Sample Design of Rural ASER 2022

Wilima Wadhwa[1]

The purpose of ASER is two-fold: (i) to obtain reliable estimates of the status of children's schooling and foundational learning (reading and math ability); and (ii) to measure the change in these basic learning and school statistics over time. Every year a core set of questions regarding schooling status and basic learning levels remains the same. However new questions are added to explore different dimensions of schooling and learning at the elementary stage. The latter set of questions can vary each year. For instance, ASER 2006 and 2007 tested reading comprehension for different kinds of readers; ASER 2007 introduced testing in English, which has been repeated in four subsequent editions of ASER (2009, 2012, 2014, 2016).[2]

Every year, ASER volunteers visit a government primary or upper primary school in each sampled village. The school information is recorded based either on direct observation (such as attendance or useability of facilities) or on information provided by the school (such as grants information). School observations have been reported in 2005, 2007, and every year since 2009. Beginning in 2010, information is also collected on schools' RTE compliance.

ASER was done annually for ten years (2005-2014). After a break of one year,[3] ASER 2016 started a new series of ASER estimates using Census 2011 as the sampling frame. In this new series of ASER starting in 2016, the nation-wide assessment of foundational learning is done every other year and competencies for other age-groups are explored in the intervening years.[4] This alternate-year cycle was broken in 2020 due to the COVID-19 pandemic which severely restricted movement in the field. ASER 2022, therefore, returns with estimates at the district, state and national levels after a gap of 4 years.

ASER has a two-stage sample design. In the first stage, for each rural district, villages are randomly selected from the Census village directory. Therefore, the coverage of ASER is the population of rural India.[5] ASER 2005-2014 uses the Census 2001 village directory as the sampling frame. The Census 2011 sampling frame became available in the public domain in 2015 and ASER 2016-2022 uses this frame. In the second stage, households are randomly selected in each of the villages selected in the first stage. This sampling strategy generates a representative picture of each district. All rural districts are surveyed. The estimates obtained are then aggregated to the division, state and all-India levels.

Sample size calculations for ASER done at the district level – the lowest geographical unit at which the estimates are representative – resulted in a sample of 600 households per district.[6] At the state level and at the all-India level the survey has many more observations, lending estimates at those levels much higher levels of precision.

Since ASER has a two-stage sample design,[7] the district level sample size of 600 households has to be allocated to the two stages of sampling. ASER samples 30 villages in the first stage. These are randomly selected using the village directory of the Census as the sample frame.[8]

[1] Director, ASER Centre

[2] For more details, see the section 'ASER domains over time' in this report.

[3] In 2015, ASER was done in only two states – Maharashtra and Punjab.

[4] For instance, ASER 2017 explored functional competencies for 14-18-year-olds.

[5] No adjustments are made to the population as given in the Census.

[6] Sample size calculations assume simple random sampling. However, simple random sampling is unlikely to be the method of choice in an actual field survey. Therefore, often a "design effect" is added to the sample size. A design effect of 2 would double the sample size. At the district level a 7% precision along with a 95% confidence level would imply a sample size of 196, giving us a design effect of approximately three. However, a sample size of 600 households gives us approximately 1000-1200 children per district.

[7] For a two-stage sample design, as explained above, sample size calculations have to take into account the design effect, which is the increase in variance of estimates due to departure from simple random sampling. This design effect is a function of the intra-cluster correlation. The greater this correlation, the larger is the design effect implying a larger sample size for a given level of precision. For a given margin of error (me), the sample size can be backed out from

$$me = \frac{2\,\sigma}{p} = \frac{2\sqrt{\frac{d\,p\,(1-p)}{N-1}}}{p}$$

where *d* is the design effect, *p* is the incidence in the population, σ is the standard error and N the sample size.

[8] Since the sampling frame is not current, sometimes sampled villages need to be replaced. As far as possible, however, villages are not replaced. There are three main reasons for replacing a village: First, if it has been converted to an urban municipality; second, due to natural disasters, like floods; or third, due to insurgency problems. Replacement villages are also drawn as an independent sample.

