Department of School Education & Literacy
Ministry of Education, Government of India

Notes on Sampling Design for National Achievement Survey (NAS) 2021

Sampling design for the NAS 2021 intends to support the predefined and agreed objectives of the national assessment. NAS 2021 intends to provide information of what India's students know and can do in key grades and subjects at national, state, district, and school type administrative levels.

Introduction

The National Achievement Survey (NAS) is a nationally representative large-scale survey of students' learning in India administered periodically in key grades and subjects of primary and secondary education. Sampling design and procedures play a crucial role in ensuring that the results obtained on a sample can be reliably applied to the entire population.

Sampling design for the NAS 2021 intends to support the predefined and agreed ***objectives*** of the national assessment. NAS 2021 intends to provide information of what India's students know and can do in key grades and subjects at ***national***, ***state***, ***district***, and ***school type*** administrative levels, as schematically depicted in Figure 1. The frame for selection of schools in NAS 2021 would be UDISE+ 2019-20. Therefore, the States, Districts, etc. used for drawing of the samples of NAS 2021 would be exactly as per the UDISE+ 2019-20.

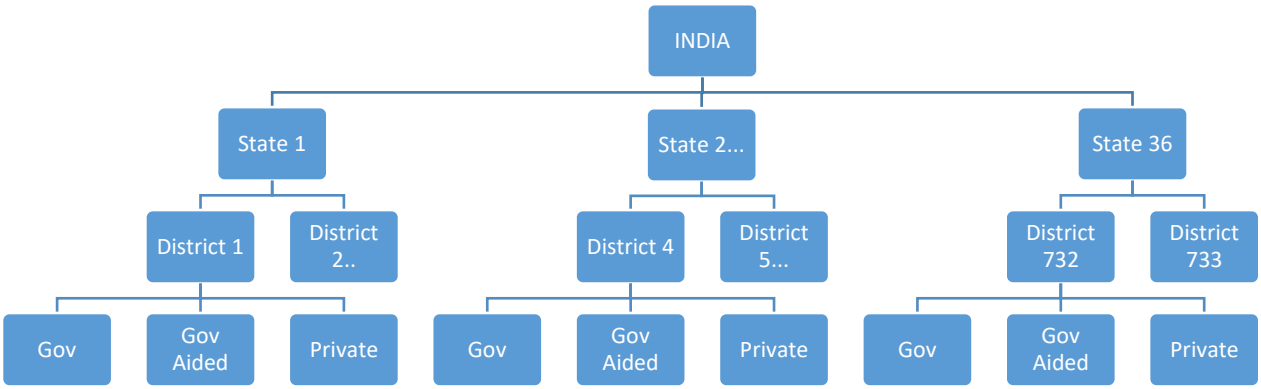

Figure 1 Hierarchical structure of the NAS 2021 sample strata

Sampling plan for NAS 2021 is designed to enable reporting of student academic performance at each of the hierarchically organized administrative levels, as presented in Figure 1, in mathematics (grades 3, 5, 8, and 10), modern Indian languages (grades 3, 5, 8, and 10), environmental science (grades 3 and 5), science (grades 8 and 10), social science (grades 8 and 10), and English language (grade 10). Thus, NAS 2021 will inform about student achievement in these key subjects and grades at India (i.e., national level), states/UTs and district levels. Moreover, within each district, student achievement will be estimated for each type of school management (Government, Government aided, and Private unaided recognised). This note is intended to explain the questions which the sampling design needs to address to meet these objectives. The issues at hand are as under:

1) What is the minimum sample size needed to obtain reliable information of students' performance in the groups at the lowest (stratum) level – school types within districts?
2) What are the resultant sample sizes at district, state, and national level?
3) What methods will be used for selection of schools within each stratum (district ×school management) and for selection of students within schools?

First, we present a rationale of the sampling design and factors relevant for answering these questions. Then, the exact methodology followed for selection of samples along with relevant input and output parameters has been discussed.

Part A - Rationale of Sampling Design

In a typical large-scale survey, data are obtained on a sample of students drawn from the population. Considering that the results based on sample data can be generalized to the population only under

certain conditions, it is important to identify required characteristics of the sample that enable attribution of the statistical information obtained on a sample of students from a certain group to all the students in that group.

