Description of sampling procedure for EGRA Plus Liberia

Background

As per the commitment to USAID and the MoE, the sampling procedure will focus on public schools only. It will select 60 control schools, 60 "treatment 1" schools (information treatment only), and 60 "treatment 2" schools.

In reality there are four types of schools in Liberia as per the EMIS database: Public, Self-Help/Community, Religious/Mission, and Private. On the advice of the MoE, we will use an expadned definition of "Public" to mean "Public" in a narrow sense plus "Community."

It has been agreed previously that in order to make this a proper experiment, allocation of schools into these three groups will be randomized. It has also been agreed that to make the schools representative of all of Liberian public schools, selection would be random but proportional to population.

In order to make the intervention cost-effective, and to make its implementation reminiscent of what a scaled up process would look like, the selection will proceed to select clusters of schools that will be similar in nature to the natural intervention or supervision area of district officers. Thus, schools will be selected in clusters. Schools will be visited and assisted in clusters of 4. This is a good compromise between the need for work efficiency and the need for representativity, and is done to minimize problems with "design effect." (A technical issue one may safely ignore for now.) One could have worked in clusters of 1 or 2, but this would have raised the cost astronomically, and does not simulate what happens in reality, since officers work with groups of schools—that is the nature of supervision. On the other hand, one could just do 2 or 3 or 4 clusters of 30 or 20 pr 15 schools, but this means that the first-stage selection (2 or 3) would not possibly be representative of the country. A wise compromise is 15 clusters of 4 schools, with random selection at both stages.

It is extremely important to note that this sampling is not for a study, but for an intervention, and has to respect the nature of such an intervention.

Selection of districts

First, 15 districts will be selected in a manner proportional to population. The selection tool is an Excel spreadsheet containing data on schools by region, county, and district. This tool will have been sent along with the e-mail containing this Word file. The database contains data to characterize schools as to which settlement they belong to, but settlements is too small unit of aggregation to permit efficient sampling selection in a first stage. Selecting settlements would not permit efficient cluster selection, as the settlements are themselves below the cluster level of aggregation. (There are half as many settlements in the database as schools, making it an ineffective level of aggregation for a first stage of sampling.)

A simple sampling program was written in Excel. The =rand() function in Excel is used to random sample. Certain tricks are used to make the random sample proportional to population. The tool allows one to sample and re-sample, given that new samples can be generated easily simply by pressing F9. This can serve a capacity-building exercise on how to sample proportional to population, using a simple, standard software and a very step-wise, transparent logic.

A recommended sample is the following:

Number	District	No. of schools to choose from: according to EMIS data
1	Left Bank St. Paul	260
2	Kolahun	87
3	Greater Monrovia I	428
4	Greater Monrovia II	256
5	Garwula	50
6	Right Bank St. Paul	114
7	Zoe-geh	59
8	Kakata	85
9	Compound # 2	61
10	Foya	59
11	Gbeapo	16
12	Foya	59
13	Gbarnga	55
14	Sanoyea	43
15	Left Bank St. Paul	260

But note that the attached software would make it possible to re-sample.

The reader may note that some districts are included twice. (In this case, Left Bank St. Paul.) That is as it should be, if one is sampling proportional to population. For example, the largest 3 districts in Liberia (Monrovia I, Monrovia II, and Left Bank St. Paul) have 26% of the student population. Since 26% of 15 is approximately 4, it makes sense that one or two districts would show up twice in many samples, and that many samples would have 4 representations of these 3 districts.

We have simulated the selection over many dozens of repeated samples. The resulting correlation between the actual share of each district's population and the resulting proportion of the time each district shows up in a sample is 0.99 (converging to 1 at the limit). In repeated samples, the proportion of times the largest three districts show up is 26%, if they are allowed to sometimes show up more than once in a sample, which is exactly proportional to the population.

Selecting the districts immediately, or as soon as possible, is an important step, as it affects the selection of the teacher trainers. We would like the teacher trainers to be selected from, or be willing to be based in, the relevant districts, or at least not too far away. And, since selection of the trainers is very much on the critical path of the project, it needs to happen soon. Thus, we would like to plan to go along with the proposed list. Note that there may be a temptation to remove on district as too distant, or something along those lines. That should be resisted, as one is indeed trying to make the sample truly representative, and the best way to do this is to make it fully random, i.e., machine-selected, not selected by people.

Selecting schools within clusters

Once having selected the districts, clusters of 4 schools each will be selected. The EMIS database has data on the X-Y coordinates of the schools. The procedure we have used is as follows, but for a brief window this could be subject to discussion.

For the selected districts, we create a distance matrix of all schools i to all other schools j, simply calculating the length of the hypotenuse between x(i), y(i) and x(j) and y(j). This is obviously not perfect, as it does not take into consideration infrastructure, but it is a good first approximation. There may be a need to substitute out some schools anyway, because of poorly entered X-Y coordinates, or other reasons, such as there being a river between schools that in terms of X-Y coordinates appear close to each other.

We then select one school at random in the district, which can be considered the centroid of that district's cluster.

We then had two choices or options, and we have opted for one of them.

A first choice is to find the 3 schools closest to the centroid. The problem with this is not what one might at first think, namely that it creates a bias towards higher population density areas. After all, if the selection of the centroid schools is truly random, some of them will be in low density areas, and the nearest schools will actually be quite far, precisely because they are in low-density areas. This option does have the advantage of minimizing the cost of intervention, and also more closely mimicking the way an actual supervisor would work, by going from school to school, taking the closest ones in sequence. In that sense, it has all the "realism" and representativity one needs. However, the option does suffer from one problem, which is excessive design effect. By clustering on the closest schools after picking one at random, one does minimize the range of schools one is dealing with, to some degree, just as one does with any type of clustering, only more so. A clustering process, relative to a pure random sample, will restrict the range of observation somewhat, because schools within clusters will tend to differ from each other less than schools selected totally at random. In the ideal world, clusters should all be "mini populations." If that were the case, then clustering would be extremely efficient. But we know that in the real world, clustering censors the observed total variance relative to the real variance, because units will tend to be similar to each other. Thus, this is somewhat of a disadvantage. But because we are talking here about a very labor intensive intervention, it is important to economize: this is why we have clusters of 4 to begin with. If, having clustered at the district level, one picks a single school in the cluster, and then finds the closest 3 schools, one is then also restricting the range of observation. The other extreme is to select the 4 schools completely at random within the cluster. But some of the districts in Liberia are big, so this creates an artificial cluster, unlike anything that anyone in real life would work. The average district has 90 schools, which is way beyond anything any one agent could truly help. In real life any improvement process most likely would require a span smaller than 90 schools.

The cost saving involved in clustering within districts by picking a school at random, and then the 3 closest, seems worth the possible sacrifice in variability between schools. Again, we repeat, this will not create an urban bias, or a bias towards areas with higher population density. It just makes the sampling a little less efficienty in a statistical sense, but a great deal more efficient in a cost and substantive sense.

An example of how this work containing the districts for Zoe-geh is included. Note that this example has not eliminated the non-public and secondary schools. An eventual sample within districts will do that.

Conclusion

Taking these two steps will have allowed the sampling to be proportional to population, will have clustered schools into reasonably natural clusters that would be more or less similar to the administrative or jurisdictional units that would occur in reality.