Ministry of Education
Government of India

सत्यमेव जयते

विद्यया ऽमृतमश्नुते

एन सी ई आर टी
NCERT

NIPUN BHARAT

National Initiative for Proficiency in Reading with Understanding and Numeracy (NIPUN)

BENCHMARK SETTING FOR FOUNDATIONAL LITERACY AND NUMERACY IN INDIA

Supported by unicef

for every child

BENCHMARK SETTING FOR FOUNDATIONAL LITERACY AND NUMERACY IN INDIA

Preface

The Foundational Learning Study 2022 represents a significant endeavour to explore and enhance the fundamental building blocks of education. Within this study, benchmarking emerges as a key component, offering a comparative lens through which to evaluate educational practices and outcomes. This preface sets the stage for a detailed examination of the benchmarking process employed in the study, highlighting its importance, challenges, and potential impact on shaping the future of foundational learning.

Benchmarking is a powerful tool that enables one to measure their performance against a scientifically identified standard. In the context of education, benchmarking allows education systems to assess their effectiveness in delivering foundational learning outcomes and identify areas for improvement. When performance is looked at from a comparative perspective, educators can gain valuable insights into their strengths, weaknesses, and opportunities for growth.

The Foundational Learning Study 2022 recognizes the transformative potential of benchmarking in the field of education. By applying this approach to the study of foundational learning, the research aims to uncover innovative practices, identify barriers to success, and develop strategies for planning interventions in order to overcome these challenges. Through a rigorous process of data collection, analysis, and comparison, the study seeks to provide a comprehensive understanding of the factors that contribute to effective foundational learning.

Through this extensive report, we aim to provide a comprehensive overview of benchmarking in the context of the Foundational Learning Study 2022, offering valuable insights for educators, policymakers, and stakeholders invested in improving educational outcomes for all learners. By sharing the findings and lessons learned from this study, we hope to inspire and guide the education community in their efforts to enhance foundational learning and ensure that every student has the opportunity to reach their full potential.

Prof. Indrani Bhaduri
Head ESD, NCERT

Table of Contents

List of Figures

List of Boxes

List of Pictures

List of Tables

ABBREVIATIONS

AIR	American Institutes for Research
DIKSHA	Digital Infrastructure for Knowledge Sharing
DoSEL	Department of School Education and Literacy
FLS	Foundational Learning Study
GPD	Global Proficiency Descriptor
GPL	Global Proficiency Level
GPF	Global Proficiency Framework
IRR	Interrater reliability
JE	Just Exceeds Minimum Proficiency
JM	Just Meets Minimum Proficiency
JP	Just Partially Meets Minimum Proficiency
MOE	Ministry of Education
NCERT	National Council of Educational Research and Training
NEP	National Education Policy
NIPUN	National Initiative for Proficiency in Reading with Understanding and Numeracy Programme
ORF	Oral Reading Fluency
PTR	Pupil teacher ratio
RC	Reading Comprehension
SCERT	State Council of Educational Research and Training
SDG	Sustainable Development Goal

1. NIPUN Bharat – National Mission for Foundational Learning in India

With India's commitment to the targets set for SDG 4, the National Education Policy (NEP) 2020 acknowledges that the *"lofty goal"* will require the entire education system to be reconfigured. The NEP acknowledged that a large number of students in elementary school have not achieved foundational literacy and numeracy. Duly noting India's learning crisis at foundational years, the NEP clearly states that – *"the rest of this Policy will become relevant for our students only if this most basic learning requirement (i.e., reading, writing, and arithmetic at the foundational level) is first achieved"* One of the key recommendations of the policy is a modified pedagogical and curricular restructuring of the Indian education system. The existing 10+2 structure is revised to a new academic structure of 5+3+3+4 covering children from ages 3 to18. This has resulted in inclusion of three years of early childhood care and education in the first stage of 5-years, also referred to as the foundational stage of learning. All children attaining foundational literacy and numeracy by the end of grade 3 and no later than grade 5, by the year 2026-2027 has become an urgent national mission and is being given the highest priority. To this end, in July 2020, the Ministry of Education (MoE) launched the National Initiative for Proficiency in Reading with Understanding and Numeracy Programme (NIPUN) Bharat[1].

