hola	
escoba	jugo	
escribir	leon	
estar	limpio	
familia	lluvia	
fiesta	luna	
guagua	luz	
helado	mesa	
huevo	muneca	
iglesia/templo	nariz	
jabon	no toques eso	
jugar	noche	
lavabo	oscuro	
leche	Otro	
llaves	Pajaro	
manana	papa/papi	
manos	Pastel	
mas	Pato	
mirar	Piedra	
mono	Piernas	
mucho	Planta	
no hay	Plato	
nuevo	Radio	
ojitos	Rapido	
osito	Raton	
para	Romper	
pollito	se acabo	
ponerse	Silla	
querer	Sofa	
quiquiriqui	Sombrero	
rana	Sonreir	
senora	Taza	
sentarse	Tortillitas	
shhh	Tutu	
si	Yo	
sol		
este		
sucio		
su		
tambien		
tambor		
tener		
tigre		
tomarse		
tortilla		
tren		
uno dos tres		
vaca		
vaso		
ver		

Virginia Marchman, another author of the CDI, ran the correlations between the overlapping items, the Baby FACES short form, and the official infant short form using their norming data. The CDI long form contains 86 of the 89 items on the Baby FACES form; Ms. Marchman created virtual scores for the Baby FACES form using these 86 items. The results suggest that scores generated using only the subset of the overlapping items and the virtual scores for the Baby FACES form are highly correlated to scores using the official short form (with correlations ranging from 0.96 to 0.99, see Tables D.8 and D.9).

Table D.8. Spanish Infant Version Vocabulary Comprehension Between 34 Overlapping Items Shows a Strong Correlation Between the Baby FACES Form and the Official Form Based on Norming Data

	34-Item Comprehension	Baby FACES Short Form Comprehension	Offical Short Form Comprehension
34-item[a] Comprehension	--		
Baby FACES Short Form[b] Comprehension	.974	--	
Official Short Form Comprehension	.971	.984	--

Source: Analyses conducted by Virginia Marchman, personal communication, October 2009.

[a] Overlapping between Baby FACES form and the official Spanish short form.

[b] Using 86 items that are on the CDI long form (3 Baby FACES items are not on the long form).

Table D.9. The 34 Overlapping Vocabulary Production Items Between the Baby FACES Form and the Official Form Show Strong Correlations Based on Norming Data

	34-item Production	EHSREP Short Form Production	Official Short Form Production
34-item[a] Production	--		
Baby FACES Short Form[b] Production	.961	--	
Official Short Form Production	.964	.987	--

Source: Analyses conducted by Virginia Marchman, personal communication, October 2009.

[a] Overlapping between Baby FACES form and the official short form.

[b] Using 86 items that are on the CDI long form (3 Baby FACES items are not on the long form).

In light of these issues and need for information about the language development of DLL children by the Office of Planning, Research & Evaluation (OPRE), we have considered a number of options for future Spanish CDI data collection rounds for both Newborn Cohort children when they are 12 months of age in spring 2010 and for 1-year-old Cohort children when they are 2 years old in spring 2010. For the then 1-year-olds, we will administer the 89-item Spanish CDI (the EHSREP version). We opted to administer the wrong version again because there will be so few children in that cohort that will meet the criteria for administering the Spanish CDI (we estimate 30 completed CDIs for this group) that they will not allow analysis on their own.

However, to allow comparison of the EHSREP Spanish form with the official 104-item version in spring 2010, we will also ask the teachers and home visitors to rate the then 1-year-olds on the official CDI Spanish short forms. If the correlation between the two forms is high, it will provide further support for the use of the EHSREP version in analyses. This dual administration provides the best method for determining the validity of the EHSREP short form data gathered in spring 2009. Given that there are 34 items overlapping between the EHSREP and official versions, if we get a similar range of item difficulties on both versions, we could anchor on the item difficulties and put the two versions on the same scale. Examination of the distribution of the 34 items on the EHSREP form indicates a wide range of difficulties for Vocabulary Comprehension; however, the items are clustered at the difficult end for Vocabulary Production. We will have to return to considering the feasibility of this approach for incorporating the 2009 data after we analyze the infant data from Cohort 0 and from the Toddler form administered to Cohort 1 in 2010.

