Weighting in the Dark: What to Do in the Absence of Benchmarks

Barbara Lepidus Carlson[1] and Jerry West[2]
[1]Mathematica Policy Research, 955 Massachusetts Avenue, Suite 801
Cambridge, MA 02139
[2]Mathematica Policy Research, 1100 1st Street, NE, 12th Floor
Washington, DC 20002-4221

Abstract
The 2009 Head Start Family and Child Experiences Survey (FACES) involved four stages of sampling: Head Start programs, centers, classrooms, and children. At the time of sampling, eligible children were those who were one or two years away from kindergarten and were new to Head Start in fall 2009. These children were followed through their first year of Head Start, and then followed for one or two more years, depending on their age, through kindergarten. Children who left Head Start after fall 2009 but did not go to kindergarten were considered ineligible for followup. There existed no published population counts for the study's baseline population, nor were there existing benchmarks for the Head Start retention and kindergarten transfer rates needed to define the study population at followup. This paper shows the steps we took to make use of an earlier cohort of FACES to ensure that the baseline and followup weights, which adjusted for sampling and response patterns, appropriately reflected their respective target populations. We also show how different assumptions about eligibility among those with undetermined status can substantively affect estimated totals and mean estimates.

Key Words: Weighting, FACES, Benchmarking

1. Introduction

The Head Start Family and Child Experiences Survey (FACES) is funded by the U.S. Department of Health and Human Services, Administration for Children and Families as a repeated longitudinal study of Head Start program quality and child outcomes (West et al. 2007, 2011). Nationally representative cohorts of Head Start children have been selected every three years since 1997. Mathematica Policy Research has designed and conducted the last two FACES cohort studies (FACES 2006 and 2009).

In FACES, children are followed from their entry into Head Start through one or two years of program participation, with a followup in the spring of kindergarten. As shown in Table 1, data are collected in the fall of the first Head Start program year, the spring of that program year, the spring of the second Head Start program year (for the younger cohort only), and the spring of kindergarten.

Table 1. FACES 2009 Data Collection Schedule by Age Cohort

Cohort	Fall 2009 (Baseline)	Spring 2010	Spring 2011	Spring 2012
Younger[a]	Head Start Year 1	Head Start Year 1	Head Start Year 2	Kindergarten
Older[b]	Head Start Year 1	Head Start Year 1	Kindergarten	

[a]3-year-olds at baseline, two years from kindergarten
[b]4-year-olds at baseline, one year from kindergarten

The target population at baseline is comprised of children entering Head Start, and excludes children returning to Head Start for a second year. The children must be one or two years away from kindergarten at baseline, based on the child's date of birth and the local kindergarten cutoff date, meaning that most of them are 3 or 4 years of age that fall. Children must remain in the Head Start program to remain eligible for the spring Head Start data collections. To be eligible for the kindergarten followup, children had to have completed one or two years of Head Start and attended kindergarten the year after leaving the Head Start program.

FACES data are used to describe the population of children and families served by Head Start; staff qualification and credentials; Head Start classroom practices and quality; and children's cognitive and non-cognitive skills and abilities at program entry, exit, and in kindergarten (West et al., 2010, Hulsey et al., 2011, Aikens et al., 2010). The study is designed primarily to estimate population means (for example, mean child outcomes) and percentages (for example, the percent of children's mothers who have different levels of education), and associations between child, family, classroom, and program characteristics. However, from time to time, the sample is used to estimate population totals such as the number of children who are entering the program for the first time. Although FACES baseline child-level weights can be used to produce estimates of totals for the population of Head Start children at program entry, there are challenges in using weights in this way.

This paper describes the difficulties associated with developing sampling weights that can be used to estimate population totals when there are no published population counts for the study's baseline population, nor existing benchmarks for the Head Start retention and kindergarten transfer rates needed to define the study population at followup. This paper shows the steps we took to make use of an earlier cohort of FACES to ensure that the baseline and followup weights, which adjusted for sampling and response patterns, appropriately reflected their respective target populations. We also show how different assumptions about eligibility among those with undetermined status can substantively affect estimated totals and mean estimates.

