DOCUMENT RESUME

ED 107 725 TM 004 583

AUTHOR	Shaycoft, Marion F.
TITLE	The "Closed School-Cluster" Method of Selecting a Probability Sample.
PUB DATE	[Apr 75]
NOTE	13p.; Paper presented at the Annual Meeting of the American Education Research Association (Washington, D. C., March 30-April 3, 1975)
EDRS PRICE	MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS	*Cluster Grouping; Cost Effectiveness; Cross Sectional Studies; Educational Research; Longitudinal Studies; Matched Groups; *Probability; Research Design; *Sampling; *Schools

ABSTRACT

In some educational research studies--particularly longitudinal studies requiring a probability sample of schools and spanning a wide range of grades--it is desirable to so select the sample that schools at different levels (e.g., elementary and secondary) "correspond." This has often proved unachievable, using standard methods of selecting school samples. The "closed school-cluster" method, a way of selecting a probability sample with the desired characteristics, has now been developed. Its description and the instructions for implementing it are presented, along with the necessary formulas, including one for determining differential weights for the students in the sample. (Author)

APPENDIX C

THE "CLOSED SCHOOL-CLUSTER" METHOD OF SELECTING A PROBABILITY SAMPLE*

Marion F. Shaycoft
American Institutes for Research
Palo Alto, California

A. The problem

This paper is concerned with a problem that applies to many educational research studies and that hardly ever (or possibly never) has been handled optimally. The problem is how to select a probability sample of schools in such a way that all grade levels to be included in the study are represented either by the same schools or by corresponding schools; for instance the high schools that provide the grade 10 sample should be those high schools whose students come from the elementary schools that provide the grade 3 sample. The sample to be selected should not only meet this requirement but it should do it in a relatively efficient way; in other words in such a way that the sampling errors are kept as small as practical considerations will permit. Practical considerations also dictate that the procedure should be cost-effective.

B. Applicability of the problem

The problem is an important one because the studies to which it is particularly applicable are the large-scale expensive-to-execute ones that are either ambitious longitudinal studies or cross-sectional studies covering a fairly wide range of grades or both. The kind of longitudinal studies to which the sampling problem we are here concerned with applies are those that call for following the same cohort of students over a period of several years, in the course of which two or more school levels (e.g., elementary school, junior high school, senior high school) are involved. The kind of cross-sectional studies to which the sampling problem applies are those for which the samples for all grades should be composed of either the identical schools or insofar as different levels of schools (e.g., both elementary and secondary

*This is a slightly revised version of a paper presented at the Annual Meeting of American Educational Research Association in Washington, 1 April 1975. The author is indebted to Dr. Charles Stegman for some helpful suggestions concerning the earlier version.

schools) are required, the corresponding schools in the same school system. The type of design that Beaton refers to as a "pseudo longitudinal" design* would fall in this category. For such studies, as well as for true longitudinal studies, it is generally desirable that rather than selecting an entirely separate sample of schools at each school level (for instance a probability sample of elementary schools and a separately selected probability sample of secondary schools) the schools at the different levels should be selected in such a way as to "correspond." This requirement may be necessary for reasons of administrative convenience and economy--in other words fewer school systems to deal with--or it may be necessary because the research plan calls for comparison of different grade cohorts at the same time within the same (or corresponding) schools.

Project TALENT is an exemplar of a study that is both cross-sectional, in the manner we have been talking about, and longitudinal. It cuts across several grades (grades 9-12) and involves corresponding schools at two levels--primarily senior high schools (for grades 10-12 and for some grade 9 students) but some junior high schools (for those grade 9 students not in four-year high schools). The Project TALENT sample was not selected in precisely the manner to be proposed in this paper--a manner which utilizes a concept we are terming "closed clusters" of schools. (We hadn't developed the concept back in 1959, when selection of the sample was being planned.) As a matter of fact, although the Project TALENT sample is generally regarded as excellent not merely by those of us who were involved in its design and selection but by researchers entirely unconnected with the project, nevertheless we have been aware that like most products of human effort it is imperfect; the chief difficulty being that because of problems in matching junior high schools with senior high schools the grade 9 sample did not entirely correspond to the sample of students in grades 10, 11, and 12. Now we know a solution to the problem; if we were selecting the Project TALENT sample today, we would use the "closed-cluster" concept, and thereby would obtain not just a very representative sample of 10th, 11th, and 12th-grade students, which we also acheived in 1959, but also a

*Beaton, A. E. Pseudo Longitudinal Model, 1974 (unpublished paper) grade 9 sample that would correspond perfectly to the grade 10-12 sample in terms of school composition.