An Incorrect 12-Month ASQ-3 Precluded Analysis for This Age Group

As part of the spring 2009 Baby FACES data collection for the 1-year-old Cohort, we administered the ASQ-3 to parents as part of the telephone interview. Depending on the age of the child on the day of the parent interview, this required administration of the ASQ-3 10-, 12-, 14-, 16-, or 18-month form. The ASQ-3 was not published at the time we started the data collection, and we obtained advance versions of the questionnaires from the ASQ development team. However, in error we were sent the wrong 12-month questionnaire, which included 10-month items instead of 12-month items in all areas except Communication. Thus, in spring 2009, the 12-month age group has incorrect items in the other four areas (Gross Motor, Fine Motor, Problem Solving, and Personal-Social). The 12-month questionnaire is the same in the ASQ-2 and ASQ-3, and both are different from the Baby FACES 12-month version. For each area, the Baby FACES questionnaire has only three items that are the same as those in the official versions. The Baby FACES 12-month questionnaire is the same as the 10-month ASQ-2 and ASQ-3 questionnaires in these areas.

We scored the resulting (incorrect) 12-month data using the 12-month cutoffs. Not surprisingly, the percentages of children identified as at risk in the 12-month group are substantially lower than those for other age groups. Table D.10 presents the percentages of children at or below three different cutoffs (2, 1.5, and 1 standard deviation) by age interval.

We contacted the ASQ developers about the problem and are following their recommendation to use only the scores for Communication for the 12-month-olds who incorrectly received the 10-month version for the other areas. We adopted this approach to avoid skewing the percentage of at-risk children and report *only* the Communication scores for this group of 214 children.

One further issue arose in analysis of the ASQ data. As noted earlier, there are 21 ASQ-3 questionnaires for use at different ages. It is critical to determine the child's age at the time of screening and select the correct questionnaire for the child. In five cases we were given inaccurate birth dates by programs and this resulted in us administering the wrong version of the ASQ. We also identified an issue in how age in months was calculated in the telephone survey instrument (using the ASQ-recommended 30 days rather than the more accurate 30.4 days as the divisor). The Blaise programming methodology used as part of the telephone interview calculated the child's age and routed the interviewer through the correct items. Using 30 days resulted in children being categorized as six days older than they actually are and parents potentially completing the ASQ form that was above the one they should have received. This happened in approximately 70 cases. These children are more likely to be identified as at risk because the items are more difficult. We checked whether this is also true of children who are legitimately within the first week for the age on a given form and they are also more likely to be identified as at risk. For the next round of data collection, we have changed the Blaise calculation and will use 30.4 as the divisor to calculate age. We may also use the data from the spring 2010 data collection to determine whether a child being at the lower range of age for a form affects his or her probability of being identified as at risk.

Table D.10. Fewer Children Fall Below the Cutoffs at 12 Months Old Due to the Mistaken Administration of the 10-Month Form

	Number	Percentage ≤ 2 SD	Percentage ≤ 1.5 SD	Percentage ≤ 1 SD
Communication				
10 months	37	13.51	16.22	32.43
12 months	214	2.34	6.07	21.03
14 months	248	8.47	12.5	29.03
16 months	167	10.18	17.96	41.92
18 months	9	0	0	11.11
Gross Motor				
10 months	37	21.62	21.62	24.32
12 months	214	3.74	4.21	10.75
14 months	248	12.90	16.53	18.15
16 months	167	5.39	12.57	19.16
18 months	9	0	11.11	11.11
Fine Motor				
10 months	37	18.92	29.73	43.24
12 months	213	8.45	13.15	22.54
14 months	248	8.06	15.32	22.58
16 months	167	21.56	30.54	40.72
18 months	9	0	11.11	44.44
Problem Solving				
10 months	37	16.22	27.03	35.14
12 months	214	4.67	10.28	14.02
14 months	248	14.52	21.77	33.47
16 months	167	28.74	37.13	53.89
18 months	9	22.22	33.33	44.44
Personal-Social				
10 months	37	8.11	13.51	29.73
12 months	214	2.34	5.61	11.21
14 months	248	6.85	14.92	32.66
16 months	167	11.98	21.56	33.53
18 months	9	0	0	11.11

Source: Spring 2009 Parent Interview.

