EFFECTIVE STRATIFICATION WITH FEW STRATA AND OTHER SAMPLE DESIGN ISSUES IN A CHILD CARE SURVEY

Gary M. Shapiro, Westat, Inc.[1]
Gary M. Shapiro, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Keywords: **Multi-stage sampling, Early Childhood and Child Care Study, Sampling states**

1. Introduction

The major topic of this paper is to describe the stratification methodology used for the primary sampling units (PSUs) in the Early Childhood and Child Care Study (ECCS). Only twelve strata were formed for the survey. With such few strata, ordinarily only one or two stratification variables would be used. For this study, however, there were four stratification variables that were considered important. There was concern that ignoring two of the stratification variables in forming the strata would lead to heterogeneous strata and large between PSU variances. The paper describes how all four stratification variables were used in forming the strata. The methodology for using all the variables is ad hoc and simple to apply. This methodology is potentially useful for other surveys in which there are few strata and a desire to use more than one or two stratification variables.

The ECCS has a very complex design with a number of interesting aspects. However, this paper only discusses in detail two topics that are likely to have applications to other surveys. A brief overview of the ECCS is first presented. This study was carried out by Abt Associates Inc. for the Food and Consumer Service of the Department of Agriculture. The objectives are to provide descriptive information on the institutions and children that participate in the Child and Adult Care Food Program, descriptive information on the nutrient content of meals offered under the program, and an assessment of the contribution of foods consumed while in the child care setting to the total daily diet of participating children.

The Child and Adult Care Food Program provides meals and snacks in child and adult care facilities. Only the child care component is covered by the ECCS. Funds are provided for meals and snacks for children in three types of non-residential day care facilities, called providers: Family day care homes (FDCH), head start centers (HS), and child care (CC) centers. CCs are all day care facilities that are not HS and that are not operated as a home-based day care facility. Head start centers and child care centers are described as a combined group as simply "centers". See Glantz et al (1996) for more details on both the program and the study objectives.

There are several stages of sample selection and several sets of data collection for the ECCS. The PSUs for the survey are states, with 20 sampled. Details on this selection stage are given in the next section. Sponsors were sampled at the second stage of selection. Each FDCH and center has a sponsoring agency, although in some cases a center is its own sponsor. A sponsoring agency can have as many as 2,000 providers or at the other extreme only a single provider. Some sponsors have two or three types of providers, though most only have a single type. Sponsors were sampled by type of provider. A sponsor with two (or three) types of providers was treated as two (or three) separate sponsors for purposes of sampling. Sampled sponsors were mailed a questionnaire regarding characteristics of the sponsoring agency.

Providers were sampled at the third stage of selection. Cluster sampling among sampled sponsors was used. Sampled providers were administered questionnaires regarding characteristics of the provider, nutritional knowledge and food preparation practices of the food preparer, and menus of food offered.

Children were sampled at the fourth stage of selection. Children were selected only from a subsample of selected providers. Thus, a subsample of providers was selected prior to the selection of children. A cluster sample of children was taken. Sampled children were observed for two days while at the provider, and their parents received a telephone interview on food consumed by the child while at home. See Abt Associates Inc.(1994) for more details on the sample design and data collection plan.

The second topic of this paper is the selection methodology for sponsors. Efficient sampling of providers and children was of much greater importance than efficient sampling of sponsors. Thus, sponsors were selected to obtain an approximately self-weighting sample of providers (within provider type) with a pre-determined sample size for providers, without regard to optimal sampling for estimates obtained from sponsor questionnaires. The methodology is an unusual application of standard methodology used in sampling households and addresses.

2. Stratification of First Stage Units

The primary sampling units (PSUs) in the first stage of sample selection for this survey are the 48 continental states and Washington D.C. The original

[1] The author designed this survey while at Abt Associates Inc.

design specifications agreed to in the contract between the Food and Consumer Service and Abt Associates Inc. were for 20 of these 49 PSUs to be selected for sample. State is not usually a good choice for a PSU definition. For this survey, however, the list of sponsors from which a sample was to be selected was only available from state governments. If we could have easily obtained the information from each state, we would have skipped state as the first stage of selection and selected sponsors across all states. Obtaining the list of sponsors was time-consuming and difficult for both the states and for Abt Associates Inc. Thus, it was felt that it was crucial to confine the sample to a subset of states.

