SAMPLE DESIGN ISSUES FOR THE BASE YEAR OF A LONGITUDINAL SURVEY OF KINDERGARTEN CHILDREN

John Burke, Thanh Lê, John Michael Brick, Westat Inc.,
Thanh Lê, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Longitudinal survey, multi-stage sampling, oversampling, coverage, precision of estimates.

1. Introduction

The Early Childhood Longitudinal Study: Kindergarten Class of 1998-99 (ECLS-K) is sponsored by the U.S. Department of Education, National Center for Education Statistics (NCES). It will provide national data on children's characteristics as they progress from kindergarten through the fifth grade. It will also provide information on key analytical issues such as school readiness; transition to kindergarten and subsequent grades; kindergarten and first grade student performance; and, cognitive growth and student progress.

The ECLS-K will collect data on a nationally representative sample of approximately 20,000 children enrolled in about 1,000 kindergarten programs beginning with the 1998-99 school year. During this school year there will be two data collections, one at the beginning (fall) and one near the end (spring). Thereafter, most follow-up studies will be conducted in the spring, beginning with spring 2000. In the fall of 1999, data will be collected on a 25 percent subsample of first graders that will provide information to assess children's growth over the summer vacations. Data collection will consist of direct assessments of the students themselves, interviews with their parents, as well as abstracts of school records. Teachers and school administrators will complete self-administered questionnaires. In the base year, the sample of children is selected using a multi-stage probability design. The first-stage or primary sampling units (PSUs) are geographic areas that are counties or groups of counties. In the second stage, samples of public and private schools with kindergarten programs are selected within the sampled PSUs. Both PSUs and schools are selected with probability proportional to measures of size that take into account the desired oversampling of Asians and Pacific Islanders (APIs). The third stage sampling units are children of kindergarten age, selected within each sampled school.

In this paper, we discuss the evaluation of alternative designs for sampling PSUs, the method of sampling within PSUs, and the sampling of students within schools. The school sampling frames are also described, as well as procedures adopted to improve the school coverage, separately for public, Catholic and non-Catholic private schools. Other features of the design, such as procedures used to include students in the follow-up collections, are not discussed here.

2. Sampling PSUs for the ECLS-K

2.1 Issues Under Consideration

In the base year, the design for the ECLS-K involves a clustered sample of PSUs that are counties or groups of counties; in the second stage, about 800 public and 200 private schools are selected from the sampled PSUs; the final stage is the selection of a fixed number of about 24 students from each sampled school. In subsequent years, students will be followed as they move to first grade and beyond, with subsampling of students that will largely be determined by probabilities that are a function of how many sampled students move into the same school so that the cost of data collection can be contained. The number of study schools in the subsequent years is expected to be substantially larger than the number in the base year because of this migration. The clustered design is necessary to limit the costs of data collection that are highly related to the dispersion of the children.

The primary focus of the analysis of the ECLS-K data will be at the student level, as indicated by the issues of interest such as school readiness and transition to kindergarten and subsequent grades. The optimal sample design for student level estimates is to sample students with probabilities that are approximately the same for each student. In most designs, this is achieved by sampling PSUs and schools with probabilities proportional to the number of students and selecting a fixed number of students per school. An equal probability student sample in the subsequent years would also be optimal if the data collection costs were roughly equal, but unequal probabilities may be necessary to account for the cost efficiency associated with sampling students clustered in the same first grade school.

On the other hand, school level estimates in the base year are more efficient if the schools have equal probabilities of selection, irrespective of the number of students in the school. A compromise scheme that is very useful when both school and student estimates are of equal importance is to select schools with probabilities proportional to the square root of size

(this is between equal probability and probability proportional to the number of students). However, if this procedure is followed and an epsem sample of students is to be achieved, students have to be subsampled at rates that equalize their probabilities of selection. This results in a different number of students being sampled per school. In the ECLS-K, a fixed student sample size per school is highly desirable for burden and cost reasons, so that this option is not viable.

This brief review of the relationship between the sampling probabilities and the analysis issues for the ECLS-K is intended to provide some background for subsequent discussions of the alternatives for sampling PSUs, schools, and students. Two design options, A and B, are considered below.