APPENDIX E

SUPPLEMENTAL TABLES

APPENDIX E: SUPPLEMENTAL TABLES AND METHODOLOGY

This appendix presents tables that contain additional data cited in Chapters III and VIII and an explanation of the methodology of the logistic regressions used in Chapter IX. The table numbers indicate which chapter they relate to, for example, tables for Chapter III are numbered E.III.1, E.III.2, and so forth.

Chapter III Supplemental Tables

Table E.III.1 Proportion of Families Receiving Each Service in Multiple-Approach Programs

Services	Number of Programs	Percentage of Programs
60% or more of families in home-based care (Max = 91%)	22	37%
60% or more of families in center-based care (Max = 90%)	21	35%
60% or more of families in combination care (Max = 86%)	1	2%
Less than 60% in any one form of care	16	26%
Total	**60**	**100%**

Source: Spring 2009 Program Director Interview.

Table E.III.2 Service Receipt Classification at Family Level

	Number of Families in Each Approach				
Approach According to Directors	Center-Based	Home-Based	Combination	Family Child Care	Total
Center-Based	141	0	0	0	141
	100%	0%	0%	0%	100%
Home-Based	0	107	0	0	107
	0%	100%	0%	2%	100%
Multiple	194	260	29	5	487
	40%	54%	6%	1%	100%
Total	**334**	**367**	**29**	**5**	**735**
Total Percentage	**44%**	**52%**	**4%**	**1%**	**100%**

Source: Spring 2009 Program Director Interview. Spring 2009 Parent Interview.

Note: There are 46 missing values; these values reflect all families in multiple-approach programs that did not complete service receipt questions in the Parent Interview. Due to rounding, totals may not sum to 100 percent.

Table E.III.3 Reasons for Changing Service Options

Reason	Weighted Percentage of Programs[a]
Families can change service options	92.9 (4.4)
Reasons for change	
Changes in the family's needs or preferences	95.0 (2.6)
Changes in the availability of slots in each service option	88.9 (4.2)
Changes in the parent's employment status	76.9 (6.1)
Staff reassessment of the family's needs	75.1 (6.5)
The age of the child or pregnancy	73.5 (6.7)
Other	23.3 (5.7)
Sample Size	**60**

Source: Spring 2009 Program Director Interview.

[a] Among programs with multiple service options.

Table E.III.4 Frequency of Service Receipt by Approach, According to Directors

	EHS Services Reported by Directors	
Approach	Average Home Visits per Year	Average Center Days per Week
Center-Based	2.7	5.0
Home-Based	52.0	n.a.
Multiple-Approach	29.1	4.8[a]
Center-based families	3.4	5.0
Home-based families	51.0	n.a.
Combination families	18.0	3.3
Sample Size		**89**

Source: Spring 2009 Program Director Interview. Spring 2009 Parent Interview.

Note: There are 0 missing values.

[a] Excludes home-based families because the question was not asked of these families.

n.a. = not applicable because the question was not asked of home-based families.

Table E.III.5: Characteristics of Child Care Partnerships

Characteristic	Weighted Percentage or Mean (standard error)
Percentage of programs with formal written partnership with child care provider	35.1 (5.1)
Mean number of formal written partnerships[a]	6.3 (2.2)
Mean percentage of children served through partnerships[a]	23.3 (3.7)
Percentage of programs having scheduled contacts with at least one child care partner: [a]	
Annually	5.7
Every few months	3.4
Monthly	15.5
More than once a month	75.3
Percentage of programs with at least one inactive partnership	6.7
Percentage of programs with one or more inactive partnerships	1.3 (0.1)
Reasons for inactive partnerships (percentage of programs reporting reason):[b]	
Inadequate quality	20.0
Lack of slots available	20.0
Funding issues	
Other	80.0
Sample Size	**89**

Source: Spring 2009 Program Director Interview.

[a]Among programs with formal written partnerships.

[b]Among programs with an inactive partnership.