C. Purpose of this paper

It is the purpose of this paper to describe a sampling procedure utilizing the concept of closed clusters of schools--a concept alluded to in the preceding paragraph--and to explain how it can be applied to solve, in a cost-effective way, the sampling problem that has been described.

D. The necessity of, and difficulty in, identifying corresponding schools at different levels

A basic question that has to be answered very early in the planning of any probability sampling for use in educational research is what the primary sampling unit should be. Should it be students? classrooms? schools? school districts? school systems? counties? states? In general, the larger the primary sampling unit the easier and more economical it is to handle administrative arrangements; on the other hand the larger the primary sampling unit the larger the sampling errors. Balancing the pros and cons, in a large-scale study to take place in the schools it is generally not feasible from a practical viewpoint to use students as the primary sampling unit. Any unit smaller than the school is likely to be too costly to be practicable.

Assuming that in the case of a particulär longitudinal study the exigencies of the situation are such that the basic unit must be the school (or something larger) rather than the individual student, how should the schools be selected? This problem presents unique difficulties in longitudinal studies that span a time period when some members of the sample may move from one school to another (from elementary to high school, for instance). The problem of how to handle the shift from elementary to high school--or in districts that have junior high schools, the shift from elementary to junior high school and then to senior high school--would be no problem at all if there were a one-to-one correspondence between elementary schools and high schools (and junior high schools where applicable). By one-to-one correspondence we mean that all graduates of elementary school a go to high school A (if they go to high school at all); similarly that all graduates of elementary school _b_ go to high school _B_; that the graduates of elementary school _c_ go to high school _C_; and conversely that high school _A_ normally draws students from no elementary school but _a_; high school _B_ from no elementary school but _b_, etc. These rules refer, of course, just to normal school-to-school transitions; exceptions resulting from moving out of the neighborhood or from abnormal circumstances of one kind or another that occur rarely and just affect one or two individual children don't count in this context of defining one-to-one correspondence. But unfortunately (from the viewpoint of sampling simplicity), although one-to-one correspondence _is_ the situation in many communities there are many others where it isn't. One way of coping with this problem is by means of the "closed cluster of schools" concept.

E. The "closed cluster of schools" concept

To begin with, let's consider the situation where we are concerned only with elementary schools and senior high schools--not with junior high schools. Such a situation might be a longitudinal study whose design involved contacting the sample members initially in grade 1 and then again in grade 5 and then in grade 11, so that none of the two junior high school grades would be involved.

In terms of the relationship between elementary and high schools (what elementary schools each high school draws from and what high schools each elementary school sends its graduates to), we may classify schools into four kinds of elementary-and-high-school cluster, as follows:

Type A cluster

This occurs where there is a one-to-one correspondence (as described above) between elementary and high schools. In this case each elementary school and its corresponding high school constitute a two-school cluster.

Type B cluster

This occurs where there is an m-to-one correspondence between elementary schools and high schools. This is the situation where several elementary schools send all their graduates to one high school and that high school draws from no other elementary schools. In this case the m elementary schools and one high school constitute an m+1 school cluster.

Type C cluster

This occurs when there is a one-to-n correspondence between elementary and high schools. It occurs when one elementary school sends its graduates to several high schools and each of those high schools draws from no other elementary school. In this case the one elementary school and the n high schools constitute a 1+n school cluster.

Type D cluster

A Type D cluster is an m+n school cluster consisting of m elementary schools and the n high schools they send their graduates to, these n high schools drawing from no other elementary schools than those m. Type D clusters, by definition, consist only of schools that do not fit into any Type A, B, or C clusters. Furthermore each Type D cluster should contain the fewest possible schools that will fit the definition. In other words no Type D cluster should consist of the combination of two or more smaller Type D clusters nor should it consist of two or more smaller clusters of any other type than Type D.

As defined above, every elementary school and every high school in the United States belongs to one and only one cluster. The classification principle can of course be readily generalized to include junior high schools. Then, clearly, every elementary school, every junior high school, and every high school in the United States belongs to one and only one cluster. We are calling these groups of schools "closed clusters" because nobody enters or leaves them except through entering grade 1, graduating from high school, moving into or out of the geographic area covered by the cluster, or becoming a dropout.