The relationship between sample and population is based on a statistical probability, which means that descriptions and inferences about the characteristics of the student population based on a sample data are never fully accurate, they are made only to some degree of certainty. It is just a matter of statistical conventions to determine what is considered a sufficient ***degree of certainty*** at which we should accept that the characteristic observed on a sample can be applied to the population. This convention is commonly called "statistical significance".

Thus, the focal points of sampling are the factors associated with this degree of certainty and how do we ensure that a sufficient level of certainty will be reached.

The answers are derived from the "laws of large numbers", properties of "normal distribution" and the concept of "standard error" associated with every statistical indicator derived from sample data. The size of standard error is proportionate to the variability of measured phenomenon and inversely related to the sample size. Thus, with increase of sample size the standard error decreases, which further leads to the main sampling question – ***how large should be sample size*** so that standard error becomes sufficiently small to enable the ***desired level of certainty*** in making conclusions about student performance in the population? The answer to this question is dependent on two categories of factors:

a) Statistical conventions and deliberations that we accept to abide.
b) Objectives of the assessment, i.e., what information do we intend to derive from data.

Statistical conventions and deliberations

The main statistical conventions and deliberations that need to be defined and agreed upon when deciding about sample size include:

1) ***Statistical significance***: Defines the probability at which we decide that results observed on a sample can be generalized to the population. This probability can be interpreted as the level of risk associated with accepting the hypothesis that the statistical information derived from a sample can be generalized to the population. This probability is also called Type I error and it is conveniently set to 0.05 (symbolized as alpha), which means that we accept a risk up to 5% that we are wrong when rejecting a hypothesis of chance, and generalizing the effect observed on a sample to the entire population.
2) ***Statistical power***: Defines the probability at which we can decide that "no-effect" observed on sample can be generalized to the population. The complement of power (1 – power) is also called Type II error and it is typically set to 0.20 (symbolized as beta) in educational surveys, which means we take a risk of up to 20% that we are wrong when stating there is "no-effect" and generalizing this finding to the population.
3) ***Minimum detectable difference***: Size of the difference that we can detect based on data collected from our sample. This difference may refer to either the difference between a sample statistic and a population parameter (e.g., average student performance), or the difference between two or more groups of students (e.g., males vs. females). Minimum detectable difference (MDD) can be standardized, i.e., expressed as a fraction of standard deviation, in which case the value is called ***minimum detectable effect size*** (MDES=MDD/s). Defining a desired MDES affects the sample size estimation because larger samples are needed for detection of smaller differences. The MDES for the NAS 2021 sampling is set to be little less than 0.30, which means that the sample size needs to be sufficiently large to detect a difference of the size equal to 30% of standard deviation.
4) ***Cluster sampling method***: In educational research, like many other practical surveys, the ultimate sampling units (students) are grouped in clusters (schools). Since the utilization of

simple random sampling (SRS) would be associated with greater logistic challenges, it is customary to use a cluster sampling (CS) method. In CS we first select a sample of clusters (schools) and then a sample of students within each school. Since the schools share common contextual factors, there is a certain degree of similarity or uniformity of student performance within schools. Consequently, increasing the number of students within schools does not bring much of new information. As a result, cluster sampling usually requires a larger sample size compared to SRS method. Thus, depending on the level of uniformity of students within schools (and heterogeneity across schools), the sample size needs to be increased by a factor called ***design effect (DEFF)***. This adjustment is based on the level of similarity between the members within clusters, technically called ***intraclass correlation (ICC)***, which is calculated from existing data collected by similar instruments in similar conditions. For NAS 2021 sample size estimation, the ICCs and corresponding DEFFs were calculated from the NAS 2017 data, and it was decided that the average obtained DEFF of 7 will be applied. This means that the estimated sample size for SRS will be multiplied by factor 7 to obtain sample size for CS.