2.2 Two Alternative Designs

Option A involves using the existing Private School Survey[1] (PSS) first stage sample of PSUs. In PSS, about 124 PSUs are sampled with probabilities proportional to the square root of total population. Let f be the overall student-sampling fraction in the ECLS-K. Since the PSU selection probability P_1 is known from the PSS (PSUs are already selected), and the conditional probability of sampling students within a sampled school P_3 is fixed (24 per school), the conditional probability of selection of a school in a PSU (P_2) can be calculated as $P_2 = f/(P_1 P_3)$. It is easy to show that the number of schools sampled per PSU will vary if the PSS PSUs are used in order to obtain an equal probability selection of students. Also, the overall probability of selecting a school under this option is the product of P_1 and P_2 and this is not an equal (or a square root of size) probability sample of schools.

Option B involves selecting a new sample of 100 PSUs, with probabilities proportional to the number of kindergarten students (or the count of five-year-old children in a PSU, which is a close surrogate for this). The second stage sampling of schools is proportional to the number of kindergarten children in the school and, in the third stage, 24 students are sampled per school. The number of schools sampled per PSU should be approximately constant (on average 8 public and 2 private schools would be sampled per PSU if 100 PSUs and 1,000 schools are sampled). The overall probability of selection for a school is close to proportional to the number of students in the school.

We evaluate the two options by looking at the student coverage and the precision of the estimates. These are two factors that are most likely to have different characteristics under the two options. There are also cost differences that are discussed at the end of this section.

2.2.1 Coverage

Coverage of all students in the base year sample is of great importance to the ECLS-K. In a longitudinal survey, biases in the base year are often carried throughout the multiple years of the study. The most important coverage concern in the ECLS-K is the coverage of students in private schools. Private schools have much greater rates of both openings and closings than do public schools so that using old sampling frames can lead to coverage bias.

Under both options, the most current Common Core of Data[2] (CCD) sampling frame is used for the public school sector and the most recent PSS list frame is used for the private school sector. These sampling frames can be partitioned into the sampled PSUs so that samples of schools within each PSU can be selected.

Data from previous PSS samples indicate that updating the area frame increases the estimated number of all private schools by about 8 percent and the estimated number of all private school students by about 3 percent. For kindergarten, the percentage increases are probably greater. If nothing is done about the coverage, the ECLS-K could exclude up to 5 to 7 percent of children in private kindergarten programs and about 1 percent of all children in kindergarten (5 to 7 percent of about 15 percent of children in private kindergarten). While the overall loss is small, the loss within the private school sector is too large to ignore.

Option A was thought to be better because the PSS is an on-going survey that includes an area sample to improve coverage of private schools and students, and because of its potential integration with the ECLS-K. The search for new private schools is conducted every two years, so the frame is more

[1] The PSS, conducted by the Bureau of the Census for NCES, is designed to build a universe of private schools in the United States. The main component of the PSS is the list frame. Data sources for building the list frame are commercial lists, state lists and private school association lists. An area frame is included to identify schools overlooked in the list frame. This area search for additional schools is conducted in randomly selected counties. For more details, see NCES (1998) and McMillen (1993).

[2] The CCD is the NCES database of elementary and secondary public schools in the United States and its territories. It collects data on schools and state and local school districts (or education agencies), mostly from administrative records. The database contains information on schools, school districts, students and staff, as well as fiscal data. For more details, see McMillen (1993).

complete within these PSUs. However, the private school sampling frame in the PSS is outdated because no fieldwork has been done since 1995. The search for new schools in PSS for 1997-98 was not completed in time for ECLS-K school sampling. As a result, the same work would be required for both options, and there would be no coverage advantage if the PSS PSUs were used.

2.2.2 Precision of the Estimates

The precision of the estimates is affected by the sample design in various ways. For the ECLS-K, the two main factors that cause losses in precision relative to a simple random sample design are the clustering of schools and students within the sampled PSUs and the variability in the sampling rates or weights of the units. These two issues are discussed below using estimates of the number of sampled schools and students.