The sample of states was to be selected in traditional fashion with probability proportional to measure of size. For the measure of size, it was decided to use a weighted average of the number of meals for Family Day Care Homes (FDCH) and of meals for all centers (both Head Start(HS) centers and Child Care (CC) centers). It was also decided to select one sample PSU per stratum rather than two sample PSUs per stratum. This was an important decision since the number of sample PSU's is small. See U. S. Census Bureau (1982) for an empirical comparison of one vs. two PSUs per stratum.

As is normal in such sampling, the largest states would be self-representing, with the remaining smaller states to be grouped into strata. An initial decision was made for the 8 largest states (in terms of measure of size) to be self-representing. It was desired to have an even number of strata for the selection of the non-self-representing states to facilitate variance estimation. I was open to increasing the number of self-representing states during the stratification process. If there had been two relatively large states that were particularly difficult to combine into homogeneous strata, these would have been made self-representing. The average measure of size for the non-self-representing strata was about 30,000,000. States that would have had very high probabilities of selection as non-self-representing were made self-representing. The seventh largest state had a measure of size of almost 22,000,000 and was thus made self-representing by this criteria. The eighth largest state had a measure of size of about 18,500,000. It might have been classified as non-self-representing if not for the desire to have an even number of non-self-representing strata.

Program staff stated that there were four important variables that should be used in the stratification. Three of the variables were quantitative and the fourth variable was region. With only twelve strata to be formed, a traditional approach to stratification might have been to ascertain the two most important of these variables and to ignore the other two variables in the stratification. Had this approach been used, region and one of the quantitative variables would probably have been selected. Region is important in terms of face validity but is probably less important than the other three variables in terms of affecting between PSU variance for national data. For example, if a sample had been selected that included no states from the Southeast (which contained no self-representing states), it would be difficult to defend the sample selection to data users no matter how good a statistical argument there was that the strata were homogeneous on the quantitative variables. Thus, a traditional approach might have totally ignored two of the three most important variables in the stratification. Since this was undesirable, stratification was done so that attention could be paid to all four variables. The methodology used was ad hoc and partly subjective rather than statistically elegant. It was effective in achieving the stratification goals and did not require a great deal of staff time to implement.

The stratification variables were as follows:

1. Seven Food and Consumer Service geographical regions;
2. Relative importance of FDCH vs. Centers: [FDCH meals served]/[All meals served] as %;
3. Relative importance of HS: [Number of HS providers]/[All providers] as a %;
4. Relative importance of subsidized meals:
   [Number of subsidized meals]/[Total number of meals].
   (Providers generally receive federal government subsidies for most of the meals they serve.)

Estimates by state were available for each of the variables needed for the stratification variables.

Table 1 provides definitions of the twelve non-self-representing strata used for the survey, including values of each stratification variable for all the states.

For the regional variable, it was not deemed vital to have each stratum consist of a single region. It was even permissible to have more than two regions represented in some strata. The main goal was to ensure close to the expected number of sample PSUs from each region. For example, the Mountain Plains region had an expected 2.9 sample states before selection of self representing states. One of the self representing states was from this region. Re-calculating expected sample sizes after removing the self representing states yielded an expected 2.5 non-self-representing sample states for the Mountain Plains region. It was decided that the permissible

deviation from the expected 2.5 was 1.5, resulting in a permissible sample of between 1 and 4 non-self-representing states from this region. To ensure this, I made sure that at least one stratum consisted only of states from this region. Stratum 4 in Table 1 is the only such stratum. To guarantee that there not be more than 4 sample non-self-representing states from the region, the Mountain Plains states were confined to only 3 other strata (#10, #11, and #12). Otherwise, I allowed states from this region to be combined with other states in any manner.

