Sampling with Uncertain Frame Counts: Challenges in Sampling Head Start Children for the FACES Study

Barbara Lepidus Carlson
SIS, Mathematica Policy Research, P.O. Box 2393, Princeton, NJ, 08543-2393

Abstract

The 2006 Head Start Family and Child Experiences Survey (FACES) involved four stages of sampling: Head Start programs, centers, classrooms, and children. Eligible children were those who were one or two years away from kindergarten and were new to Head Start in the fall of 2006. Because only a list of Head Start programs was available as a sampling frame, we relied on selected programs to provide lists of centers, and relied on selected centers to provide lists of classrooms and eligible children. Sample selection at each level was conducted on a rolling basis, as sample frames were provided by the programs. This paper will describe the challenges in implementing a sampling strategy that met the sample design goals, including an oversample of children who were two years away from kindergarten, and one that was flexible enough to adapt to the actual child counts when they were lower than estimated.

KEY WORDS: FACES, Head Start, oversampling

1. Background

Head Start is a national program administered by the Administration for Children and Families (ACF) in the U.S. Department of Health and Human Services to promote school readiness among economically disadvantaged children. Its goal is to enhance the social and cognitive development of these children through educational, health, nutritional, social, and other services. Agencies receiving grants from Head Start provide comprehensive child development services, with a focus on helping preschoolers develop the early reading and math skills they need to be successful in school.

The Head Start Family and Child Experiences Survey (FACES) is a repeated longitudinal study of Head Start program quality and child outcomes. FACES gathers comprehensive data on the cognitive and social-emotional development of Head Start children along with detailed information about their families and Head Start programs. Previously, FACES had three nationally representative cohorts, beginning data collection in 1997, 2000, and 2003, respectively, and following sampled children through their kindergarten year. Mathematica Policy Research, Inc., is the prime contractor for carrying out the 2006 cohort of FACES.

2. Sampling Overview

In FACES, each cohort of children is followed from entry into Head Start through one or two years of program participation (for four-year-olds and three-year-olds, respectively) with followup in the spring of the kindergarten year. In the current round of FACES, the target population is children in the United Sates who were receiving Head Start services for the first time in the fall of 2006. In the 2006 cohort, there are four sample selection stages. The first stage is Head Start programs, with a target of 60 participating programs. The second stage is centers within programs, with a target of 2 centers per program. The third stage is classes within center, with a target of 3 classes per center. And the fourth and final stage is children within classes, with a target of 10 children per class. Because some programs have fewer than two centers, some centers have fewer than three classes, and some classes have fewer than ten children, we expected to sample approximately 3,500 children.

The first three stages of sampling were selected with probability proportional to size (PPS), with the estimated number of eligible children in the sampling unit being the measure of size. The fourth stage was selected with equal probability.

3. Study Eligibility

For the first stage of sampling, the Head Start programs, we included programs located in all 50 states and the District of Columbia, and excluded those in U.S. territories. We included programs that were providing direct services to children in the target age group, and excluded those with only administrative functions. We excluded programs that were administered by the two ACF regional offices that serve American Indian and Alaskan Native children and families and seasonal and migrant workers and their children. We also excluded any programs that were about to lose Head Start funding or that were under transitional management because the previous grantee lost its funding. One such program was sampled and subsequently excluded, but most were eliminated from the sample frame a priori.

For the second and third stages of sampling (centers and classes), the only eligibility criterion was that the sampling unit have at least one eligible child. On occasion, a center or class was listed on the sampling frame, but after selection turned out to be ineligible.

For the final sampling stage, selecting children within classes, we classified as eligible any child who was new to Head Start in the fall of 2006, and excluded any four- or five-year olds who had been in Head Start the prior school year. (Three-year-olds who had been enrolled in the *Early* Head Start Program the prior year were not disqualified.) The child had to be one or two years away from kindergarten in the fall of 2006. We asked each program for their local kindergarten cutoff month and day, and kept only those children who were ages 3 or 4[1] as of that day in 2006. Finally, the child had to be enrolled and actively attending a selected Head Start center at the time of the site visit.

