# Sample Design of Rural ASER 2014

**Wilima Wadhwa,** Director, ASER Centre

The purpose of rural ASER 2014 is twofold: (i) to get reliable estimates of the status of children's schooling and basic learning (reading and math ability); and (ii) to measure the change in these statistics over time. Every year the core set of questions regarding schooling status and basic learning levels remains the same. However a set of new questions is added for exploring different dimensions of schooling and learning at the elementary stage. The latter set of questions is different each year.

ASER 2006 and 2007 tested reading comprehension for different kinds of readers. ASER 2007 introduced testing in English and asked questions on paid tuition, which were repeated in 2009. ASER 2008 for the first time had questions on telling time and oral math problems using currency. In addition, ASER 2008 incorporated questions on village infrastructure and household assets. Investigators were asked to record whether the village visited had a pukka road leading to it, a bank, a ration shop, etc. In the sampled households, information on household assets (availability of television, type of house etc.) was recorded. These questions were repeated in 2009 and in addition father's education was also recorded. ASER 2010, while retaining the core questions on parents' education, household and village characteristics, introduced higher level testing tools for the first time. Questions on critical thinking were introduced, based on simple mathematical operations that appear in Std V textbooks. These were further refined in ASER 2011. ASER 2012 included testing of reading and comprehension of English that was first introduced in 2007 and repeated in 2009. ASER 2013 added expenditure on private tuition to the household questionnaire.

ASER 2014 brings together elements from various previous ASER rounds. The core questions on school status and basic reading and arithmetic remain. Children have been tested in English again, after 2012. In addition, parents' education, and household and village characteristics continue to be surveyed.

Every year, ASER surveyors visit a government primary or upper primary school in each sampled village. The school information is recorded through observations (such as attendance and usability of the facilities) and using information provided by the school (such as grants information). School observations have been reported in 2005, 2007 and 2009-2013, and are also reported in ASER 2014. Beginning in 2010, school information is collected on RTE indicators. In ASER 2014 grant information for the 2013-14 and current fiscal year has also been collected.

Finally, ASER 2014 continues the process of strengthening and streamlining started in 2008. Recheck of 4 or more villages in each district was introduced in 2008. This process was further strengthened in 2009. In ASER 2010, special attention was focused on improving training. In ASER 2011, in addition, master trainers monitored the survey process in the field. In ASER 2012, phone recheck was used on a large scale during the survey. During the survey, master trainers were called from a state specific call centre to get feedback on a daily basis. ASER 2013 incorporated all of these procedures and further streamlined processes in the field. ASER 2014 adds external rechecks to the process.

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 the change. However, given the large sample size of the ASER surveys and cost considerations, we adopted a rotating panel of villages rather than children. In ASER 2013, we retained the 10 villages from 2011 and 2012 and added 10 new villages. In ASER 2014 we dropped the 10 villages from ASER 2011, kept 10 villages each from 2012 and 2013 and added 10 more villages from the census village directory.

The sampling strategy used generates a representative picture of each district. Almost all rural districts are surveyed. The estimates obtained are then aggregated to the state and all India levels.

Since estimates are generated at the district level, the minimum sample size calculations are done at the district level. The sample size is determined by the following considerations:

- Incidence of what is being measured in the population. Prior to ASER 2005, a survey of learning had never been done in India. Therefore, the incidence of what we were trying to measure was unknown in the population. However, now we can use estimates from previous ASER rounds for sample size calculations.
- Confidence level of estimates. The standard used is 95%.

- Precision required on either side of the true value. The standard degree of accuracy most surveys employ is between 5 and 10 per cent. An absolute precision of 5% along with a 95% confidence level implies that the estimates generated by the survey are within 5 percentage points of the true values with a 95% probability. The precision can also be specified in relative terms - a relative precision of 5% means that the estimates will be within 5% of the true value. Relative precision requires higher sample sizes.

Sample size calculations can be done in various ways, depending on what assumptions are made about the underlying population. With a 50% incidence, 95% confidence level and 5% absolute precision, the minimum sample size required in each strata[1] is 384.[2] This derivation assumes that the population proportion is normally distributed. On the other hand, a sample size of 384 would imply a relative precision of 10%. If we were to require a 5% relative precision, the sample size would increase to 1600.[3] Note that all the sample size calculations require estimates of the incidence in the population. In our case, we can get an estimate of the incidence from previous ASER surveys. However, incidence varies across different indicators - so incidence of reading ability is different from incidence of dropouts. In addition, we often want to measure things that are not binary for which we need more observations.

