CHAPTER 7: ASSESSMENT DESIGN

Introduction

The NEA is designed to provide information about students' achievement in Reading Literacy, Writing Literacy, Listening and Speaking Literacy, Mathematical Literacy, and Scientific Literacy, which can be used for evidence-based decision making by policymakers and education leaders to improve students' achievement levels and practices as set out in the national goals.

A key aspect of the NEA is to monitor trends in performance between grades and track changes in students' performance from one assessment cycle to the next. This ability to track changes in students' performance across years will enable the Royal Government of Bhutan to develop interventions that support improvement in student performance. The central idea is to collect information about student performance at regular intervals of three years on achievement in Reading Literacy, Writing Literacy, Listening and Speaking Literacy, Mathematical Literacy, and Scientific Literacy, along with contextual information that could have an impact on learning.

Instrument Design

Key considerations at the time of designing the instruments are – what sort of data will be collected, how will data be analysed and reported, and how the findings might be used in policy making and improving the education system.

Different instruments will be developed for each grade level. At each grade level, domain- specific questions will be included in the instruments. The usefulness of each variable will be described within the analysis plan, including how to analyse the variables. Variables that are critical to inform education policy development and will be able to produce reliable data will be selected for the instruments.

The instruments will be designed in the English language except for Dzongkha domain, however, care will be taken to ensure that respondents are able to understand and respond appropriately. The instruments will contain instructions on how questions should be answered. The two types of items that shall be used to assess the cognitive domains are Multiple Choice Question (MCQ) and

Constructed Response Task (CRT). Given the nature of large-scale assessment, majority of the items will be MCQ. Item difficulty and complexity shall also be considered.

Annexure 4 provides an example of the steps in questionnaire development, and the person who should be responsible to complete the task.

Finalisation of the Instruments

The instruments will be reviewed and revised by a panel of experts. The instruments will undergo field trialling to establish the suitability of questions, appropriateness and clarity of language used, and response categories. Post the field trial, the data will be analysed and the instruments will be modified, if necessary.

Questionnaire layout- While designing the questionnaire layout for NEAF, the following points will be considered –

1. Questionnaires that are easy to use for the respondent will have
   - a simple, consistent way of answering items
   - an uncluttered presentation
   - response categories that are clearly associated with each question
   - explanations for codes used, if any
2. Layout of questionnaire will be easy to use for data procession after the administration.

Data Analysis Plan

An outline of the proposed contextual data analysis will be ready before the finalisation of the key variables included in each instrument. The data analysis plan will specify the following:

- The kind of information will be provided by each question in an instrument
- The way of the information will be used in the analysis
- The method of meaningfully analysing the variables included in the instruments, while ensuring that there is no redundancy in the analysis.

It may not be possible to capture some variables that may have an impact on the learning of students, independently. For example, a composite index of socio-economic variables will be constructed by combining various background factors of parents of the sample students such as their economic background, their education and occupation, etc. in a weighted manner. Similarly, a composite index of essential learning attributes will be constructed by using cognitive instrument data and contextual questionnaires. The data analysis plan will describe how these variables will be aggregated to produce the composite index (if required), and how the composite index will be used.

The information captured by the instruments will be analysed with various statistical techniques such as descriptive statistics, graphical analysis, comparing means of two variables, correlation, and regression analysis. The choice of data analysis methodology will be dependent on the types of variables, research model and their hypotheses.