Table E.III.6: Most Programs Have Contact with a Part C Provider Monthly or More Frequently

Characteristic	Weighted Percentage or Mean
Has a formal written partnership with a Part C agency (percentage of programs)	93.0
Provides Part C services directly (percentage of programs)	3.4
Frequency of contacts with Part C partner	
Annually	1.3
Every few months	26.5
Monthly	37.5
More than once a month	24.3
Weekly	5.5
Daily	2.3
No regularly scheduled contacts	2.7
Sample Size	**89**

Source: Spring 2009 Program Director Interview.

[a]Among programs with formal written partnerships.

Chapter VIII Supplemental Tables

Table E.VIII.1 Child Cognitive and Language Development

	Possible Range		Reported Range				
Outcome	Min.	Max.	Min.	Max.	Mean/ Percentage	Standard Deviation	Cronbach Alpha
ASQ-3[a] Raw Score							
Communication	0	60	0	60	40.32	14.00	0.65 - 0.73
Gross Motor	0	60	0	60	50.67	13.71	0.79 - 0.85
Fine Motor	0	60	5	60	43.40	13.03	0.69 - 0.73
Problem Solving	0	60	0	60	40.23	14.24	0.68 - 0.77
Personal-Social	0	60	0	60	42.96	12.90	0.61 - 0.70
Total Score	0	300	15	300	216.23	50.54	0.78 - 0.84
ASQ Cut-Off Score (2SDs below the mean or lower)							
Communication	0	1	0	1	7.12	25.74	.
Gross Motor	0	1	0	1	10.65	30.88	.
Fine Motor	0	1	0	1	13.70	34.42	.
Problem Solving	0	1	0	1	20.00	40.04	.
Personal-Social	0	1	0	1	8.70	28.21	.
ASQ in the Monitoring Zone (1 to 2SDs below the mean)							
Communication	0	1	0	1	22.55	41.82	.
Gross Motor	0	1	0	1	8.26	27.56	.
Fine Motor	0	1	0	1	17.39	37.94	.
Problem Solving	0	1	0	1	21.30	40.99	.
Personal-Social	0	1	0	1	23.70	42.57	.
CDIb (English) Raw Score							
Vocabulary Comprehension	0	89	0	89	30.34	20.90	0.98
Vocabulary Production	0	89	0	72	2.86	6.31	0.95
CDIb (Spanish) Raw Score							
Vocabulary Comprehension	0	89	0	89	35.86	22.49	0.98
Vocabulary Production	0	89	0	20	2.16	3.66	0.87
Sample Size							
Parent Interview	**674**						
Parent Interview[c]	**460**						
SCR English CDI	**691**						
SCR Spanish CDI	**113**						

Table E.VIII.1 *(continued)*

Source: Spring 2009 Parent Interview and Staff-Child Report.

Note: ASQ-3 = Ages & Stages Questionnaires (Third Edition); CDI = MacArthur Communicative Development Inventories; SCR=Staff Child Report. Sample restricted to Cohort 1. Depending on the age of the child on the day of the parent interview, the age range of children at the baseline required administration of the ASQ-3 10-, 12-, 14-, 16-, or 18-month questionnaire. In error, we administered the wrong version of the ASQ to parents of children aged 11 and 12 months in all domains except Communication, and therefore report only Communication scores for this group of children.

[a]Parent report.

[b]Teacher/home visitor report.

[c]Pertains to ASQ Gross Motor, Fine Motor, Problem Solving, and Personal-Social. Excludes 12-month group.

Table E.VIII.2 Child Social-Emotional Development

	Possible Range		Reported Range				
Outcome	Min.	Max.	Min.	Max.	Mean/ Percentage	Standard Deviation	Cronbach Alpha
Parent Interview BITSEA Raw Score							
Problem Domain	0	62	0	40	10.57	6.31	0.79
Competence Domain	0	22	5	22	16.16	3.38	0.66
SCR BITSEA Raw Score							
Problem Domain	0	62	0	27	6.29	4.70	0.78
Competence Domain	0	22	0	22	12.82	3.53	0.73
Parent Interview BITSEA Cut-Off Score							
Problem Domain	0	1	0	1	26.80	44.33	
Competence Domain	0	1	0	1	9.81	29.76	
SCR BITSEA Cut-Off Score							
Problem Domain	0	1	0	1	13.62	34.33	
Competence Domain	0	1	0	1	14.65	35.39	
Parent Interview BITSEA Screen Positive	0	1	0	1	32.99	47.05	
SCR BITSEA Screen Positive	0	1	0	1	24.77	43.20	
Sample Size							
Parent Interview	**673-679**						
SCR	**628-739**						

Source: Spring 2009 Parent Interview; Staff-Child Report.