4. Sampling Frames

The sampling frame for the program sample was the Head Start Program Information Report (PIR). Each year, ACF requires each Head Start and Early Head Start program to submit detailed data on its enrollment, staff, services, and equipment. A Head Start grantee with more than one program is required to submit a separate record for each program (including delegate agencies). The data are compiled into an Access database, with one record per program. For the FACES 2006 sample of programs, we used the most recent PIR available at the time, from the 2004-05 program year.

Sampling frames for the next two stages of selection (centers and classes) were obtained from the selected programs and centers, respectively, on a rolling basis, as they were recruited and agreed to participate, and as they were able to provide these lists. When providing the lists of centers and classes, they were asked to also provide their best estimate for each of the number of new 3- and 4-year-olds they would have enrolled in the fall of 2006.

The lists of children in each selected class were provided as classroom rosters on a rolling basis, two weeks before the scheduled site visit for that program. The site visits were scattered over a nine week period in the fall of 2006.

5. Sampling Issues

5.1 Program Level

As mentioned earlier, the PIR that was available at the time of program sampling (which took place in the spring of 2006) was the one reflecting the 2004-05 program year and was submitted in August of 2005, so its data were a year old by the time we went into the field in the fall of 2006. Furthermore, the PIR did not contain the measure of size needed to do the PPS sampling: the number of enrolled 3-, 4-, and 5-year-olds new to Head Start. The PIR has the number of children in each of those age groups, and has the percent of children who were in Head Start last year, but not the combination of the two. We estimated the number by applying the proportion that were not in Head Start last year to the number of children in our targeted age group.

We selected a stratified PPS sample of programs, using control variables to sort the frame within the explicit strata before sampling. The stratification variables were census region, urbanicity (metro or non-metro), and proportion of children who were black or Hispanic (less than 40 percent of both, or more than 40 percent of either). The control variables were: whether the program was a public school; the percent of children who whose primary home language is English (10 categories); and the percent of children with disabilities (IEPs or Individual Education Plans). To select the sample, we used a sequential sampling technique developed by Chromy (1979) and available in SAS SurveySelect.

We selected 120 programs, twice our target, and then paired adjacent selections within stratum, which were likely to be similar in terms of the control variables. We randomly selected one from each pair to be part of the main sample release. The other member of the pair was released only if the main release turned out to be ineligible (n=1) or a refusal (n=3). When both members of the pair were released, this was reflected in their probability of selection. We then accounted for the refusing programs in the nonresponse adjustment to the program-level weights. This procedure, which we have used in a number of similar projects, works well to give us control over the final sample size while still allowing us to select a probability sample with quantifiable selection rates.

[1] Note that some children are enrolled in Head Start even though they are 5 years old as of the kindergarten cutoff date, delaying kindergarten enrollment by one year. These children are combined with the 4-year-old cohort for purposes of this study.

5.2 Center and Class Level – Measures of Size

Some children receive Head Start services through a home visitor, rather than at a center. This was our first sampling challenge—how to sample the home visitors. Should we treat them like centers or like classes? In some (but not all) programs, the home visitors are associated with particular centers, in which case it would make sense to treat home visitors like classes or teachers, which we ultimately decided to do. Even if a home visitor was not directly associated with a particular center within a program, we asked the program director what center the home visitor's families tended to go to for socialization and other services. In the remainder of this paper, references to class-level sampling also include the sampling of home visitors.

We obtained the lists of centers and classes on a rolling basis over the course of the summer of 2006, which was often before the fall enrollment counts were known. In particular, many programs and centers were unable to predict the fall enrollment by age (3-year-old vs. 4- or 5-year-old). We asked each program and center to give us their best estimates of the number of 3s and 4s new to Head Start so that we could use these as the measures of size in the sampling.

It turned out that the size estimates were often quite different from actual sizes, in both directions---sometimes they were too high and sometimes too low. Large discrepancies between predicted and actual measures of size can introduce variability into the sampling weights. Suppose PMOS$_i$ is the measure of size for program *i*, CMOS$_{ij}$ is the measure of size for center *j* in program *i*, TMOS$_{ijk}$ is the measure of size for class *k* in center *j* in program *i*, and KMOS$_{ijk}$ is the actual number of children in class *k*. Suppose P_h is the number of programs selected in stratum h, and that we are selecting 2 centers per program, 3 classes per center, and 10 children per class. The probability of selection is calculated as:

$$\frac{P_h PMOS_i}{\sum_h PMOS_i} \cdot \frac{2CMOS_{ij}}{\sum_{hi} CMOS_{ij}} \cdot \frac{3TMOS_{ijk}}{\sum_{hij} TMOS_{ijk}} \cdot \frac{10}{KMOS_{ijk}}$$