Objectives of the assessment

In addition to statistical conventions and deliberations, sample size also depends on objectives and nature of the assessment information that is intended to be derived from data, i.e., formulation on research questions, stratification decisions, etc. For NAS 2021, the main objectives considered for determination of sample size include:

- Measure student achievement at national level
  - Overall national
  - By gender
  - By rural/ urban
  - By social group
  - By school management
- Measure student achievement at each state level
  - Overall state
  - By gender within state
  - By rural/ urban within state
  - By social group within state
  - By school management within state
- Measure student achievement at each district level
  - Overall district
  - By gender within district
  - By rural/ urban within district
  - By social group within district
  - By school management within district

When research questions require deep disaggregation of the sample down to the district level, and moreover, within districts into sub- groups (e.g., gender, social group, school management), then several key questions need to be considered as they will substantially affect the sample size at all levels:

A. What is a ***minimum size of the group that can be meaningfully reported*** with a reasonable level of precision? Examples: estimating average district performance score, estimating how this score relates to some standard (performance level cut score, or national average) or any other fixed value. This requires sample size estimation method for one group vs. population (or fixed value).
B. What are ***minimum sizes of the groups that can be compared*** at a reasonable level of power? Examples are comparisons between girls and boys, urban and rural students, students from different social groups, or students from different school management. This requires sample size estimation method for two independent groups.

Method A: one group vs. population comparison

The estimation of sample size based on method A (one group vs. population or fixed value) relies on evaluation of standard error of the mean and associated confidence intervals based on central limit theorem (CLT) and sampling distribution of the sample mean. In the case of one group, there is one standard error as the other value being compared is fixed (for example cut score) or associated with negligible error (group population mean, national mean).

A common formula for estimation of sample size for SRS in the case of one group vs population is derived from the expression for Confidence Interval:

$n = 4\left(z_{1-\frac{\alpha}{2}}\right)^2 \frac{s^2}{(2\ MDESs)^2}$ which simplifies into $n = \left(z_{1-\frac{\alpha}{2}}\right)^2 \frac{1}{MDES^2}$

Method B: two groups comparison

Commonly used formula takes the following inputs: Type I and II error rates, variance, minimum detectable difference, and the relative sizes of the compared groups (most commonly they are treated of being equal). Total sample size for both groups is estimated by the following expression:

$$N = \left(\frac{1}{q_1} + \frac{1}{q_2}\right)\left(z_{1-\beta} + z_{1-\frac{\alpha}{2}}\right)^2 \frac{1}{MDES^2}$$

Where q_1 and q_2 are proportions of cases in group 1 and group 2 (typically 0.5 for both if the groups are of the same size). Then, a sample size for each group equals:

$n_1 = q_1 N$ and $n_2 = q_2 N$

It is important to consider this formula for NAS 2021 in situations where two compared groups are of unequal size, for example by school management, government and government aided school groups may be of substantially different size if proportionate sampling is used. The alternative is to use non-proportionate approach (which in the stage of data analysis needs to be compensated by sampling weights) and to set the groups to equal size.

Adjustments of sample size

Based on size of finite population

Further adjustments of sample size can be done in case the size of finite population is known. This adjustment does not have effect for larger populations, but for smaller populations the estimated sample size will be progressively decreased. The following formula can be used for adjustment based on the known finite population size:

$n_{adj} = N*n / (N + n - 1)$

where n_{adj} is adjusted sample size, n is original sample size, and N is size of finite population.

For example, if the original sample is estimated to be 500, for the finite population size of 5000 the adjusted sample size will be 455, and for the finite population size of 1000 it will be adjusted to 334. This adjustment for finite populations for NAS 2021 will be made at the district level for the samples of students in schools by administration types.

Based on design effect

The rationale for adjustment by DEFF is described earlier in this document and is normally used in practice whenever cluster sampling method was applied. DEFF is typically calculated from similar data administered in similar conditions. Researchers are cautioned to be careful in making decisions about the size of DEFF as it is not certain to which degree are old data alike the new data to be collected.

Analysis of NAS 2017 data to obtain the ICCs at district levels showed substantial variation among districts and subjects in estimating ICCs and subsequently DEFF. The range of DEFF was between 2 at lower end and 20 at high end, with most of values being in the range between 5 and 10. Overall, these are large DEFF factors, and their application sets very demanding sample sizes.

Based on assessment design

The number of students to be sampled also needs to take into consideration the assessment design. For example, if the assessment design decides that a sampled student will be tested in two subjects; the calculation of sample number of students needs to take into consideration this aspect as well. Based on the total number of subjects in each grade and the fact that each student takes two subject tests, it is relatively easy to determine adjustment factors for sample sizes at each grade. In NAS 2021, all students shall be appearing in all subjects. Therefore, for grades 3 and 5, the adjustment factor is 1; in grade 8 there are 4 tests, and each student will be appearing in 2 tests, so the adjustment factors is 4/2 = 2; and in grade 10 there are 5 tests, and each student will be appearing in 2 tests, so the adjustment factor is 5/2 = 2.5.