In the second stage 20 households are randomly selected in each of the 30 selected villages in the first stage.[9]

Villages are selected using the probability proportional to size (PPS) sampling method. This method allows villages with larger populations to have a higher chance of being selected in the sample. It is most useful when the first stage sampling units vary considerably in size, because it ensures that households in larger villages have the same probability of getting into the sample as those in smaller villages.[10,11]

There are various issues that complicate the second stage sampling. First is the issue of sparse populations of interest, namely that the sampling strategy may not result in sufficient sample sizes of the target population. The best solution to this problem is to create a listing of the target population (for a particular cluster) and sample from that, thus, employing a stratified sample. However, given the rapid assessment nature of ASER and several resource constraints (time, people, money), ASER does not stratify at the second stage – houselisting is not done at the village level.

Second, the absence of a houselisting creates additional problems in surveys that are representative at multiple levels of aggregation. In these surveys estimates have to be weighted[12] with appropriate weights to account for different underlying population sizes – a more populous state like UP will have a higher weight in the national estimate than a state like Himachal Pradesh. The calculation of these weights requires the underlying population proportion of the target group of interest. So, if the household were the unit of sampling, then we would need the number of households in the village to calculate the weights. On the other hand, if children in the age group of 3-16 years were our target population, we would need the total number of such children in the village to calculate the weights. A houselisting of the village would provide not only the frame for sampling these children, but also the total number of such children in the village.

ASER resolves both these problems by sampling households. Household weights are easy to calculate since the Census provides the village population of households. Therefore, the sample in ASER is defined in terms of households and not children. In ASER, all children in the age group of 3-16 years living in the sampled households are surveyed. So as to get a representative sample of the household distribution, households with no children in the target age group are counted as part of the sample. Given the scale of ASER and large household sizes in rural India, this strategy yielded large enough samples to do age-wise or grade-wise analysis at the state level.

However, while the number of households and villages in ASER has remained more or less unchanged since 2006, the number of children surveyed has been falling steadily. Between 2006 and 2018, the number of sampled children in ASER has fallen by about 30%.[13] With this secular decline, granular analysis for some smaller states and the less populous southern states was posing a problem.

ASER 2022, therefore, employs a sampling strategy that modifies the ASER approach, so as to get sufficient sample sizes and be able to calculate weights without creating a houselisting in the village. The standard ASER sampling approach in the village is to mimic simple random sampling without doing a houselisting. Volunteers walk around the village, make a map,

[9] This allocation of the total sample size to the different sampling stages is often based on logistical and cost considerations. For instance, a sample size of 600 households per district could have been allocated into 40 villages per district and 15 households per village; or 20 villages per district and 30 households per village. The first allocation would yield higher precision but cost more. Precision increases with a larger number of first-stage units since that reduces the adverse effect of a large intra-cluster correlation; however, cost also increases with a larger number of first-stage units, since that entails travelling to more villages (the marginal cost of surveying additional households in a given village is negligible). Therefore, there is a tradeoff between precision and cost.

[10] Probability proportional to size (PPS) is a sampling technique in which the probability of selecting a sampling unit (village, in our case) is proportional to the size of its population. The method works as follows: First, the cumulative population by village calculated. Second, the total household population of the district is divided by the number of sampling units (villages) to get the sampling interval (SI). Third, a random number between 1 and the SI is chosen. This is referred to as the random start (RS). The RS denotes the site of the first village to be selected from the cumulative population. Fourth, the following series of numbers is formed: RS; RS+SI; RS+2SI; RS+3SI; …. The villages selected are those for which the cumulative population contains the numbers in the series.

[11] Most large household surveys in India, like the National Sample Survey and the National Family Health Survey also use this two-stage design and use PPS to select villages in the first stage.

[12] The weight associated with each sampling unit, household in ASER, is the inverse of the probability of it being selected in the sample.