In order to achieve the objectives of NIPUN Bharat, it was necessary to understand the current status of foundational learning of students enrolled in primary grades. To this the national guidelines of NIPUN Bharat mention that - "a study will be undertaken by NCERT which will be the first large scale assessment and benchmarking study for foundational literacy including oral reading fluency across different languages in India". MoE, along with the National Council of Educational Research and Training (NCERT) initiated the planning of the foundational learning study The objectives of the foundational learning study (FLS) is to set benchmarks for oral reading fluency (ORF) with reading comprehension, for all languages used as a medium of instruction in schools in India and numeracy. **This document is a description of the process adopted for setting benchmarks for ORF and numeracy.**

[1] https://www.education.gov.in/sites/upload_files/mhrd/files/nipun_bharat_eng1.pdf

1.1. Aims and Objectives of NIPUN

NIPUN Bharat programme will focus on:

1. Curtailing drop-outs and providing access to quality education for children in foundational years.
    a. Aspects related to health and nutrition will be parallelly integrated to ensure good child-health for improved school attendance and cognitive development.
    b. Pupil teacher ratio (PTR) will be maintained under 30:1. Areas having large numbers of socio-economically disadvantaged students will aim for a PTR of under 25:1.

2. Filling teacher posts and teacher capacity building.
    a. Priority will be accorded to local teachers who know local languages.
    b. Professional development programmes for teachers will be accelerated and regularised.

3. Development of high quality and diversified student and teacher resources and learning materials.
    a. NIPUN Bharat will ensure that teachers focus on developing – (i) phonological awareness and sound discrimination (ii) visual perception and visual association (iii) abstract thinking (iv) play and activity-based approach (including toymaking, art integration, sports integration, story-telling based learning, ICT integration, groupwork, role plays, project work etc).
    b. A national repository of high-quality resources on foundational literacy and numeracy will be made available on the Digital Infrastructure for Knowledge Sharing (DIKSHA)
    c. Technological interventions to serve as aids to teachers and to help bridge any language barriers that may exist between teachers and students, will be piloted and implemented

4. Tracking the progress of each child in achieving the set learning outcomes

Box 1: What is a benchmark and why it is required

Academic benchmarks refer to assessments that measure students against institution standards and learning goals. Benchmarking allows educators to identify students' strengths and weaknesses, which can then inform their future instruction.

Secondly, comparing cross-country learning outcomes and assessment-results and aggregating those results for global reporting is a challenge. The main challenge is because each country uses different assessment tools with varying levels of difficulty for the same grade. The linking of different assessments to a common scale is also done through the process of **benchmarking.**

Benchmarking may be done statistically or non-statistically., A non-statistical, judgmental method called **policy linking** has been developed for setting benchmarks on national and international assessments. This policy-linking method facilitates reporting on key global indicators related to grade-level reading and mathematics and also makes it possible for countries to set learning targets for long-term improvement of learning outcomes.

The policy-linking method is based on the **Global Proficiency Framework** (GPF). The GPF describes the global minimum proficiency levels (GPL) through a common set of global proficiency descriptors (GPDs) (also called performance standards) by grade level and subject area. Countries can link their national assessments to the GPF for global reporting. Using a standardized benchmarking approach, results from different countries and global/regional large-scale assessments that are linked to the GPF standards can then be compared.

To set the benchmarks, the policy linking method uses an internationally recognized, standardized, test-centred, **Angoff-based benchmarking** procedure. The Angoff procedure requires groups of subject matter experts, called panelists, to make individual judgments on the assessments. The panelists include teachers who are teaching grade-3 students and curriculum experts from the country who understand the performance of learners for specific grades and subjects. The Angoff procedure includes 1) examining the country's assessment instrument(s) in relation to the GPDs and 2) estimating how learners in each of the GPL categories would perform on the assessment.

Planners and facilitators organize and conduct separate workshops by grade, subject, and language with different groups of panelists to set the equivalent benchmarks for those assessments (USAID, 2020, Page 6). The same has been followed in India as well.

Governments have utilized this GPF and policy linking method successfully in more than a dozen countries. This includes seven countries (Angola, Djibouti, Kenya, Morocco, Nigeria, Rwanda, and Senegal) that have also set ORF and comprehension benchmarks.

For more details: https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2020/10/WG-GAML-5-Policy-Linking-for-Measuring-Global-learning-Outcomes-Toolkit.pdf

2. Key considerations for Benchmarking

Many countries have conducted foundational learning assessments and set benchmarks for oral reading fluency and numeracy. However, in India, the scale of the study in terms of number of languages covered for ORF and the sample size was unprecedented. The study sample covered approximately 86,000 children enrolled in grade 3, in state government schools private recognized schools and Kendriya schools. Given the language diversity in India, the FL study was administered in 20 languages, covering all mediums of instruction in the country. For each language the sample included at least 340 students from each of the category of schools. When a certain medium of instruction was used across states, the sample was representative.

This one-on-one administered assessment was done using four test booklets which included test items for both foundational literacy and numeracy. Each student was administered only one of the four booklets.