Based on non-response rate

The sample size needs to be increased further based on the anticipated non-response rate, which happen due to natural absenteeism of students on the date of test administration. The anticipated nonresponse rate is computed based on experience. For NAS 2021 it is determined that the sample size will be increased 4% to compensate for the anticipated non-responses.

Part B - Methodology followed for Sampling in NAS 2021 Approved by Steering Committee of NAS 2021 on 5th April 2021

Step 1: Finalise coverage of Schools:

Frame (all classes): UDISE+ 2019-20 is used

Target population: Students of Class 3, 5, 8 and 10 of all recognised schools

Mediums of instructions under coverage for each class: 22

(Considering all mediums of instructions found in the sample of class 3, 5, 8 or 10 of NAS 2017)

coverage in NAS 2017		Coverage in NAS 2021 (all classes 3, 5, 8 and 10)
Common for Class 3, 5, 8 and 10	16	Assamese, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Manipuri, Marathi, Mizo, Odiya, Punjabi, Tamil, Telugu, Urdu, Bodo
for class 3, 5, 8	3	Garo, Khasi, Konkani
for class 10	3	Nepali, Bhutia, Lepcha

Exclusion of small schools for administrative convenience: If enrolment in a class is less than 6, school will be excluded for sample selection of that class.

Computation using UDISE+ 2019-20	Grade-3	Grade-5	Grade-8	Grade-10
Desired national target population	2,36,81,606	2,32,88,391	2,12,88,796	1,83,59,603
Exclusions:				
Small schools (Enrolment less than 6 in grade 3/ 5/ 8/ 10)	7,73,032	7,19,624	1,41,774	21,311
Enrolment in mediums of instructions not covered in NAS 2017	3,558	3,376	4,636	7,363
Total excluded enrolment	7,76,590	7,23,000	1,46,410	28,674
% of excluded enrolment	3.20	3.04	0.68	0.16
Defined national target population	2,29,05,016	2,25,65,391	2,11,42,386	1,83,30,929

1.1 Sampling frame structure: Sampling frame will contain the following columns:

- State Code
- State Name
- District Code
- District Name
- Block Code
- Block Name
- School Code
- School Name
- Management Code
- Management (Govt./Aided/Recognized Private Unaided)
- Category Code
- Location (Rural/Urban)
- Enrolment Boys
- Enrolment Girls
- Enrolment Total
- Number of teachers (total, by classes taught)
- Whether selected grade have more than one medium of instruction (Yes/No)
- Enrolment of students with mediums of instruction under coverage of NAS
- Whether selected grade have more than one section (Yes/No)
- Whether selected grade have any children with special needs (Yes/No)

Step 2: Make separate frame of each broad type of school management as under: (school management codes as per UDISE+ Data Capture Format)

Frame 1: State Government schools (coverage of NAS 2017)	School management 01, 02, 03, 06, 90 (01-Department of Education, 02 – Tribal Welfare Department, 03 – Local Body, 06 – other Government managed, 90 – social welfare department)
Frame 2: Government Aided schools (coverage of NAS 2017)	School management 04 (04 – Government Aided)
Frame 3: Private Unaided recognised schools (New coverage of NAS 2021)	School management (05- Private Unaided Recognised, 97 – recognised madrasa)
Frame 4: Central Government schools (New coverage of NAS 2021)	School management 91 to 96 (91- Ministry of Labour, 92 – KV/ Central School, 93 – JNV, Sainik school – 94, Railway School – 95, Central Tibetan School - 96)

The objective is to provide separate estimate of each of these types of schools at district level. The districts for the purpose of sample selection will be the 733 districts as per UDISE+ 2019-20.

Step 3: Estimating target sample size per subject per grade

Input: Minimum Detectable Effect Size = 0.29, Design Effect = 7 and non-response 4%
Output: Estimated sample size per subject per subject assessment: 333
The formulae and algorithm have been given in separate word/ excel document.