For the other three stratification variables, the goal was to ensure that no stratum was particularly bad for any of the three variables. This amounted largely to ensuring that no stratum would contain a state that was at the high end of a variable as well as a state that was at the low end for the same variable. I attempted to do much better than this, but with so few states and strata it was not always possible. I was less concerned with small states with low probabilities of selection, since they contribute much less to the between PSU variance. The subsidized meal variable was treated as less important than the others, and so some undesirable strata with respect to this variable were permitted when necessary to ensure reasonable homogeneity for the other variables.

An attempt was also made to have approximately equal measures of size for all non-self-representing strata. However, since there are no operational problems with a sample state having an unusually large sample or an unusually small sample, considerable variation in stratum size was allowed if it improved homogeneity. The range was from 20,720,000 (stratum 12) to 42,020,000 (stratum 9).

My approach was to first try to form homogeneous strata within a single region. For example, for the Mountain Plains, Colorado and Iowa are reasonably close to each other for all three variables. They only have a combined measure of size of 22,200,000, so I looked for other states in the region to combine with them. I considered including Nebraska with these two states in stratum 4, but finally decided to include Nebraska in stratum 11. One consideration in this decision was that stratum 11 would only have a measure of size of 16,820,000 if it did not contain Nebraska or some other fifth state.

When it was impossible to form a stratum within a single region that had a reasonable measure of size and that was satisfactorily homogeneous (and after enough single region strata were formed to satisfy the face validity requirements for regions), strata were formed without respect to region. Stratum 9 is an example of a stratum containing states from more than 1 region in which it was possible to do very well for the other 3 stratification variables.

Stratum 2 is a situation where I initially formed an all Midwest stratum consisting of only Indiana and Wisconsin. However, as I proceeded to form additional strata, I had difficulty finding a stratum for which Arizona was similar with respect to the stratification variables. Arizona seemed to fit much better with Indiana and Wisconsin than with anything else. The Midwest had an expected sample of 0.8 non-self-representing PSUs, and thus it was not necessary to have a stratum that was all Midwest. (There were four self-representing states from the Midwest.) Thus, I added Arizona to the stratum. The small state of New Hampshire was added as well, because this seemed the best place for it.

At one point, I had a preliminary stratum consisting of Connecticut, Maine, Massachusetts, and Vermont (see Table 2). This stratum had a fairly large range in values for the FDCH variable. I finally decided to combine Connecticut and Massachusetts with two other states to form stratum 10, and to combine Maine and Vermont with three other states to form stratum 11. Both of these strata are quite homogeneous with respect to all three variables.

Stratum 8 is an example of a stratum which is relatively homogeneous with respect to the first two stratification variables, but not very homogeneous for the third. For FDCH, all the states have low percentages. For HS, this is a high percentage HS grouping. Kentucky has the highest percentage of any state. Only Pennsylvania and Tennessee (both in stratum 9) have high percentages and are not in this stratum. If this was the only stratification variable, I would probably have made a stratum consisting of Kentucky, Pennsylvania and Tennessee. However, this would not have been very homogeneous with respect to the FDCH variable. For the subsidized variable, Mississippi has the highest percentage of any state and New Jersey is also relatively high. Kentucky has a low percentage and ideally would not be in the same stratum as Mississippi. I was unable to determine a way of switching either Kentucky or Mississippi that would have worked well for the other stratification variables.

As is apparent from the discussion above, the methodology was strictly low technology - manually combining states in different ways into strata until there was reasonable adherence to the goals. This was feasible because of the small number of strata and states. I was able to accomplish the stratification without spending a great many hours on it and without any programming assistance. A precise stratification algorithm might have achieved superior stratification,

but it could not have been accomplished without expending considerably more resources. Clearly a more automated procedure would be needed when the number of PSUs are large. There was complex logic and considerable flexibility in the approach, however, which would be time-consuming to program.