Given these considerations, the sample size was decided to be 600 households in each district.[4] Note that 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.

ASER has a two-stage sample design.[5] In the first stage, 30 villages are randomly selected using the village directory of the 2001 census as the sample frame.[6] [7] Therefore, the coverage of ASER is the population of rural India.[8] In the second stage 20 households are randomly selected in each of the 30 selected villages in the first stage.

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 sampling units vary considerably in size because it assures that those in larger sites have the same probability of getting into the sample as those in smaller sites, and vice verse.[9] [10]

[1] Stratification is discussed below.

[2] The sample size with absolute precision is given by $\frac{z^2 pq}{d^2}$ where *z* is the standard normal deviate corresponding to 95% probability (=1.96), *p* is the incidence in the population (0.5), *q=(1-p)* and *d* is the degree of precision required (0.05).

[3] The sample size with relative precision is given by $\frac{z^2 q}{r^2 p}$ where *z* is the standard normal deviate corresponding to 95% probability (=1.96), *p* is the incidence in the population (0.5), *q=(1-p)* and *r* is the degree of relative precision required (0.1).

[4] 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, note that a sample size of 600 households gives us approximately 1000-1200 children per district.

[5] 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 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, s its standard error and N the sample size.

[6] Of these 30 villages, 10 are from ASER 2012, 10 from ASER 2013 and 10 are newly selected in 2014. They were selected randomly from the same sample frame. The 10 new villages are picked as an independent sample.

[7] Since the sampling frame is more than 10 years old 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.

[8] No adjustments are made to the population as given in the Census 2001.

[9] 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 is 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 cumulated 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.

[10] 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.

In each selected village, 20 households are surveyed. Ideally, a complete house list of the selected village should be made and 20 households selected randomly from it. However, given time and resource constraints a procedure for selecting households is adopted that preserves randomness as much as possible. Field investigators are asked to divide the village into four parts. This is done because villages often consist of hamlets and a procedure that randomly selects households from some central location may miss out households in the periphery of the village. In each of the four parts, investigators are asked to start at a central location and pick every 5th household in a circular fashion until 5 households have been selected. In each selected household, all children in the age group of 5-16 are tested.

The survey provides estimates at the district, 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.

The advantage of using PPS sampling is that the sample is self-weighting at the district level. In other words, in each district the weight assigned to each of the sampled households turns out to be the same. This is because the inflation factor associated with a household is simply the inverse of the probability of it being selected into the sample times the number of households in the sample. Since PPS sampling ensures that all households have an equal chance of being selected at the district level, the weights associated with households within a district are the same.[11] Therefore, weighted estimates are exactly the same as the un-weighted estimates at the district level. However, to get estimates at the state and national levels, weighted estimates are needed since states have a different number of districts and districts vary by population.

Even though the purpose of the survey is to estimate learning levels among children, the household was chosen as the second-stage sampling unit. This has a number of advantages. First, children are tested at home rather than at school, allowing all children to be tested rather than just those in school. Further, testing children in school might create a bias since teachers may encourage testing the brighter children in class. Second, a household sample generates an age distribution of children that can be cross-checked with other data sources, like the census and the NSS. Third, a household sample makes calculation of the inflation factors easier since the population of children is no longer needed.

Often household surveys are stratified on various parameters of interest. The reason for stratification is to get enough observations on entities that have the characteristic that is being studied. The ASER survey stratifies the sample by population in the first stage. No stratification is possible at the second stage. In order to stratify on households with children in the 3-16 age group, in the second stage, we would need the population of such households in the village, which is not possible without a complete house list of the village.

[11] 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:
$p_{ij} = p_i \ p_{j(i)} = \frac{30\,vpop_i}{dpop}\frac{20}{vpop_i} = \frac{600}{dpop}$, where $vpop_i$ is the household population of village i and dpop is the number of households in the district. Therefore, the weight associated with each sampled household within a district is the same and is the inverse of the probability of selection.