Note: BITSEA = Brief Infant-Toddler Social & Emotional Assessment; SCR=Staff Child Report. Sample restricted to 1-year-old Cohort.

Table E.VIII.3 Parent Mental Health

	Possible Range		Reported Range				
Outcome	Min.	Max.	Min.	Max.	Mean/ Percentage	Standard Deviation	Cronbach Alpha
CESD-SF Raw Score	0	36	0	35	5.46	5.64	0.84
CESD-SF: Severe Depressive Symptoms	0	1	0	1	7.89	26.97	.
CESD-SF: Mild Depressive Symptoms	0	1	0	1	24.76	43.19	.
CESD-SF: No Depressive Symptoms	0	1	0	1	57.65	49.44	.
PSI: Parental Distress	5	25	5	25	10.86	4.64	0.73
PSI: Parent-Child Dysfunctional Interaction	6	30	6	30	8.79	4.15	0.78
Sample Size							
Parent Interview	**649-825**						

Source: Spring 2009 Parent Interview.

Note: 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 E.VIII.4 Family Functioning

	Possible Range		Reported Range				
Outcome	Min.	Max.	Min.	Max.	Mean/ Percentage	Standard Deviation	Cronbach Alpha
FES-Family Conflict	1	4	1	4	1.58	0.52	0.70
Social Support	13	39	13	39	30.91	7.40	0.93
Parenting Alliance Measure	10	50	10	50	45.96	5.81	0.94
Sample Size							
Parent Interview	**374-825**						
FES-Family Conflict[a]	**155**						

Source: Spring 2009 Parent Interview.

Note: FES = Family Environment Scale.

[a] Only asked of the Newborn Cohort in Spring 2009.

Table E.VIII.5 Parenting Outcomes

	Possible Range		Reported Range				
Outcome	Min.	Max.	Min.	Max.	Mean/ Percentage	Standard Deviation	Cronbach Alpha
Parental Modernity Scale							
Traditional Attitudes	5	25	5	25	19.78	3.55	0.59
Progressive Attitudes	5	25	5	25	20.07	3.45	0.58
Sample Size							
Parent Interview	**648-650**						

Source: Spring 2009 Parent Interview.

Chapter IX Supplemental Tables

Table E.IX.1 Many Family Needs Are Associated with Service Option in Multiple-Approach Programs

	Odds Ratio (Std Error)	
Predictors	Combination	Home-Based
Constant	-3.25 (0.30)***	-0.69 (0.14)***
Control Variables		
Household language (non-English)	1.84 (0.22)***	0.49 (0.11)***
Community Characteristics		
Rural	0.26 (0.21)	0.69 (0.11)***
Urban (reference)		
Child Health Needs		
Child birth weight		
Normal (reference)		
Low or very low birth weight	-0.31 (0.39)	-0.30 (0.22)
Born prematurely	-0.44 (0.52)	1.66 (0.23)***
Child in fair or poor health	0.19 (0.30)	-1.35 (0.22)***
Child with a disability diagnosis	0.44 (0.58)	0.71 (0.28)*
Child Developmental Needs (at risk on any ASQ-3 domain)	0.28 (0.22)	0.50 (0.11)***
Child Social-Emotional Needs		
Parent-reported BITSEA screening positive	0.24 (0.21)	0.39 (0.11)***
Staff-reported BITSEA screening positive	-0.57 (0.22)**	-0.72 (0.11)***
Maternal Demographic Risk Factors		
Single mother	-0.82 (0.21)***	-1.06 (0.10)***
Teenage mother	-0.78 (0.20)***	-0.19 (0.10)*
Mother has no high school credential	0.75 (0.20)***	0.39 (0.10)***
Family receives public assistance	0.70 (0.21)***	0.76 (0.11)***
Mother not employed, in school, or in training	0.00 (0.21)	0.86 (0.11)***
Family Economic Risk[a]		
Low (reference)		
Medium	0.53 (0.22)*	0.41 (0.12)***
High	-0.22 (0.25)	-0.32 (0.13)*
Parent Health Needs		
Parent in fair or poor health	1.40 (0.30)***	1.72 (0.21)***
Parent not insured	0.35 (0.32)	-0.59 (0.22)**
Family Psychological Risk Factors		
Moderate or severe depressive symptoms	-0.30 (0.28)	-0.20 (0.14)
Substance use[b]	-0.32 (0.48)	0.71 (0.17)***
Parenting stress[c]	0.56 (0.21)**	0.02 (0.13)