> "Setting benchmarks in 20 languages is exponential from any country that we have done till now, globally. At max, we have done three languages in a country till now. So the sheer number and regional coverage was a massive effort. The complexity of scale, size and cultures was at a different level in India"
>
> **Dr. Jeff Davis, Lead Facilitator, American Institutes of Research**

The literacy assessment in the FL study covered the following sub-skills:

- Oral language comprehension
- Phonological awareness
- Decoding
- Reading comprehension
- Oral reading fluency with comprehension

For assessing students oral reading fluency there was a passage of 60-70 words. There were five reading comprehension questions associated with the passage. Passages were adapted from English into 20 languages based on specific guidance for adaptation.

The process of administration included the following:

- The ORF passage was presented to the student, and they were asked to read it out aloud.
- The key data noted by the field investigator included, words read in a minute; correct words read in a minute.
- Children could continue to read the passage at their own pace.
- Once the passage was read, the field investigator asked the five reading comprehension questions.

The benchmark for oral reading fluency was established as the number of correct words read per minute and reading comprehension.

The numeracy assessment in the FL study covered seven sub-skills which included skills in both numeracy and mathematics. But for benchmarking the following four sub-skills in numeracy were taken:

- Number and operations
- Measurement
- Data handling
- Patterns

"We have not seen this length in assessment in any country. Most countries used one instrument for assessment, while in India there were four booklets. These four instruments were further prepared in each of the 20 languages. And within each instrument there were 19 subtests, nine for literacy and ten for numeracy. So, test development was very complex.

Then furthermore, only one passage is used globally for ORF per language. But in India, four paras were used for setting benchmarks. Before doing benchmarking, it had to be ensured that if for instance student 1 took para-A&B then how would student 2, who took para C&D, perform on para A&B. So, bringing all four passages on same measurement scale was required to equate the study results. So not just the scale but the complexity of the test design was also different in India when we talk of benchmarking"

Dr. Abdullah Ferdous, Lead Facilitator, American Institutes of Research

The results of the FL study for literacy and numeracy and the benchmarks set will be the baseline for the NIPUN Bharat Mission and help states in designing their plans to achieve the goals set for 2026-2027.

The process of benchmarking followed in India was as per the global prescribed procedures i.e. the benchmarks were set in workshop-mode, using the policy-linking method and Angoff benchmarking tool (Refer Box 1). However, the overall process of benchmarking was at a much larger scale as benchmarks were being set for 20 languages. As mentioned earlier, the global practice is to organize separate workshops for benchmarking for each language, however given the number of languages that were covered in the study in India, **five regional workshops covering four languages each were organized for ORF benchmarking between June-July 2022. A separate five-day workshop was organized in August 2022 for setting benchmarks for numeracy.**

The technical facilitation of the workshop was done by the **American Institutes of Research (**AIR) team. The participants from the states included teachers and pedagogy experts who were referred to as 'panelists'.

The sections below describes the process adopted to set the benchmarks for oral reading fluency for the languages covered in FLS and numeracy. The process has been described across three key phases – (i) preparatory phase which covers pre-workshop activities (ii) implementation phase which describes the activities during the workshop and (iii) post workshop activities.