In almost all clustered samples, the precision of the estimates is reduced relative to simple random sampling because units within the same cluster tend to be more homogeneous than units across the entire population. This will be true in the ECLS-K since the clusters are geographic areas, and schools and students within the same geographic area are almost always more homogeneous. In most multi-stage samples the effect of clustering (on the variance of the estimate) at the PSU level is approximated well by the expression (Kish, 1965)

$$D_1 = 1 + \rho(b-1)$$

where ρ is the intra-class correlation coefficient indicating the degree of homogeneity within the PSU and b is the average sample size in the cluster (in this case, the number of sampled students).

This formulation breaks down if the average sample size per cluster is not constant. This is exactly what occurs in Option A. In this case, a better approximation (Holt, 1980) is given by

$$D_2 = 1 + \rho(b'-1)$$

where $b' = \frac{\sum b_i^2}{\sum b_i}$. Note that if the sample size per cluster is a constant across PSUs, then the two expressions are equal.

Because of the variation in the cluster sample sizes, we used D_2 to compute the effect of varying the student cluster sample size for both options. The ratio $R = D_2 / D_1$ is the expected increase in the variance of public school student level estimates. The results are shown in Table 1 for the two options and different values of ρ. The variability in the Option A student sample sizes by PSU results in a substantial increase in the variance of estimates of public school students. This increase may be understated because the comparison assumes equal values of ρ under the two designs while under Option A the PSUs are smaller and likely to have larger values of ρ. Similarly results apply to private schools.

Table 1. D_2 and $R = D_2 / D_1$ for characteristics of students from public schools

	Option A		Option B	
ρ	D_2	$R = D_2/D_1$	D_2	$R = D_2/D_1$
.01	3.13	1.23	2.84	1.03
.02	5.25	1.28	4.68	1.03
.03	7.38	1.31	6.52	1.04
.05	11.63	1.33	10.20	1.04
.10	22.26	1.35	19.41	1.04

The discussion of the increase in the variance for student level estimates under both options does not recognize the effect of differential weights on the estimates. This is appropriate because both options have approximately self-weighting samples of students. The same approach is not appropriate for school level estimates because neither option results in a self-weighting sample of schools. The effect of weights on the school sample was computed for both options using the following formula (Kish, 1976)

$$D_W = 1 + L = (\sum U_i k_i)(\sum U_i / k_i)$$

where U_i is the size of unit i and k_i is the base weight of unit i.

Using this formula, the effect due to weighting for school level estimates D_W is 2.17 for Option A and 2.63 for Option B for characteristics of public schools. This factor must then be multiplied by the appropriate D_2^* factor (computed using the number of schools sampled per PSU rather then the number of students per PSU) for each of these options to get the combined effect of the design on the variance of public school estimates. The results of this multiplication, which include the effect associated with the different numbers of PSUs in the two designs, are shown in Table 2 for the two options and different values of ρ.

Table 2. D_2^* and $D_s = D_2^* D_W$ for characteristics of public schools

ρ	Option A		Option B	
	D_2^*	$D_s = D_2^* D_W$	D_2^*	$D_s = D_2^* D_W$
.01	1.08	2.34	1.07	2.81
.02	1.16	2.51	1.13	2.98
.03	1.24	2.68	1.20	3.16
.05	1.40	3.03	1.34	3.51
.10	1.79	3.88	1.67	4.40

The increase in the variance is less under Option A than Option B for school level estimates. As mentioned before, the value for ρ is probably smaller for Option B than for Option A. Because student estimates are more important to the study objectives, Option B was used for ECLS-K.

2.3 The ECLS-K PSU Sample

Following the parameters of the Option B design, the ECLS-K sample consists of 100 PSUs which are counties or groups of counties. The distribution of five-year-olds based on 1994 population estimates by race/ethnicity was used to form PSUs with a minimum size of 320 five-year-olds and to construct a measure of size that took into account the oversampling of API children. The PSUs were stratified into self-representing and non-self-representing. There are 24 self-representing PSUs. For the non-self-representing PSUs, 38 strata of roughly equal measure of size were created, and two PSUs were selected in each stratum, yielding 76 non-self-representing PSUs. The variables used for stratifying the non-self-representing PSUs were MSA/non MSA status, and region. In the next level of stratification, size class, race/ethnicity (high concentration of API, Black or Hispanic) and per capita income were used for MSAs, and race/ethnicity and per capita income were used for non-MSAs.