If the estimated measure of size in one stage of sampling matches the corresponding measure of size in the next stage (that is, if PMOS$_i$ = $\sum CMOS_{ij}$ and CMOS$_{ij}$ = $\sum TMOS_{ijk}$ and TMOS$_{ijk}$ = KMOS$_{ijk}$) then this formula reduces to:

$$\frac{60P_h}{\sum_h PMOS_i}$$

which is constant within stratum. To the extent that these equalities do not hold, the probabilities vary within stratum, which means the weights will vary and will increase the variance of estimates.

The size estimates were often inaccurate. This is shown in the two tables immediately below. The first one shows the ratio of (a) the estimated number of children per center (summing over all classes) at the time we selected the classes to (b) the estimated number of children per center at the time we selected the centers:

Ratio of Two Estimates of Number of Children Per Center

	3-yr-olds	4-yr-olds	Total
Mean	1.05	1.32	1.06
Range	0.06-5.58	0.06-16.00	0.06-5.00

Constructing a similar ratio for the number of children per class, we have in the numerator (a) the actual number of children per class at the time we selected the children, and in the denominator (b) the estimated number of children per class at the time we selected the classes.

Ratio of Estimated to Actual Number of Children Per Class

	3-yr-olds	4-yr-olds	Total
Mean	1.07	0.92	0.95
Range	.08-4.25	.06-5.50	.06-3.50

While these ratios were close to one, on average, the wide range of ratios shows how unreliable the size estimates could be.

For the child sampling, we did not want to get the classroom rosters more than two weeks before the site visit because of the dynamic nature of classroom composition, especially during the first several weeks of the school year. Throughout the year, children enter Head Start programs, drop out of them, and change classes within programs. By gathering the roster, which serves as the sampling frame, two weeks before the time of the visit, we hoped to have minimal changes in the classroom composition between the time of sampling and the site visit, while still allowing time to process the selected sample (generate and send the needed paperwork) and obtain parental consent for them. Yet a number of discrepancies in classroom composition still occurred. Some were actual changes occurring during those two weeks, while others were likely due to errors on the roster at the time it was generated.

5.3. Center and Class Level – Grouping

There were some other complications in the sampling process. To ensure that enough children would be in the selected units (and in the sample as a whole), we needed to group some small centers and classes before sampling. If there were fewer than 10 estimated children in a center, we grouped it with a geographically proximate center in the same program. (For one program, we had to form a triple of centers to get a sufficient sample size.) Similarly, if a class had fewer than 10 estimated children, we grouped it with another class. Within a center group, we grouped the largest class with the smallest, the second largest with the second smallest, etc., until all class groups had at least 10 children or until all possible class pairs were formed. (No grouping was done at the program level.)

Once a center or class group was formed, it was treated as a sampling unit for that stage and all subsequent stages of sampling. However, this has budgetary and logistical implications, because when a center group containing two centers is sampled, it requires extra travel for the field staff, as well as obtaining the consent and interviewing of two center directors. A class group requires that there be two teacher interviews and two classroom observations.

Unfortunately, more grouping was needed than we expected. There were often too few eligible children per center or per class, and we sometimes selected all centers in a program, once we grouped them and selected two center groups per program. There were 4 out of 60 programs in which all centers were selected, and 1 of these 4 had all its classes and all its children selected. More commonly, we selected all classes in a center, once we grouped them and selected three class groups per center group. Seventy of the 121 selected center groups had all their classes selected, and 60 of these 70 had all their children selected. Eighty percent of the 284 selected class groups had all of their children selected.

And even with all of this grouping, we still ended up with an initial sample shortfall at the child level, but with more centers and classes than targeted. Because of the rolling nature of the sample and the quick turnaround required after receiving the roster, we did not identify the child sample shortfall until a few weeks into the nine-week sampling and data collection period.

Final Fall 2006 Sample Sizes at the Program, Center, and Class Levels

	Program	Center	Class
Sampled/Released	64	140	415
Eligible	63	135	410
Participating	60	135	410
Target	60	110-120	300-350
Sum of Weights	1,630	14,128	42,973

We did come in on target with the program sample size, but that was due to our replacement sampling scheme for this first stage of sampling described above.