Figure 1: Flowchart of the Workshop Process

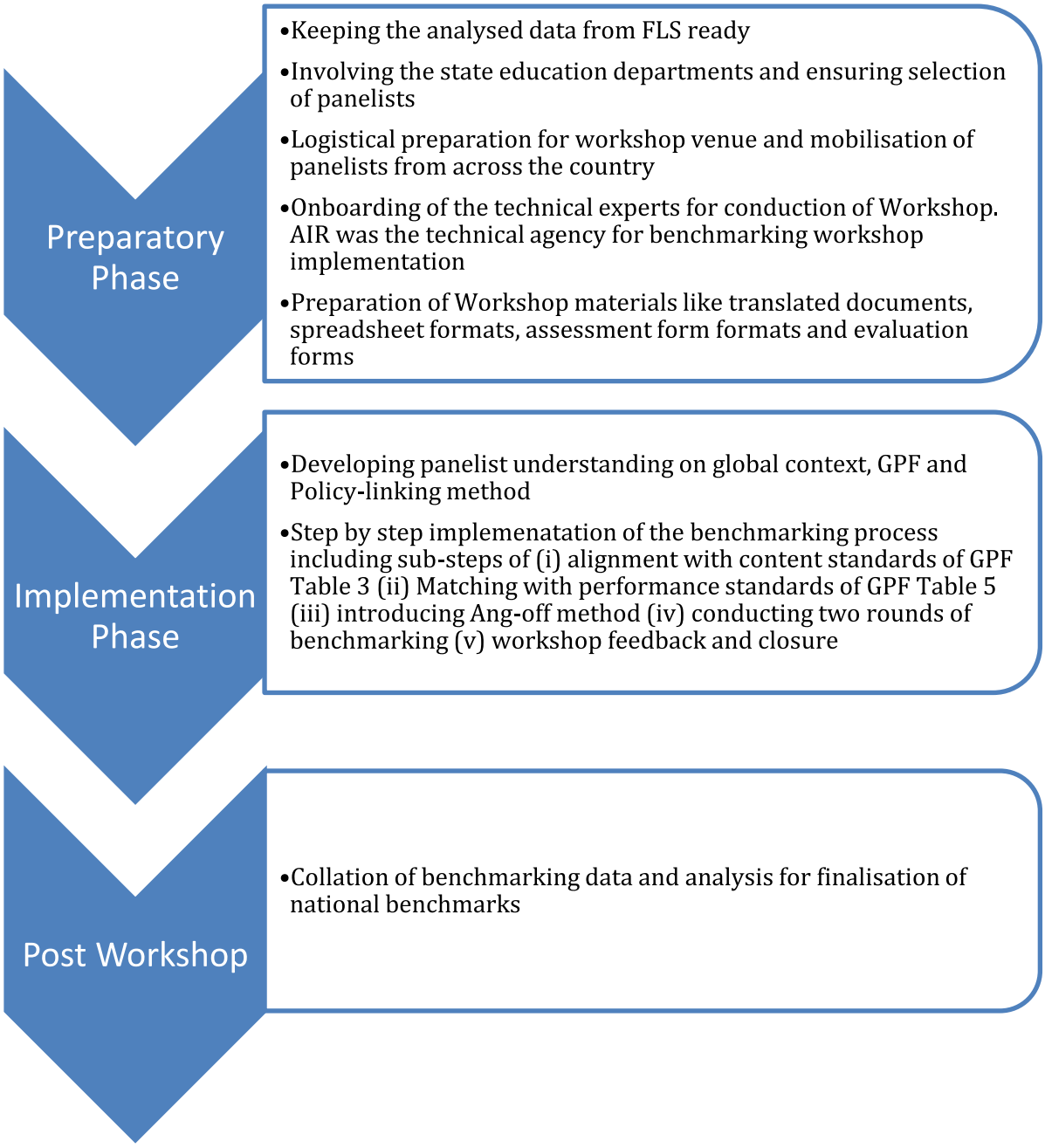


Table 1: Roles and responsibilities of partners

Partners/Stakeholders	Roles
Department of School Education and Literacy, Ministry of Education	- Overall guidance and monitoring of the benchmarking setting exercise
NCERT	- Overall coordination for benchmark setting exercise and provide logistics support - Liasion with MoE, UNICEF, AIR and SCERTs and - Approve the benchmark setting process - Issue necessary directions to the SCERTs - Provide assessment instruments, answer keys, and data sets at the beginning of the workshops to panelists - Finalize benchmark results
UNICEF and its technical agency and its technical agency - American Institutes for Research	- Support in planning and finalizing benchmark setting process - Provide technical resources for benchmark setting process - Orient panelist in benchmarking including understanding of GPF, policy linking method and Angoff method. - Development of benchmarks in oral reading fluency and comprehension for each of the 20 languages - Develop or adapt training materials (training slides and rating forms, as well as pre-programmed spreadsheets to calculate benchmark, evaluation forms). - Estimate benchmark and impact data, standard error of the benchmarks (SE) and panelists' inter-rater consistency and share with the panel between the two benchmarking rounds. - Review data of both rounds and provide suggestions for adjustment of the results. - Share the results for final approval
State Council of Educational Research & Training (SCERT)	- Selection of panelist as per direction of the NCERT and ensure their participation

3. Benchmarking oral reading fluency in India

3.1. Preparatory Phase

Selection of panelists

As indicated in the global policy linking workshop guidance, a team of 15 panelists was planned for each language. This included twelve grade-3 language teachers and three pedagogy experts who were teacher educators from district and state level. The teachers and pedagogy experts were chosen in a manner that there was representation from all states in which the particular language was used as the medium of instruction in schools. For example - if eight states use Bengali as a medium of instruction, then the workshop participants were proportionally representative from all eight states.

The pedagogy specialists selected had university degrees and several years of experience of teaching the language in primary classes at school. The teachers were also qualified and with years of experience of teaching in primary classes especially teaching language in grade 3. Amongst the participants there were teachers from state government schools.