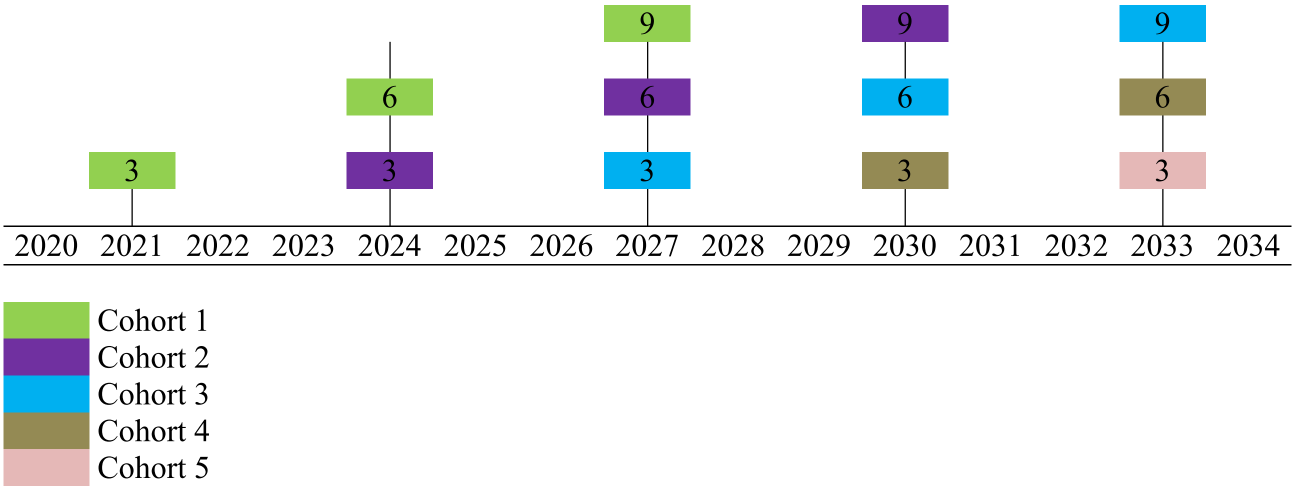
Assessment Cycle

The NEA for the grades III, VI, and IX will be conducted in a three-year cycle model is used. This will ensure that the gap between the grades will be same as the number of years of interval in each cycle. This model will serve two fundamental purposes, namely, tracking the same cohort across the school years and identifying the impact of long-term interventions in school education in Bhutan.

In this model (approach), policy changes can be introduced at grade 3 (entry level grade) and effect of the changes can be monitored in a phased manner. It reduces the load of introducing changes to cohorts accustomed to one model of education. The diagram below shows the design in which different cohorts can be tracked and their progress monitored over a period of time. The complete cycle for cohort 1 to cohort 3 are visible and the same can be replicated in the model.

In 2021, Grade 3 will be assessed in Reading Literacy and Mathematical Literacy domains. After three years, along with grade 3, grade 6 will be assessed in Reading Literacy, Writing Literacy, Mathematical Literacy and Scientific Literacy. In 2027, along with grades 3 and 6, grade 9 will be assessed in Reading Literacy, Writing Literacy, Mathematical Literacy and Scientific Literacy.

Figure 7.1: Tracking student performance in the NEA



Sampling

The NEA is a sample-based assessment to monitor student performance in Reading Literacy, Writing Literacy, Listening and Speaking Literacy, Mathematical Literacy, and Science. The assessment is for school going students in grades III, VI, and IX across Bhutan. For the purpose of the assessment, it is proposed that the target population is defined as all students studying in the target grades in Bhutan. To achieve the level of precision and accuracy necessary in the results, it is recommended that systematic sampling is used to identify a representative sample and report on subgroups of interest.

The NEA will follow a systematic multi-level sampling similar to renowned large-scale learning assessments, including PISA and TIMSS. The sampling will have two levels − school level and student level. Schools will be sampled at the first level by the method called Probability Proportional to Size, and students will be sampled at the second level by Simple Random Sampling. For consistency, the NEA needs to use a single data source across all sampling units when constructing the sample frame.

1.1.1 *Level 1 - School sampling within each sampling unit*

The sampling frame for the sample for each grade would list all schools where the target population is studying, in that grade. A separate sampling frame for each grade will be developed.

The details of sampling requirement will be decided during sample planning of the NEA based on the suggestions provided by an international expert agency. The details will include measure of size (MOS), stratification variables, use of replacement schools and method for identification and non-response adjustments at the school and student levels.

1.1.2 *Level 2 - Students sampling within each sample school*

Within school sampling procedures involve the selection of students prior to testing. When deciding which within school sampling approach to use, it is important to consider how best to balance technical and logistical demands and to ensure that every student has an equal chance of being selected in the sample. Stratification or clustering at this level will decrease the efficiency of the sample, in order to achieve the same level of precision. If this is the case, more students are required than a simple random sampling scenario.

Reporting Student Achievement

Student response data collected will be used to develop a scale for mapping student performance, comparing performance between sub-groups of population, and also for comparisons over time. Development of the scale will be followed by describing the achievements at various levels on the scale to meaningfully draw a composite picture of student performance in the respective domains. These descriptions will form the learning progression or metrics. It must be pointed out that different domains will have different scales and the scales and scores are not comparable between domains. This means that a scale score of 250 in the mathematics domain does not have the same substantive meaning as a scale score of 250 in the science domain.

In 2020, a learning progression will be developed for grade III. In the subsequent cycles, when other grades are included in the assessment, this learning progression will be further enriched to have a learning progression from grade III to the highest grade being assessed under the NEA for each cognitive domain. This learning progression will cover domain specific competencies and content to present a composite picture of achievement on a scale. Bhutan will be, then, able to compare students' performance of various subgroups on this learning progression with the help of the scale score, and will also be able to describe the achievements using the learning progression.

The scale will have a midpoint of 250 for grade III, and 50 scale score points will be equivalent to one standard deviation of all students' distribution in NEA 2020. Other grades, when included, will use the same scale and map the progression accordingly on the same scale. Some items from the cycle will be used in the next cycle to link the assessment results from the two cycles and report the performance of both the assessments on the same scale. This will enable comparisons between cycles and monitoring of performances over time and between grades.

Development of a learning progression to describe performance of the students or groups requires collecting data on a large number of items. There are limitations on the number of items a student can attempt before fatigue sets in, and also on the time for which a student can sit for the test before they start losing interest in it. Both of these limiting factors have an influence on gathering reliable information about student performance. The next section describes a model in which a large number of items can be assessed and also addresses the issues of students fatigue and interest.