5.4. Oversampling Three-Year-Olds

The initial sample design, which was modeled on previous cohorts of FACES, began with the selection of 3,274 children in the fall of 2006. We expected a parental consent rate of 90 percent (n=2,947). We expected a 95 percent completion rate (n=2,799) for child assessments and parent interviews in fall 2006 among the children with parental consent.

We would follow the 3-year-old cohort for 2.5 years until the spring of 2009, and would follow the 4-year-old cohort for 1.5 years until the spring of 2008. By the time the two age cohorts got to the spring of their kindergarten year, we expected to get 1,937 completed parent interviews. (Study protocol dictated that we not follow children who left Head Start before the start of their kindergarten year, so we factored in this type of attrition.)

Furthermore, we had assumed that 45 percent of the study-eligible children would be in the 3-year-old cohort (based on figures in an OMB submission from a prior round of FACES). If we sampled children according to that 45:55 proportion, we would start with 1,473 3-year-olds and 1,801 4-year-olds selected.

According to our assumptions, this would then yield 766 parent survey completes for the 3-year-old cohort and 1,171 completes for the 4-year-old cohort in the kindergarten year. This disparity in sample sizes would be due to the smaller proportion of 3-year-olds at the outset, and the extra year of followup and attrition for the 3-year-olds.

There were obvious advantages to having comparable sample sizes between the two age cohorts in the kindergarten year. To optimize contrasts between the two groups, it was necessary to oversample the 3-year-olds and select a larger sample overall. Instead of selecting 3,274 children, we would select 4,051. Among these, 1801 would still be 4-year-olds, but we would select 2,250 3-year-olds instead of 1,473. This would give us 1,171 children in each of the two age cohorts at the time of the kindergarten year, according to our assumptions.

Data Collection Period		Fall 2006			K year
Type of Complete		Selected	Consented	Parent Interview	
No Over-sampling	3s	**1,473**	1,326	1,260	**766**
	4s	**1,801**	1,621	1,540	**1,171**
	Tot.	3,274	2,947	2,799	1,937
Over-sampling	3s	**2,250**	2,025	1,924	**1,171**
	4s	**1,801**	1,621	1,540	**1,171**
	Tot.	4,051	3,646	3,464	2,342

When oversampling the units in a particular sampling stratum, one usually knows the population size in each stratum—in this case, the 3-year-old vs. 4-year-old stratum. Unfortunately, this population distribution was not known for the 2006-07 study year. Because prior cohorts of FACES did not attempt to explicitly stratify by age cohort, the age mix of Head Start classes was irrelevant to the previous sample designs. We had the 45:55 proportion based on estimates from the prior round of FACES. Anecdotally, we knew that the age-mix in Head Start was changing over time, with proportionally more 3-year-olds due to the increasing availability of state-funded pre-kindergarten for 4-year-olds in some areas.

We initially planned to oversample 3-year-old classes, under the assumption that most classes were single-age classes. It turned out that the vast majority of Head Start classes were mixed-age, containing both 3- and 4-year olds. With mixed-age classes we obviously could not oversample 3-year-old classes as our main vehicle for oversampling 3-year-old children. But we could at least give classes with all (or more) 3-year-olds a somewhat higher chance of selection. We created a synthetic measure of size for our PPS sampling of classes that was equal to:

$MOS_{class} = (1.5 \cdot N_3) + N_4$ rather than

$MOS_{class} = N_3 + N_4$

where N_3 and N_4 are the number of 3- and 4-year-olds, respectively in the class group. (If the center was not able to break down the number of children by age group, but knew it would be mixed age, we imputed .46 to be 3-year-olds and .54 to be 4-year-olds based on the distribution among the mixed age classes with known age breakdown.)

Knowing this approach was not enough to reach our oversampling targets for 3-year-olds, we decided to oversample them at the child-sampling stage. As shown earlier, we were aiming for initially sampling 2,250 3-year-olds and 1,801 4-year-olds, a 56:44 percent split, to end up with a 50:50 split at the kindergarten data collection point. Our oversampling plan had two phases. First, we looked at all the class rosters within a center and determined the proportion of children that were 3-year-olds. If this proportion was 56 percent or greater, we used a proportional sample allocation across the two age groups. If not, then we calculated what was needed to get to 56 percent within the center (subject, of course, to the total number of 3-year-olds available), and oversampled to get that number. The 4-year-olds comprised the remainder of the sample.