3. Selection of Second Stage Units

Sponsors of providers are selected at the second stage within sample states. Although data is collected from sponsors as well as from providers and children, the sponsor data was considered less important. Thus, it was desired to sample sponsors in a manner that produced the desired sample size and sampling rate for providers. Conceptually, sampling was identical to the common practice in area probability sampling, where blocks or enumeration districts are an intermediate stage of selection to the ultimate sampling unit, a cluster of households. (See, for example, Hanson, 1978.) Sponsors correspond to blocks and providers to households. Conceptually, clusters of providers were systematically sampled, with the sponsor sample being determined by which providers were selected. For each sponsor, we had an estimate of number of providers. Placeholders for the providers were sorted by sponsor, and clusters of placeholders were selected. In the mechanics of the operation a sample of sponsors was determined prior to the actual sampling of clusters of providers, since we only obtained a list of actual providers after the sample of sponsors was selected.

If sponsor data rather than provider and child data had been most important, I would have selected sponsors with equal probability, using a conventional systematic sampling procedure. This would have been inefficient, however, for provider and child data.

Table 3 illustrates the procedure. The data in the table is part of the universe of CC centers in one of the states. The sampling rate for CC providers in this state was fixed at 1/6.4. The intended cluster size per sample sponsor was 4. Thus, for example, sponsor A has an estimated 79 CC providers and therefore a measure of size of 79/4 = 19.75. With the 1/6.4 sampling rate, this sponsor is in sample with certainty. Providers from the sponsor will be selected at a rate of 1/6.4.

Sponsor B has a measure of size somewhat smaller than 6.4. However, I decided to make it self-representing because it had a high probability of being selected. If sponsor B actually turned out to have 19 providers, one would expect a sample of 19/6.4 = 2.97 providers from it.

Sponsor C was the first sponsor that was not self-representing. The random number selected between 0 and 6.4 for the sample selection was 0.180. Since the measure of size, 3.5, is greater than 0.180, sponsor C was selected for sample. Providers were sampled at a rate of 1/3.5 from this sponsor.

We added 6.4 to the random number of 0.180 (equal to 6.58) to determine the next sponsor to be sampled. Since the cumulative measure of size from sponsor D, 6.75, is greater than 6.58, this sponsor is also in sample. Sponsor E is not in the sample, however, because the cumulative measure of size, 9.5, is less than the hit sequence figure of 12.980.

The sampling procedure is further illustrated in the table for sponsors F, G and H.

4. Conclusion

Two aspects of the sample design of the Early Childhood and Child Care Study have been discussed. The general methodology described on the use of four stratification variables in the formation of only 12 strata has applications to any survey situation where the number of strata is small and the number of potential stratification variables is relatively large.

The second topic discussed was the sample selection of second stage units, sponsors, in a manner that was optimal for the selection of third stage units, providers. The methodology resulted in differential sampling of sponsors, which was not desirable for data collected from the sponsors, in order to produce an efficient sample of providers and children. The methodology is applicable for a wide range of multi-stage sample surveys.

5. Acknowledgments

I wish to thank Michael P. Battaglia and K. P. Srinath for their helpful comments on this paper.

REFERENCES

Abt Associates Inc. (1994), *Early Childhood and Child Care Study Data Collection Plan*, submitted to Food and Nutrition Service May 13, 1994.

Glantz, Federic B., David T. Rodda, Mary Jo Cutler, William Rhodes, Marian Wrobel (1996), *Early Childhood and Child Care Study Profile of Participants in the CACFP: Final Report Volume I*, Abt Associates Inc., report for Food and Consumer Service.

Hanson, Robert (1978), *The Current Population Survey, Design and Methodology, Technical Paper 40*, Appendix B, Census Bureau, Washington, D.C.

U.S. Census Bureau (1982), *One Versus Two PSU's Per Stratum*, paper prepared for agenda topic C for the Meeting of the Panel on Redesign Issues which was held on October 15, 1982.