In addition to the student performance in each domain, the NEA will also collect contextual information about the students and other factors that influence students learning, as stated in the previous chapter. Analysis of contextual variables and their association with the students' performance will provide meaningful implications in understanding students' learning process and ultimately developing education policies, interventions, and reforms.

Assessment Booklet Design

To correctly measure a learning progression, the number of items required to be tested in the NEA will usually be more than the items that can be possibly answered by one student within the available testing time. To mitigate this problem, the NEA will use multiple booklets which involve assigning all assessment items to at least one assessment booklet for each domain. There will be some items which would be common between booklets to enable linking of student responses from different booklets. Each student will be required to complete only one booklet in a domain.

Selection of link items will be undertaken using the two basic criteria:

- link items have a range of difficulties
- link items have a coverage of all strands in the domain.

This implies that link items will behave somewhat like a mini test. However, the exact number of link items will depend on the number of strands and also the range of difficulties covered. Usually, 10-15 items that are statistically sound and meet the above mentioned criteria are used as link items between two booklets.

For vertical linking between two grades, care must be taken to choose items that are appropriate for both grades and that all students of any particular grade do not either get a zero or full credit in those items. For example, if item A is a link item between grade III and VI, it should not have a situation where all students of grade III or grade VI get a zero or full credit in item A.

Considering the complexity of measuring students' learning, it is required for the NEA to develop a booklet design which enables to collect sufficient and reliable information through two sets of booklets per domain. Some design considerations for designing booklets are:

- statistical objectives to be met
- reporting student performance
- test administration format (for example, single domain, multiple domains, all domains administered to one student)
- student testing time
- item positioning effect
- item linking (horizontal, vertical, and historical)
- number of items in the pool
- number of items to be released to the public.

There are multiple approaches to booklet design, like rotated design, matrix design and separate booklet design. The NEA will use simpler version of separate booklets in which all of the items in a pool can be grouped into clusters. The clusters are then assigned to different booklets. However, there would be a cluster or a number of clusters which would be assigned to multiple booklets for the purpose of linking.

The items in the pool will be grouped into clusters, the size of which, can depend on one of the two factors namely number of items in each cluster or time required to answer the questions in each cluster. When the number of items in each cluster is used as the deciding factor of cluster size, the test developers need to be careful to develop clusters that require approximately same time to answer them. When the duration of time required to answer questions in a given cluster is kept for the same, then the test developers have flexibility to vary the number of items in each cluster. However, given the number of items in each cluster will be less, it may not be possible to cover all content areas. Therefore, care will be taken to maintain equivalent difficulty levels for all clusters. The difficulty level of each cluster will be decided based on the trial data available for the NEA. For the purpose of the NEA, the clusters will have equal number of items and each student will be assessed on an equal number of items.

Table 7.1 below shows a sample booklet design for a domain. In the design, all test items have been assigned to one of the seven clusters. Among the seven clusters, cluster 2 is assigned to all sets. This cluster 2 is the link cluster and the remaining six clusters are assigned to the two booklets.

Table 7.1: An example of two booklet design for a domain

Booklet A	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Booklet B	Cluster 5	Cluster 2	Cluster 6	Cluster 7

The same model can be used for assessment design for all grades being assessed, and there could be link clusters between grades to enable comparisons between the grades. It should be noted that clusters will have same number of items in a grade. However, the number of items will increase in a cluster as grade level increases.

Student Testing Time

The testing time allocated for the administration of the items has an impact on the booklet design. It determines the number of items that can be administered to each student. For example, if testing time allocated for the cognitive instrument is 40 minutes, it will allow only a certain number of items in the domain to be administered to each student. Generally, each item will require a different amount of time to answer and, therefore, a number of items are grouped into clusters that have similar difficulty and similar testing time. These clusters can then be assigned to different booklets that can be administered to students.

In the NEA, time will be allocated for administration (instructions, etc.), for responding to the cognitive instrument, as well as for responding to the questionnaire, as shown in the table below.

Table 7.2: Assessment time for one domain

	Instructional time	Questionnaire time	Domain time	Total
Grade III	10 min	15 min	40 min	65 min
Grade VI	10 min	15 min	60 min	85 min
Grade IX	10 min	15 min	60 min	85 min

According to the table, each student of grade III will need 65 minutes overall for the assessment and students of grades VI and IX, will need 85 minutes each to assess one domain. For assessing each student in two domains, testing time would increase by 50 minutes in grade III and 70 minutes in grades VI and IX. In this case, a 15 minute break should be provided to students between the first and the second domain. Total testing time in such a case will be as given in Table 7.2.

Table 7. 3: Assessment time when student takes two domains

	Questionnaire time	Instructional time	Domain 1 time	Break	Instructional time	Domain 2 time	Total
Grade III	15 min	10 min	40 min	15 min	10 min	40 min	130 min
Grade VI	15 min	10 min	60 min	15 min	10 min	60 min	170 min
Grade IX	15 min	10 min	60 min	15 min	10 min	60 min	170 min

The detailed procedures of the administration of the domains shall be specified in the Test Administration Manual.