Step 4: Adjusting of sample size determined in step 3 (apply for less than 1,00,000 students):

$$n_{adj} = N*n / (N + n - 1)$$

Examples:

i. **if enrolment of class 3 of Government schools in a district is 9000, target sample size for mathematics of class 3 will be = 9000 * 333 / (9000 + 333 – 1) = 322**
ii. **if enrolment of class 3 of Government schools in a district is 1000, target sample size for mathematics of class 3 will be = 1000 * 333 / (1000 + 333 – 1) = 250**

Step 5: Deciding target sample number of students per district per grade for frame 1 to 3

(Number of subjects to be assessed at any grade divided by number of subject tests taken by each student)

	Grade 3	Grade 5	Grade 8	Grade 10	Total
Multiplier based on assessment design	**1**	**1**	**4/2 = 2.0**	**5/2 = 2.5**	
Sample size per stratum after Step 4= 333[@]× grade-wise multiplier	**333**	**333**	**666**	**832**	**2,164**

[@] This (333) will be adjusted as per Step 4 above

If number of sample students is less than 200, then the ceiling of 201 will be fixed.

Step 6: Selection for Frame 1/2/3: District-wise

Selection is to be made independently for each class 3, 5, 8 and 10

Descriptions below are for Class 3. Same steps will be followed independently for class 5, class 8 and class 10 (333 students for class 3/ 5, 666 students for class 8, 832 students for class 10, adjusted suitably for finite population correction, as computed in step 5 above)

Sample size requirements:

A. Check the minimum student requirement for the specific grade for each district. Thus, if number of students in the district is less than or equal to 333 as per frame, all schools in the frame will be selected for survey.

B. Maximum number of students to appear in test from a school is 30. Therefore, compute maximum effective sample students from all schools, i.e., Minimum (grade enrolment, 30). If sum of maximum effective sample student is less than or equal to 333, select all schools in the frame.

For the districts × Grade where enrolment is higher than those described in steps 6-A and 6-B above, sampling will be necessary.

Size measure: Total enrolment in the target class will be considered as size measure (i.e., when we select schools for Class 3, enrolment for class 3 as per UDISE+ 2019-20 will be considered as size measure, when we select schools for class 5, enrolment of class 5 as per UDISE+ 2019-20 will be considered as size measure and so on).

Step 6-C: (for districts not covered in Step 6-A or 6-B)

i. Suppose E is total enrolment and S is total schools. Then, compute round(E/S,0) as average enrolment of school (AES).
   a. If AES <=30, find target number of sample schools TS_1= (333/AES). To select TS_1 schools, take sampling interval (SI)=round (E/ TS_1 , 0)
   b. If AES >30, find target number of sample schools TS_1= integer (333/30)+1 = 16. To select TS_1 schools, take sampling interval (SI) = round (E/ 16 , 0)

ii. Select all schools with enrolment > SI with probability 1. This is done to ensure that such large schools do not get re-selected in the systematic sample. Then, re-compute SI with reduced number of schools and enrolments.

iii. Repeat Step 6-C-i and 6-C-ii till enrolment of all remaining schools are less than or equal to SI (or modified SI as per re-computation done in step 6-C-ii).

iv. Suppose S_1 schools have been selected in Steps 6-C-i to 6-C-iii.
   a. Enrolment of these S_1 schools is E_1.
   b. Suppose some of these schools have more than 30 students in a class, while remaining schools have less than 30 students. Assume, from these S_1 schools, we get a target sample of n_1 students.

c. Therefore, we must further select (333 - n_1) target students from (S - S_1) schools.

v. Now, AES_1 = (E - E_1) / (S - S_1), the average class size of (S - S_1) schools. Modified target number of sample schools will be TS_2 = (333 - n_1) / AES_1 . Compute SI_1 = round ((E - E_1) / TS_2 , 0)`

vi. From the remaining schools of the district, select TS_2 schools by PPS circular systematic sampling with SI_1 as sampling interval.

vii. For circular systematic sampling, the schools in the frame will be arranged by enrolment in the class (descending), medium of instruction (if a school has enrolment in more than one medium of instruction, the medium in which grade enrolment is maximum will be considered) and area (rural/urban). After that, prepare a column with cumulative frequency (using enrolment in the specific class/grade). To draw the samples using circular systematic sampling, draw a random number between 1 to (E - E_1). Let this be RN_1. Select the school against which RN_1 falls as per the cumulative frequency table (example given below). Then, compute (RN_1+ SI_1), (RN_1 + 2*SI_1), …, (RN_1 + (X_1 - 1) * SI_1), to earmark the remaining sample schools. During this phase, if (RN_1 + i * SI_1) becomes higher than (E - E_1), compute ((RN_1 + i * SI_1) - (E - E_1)) to get the number for school selection.

viii. After selection through Steps 6-C-v to 6-C-viii, if it is found that the effective number of sample students is less than (333 - n_1), then Steps 6-C-v to 6-C-viii will be done afresh, starting with increasing the target number of sample schools from TS_2 to TS_2 + 1.