Source: Spring 2009 Parent Interview and Staff-Child Report.

Note: Multinomial logistic regression is performed on a sample of 355 children and families in the 1-year-old Cohort. Table shows regression coefficients (log odds).

[a] The family economic risk index aggregates financial difficulties and food security difficulties. Parents with fewer than two financial difficulties and fewer than two food security difficulties were classified as low economic risk. Parents with more than two financial difficulties or more than two food security difficulties (but fewer than four difficulties in across both categories) were classified as medium economic risk. Parents with at least four difficulties in either category were classified as high economic risk.

[b] Parent reports of drug use in the past year or ever having a drug or drinking problem.

[c] A score of one standard deviation above the mean on either of the Parenting Stress Index subscales (Parental Distress or Parent-Child Dysfunctional Interaction).

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

ASQ-3 = Ages & Stages Questionnaire, third edition; BITSEA = Brief Infant Toddler Social Emotional Assessment.

Table E.IX.2 Proportion of Families Receiving Health Services, by Family Health Needs

	Proportion of Families Receiving Health Services
Child Birth Weight	
Low or very low birth weight	27.0 (8.8)
Normal birth weight	14.7 (1.5)
Birth of Child	
Premature birth	23.7 (6.3)
Full-term birth	15.1 (1.7)
Child Physical Health	
Poor or fair health	17.3 (5.8)
Good to excellent health	15.6 (1.8)
Child Health Insurance	
Not insured	14.8 (5.8)
Insured	15.6 (1.7)
Child Disability Diagnosis	
Yes	12.8 (7.6)
No	16.1 (1.9)
Parent Physical Health	
Poor or fair health	23.6 (4.7)†
Good to excellent health	14.5 (1.7)
Parent Health Insurance	
Not insured	16.5 (5.3)
Insured	15.5 (1.8)
Sample Size	**649-822**

Source: Spring 2009 Parent Interview.

Note: We conducted chi-square tests to test significance. Standard errors in parentheses.

†*p* < .10; **p* < .05; ***p* < .01; ****p* < .001. Symbols next to the higher percentages.

Table E.IX.3 Proportion of Children Receiving Disability Services, by Child Needs

	Proportion of Children Receiving Disability Services
Child Disability Diagnosis	
Yes	19.1 (10.5)
No	2.8 (0.5)
Child Developmental Needs	
Scoring in the at-risk range on any of the ASQ-3 domains	3.8 (1.3)
Scoring in the normal range on all of the ASQ-3 domains	2.9 (0.7)
Sample Size	**650-671**

Source: Spring 2009 Parent Interview.

Note: We conducted chi-square tests to test significance. Standard errors in parentheses.

†*p* < .10; **p* < .05; ***p* < .01; ****p* < .001. Symbols next to the higher percentages.

Table E.IX.4 Proportion of Parents Receiving Mental Health Services, by Parents' Psychological Risks

	Proportion of Parents Receiving Mental Health Services
Depressive Symptoms	
Moderate or severe	10.2 (2.5)*
Mild or no	4.5 (0.8)
Substance Use Problems	
Yes	7.2 (2.9)
No	5.5 (0.9)
Parenting Stress	
Yes	7.2 (2.4)
No	4.9 (1.0)
Sample Size	**650-824**

Source: Spring 2009 Parent Interview.

Note: We conducted chi-square tests to test significance. Standard errors in parentheses.

†*p* < .10; **p* < .05; ***p* < .01; ****p* < .001. Symbols next to the higher percentages.