2. The Sample

FACES has a multistage sample design. The first stage selects 60 Head Start programs (grantee or delegate agencies); the second stage selects 2 centers per program; the third stage selects 3 classrooms per center; and the fourth stage selects about 10 children per classroom. Programs, centers, and classrooms are selected with probability proportional to size (PPS), with the measure of size (MOS) being our best estimate of the number of study-eligible Head Start children enrolled. Some centers and classrooms are grouped prior to sampling if they are too small to yield a sufficient number of children from which to sample.

When selecting the programs, we use as the sampling frame the Head Start Program Information Report (PIR), which is an administrative database updated on an annual basis that contains program-provided information for each Head Start and Early Head Start[1] grantee and delegate agency. Centers, classrooms, and children are then selected from lists or rosters provided by the program or center. Head Start programs are not obligated

[1]Early Head Start provides services to pregnant mothers and infants and toddlers through age 3.

to participate, but do respond at a high rate. Nonparticipation among selected and eligible centers and classrooms within participating programs has not been an issue.

3. Weighting Steps

We construct nonresponse-adjusted sampling weights at the program, center, classroom, and child levels. The program weights account for the PPS probability of selection, as well as program eligibility and participation, and are poststratified to the PIR counts. The conditional center weights account for the PPS probability of selection within program, which are then multiplied by the final program weight to obtain the cumulative center weight. Similarly, the conditional classroom weights account for the selection probability within center, which are then multiplied by the final center weight to obtain the cumulative classroom weight.

The conditional child weights at baseline account for the child's selection probability (with an adjustment for sibling subselection), and are then adjusted for child eligibility and parental consent. The conditional child weight is multiplied by the final classroom weight to obtain the child's base weight. Because of challenges outlined in the next section, we cannot poststratify the child base weight to any known benchmarks. The base weight is then adjusted for various combinations of completed data collection instruments (such as child assessments, teacher reports, and parent interviews) to produce a set of analysis weights at the child level.

For each of the followup data collection periods, we generate cross-sectional and longitudinal weights in a similar manner, allowing the children known to have become ineligible (no longer in the study population) to drop out of the sample, and making assumptions about those children with unknown eligibility status – also discussed further in the next section.

4. The Challenges

The PIR contains program-level information, including the total number of enrolled children by age, but does not identify the total number of newly enrolled children. And while the PIR reports the number of children returning for a second year, it does not break this down by age, nor does it use the same definition of "returning" as is used for FACES. If a child was in Early Head Start the prior year, the PIR instructions indicate that the child should now be considered in his or her second year, but FACES considers this child to be newly entering Head Start. Nor does the PIR quantify program attrition. We therefore must estimate the program MOS from the PIR based on the reported number of 3-year-olds (assuming all are newly entering the program) and the number of 4-year-olds (arbitrarily assuming that half are new and half are returning for a second year).

As noted earlier, FACES was not primarily designed to produce population totals, though the sum of the baseline weights is a good estimate of the baseline population. But, as mentioned earlier, the sum of the weights has its own associated sampling (and nonsampling) error. While the sampling error of a mean estimate is $\frac{\sigma}{\sqrt{n}}$, the sampling error of an estimate of a total is $\sigma\sqrt{n}$.

Had the primary goal been to estimate the total number of children in the study population, the study might have been designed differently. For example, one could have obtained a census of all enrolled children (along with date of birth and Head Start entry date) from each sampled program, or otherwise design the study in such a way as to increase the precision of the estimated total, which may not have involved sampling children.

When children leave the Head Start program from which they were sampled, they are considered eligible (and part of the target population) only if they stay in Head Start (say, moved to a non-sampled center in the same program, or moved to a non-sampled Head Start program) or go off to kindergarten. While data collection is not attempted for children moving to a non-sampled Head Start center or program, these children are considered to be eligible nonrespondents for purposes of weights and response rates. Children are considered ineligible (and not part of the target population) if they leave Head Start and go to another type of preschool (such as a state-sponsored or other public pre-kindergarten program for 4-year-olds), to first grade, or are no longer in any kind of school program. But sometimes we do not know the eligibility status of a child who has left the sampled Head Start program, and what we assume about these children with undetermined status impacts both response rates and weights. These issues contributed to two issues that arose while producing analysis weights for FACES 2009.