[13] The drop in number of sampled children is probably due to the increase in the number of rural households since 2006. Census 2011 notes that there was a 24% increase in rural households since Census 2001. Yet, the rural population increased by only 12% during the same period, implying that the average rural household size has gone down, implying fewer children per household. In addition, declining fertility rates, especially in the south, have resulted in fewer children per family, which coupled with more nuclear households in rural India, has led to declining samples of children in ASER.

divide the village into four parts, and sample 5 households using the fifth household rule in each part to get 20 households in the village. Households with no children in the target age group count as part of the sample since the aim is to get a representative picture of the household distribution.

In the ASER 2022 survey this approach was modified so as to capture sufficient numbers of 3-16-year-old children. The process is described below:

1. Walk around the village and make a map and divide the village into four parts.
2. In each part go to a central location and use the fifth household rule starting from the left to sample households.
3. If the household has children in the 3-16-year age group currently residing the household, record the household number, and the number of such children. Administer the survey to all children in the target age group in the household and collect information on the household. Proceed to the next fifth household.
4. If the household has no children in the 3-16-year age group, record the household number and the fact that it has no children in the target age group and move to the next household.
5. If the household is locked or does not want to participate in the survey record the household number and the fact that it was locked or a non-response household and move to the next household.
6. Continue this procedure until you have administered the survey in 5 households in each of the four sections of the village.

At the end of the survey in the village this procedure will yield 20 households with completed survey information, as well as the total number of households visited to achieve this. The latter is needed for the calculation of correct weights.

To summarise, ASER 2022 employs a two-stage clustered design. In the first stage 30 villages are sampled from the Census 2011 village directory using PPS. In the second stage, 20 households with resident children in the age group of 3-16 years are surveyed in each sampled village.

Since one of the goals of ASER is to generate estimates of change in learning, a panel survey design would provide more efficient estimates of change. However, given the large sample size of the ASER surveys and cost considerations, we adopted a rotating panel of villages rather than children. For ASER 2008-2014, each year 10 villages from three years ago were dropped, 20 villages from the previous two years were retained and 10 new villages were added.[14] Given the sample size of 30 villages per district, this procedure created a 3-year cycle in which the entire village sample is replaced. For instance, in ASER 2014 we dropped the 10 villages from ASER 2011, kept the 20 villages from 2012 and 2013 and added 10 more villages from the 2001 Census village directory. However, for ASER 2016 a fresh sample of 30 villages was drawn for each district because we were using a new sampling frame – Census 2011. In ASER 2018, we randomly dropped 10 villages from the 2016 sample, and added 10 new villages. In ASER 2022, an additional 10 villages were dropped from the ASER 2016 sample, the 10 villages from 2018 were retained and 10 new villages were added. Like before, these 10 new villages are drawn as an independent sample from the Census 2011 frame.[15]

The survey provides estimates at the district, division, state and national levels. In order to aggregate estimates up from the district level households have to be assigned weights – also called inflation factors. The inflation factor corresponding to a particular household denotes the number of households that the sampled household represents in the population. Given that 600 households are sampled in each district regardless of the size of the district, a household in a larger district will represent many more households and therefore, have a larger weight associated with it than one in a sparsely populated district.[16]

[14] The 10 new villages are drawn as an independent sample from the same sampling frame.

[15] Since the new series of ASER that started in 2016 visits all rural districts and assesses all children in basic foundational reading and arithmetic in alternate years, rather than every year, the entire village sample will be replaced in 6 rather than 3 years.

[16] The probability that household j gets selected in village $_i$ (p_{ij}) is the product of the probability that village$_i$ gets selected (p_i) and the probability that household $_j$ gets selected ($p_{j(i)}$). This is given by:

$$p_{ij} = p_i \, p_{j(i)} = \frac{n_v \, vpop_i}{dpop} \, \frac{n_{hi}}{vpop_i} = \frac{n_v \, n_{hi}}{dpop}$$

where n_v is the number of villages sampled in the district, $vpop_i$ is the household population of village i, *dpop* is the number of households in the district, and n_{hi} is the number of households visited in the village (to get the 20 sampled households). The weight associated with each sampled household within a district is the inverse of the probability of selection. Note that, in each district, the sum of the weights of the households will give the district population and the sum of the weights for all children in the sample will approximate to the population of children in the 3-16 age group in the district.