This classification of schools into clusters is at the heart of the proposed solution to the problem of sampling in such a way as to solve the problems caused by irregular patterns of movement from elementary school to secondary school, and from junior high school to senior high school.

F. Using closed clusters of schools in sample selection

It is probably obvious that the proposed solution uses closed clusters of schools as the primary sampling unit. Ideally this would involve sampling from a directory that contains a list of all such clusters and identifies the schools in each. Unfortunately, however, no such directory exists.

It is necessary, therefore, to improvise a method that will permit obtaining essentially the same results in the absence of a comprehensive list of clusters. The suggested procedure may be described as a one-stage or multi-stage stage sampling procedure, the first stage of which is a modified form of stratified or simple random cluster-sampling with differential sampling ratios.

The notation used in the description that follows is summarized in Figure 1.

The steps are as follows:

1. If the sample is to be stratified select the stratification variables and define the strata. Set the sampling ratio at approximately the level that might be expected to give about the desired number of schools if the sampling ratio were to be applied to entire clusters rather than to individual schools. (If, for instance, you surmise that the average closed cluster in the segment of the population of schools you are interested in contains about five schools, you would set your initial sampling ratio at only about one-fifth of what you would normally have used to get the number of schools you would like included. If differential sampling ratios are to be used for different strata, these, too, should be set as part of this initial step.

2. Using the stratification (if any) and the sampling ratio (or ratios) decided upon and using the Office of Education source data on public schools (or any other source you may prefer), select* the initial sample of schools. These are the potential "nucleus schools." A nucleus school is the first school selected for a cluster--in other words the nucleus of the cluster.

*with replacement

3. For each of the potential nucleus schools selected, determine what other schools are in the same closed cluster. This will necessitate determining, by interview, questionnaire, or delving into official records, what schools it gets its students from (if it is a secondary school) and what schools it sends its graduates to (if it is an elementary or junior high school). This information will have to be secured not only from the initial school selected (the nucleus of the cluster) but also from each school from which it receives or to which it sends students, and in turn from each of those schools, and so forth until no new schools enter the picture.

4. If multi-stage sampling has been decided upon, for instance a second-stage sampling to select a probability sample of students within the sample schools and in the appropriate grades, this should be done at this point. It is often desirable to set sampling ratios which will make the weights (w_{jk}) turn out to be at least roughly equal, since for a sample of given size this reduces the samplirg errors of statistics.

5. After the sample has been selected it is necessary to determine the weight to be attached to each student in the sample. Normally the weight would be the reciprocal of the product of all the sampling ratios involved (first-stage, second-stage, etc.). With the suggested closed-cluster procedure, however, it is necessary to remember that p'_{jk} is an understatement of the probability that school i will be in the sample, because it will be in the sample not only if it is the first school selected in the sample of potential nucleus schools but also if it is a subsequ'nt selection in the nucleus school sample; and also because it will be included in the sample if any other school in its cluster is so selected.

   The relevant formulas for determing the weights are shown in Figure 1.

Hypothetical data showing how sets of schools break down into closed clusters, and how the proposed procedure works, are shown in Tables 1-4.

Table 1 shows hypothetical percentage distributions corresponding to the transition between junior high school and senior high school in a particular community. This hypothetical community has 12 junior high schools and 10 senior high schools, and it just happens to include closed clusters of all four types--Type A, Type B, Type C, and Type D--although the researchers wouldn't have to know that in order to proceed.

Table 2 shows diagrammatically how the first-stage sampling would be done, assuming it had been decided to start with a potential-nucleus-school sample consisting of three schools from the community.

Table 3 shows how the weights would be determined, assuming a second-stage sampling is also carried out.

Table 4 is not relevant to actual execution of the procedure since it shows, on the basis of the Table 1 data, the classification of all the schools into closed clusters--not just those schools in the sample. The 22 schools of the community break down into 5 closed clusters, having 2 to 8 schools each.

G. Importance of the procedure

It is believed that the proposed procedure provides a definitive solution to the knotty problem of selecting a probability sample of schools that include all corresponding schools at different levels. This will be particularly helpful in longitudinal studies requiring a probability sample; in quasi-longitudinal studies; and in some large-scale cross-sectional studies (those that require, for administrative or research reasons, that corresponding schools at different levels be included).