Table E.IX.5 Proportion of Families Receiving Literacy and English Classes, by Household Language

	Proportion of Families Receiving Classes to Learn English	Proportion of Families Receiving Training on How to Read and Write
Household Language		
DLL	0.8 (0.4)	4.6 (1.4)*
English	15.8 (3.0)***	1.5 (0.5)
Sample Size	**856**	**856**

Source: Spring 2009 Parent Interview.

Note: We conducted chi-square tests to test significance. Standard errors in parentheses.

†*p* < .10; **p* < .05; ***p* < .01; ****p* < .001. Symbols next to the higher percentages.

Table E.IX.6 Proportion of Families Receiving Education or Job Training, by Maternal Education and Employment

	Proportion of Families Receiving Education or Job Training
Maternal Education	
Lack of a high school credential	7.6 (1.8)
High school or above	8.1 (1.6)
Maternal Employment	
Not employed, in school or training	6.3 (1.6)
Employed, in school or training	9.0 (1.3)**
Sample Size	**843-844**

Source: Spring 2009 Parent Interview.

Note: We conducted chi-square tests to test significance. Standard errors in parentheses.

†*p* < .10; **p* < .05; ***p* < .01; ****p* < .001. Symbols next to the higher percentages.

Table E.IX.7 Proportion of Families Receiving Job-Related Services, by Maternal Employment

	Proportion of Families Receiving Help Finding a Job	Proportion of Families Receiving Help Finding Good Child Care	Proportion of Families Receiving Help Getting To and From Work or Other Places
Maternal Employment			
Not employed, in school or training	9.0 (1.8)	13.9 (2.5)	12.2 (2.1)
Employed, in school or training	6.3 (1.2)	17.5 (1.7)	13.9 (1.9)
Sample Size	**843**	**839**	**841**

Source: Spring 2009 Parent Interview.

Note: We conducted chi-square tests to test significance. Standard errors in parentheses.

†*p* < .10; **p* < .05; ***p* < .01; ****p* < .001. Symbols next to the higher percentages.

Table E.IX.8 Proportion of Families Receiving Help Finding Child Care, by Maternal Age and Marital Status

	Proportion of Families Receiving Help Finding Good Child Care
Marital Status	
Single	19.5 (2.6)
Married or cohabiting	13.3 (2.0)
Teenage Mother	
Yes	13.8 (1.9)
No	19.9 (2.5)*
Sample Size	**670-851**

Source: Spring 2009 Parent Interview.

Note: We conducted chi-square tests to test significance. Standard errors in parentheses.

†*p* < .10; **p* < .05; ***p* < .01; ****p* < .001. Symbols next to the higher percentages.

Table E.IX.9 Proportion of Families Receiving Financial Support, by Family Financial Distress

	Proportion of Families Receiving Short-Term Help Obtaining or Paying for Things Needed in an Emergency	Proportion of Families Receiving Help Finding or Paying for Housing	Proportion of Families Receiving Counseling on How to Manage Money
Family Economic Risk			
Low	5.2 (1.2)	5.6 (1.2)	7.2 (1.4)
Medium	21.2 (3.8)	8.9 (2.0)	13.1 (2.6)
High	23.6 (4.1)***	11.5 (2.7)*	10.4 (2.7)
Family Receives Public Assistance			
Yes	17.0 (2.3)***	9.6 (1.5)**	11.7 (1.7)***
No	5.0 (1.4)	3.6 (1.2)	4.0 (1.2)
Sample Size	**820-823**	**821-824**	**820-823**

Source: Spring 2009 Parent Interview.

Note: We conducted chi-square tests to test significance. Standard errors in parentheses.

[a]Family economic risk is an index that aggregates financial difficulties and food security difficulties. Parents with less than two financial difficulties and less than two food security difficulties were classified as low economic risk. Parents with more than two financial difficulties or more than two food security difficulties (but less than four difficulties in either category) were classified as medium economic risk. Parents with at least four difficulties in either category were classified as high economic risk.

†*p* < .10; **p* < .05; ***p* < .01; ****p* < .001. Symbols next to the higher percentages.

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

OPRE
Office of Planning, Research and Evaluation
Administration for Children and Families

MATHEMATICA
Policy Research, Inc.