4.1 Issue 1: Weight Totals for Fall 2009

We went through virtually identical weighting steps for FACES 2006 and FACES 2009. When we produced the baseline child weights for FACES 2009, they summed to 560,392. This was a 22 percent increase from the baseline weight total for FACES 2006, where the sum of the weights was 458,473. Although we had no benchmarks in either year for the FACES study population, we knew that it did not grow by 22 percent over the course of three years. So we set off to check for sources of the increase. First, we checked the actual trends in the overall Head Start population by age group, and as expected did not find increases in the overall population of this magnitude. Thinking that this increase could be within the sampling error of the total estimates, we then checked the confidence intervals for the estimated totals in both rounds, using two methods. First, we took the total of the number of children eligible for sampling in sampled classrooms, and obtained the weighted sum using the final classroom weight. Second, we used the baseline child weight as the estimate and took the sum of that, unweighted. The table below shows the upper and lower endpoints of the 95 percent confidence intervals.

Table 2. Confidence intervals (95 percent) of estimated child totals at baseline

		Point Estimate	Lower Limit	Upper Limit
Method 1 (class weights)	FACES 2006	457,711	409,850	505,572
	FACES 2009	559,791	503,880	615,702
Method 2 (child weights)	FACES 2006	458,473	410,405	506,541
	FACES 2009	560,392	504,581	616,203

From this table, we can see that the upper ends of the FACES 2006 confidence intervals overlap ever so slightly with the lower ends of the FACES 2009 confidence intervals. This means that it is highly unlikely that the increase is due to sampling error.

We then double-checked all weighting procedures and generated diagnostic counts to determine at what step in the process the weights began to diverge in FACES 2009. All weighting steps were found to have been applied correctly, though we did determine that the weights began to increase disproportionately after the center level weighting steps. One possible "smoking gun" was the grouping of small classrooms before sampling, and their subsequent disaggregation after applying the sampling weight to the classroom group.

As a simple example, suppose a center had six classrooms: four classrooms had 16 children each and two classrooms had 8 children each. Because we wanted to sample 10 children per classroom, we grouped the two smaller classrooms into a single classroom group before sampling, then selected three of the five with PPS. Suppose we selected the paired classroom along with two of the 16-children classrooms. The sampling weight for any of the three selected 16-child classrooms would be $80/(3\times16)= 1.67$, and the sum of these weights for the three is equal to 5, which is an estimate of the number of classroom groups in the center. We assign this sampling weight to each of the two classrooms in the grouped classroom to get an estimate of the number of classrooms, which is 6.67. Because one difference in sampling methodology in FACES 2006 and FACES 2009 was the extent of classroom grouping – with more grouping done in FACES 2009 to avoid sample size shortfalls encountered in some programs in FACES 2006 – we concluded that this may in fact have been the culprit. While in expectation the sum of the classroom group weights should add up to the number of classroom groups from which the sample was selected, and the sum of the classroom weights (expanding the classroom group weight to each classroom in the group) should add up to the number of classrooms, the actual sum can vary depending on which units are sampled.

We concluded that the difference we saw between FACES 2006 and FACES 2009 was not due to any systematic error, but was likely due to the sample of classrooms that was selected by chance in conjunction with the higher level of classroom grouping (and corresponding higher chance of selection of the larger classroom groups) in FACES 2009. Because the classroom weight is the starting point of the child weight, any inflation found in the former stage carries through to the latter. While both the FACES 2006 and FACES 2009 estimates of the total were unbiased (or nearly unbiased – nonresponse adjustments can introduction bias) estimates of the population, we decided to apply a constant to all weights below the center level (that is, classroom, teacher, and child weights), calibrating them downwards so that the increase between FACES 2006 and FACES 2009 was 6.3 percent, the reasonably sized increase we saw at the center level. The revised child weight total for FACES 2009 was reduced from 560,392 to 487,541, which seems more reasonable when compared to the FACES 2006 total of 458,473. This deflation factor of .87 was also applied to all weights in the followup rounds of data collection. By applying a constant, we did not impact any of the mean, proportional, or association estimates made using the original weight.\