Table 1. Strata definitions

State and stratum	FDCH	HS	Subsidized	Region	Measure of size
Stratum 1					
Delaware	60%	1%	82%	Mid-Atlantic	3.47
Maryland	69%	2%	80%	Mid-Atlantic	12.01
Virginia	54%	6%	78%	Mid-Atlantic	10.29
West Virginia	38%	10%	77%	Mid-Atlantic	3.32
					29.09
Stratum 2					
Indiana	46%	12%	73%	Mid-West	11.58
Wisconsin	48%	5%	72%	Mid-West	11.55
Arizona	40%	6%	81%	West	11.17
New Hampshire	42%	8%	64%	Northeast	1.59
					35.89
Stratum 3					
Arkansas	47%	1%	81%	Southwest	8.16
Louisiana	56%	3%	92%	Southwest	14.25
New Mexico	65%	3%	86%	Southwest	11.73
					34.14
Stratum 4					
Colorado	65%	2%	78%	Mtn Pl	14.01
Iowa	52%	6%	68%	Mtn Pl	8.19
					22.20
Stratum 5					
Alabama	48%	7%	90%	Southeast	10.32
Georgia	52%	6%	93%	Southeast	10.72
					21.04
Stratum 6					
Nevada	56%	4%	77%	West	1.58
Oregon	74%	4%	84%	West	9.91
Washington	68%	3%	78%	West	18.18
					29.67
Stratum 7					
Florida	18%	5%	69%	Southeast	17.14
North Carolina	22%	9%	72%	Southeast	16.66
					33.80
Stratum 8					
Kentucky	13%	21%	68%	Southeast	8.00
New Jersey	17%	13%	82%	Mid-Atlantic	10.63
Mississippi	28%	10%	97%	Southeast	8.88
District of Columbia	13%	10%	79%	Mid-Atlantic	1.23
					28.74
Stratum 9					
Pennsylvania	36%	15%	79%	Mid-Atlantic	17.12
Tennessee	37%	14%	86%	Southeast	8.85
Oklahoma	30%	12%	80%	Southwest	10.04
South Carolina	35%	11%	88%	Southeast	6.01
					42.02

Table 1. Strata definitions (continued)

State and stratum	FDCH	HS	Subsidized	Region	Measure of size
Stratum 10					
Connecticut	61%	3%	82%	Northeast	5.47
Massachusetts	62%	5%	86%	Northeast	16.03
Utah	64%	1%	82%	Mtn Pl	12.54
Wyoming	63%	4%	76%	Mtn Pl	2.18
					36.22
Stratum 11					
Maine	78%	6%	83%	Northeast	4.92
Vermont	77%	2%	80%	Northeast	2.25
North Dakota	83.5%	2%	83%	Mtn Pl	5.96
Montana	71%	6%	84%	Mtn Pl	3.69
Nebraska	70%	2%	79%	Mtn Pl	11.73
					28.55
Stratum 12					
Rhode Island	27%	5%	83%	Northeast	1.31
Maryland	57%	9%	80%	Mtn Pl	13.92
South Dakota	67%	11%	79%	Mtn Pl	3.25
Idaho	67%	9%	79%	West	2.24
					20.71

Table 2. Example of a stratum that was considered but not used

State and stratum	FDCH	HS	Subsidized	Region	Measure of size
Connecticut	61%	3%	82%	Northeast	5.47
Massachusetts	62%	5%	86%	Northeast	16.03
Maine	78%	6%	83%	Northeast	4.92
Vermont	77%	2%	80%	Northeast	2.25

Table 3. Example of sponsor sample selection

Sponsor	Number of providers	Measure of size	Cumulative measure	Hit sequence	In sample?
A	79	19.25	-	-	Y
B	19	4.75	-	-	Y
C	14	3.5	3.5	0.180	Y
D	13	3.25	6.75	6.580	Y
E	11	2.75	9.5	12.980	N
F	11	2.75	12.25	12.980	N
G	10	2.5	14.75	12.980	Y
H	8	2	16.75	19.380	N