5.5. Child Sample Release and Replicates

Aside from the oversampling issue, we had planned to release the child sample in replicates to better control the final sample size. We would select 20 children per class group instead of the 10 targeted, then randomly subsample 10 of these 20 to be the main sample release. We would randomly order the other 10 selected children and fix that order in place. These other 10 children were to be released as needed, one by one, to account for ineligible and nonparticipating children among the main release.

We have used this method successfully in other studies to help regulate the sample size while still being able to quantify the probability of selection. It does require some training for the field staff in terms of when and how to "dip into" (release) the reserve sample. In terms of the sampling probabilities, all children through the last one released on the reserve list are considered to be released. However, after we completed the site visits for the first 20 programs (out of 60), we decided to modify this procedure and simplify it significantly. This was done mainly to address the emerging sample size shortfall described earlier.

For the last 40 programs, we abandoned this release process. Instead, if there were fewer than 80 children in a program, we selected them all and, if there were fewer than 40 children in a center, we selected them all. Otherwise, we used the two-step process described earlier, but selecting only 10 children per class group: (1) looking at whether proportional allocation between the age groups achieved the target of 56 percent 3-year-olds, then (2) if not, calculating how many 3-year-olds were needed within the center to reach 56 percent and oversampling accordingly.

6. Final Child Sample

Below we show the final fall 2006 frame and sample selection sizes by age cohort.

	All Children in Selected Classes*		Children Sampled and Released	
	N	%	N	%
3-yr-olds	2,460	58.2	2,256	59.1
4-yr-olds	1,765	41.8	1,561	40.9
Total	4,225	100.0	3,817	100.0

*Recall that classes with more 3-year-olds had a higher chance of selection.

We note here that the ratio of 3-year-olds to 4-year-olds in the frame is 58:42, and not the 45:55 we had assumed when we developed the data collection plan. And after the child-level oversampling process, this ratio changed to 59:41, which exceeded our goal of 56:44.

This table shows the final sample of eligible children with parental consent compared to our targeted number.

	Eligible Children with Parental Consent			
	Actual		Targeted	
	N	%	n	%
3-yr-olds	2,017	60.8	2,025	55.5
4-yr-olds	1,298	39.2	1,621	44.5
Total	3,315	100.0	3,646	100.0

We came close to our target for the 3-year-old cohort, but our sample had too few 4-year-olds and too small a sample overall compared to our target. It should be noted that, despite this initial sample shortfall, our sample sizes in spring 2007 met our expectations given higher-than-expected response rates.

7. Conclusions

So what were the lessons learned? We learned that estimates of class sizes and age mix for Head Start programs are unreliable before the program year starts. If we knew then what we know now, would we have done anything differently? Probably not. The purpose of this paper is to present the difficulties encountered when dealing with a real-life sampling situation that may not be uncommon for other survey resesarchers. For early childhood education or even school-based research, it is often necessary to select some sample stages before the school year starts, to allow for centers or schools to be recruited into the study and for other logistical steps to occur. And that means we may not have the luxury of knowing what the population will look like in the end stages of sampling. The three major consequences of this are:

(1) Lack of accurate size estimates. This refers to the measure of size used in one stage, and how it compares to the actual frame size in the next. As described above, we ideally want the measure of size and the frame size to match.

(2) With the quick turnaround for sample selection and the rolling process for obtaining sample frames, it is very hard to see the big sampling picture. It was difficult to adjust and adapt the sampling algorithms for children to account for the emerging sample shortfall, and for site-specific issues such as misspecified measures of size and certainty selections.

(3) Trying to base sample allocation, especially in the presence of oversampling, on a partial or inaccurate picture can lead to samples sizes over or under the target.

Acknowledgment

The author would like to thank Shinu Verghese for her valuable contributions to the sampling process and also for providing many of the figures used in this paper, and would like to thank Jerry West and John Hall for providing valuable comments.

References

Chromy, J.R. "Sequential Sample Selection Methods." Proceedings of the Survey Research Methods Section of the American Statistical Association, 1979, pp. 401-406.