4.2 Issue 2: Weight Totals for FACES 2009 Year 2 (Spring 2011) Follow-Up

Having encountered and dealt with Issue 1, we continued to compare the constant-adjusted weight sums at each subsequent data collection point. Much to our dismay, we saw another divergence when comparing the sum of the initial spring 2011 child weight to that that spring 2008 – the point in data collection for both cohorts during which the older cohort was in kindergarten and the younger cohort in its second Head Start program year. Table 3 shows what we found when making that comparison. By spring of the second study year, the sum of the child weights was almost 25 percent higher in the

FACES 2009 study than it was in the FACES 2006 study. The study retention rate between year one and year two was 79 percent in the earlier study and 90 percent in the later study.

Table 3. Sum of Child Weights Over Time for 2006 and FACES 2009

	FACES 2006	FACES 2009	Change
Fall of Study Year 1	458,473	487,541	1.063
Spring of Study Year 1	405,128	444,330	1.097
Year 1 Retention	0.88	0.91	
Spring of Study Year 2	321,431	400,096	1.245
Year 2 Retention	0.79	0.90	

We isolated the problem to the number of children whose eligibility status was classified as "don't know" at this data collection point. For these children, we did not know if they were eligible (in Head Start or kindergarten) or ineligible (in first grade, in another preschool, or not in school). Table 4 shows the comparison, with the main disparities in bold text.

Table 4. Comparison of Status Codes: FACES 2006 and FACES 2009 Year 2

Classified	Description	Younger Cohort	Older Cohort	Combined
FACES 2006 - Year 2 (Spring 2008)				
Undetermined	Don't know where	**38**	83	121
Participating	Head Start or Kindergarten	1,220	1,007	2,227
Nonrespondent	Other Head Start	4	2	6
Ineligible	Not in school, PreK, 1st grade	**521**	39	560
Total		1,783	1,131	2,914
FACES 2009 - Year 2 (Spring 2011)				
Undetermined	Don't know where	**345**	15	360
Participating	Head Start or Kindergarten	1,190	1,211	2,401
Nonrespondent	Other Head Start	119	5	124
Ineligible	Not in school, PreK, 1st grade	**110**	25	135
Total		1,764	1,256	3,020

We consulted the survey director to determine whether the classification methodology had changed between the two cohorts, and learned that in fact there had been a change in procedures. In the earlier study, a child was more likely to have been coded as "not in school" if the Head Start program told us that the child was not in that school. In the later study, we probed more deeply and only used this status code if we were told specifically that the child was not enrolled in a school of any kind, and classified children as being in kindergarten if we were able to verify the child's status with the school. Finally, we had relatively fewer completed parent interviews at this point in the later study and, because these interviews are one of our first sources of school information, this may have contributed to the different rates as well.

This classification of noncompleted cases matters for both weights and response rates, because a certain proportion of the "undetermined" cases essentially get treated as eligible, often based on the eligibility rate of the "known" cases. The known eligibility

rate among the younger age cohort for the earlier study for the spring of study year 2 was 70 percent, and this rate applied to the 38 undetermined cases adds 27 eligible noncompletes to the 4 such known cases. In the later study, the known eligibility rate for these younger children was calculated as 92 percent, and this rate applied to the 345 undetermined cases adds 318 eligible noncompletes to the 119 cases already coded that way.[2]

While the classification of cases was done more carefully in the later study, we reasoned that it would be highly unlikely for so many of the children in the younger cohort to have left the sampled Head Start program and still be eligible for the study - that is, for so many to be in another (nonsampled) Head Start program or already in kindergarten. Applying the determined eligibility rate to the undetermined cases was not appropriate in this situation. We decided to take the 300 youngest children in the "don't know" category for the younger age cohort, and change their status to "ineligible" to make our classification comparable with the proportions seen in the earlier study. By so doing, the assumed younger cohort eligibility rate was reduced from 92 percent to 76 percent for the undetermined cases, now which gets applied to 45 (not 345) undetermined cases, to get 34 (not 318) estimated eligible nonrespondents. We then re-weighted using these revised statuses. In the later study, the sum of the child weights for spring of study year 1 was reduced from 400,096 to 375,156, which is much more closely aligned with the comparable weight sum for the earlier study (321,431).