FIGURE 1. Notation and formulas

NOTATION

N' = no. of schools in pool from which potential nucleus schools are selected.

m = no. of strata

N'_{hk} = no. of schools in stratum h (which contains school k) in pool from which, in the case of stratified random sampling, potential nucleus schools are selected

n' = no. of potential nucleus schools selected (including those that may be rejected because another school in the same cluster has already been selected as the nucleus school)

n'_h = no. of potential nucleus schools selected (in the case of stratified random sampling) from stratum h

p'_{jk} = probability that school k, which happens to be in cluster j, will be the first school selected in the first-stage sampling to select potential nucleus schools. (In other words p'_{jk} = probability of selection of school k on a single draw.)

p'_{hjk} = probability that school k, which happens to be in stratum h in cluster j, will be the first school selected from stratum h in the first-stage sampling to select potential nucleus schools. (In other words, p'_{jk} = probability of selection of school k on a single draw.)

p_{hj} = probability that at least one school in stratum h in cluster j will be the first school selected from stratum h in the first-stage sampling to select potential nucleus schools.

p_{1jk} = first-stage sampling ratio for selecting school k in cluster j as a nucleus school

n_j = no. of schools in cluster j

n_{hj} = no of schools in stratum h in cluster j

(Note that not all schools in the cluster are necessarily in the same stratum.)

p_{ijk} = i^{th}-stage sampling ratio (i 1) applied in school k in cluster j to select a probability sample of classrooms or students (or some other entities) within school. (This is applicable only if multi-stage sampling is being done.)

p'_{1j} = probability that schools in cluster j will be in the sample (on the basis of first-stage sampling)

w_{jk} = weight for each case in school k in cluster j

FORMULAS*

$$p'_{hjk} = \frac{1}{N'_{hk}} \quad \text{(assuming all schools in the stratum have equal a priori probabilities)} \tag{1}$$

$$p_{hj} = \sum_{k=1}^{n_{hj}} p'_{hjk} \tag{2}$$

$$P_{1j} = 1 - \left(1 - p_{1j}\right)^{n'_1} \left(1 - p_{2j}\right)^{n'_2} \ldots \left(i - p_{mj}\right)^{n'_m} \tag{3}$$

$$w_{jk} = \frac{1}{P'_{1j}\, p_{2jk}\, p_{3jk} \ldots} \tag{4}$$

*When the first-stage sampling is an unstratified samp.ing of schools, formulas 1 and 3 simplify to formulas 5 and 6 respectively:

$$p'_{jk} = \frac{1}{N'} \quad \text{assuming all schools have equal a priori probabilities} \tag{5}$$

$$P'_{1j} = 1 - \left(1 - \sum_{k=1}^{n_j} p'_{jk}\right)^{n'} \tag{6}$$

TABLE 1. Hypothetical joint percentage distributions for the transition from junior high school to senior high school in Community X

		% of JHS's 9th-grade enrollment												% of SHS's 10th-grade enrollment														N	% of Total
		Junior high schools												Junior high schools												Transfer in**	Total		
	School* → ↓	J11	J12	J13	J14	J15	J16	J17	J18	J19	J20	J21	J22	J11	J12	J13	J14	J15	J16	J17	J18	J19	J20	J21	J22				
Senior high schools	S1	-	93	-	86	-	85	-	-	24	-	-	-	-	32	-	22	-	27	-	-	7	-	-	-	12	100	338	25
	S2	-	-	-	-	-	-	-	-	-	65	-	-	-	-	-	-	-	-	-	-	-	96	-	-	4	100	132	10
	S3	-	-	45	-	8	-	-	-	-	-	-	-	-	-	85	-	11	-	-	-	-	-	-	-	4	100	74	6
	S4	88	-	-	-	-	-	-	-	-	-	-	80	49	-	-	-	-	-	-	-	-	-	-	36	15	100	206	15
	S5	-	-	24	-	25	-	96	-	-	-	6	-	-	-	17	-	13	-	62	-	-	-	2	-	6	100	193	14
	S6	-	-	25	-	-	-	-	-	-	-	80	-	-	-	35	-	-	-	-	-	-	-	57	-	8	100	99	7
	S7	-	-	-	-	-	-	-	-	70	-	-	-	-	-	-	-	-	-	-	-	85	-	-	-	15	100	79	6
	S8	-	-	-	-	60	-	-	-	-	-	9	-	-	-	-	-	88	-	-	-	-	-	9	-	3	100	67	5
	S9	-	-	-	-	-	-	-	92	-	-	-	-	-	-	-	-	-	-	-	90	-	-	-	-	10	100	87	7
	S10	-	-	-	-	-	-	-	-	-	30	-	-	-	-	-	-	-	-	-	-	-	92	-	-	8	100	64	5
	Transfer away** or drop out	12	7	6	14	7	15	4	8	6	5	5	20																
	Total	100	100	100	100	100	100	100	100	100	100	100	100															1339	100
	N	115	116	140	85	98	110	124	85	96	196	70	92														1327	1212***	
	% of Total	9	9	11	6	7	8	10	6	7	15	5	7														100		