The measure of size used for selecting PSUs takes into account the oversampling of APIs. The weighted measure of size is calculated as $2.5 \times n_{API} + n_{other}$, where 2.5 is the oversampling rate for APIs, n_{API} and n_{other} are the counts of five-year-old APIs, and all others, respectively.

3. Sampling Within PSUs

3.1 School Sampling Frames

In the second sampling stage, public and private schools offering kindergarten programs were selected. The target number of schools was set at 800 public and 200 private schools from within the ECLS-K sampled PSUs. The number of schools selected is the target number of schools adjusted upward by an expected school response and eligibility rate. In total, 934 public schools and 346 private schools were selected with probability proportional to the measure of size described below.

The school frame for the ECLS-K was built using several data sources: the 1995-96 CCD, the 1995-96 PSS and the 1996 lists of schools run by the Bureau of Indian Affairs and the Department of Defense. Data from the 1997-98 PSS list frame and the Quality of Education school and district files were used to update school location information. The constructed ECLS-K school frame included 18,891 public schools and 12,412 private schools with kindergarten programs within the sampled PSUs. The school frame was augmented in the spring of 1998 to include schools that are operational but were not included in the frame, as discussed in Section 4.

3.2 School Measure of Size

Schools were selected with probability proportional to size. The measure of size was constructed taking into account the oversampling of APIs, separately for public and private schools. The measure of size for school j in PSU i is calculated as

$$SCHMOS_{ij} = 2.5 \times n_{API.ij} + n_{other.ij}$$

where 2.5 is the oversampling rate for APIs, $n_{API.ij}$ and $n_{other.ij}$ are the counts of API kindergarten students, and all other kindergarten students, respectively, in school j of PSU i.

3.3 Clustering of Schools

Schools with fewer than 24 students (public) or 12 students (private) were clustered together within PSUs in order to obtain a sample that is closer to self-weighting. For example, if a school with 12 students was not clustered the students from that school would be sampled at about half the probability as students in larger schools. The goal was to group small numbers of schools to form heterogeneous clusters with an aggregate number of students as close to 24 as possible. This goal was set so that if a cluster was selected we would not need to recruit many small schools; furthermore, the heterogeneity of schools improves the reliability of the estimates. We defined heterogeneity for public schools by school size and for private schools by religious affiliation and school size.

3.4 Stratification of Schools

The schools were stratified implicitly within each PSU. For public schools, (clusters of) schools were sorted by the measure of size and separated into three size classes of roughly equal size (high, medium, and low). Within each size class, they were sorted by the proportion of APIs in a serpentine manner. In private schools, each cluster was identified as religious, mixed, or non-religious. The list of clusters was then sorted by these three categories. Within each category, clusters were sorted in a serpentine manner by the measure of size.

3.5 Sampling Students

In the third stage, 24 students will be selected for the base-year study in each school (or fewer when the school does not have 24 students), with oversampling of API students. For the ECLS-K, PSUs and schools were sampled assuming that API students would be oversampled by a factor of 2.5. However, in about 40 percent of the school sample, it may not be possible to select a total of 24 students while oversampling API by a factor of 2.5. We determined that the oversampling factor would have to be as high as 5.5 in order to meet the target. Increasing the oversampling factor would have the unfavorable effect of increasing the variability of the weights and increasing variances. Therefore, we have chosen to oversample the API students by a factor of 3. In subsequent years, API students may be followed at a higher rate as they transfer to new schools to reduce the attrition in this domain.

4. Improving Coverage of Schools

The sampling frames used for the main sampling of schools offering kindergarten programs were augmented to include newly opened schools that were not included in the frame. Some schools that were in the CCD and PSS but not included in the ECLS-K frame for various reasons were also included in this process. Procedures for augmenting the frames were different for public schools, Catholic schools and non-Catholic private schools. Each is discussed below.

4.1 Public Schools

The sample of 934 public schools falls in 535 school districts in 41 states and the District of Columbia. The sampled districts were asked if any school expected to offer a kindergarten program in 1998 or any ungraded school was missing from a list sent to them (developed from the original frame). Districts that were in the sampled PSUs but were not reported in the CCD as having any schools with kindergarten programs were also contacted. If they had any schools that would offer a kindergarten or ungraded program in fall 1998, information on these schools was collected. The information obtained from the school districts was checked against the ECLS-K public school frame to confirm that they truly were new or newly eligible. Checking was not restricted to within the school district but was done within state in order to ensure that each "new" school was not already listed under a different district and that it was new in this district due to district reorganization. Through this process, 252 new public schools were identified. A sample of 19 schools was selected. Since a district identifies a new school, each school was selected with a probability conditioned on the within stratum probability of selecting that district.