Step 6-D: (for districts which return less than target students after 1 iteration of step 6 C) The entire step 6-C will be repeated after increasing target number of sample schools by 2.

Step 7: Selection for Frame 4 (Central Govt schools)

Note that total enrolment in Central Govt schools, as per UDISE+ 2018-19 is 16.86 lakhs, which is only 0.7% of enrolments in schools under NAS coverage. Total such schools are **2211** at all-India level, spread over 667 districts. In 2 districts, namely, Lucknow and Gorakhpur, number of such schools are more than 50. In 16 other districts, more than 10 such schools are available.

i. If number of schools in a district is 10 or less, all schools will be selected. All eligible classes will be surveyed. For example, in most States, JNV starts from class 6. In those cases, class 8 and 10 of the JNV will be surveyed.

ii. If number of schools is more than 10, 10 schools will be selected from the district by PPSWR, with total enrolment as the size measure. All eligible classes will be surveyed.

Step 8: Selection of Section and Students in a selected school (All the frames)

i. In a selected school, if the enrolment in target class is less than 30, all students will be selected for the test.

ii. If enrolment in target class is more than 30 and students are in a single Section and number of students present on the date of assessment is more than 30, 30 students will be selected at random using circular systematic sampling based on class attendance register. The total number of students in the class should be divided by 30 to get the interval (m, rounded off to the nearest lower integer). The first number can be decided based on draw of lot method. Suppose the class has 45 students. Then, numbers 1 to 45 can be written on small pieces of paper, folded alike, and mixed well. Then one piece will be drawn randomly. Suppose the number drawn is 35. This will be the roll number of the first random student selected for the test. Then, every m^{th} student should be selected starting from roll number 35 in the attendance register, till selection of 30 students is

complete. If some selected roll numbers are absent on the day of test, then the selection process will continue till a total of 30 students are selected for the test.

iii. If students are in multiple Sections, the following stepwise procedure will be followed:

a. One section will first be selected at random. This can be done through draw method. Names/ numbers of all sections of the class can be written on small pieces of paper, folded alike, and mixed well. Then one piece will be drawn randomly and that would be the Section of the class where students will be assessed. The Teacher Questionnaire will be canvassed for this selected section only.

b. Up to 30 students from this selected Section will be selected for the test by circular systematic sampling as described above in step 8_ii.

c. Suppose the number of students on the date of examination in the Section selected at Step 8-iii-a is n_2 (less than 30), one more Section will be selected by SRSWOR.

d. $30 - n_2$ students will be selected by circular systematic sampling from the Section selected at Step 8-iii-c.

e. Step 8-iii-c and 8-iii-d will be continued till a total of 30 students are selected from this class in the selected school. This is to be done to minimise loss in target number of sample students.

f. If the different Sections of the selected class are taught through different mediums of instruction, steps 8-iii-c to 8-iii-e shall be done for the Sections whose medium(s) of instruction(s)is (are) in the survey coverage of NAS 2021. The mediums of instructions being covered in NAS 2021 is listed in Step 1 above. This is done to ensure minimum loss of student sample during the student selection process within the selected school and class.

Step 9: Distribution of Test Forms to Students selected for the test (All the frames)

i. Suppose there are t types of test booklets for a class (e.g., Booklet type 1, booklet type 2, booklet type 3, etc.). Then, the UDISE+ 2019-20 code of the school will be divided by t and the remainder will be calculated.

ii. If the remainder is 1, distribution of booklets in the school will start with booklet type 1, if the remainder is 2, distribution of booklets in the school will start with booklet type 2, and so on. If the remainder is 0, distribution of booklets in the school will start with booklet type t.

iii. Step 9 ii will ensure that the distribution of booklets is nearly equal at the frame/ district level.

***