While the adjustment for Issue #1 was a constant and did not affect any estimates other than totals, the adjustment here had the potential to affect other estimates. Had we not caught this error, the estimates based on the original estimates could potentially have been misleading. We looked at a number of key estimates for the study, looking at the relative difference in the estimates using the original and revised weights. Table 5 shows the relative differences in these estimates. Fortunately, most of the differences we found were very small.

Table 5. Relative Difference in FACES 2009 Year 2 (Spring 2011) Weighted Estimates With and Without Adjustment

Measure	Relative Difference (Percentage)
Estimated Population (sum of weights)	6.23
Peabody Picture Vocabulary Test (PPVT)-4 Standard Score (mean)	0.05
PPVT-4 W Score (GSV) (mean)	0.16
Expressive One-Word PVT Total Std Score (English norms) (mean)	0.09
Number of Family Economic Risk factors (mean)	0.07
Percentage with Two or More Risk Factors	0.16
Percentage Male	0.36
Percentage Living with Both Parents	0.07
Percentage Below Poverty Line	0.00

[2]These 119 represent a group of children who moved from a sampled Head Start center to a nonsampled center in the same Head Start program.

N.B. Standard scores measure the child's performance relative to same-age peers. They are standardized to a mean of 100 and a standard deviation of 15. The W Score is an IRT-based score that measures absolute performance on a vertical scale. Risk factors included poverty, mothers with less than a high school diploma, and a single-parent family.

5. Lessons Learned

Not all samples have known population totals. Just like other types of estimates, estimates of the total population have associated sampling and nonsampling error. By chance, one can end up with different estimated population totals, even when using standard weighting steps – multistage samples in particular. In our study, we found that having a prior round of the study, and a longitudinal design, enabled us to calibrate our weights, and we wanted to share what we learned.

We view our experiences with these two issues as a cautionary tale. One can easily be led astray even if the methods used are solid. We learned to stand back and look at what the numbers were telling us, and not to blindly go through the weighting steps. One always wants to apply the "subject matter lens" to see if what you are generating is plausible, considering the magnitude of estimates and trends over time. It is important to think more broadly about the weights and the estimates you are generating - not just the estimates you are preparing at the time, but the entire context, including the relationship of the current weights and estimates to others generated over the course of the study.

Both issues were discovered only because we had access to weights and status codes from the earlier study, and because each study was longitudinal. Both were discovered and resolved before making data and reports available to others. Gathering and using tracking and other longitudinal information in weighting adjustments - and reassessing assumptions based on new information gathered about the sample over time - should not be overlooked. We were able to make realistic assumptions about the noncompleted cases by using other information we had collected about the sample. Without known population totals, we fortunately set up data collection in such a way as to keep tabs on the sample over time. We also learned it is important to monitor changes in survey operations and procedures, keeping a dialog open with the survey director.

References

Aikens, N. L. Tarullo, L. Hulsey, C. Ross, J. West, Y. Xue. (2010). ACF-OPRE Report: A Year in Head Start: Children, Families and Programs. Washington, DC. U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Hulsey, L. K., N. Aikens, A. Kopack, J. West, E. Moiduddin, and L. Tarullo (2011). Head Start Children, Families, and Programs: Present and Past Data from FACES. OPRE Report 2011-33a. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

West, J., L. Malone, L. Hulsey, N. Aikens, and L. Tarullo. "ACF-OPRE Report: Head Start Children Go to Kindergarten." Washington, DC: U.S. Department of Health and Human Services, Office of Planning, Research, and Evaluation, Administration for Children and Families, 2010.

West, Jerry, Louisa B. Tarullo, Nikki L. Aikens, Susan C. Sprachman, Christine M. Ross, and Barbara L. Carlson. "FACES 2006 Study Design." Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, 2007.

West, Jerry, Louisa Tarullo, Nikki Aikens, Lizabeth Malone, and Barbara Lepidus Carlson. "FACES 2009 Study Design." OPRE Report 2011-9. Washington, DC: U.S. Department of Health and Human Services, Office of Planning, Research, and Evaluation, Administration for Children and Families, 2011.