4.2 Catholic Schools

There are 117 Catholic schools in the ECLS-K sample in 59 dioceses. The procedure for contacting the dioceses and for obtaining new school information was exactly the same as for public schools. Since a diocese can cover more than one city or county and can sometimes cut across state, checking included an additional step of placing each school on the list sent by the diocese in the correct county and hence the correct PSU, before checking for new schools. A total of 117 new Catholic schools were identified, and 6 were sampled. As for public schools, the new school selection probability is conditioned on the within stratum probability of selecting the diocese that identifies the new school.

4.3 Non-Catholic Private Schools

The main source used to search for non-Catholic private schools was the telephone book Yellow Pages. In addition, local education agencies (LEAs) and local government offices were also contacted for information on non-Catholic private schools in their areas, but this was only implemented in 22 PSUs with large PSU weight (greater than 20).

For all the counties in the ECLS-K sample, electronic Yellow Pages listings of elementary schools, private and parochial schools, special education schools, preschools, nurseries and kindergartens were created. The procedures involved matching these listings in the sampled PSUs to various ECLS-K data files in order to purge, to the extent possible, schools that were already in the ECLS-K frames. Schools that were on the PSS file but were out-of-scope or did not contain any kindergarten children according to the PSS were also included. The files were matched and school names of matches and near-matches were examined in order to decide on

true matches. Non-matches were put through a screening of school names using keywords to exclude any that had 'high school', etc. in the name. The private school frame constructed using the Yellow Pages had 11,405 schools in the sampled PSUs. A sample of 279 schools was selected and then screened for eligibility. Of these, only 85 schools reported that they were private, would be open in fall 1998 and would have kindergarten or kindergarten-age students. These schools were added to the main sample.

The 22 PSUs with largest PSU weights cover 53 counties. In these counties, 135 LEAs and 218 cities/towns were identified. In each city/town, a list of local government offices was compiled using the electronic Blue Pages. The telephone interviewers contacted all LEAs. However, in cities/towns with multiple government offices, they contacted more than one only if the first call did not yield any information on private schools. Of the 135 LEAs, only 54 had information on private schools in their area. After the information collected was unduplicated against PSS and the Yellow Pages, 30 new private schools were identified. Of the 218 cities/towns, only 75 yielded information on private schools. After the information was unduplicated against PSS, LEAs and the Yellow Pages, 19 new private schools were identified. In addition to these procedures, three new private schools were reported by the field staff working in the area. Of the 52 new schools identified, 24 were sampled and screened to ensure that they are eligible for the ECLS-K.

In total, 134 new schools (public, Catholic and non-Catholic private) were added to the original sample of 1280 schools. Since the grade span of some of these schools was not known at the time of sampling, particularly the non-Catholic private schools, the actual number of productive new schools added will be smaller.

5. Conclusion

In this paper, two alternatives for sampling PSUs for the ECLS-K were examined. A new sample of PSUs with a measure of size appropriate for the study was found to provide much greater reliability than using a sample with a less efficient measure of size. The sample of schools was selected from a school frame that is somewhat dated, and then augmented to improve coverage, particularly important in the case of private schools.

6. References

Holt, D. (1980). Discussion of the paper 'Sample designs and sampling errors for the World Fertility Surveys' by Verma, Scott and O'Muircheartaigh. *Journal of the Royal Statistical Society* (Series A), 143, Part 4, p. 468.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Kish, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society* (Series A), 139, Part 1, p. 80.

McMillen, M., Kasprzyk, D., and Planchon P. (1993). Sampling frames at the United States National Center for Education Statistics. In *Proceedings of the American Statistical Association Conference on Establishment Surveys*, pp. 237-243.

National Center for Education Statistics (1998). *Private School Universe Survey, 1995-96.* NCES 98-229. Washington, DC: National Center for Education Statistics.