*S1, S2, ... S10 are the 10 senior high schools; J11, J12, ... J22 are the 12 junior high schools.

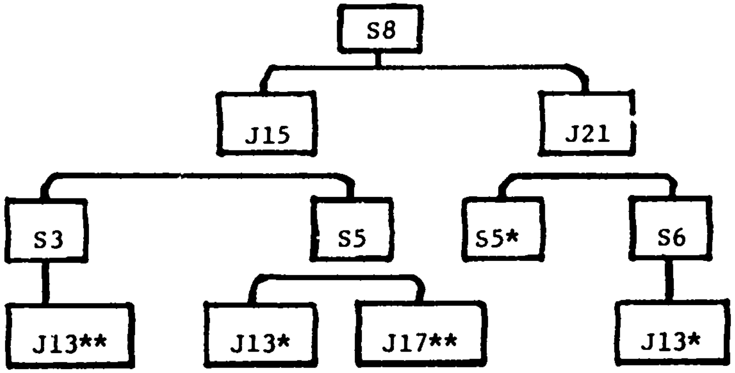
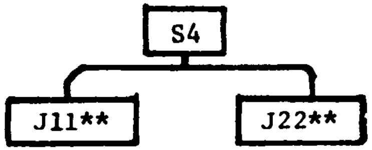
**Change of residence

***Total excluding transfers and dropouts

TABLE 2. Hypothetical 1st-stage sampling to select 3 potential nucleus schools in Community X

(n´ = 3)

Pool from which nucleus schools are selected:
All senior high schools (S1, S2, ... S10)
(N´ = 10)

Order of selection	Potential nucleus school selected by random sampling	Sequential identification of schools belonging to cluster	List of schools in cluster	Cluster #
1	S8	 S8 J15 J21 S3 S5 S5* S6 J13** J13* J17** J13*	S3 J13 S5 J15 S6 J17 S8 J21	1
2	S5	***		
3	S4		S4 J11 J22	2
****				

*Same school appears to left, in same row of chart.

**Dead end; all corresponding schools have already been identified as belonging to the cluster.

***No cluster formed; selected school was rejected as nucleus school because it was in a previously selected cluster.

****No additional school selected, since it is hypothesized that it was decided to select only 3 schools from Community X as potential nucleus schools. Note that the fact that one of them (S5) did not turn out to be a nucleus school does not modify this decision.

TABLE 3. Hypothetical 2nd-stage sampling, and determination of weights for cases in the 2 selected clusters

Cluster #	School	Probability of selection as a nucleus school	Probability of selection	2nd-stage sampling ratio*	Weight for each student in sample
(j)	(k)	(p'_{jk})	(P'_{ij})	(p_{2jk})	(w_{jk})
1	S3	.1	.784	1/3	3.83
	S5	.1	.784	1/3	3.83
	S6	.1	.784	1/3	3.83
	S8	.1	.784	1/3	3.83
	J13	0	.784	1/3	3.83
	J15	0	.784	1/3	3.83
	J17	0	.784	1/3	3.83
	J21	0	.784	1/3	3.83
2	S4	.1	.271	1	3.69
	J11	0	.271	1	3.69
	J22	0	.271	1	3.69

*Assume that these values are selected as second stage sampling ratios because they are convenient to work with (i.e., selecting every 3rd student or every student) and have the desirable feature of making the weights nearly equal.

TABLE 4. Complete set of closed clusters for hypothetical data of Community X

Cluster #	Schools JHS	SHS	Type of cluster*
1	J13 J15 J17 J21	S3 S5 S6 S8	D
2	J11 J22	S4	B
3	J20	S2 S10	C
4	J18	S9	A
5	J12 J14 J16 J19	S1 S7	D

*Definition of closed-cluster types

Type	No. of JHS's	No. of SHS's
A	1	1
B	>1	1
C	1	>1
D	>1	>1