Additionally, one lead facilitator (with strong assessment and psychometric expertise) and one data analyst (with statistics and data processing background) from NCERT, supported the lead facilitators of the workshop during the presentations and discussions with the panelists and in data entry respectively. One language or reading expert (resource person) and one data entry person were ensured for each language.

Pre-workshop exercise

Prior to coming for the benchmarking workshop, all panelists were required to administer the oral reading fluency assessment and the related comprehension questions to at least nine children in their school. A sample paragraph and reading comprehension questions was shared with them in advance, along with detailed guidance for conducting the assessment. The primary objective of this exercise was to help panelists develop a familiarity on oral reading fluency and how children generally perform on this test item. This experience was important as it enabled panelists to understand the technical processes followed for setting benchmark in the workshop.

Materials shared for the Workshop

1) ORF passage with five reading comprehension from the Literacy test booklet
2) Global Proficiency Framework (GPF) - Reading, for grade 3 (translation of this was made available for participants in their respective languages)
3) Rating Forms used by panelists
4) Workshop Evaluation Form

Agenda for the Workshop

A standard format of four-days was applied to each of the five regional benchmarking workshops for ORF. The broad agenda was as follows:

Day 1: Background and Alignment

Day 2: Matching

Day 3: Round 1 rating by panelists for setting benchmarks

Day 4: Feedback on Round 1 ratings, setting Round 2 ratings, feedback by participants and workshop closure.

The round 2 ratings were not shared with panelists.

3.2. Implementation Phase

The workshops were a sum of three key tasks. The **first task** involved orienting panelists on what is the Global Proficiency Framework (GPF), policy linking method and making judgements on the alignment of the assessment questions to the GPF. The **second task** involved discussing the GPF for reading for grade 3 and matching items with the Global Proficiency Levels (GPLs) and Global Proficiency Descriptors (GPDs) to judge the skills and abilities needed by students (hereon referred to as 'learners') to answer the items correctly. The **third task** involved introducing the Angoff method and conducting two rounds of ratings to set initial and final benchmarks on the ORF questions (hereon referred to as 'items') included in the FL study for the particular language. These key tasks are described in detail in the following sections.

Understanding GPF and Policy Linking

At the opening of the workshop, the lead facilitator shared an overview on policy linking, including a brief chronology of the development of the method of policy linking for global reporting on SDG Indicator 4.1.1.

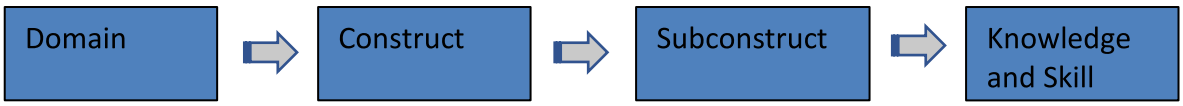
The panelists were then introduced to the GPF related to Reading for Grade 3, specifically focusing on the content standards given in Table 3 of GPF and performance standards presented in Table 5 of GPF. The panelists were provided with a translated GPF document in the language represented by them in order to aid deeper understanding on the subject.

> *"For the first time we are getting to know about global standards of learning through GPF. I would like to learn more about this, I can use it for monitoring learning progress of children in my classroom."*
>
> **- Response from a Panelist**

Content standards: Is the **WHAT** of the content that learners are expected to know and be able to do with relation to reading in grade 3. This is indicated by the knowledge or skills in the GPF (Table 3).

Table 3 connects each content standard to (a) Domain (b) Construct (c) Sub construct. Refer to Picture 1 for a snapshot from Table 3 from the main GPF document.

Figure 2: Content Standards - Table 3 of GPF



Picture 1: Excerpt for Reading from Content Standards - Table 3 of GPF

DOMAIN: R—READING COMPREHENSION (EXCERPT FROM THE READING GPF)											
Construct	Subconstruct	Knowledge or Skill	Grade								
			1	2	3	4	5	6	7	8	9
R2 Interpret information	R2.1 Identify the meaning of unknown words and expressions in a grade-level text	R2.1.1 Identify the meaning of unknown words (including familiar words used in unfamiliar ways) and idiomatic and figurative expressions in a grade-level text			x	x	x	x	x	x	x
	R2.2 Make inferences in a grade-level text	R2.2.1 Make simple inferences in a grade-level text by relating pieces of explicit and/or implicit information in the text				x					
		R2.2.2 Make inferences in a grade-level continuous text by relating pieces of explicit and/or implicit information in the text				x	x	x	x	x	x
		R2.2.3 Make inferences in a grade-level non-continuous text (e.g., tables, diagrams, graphs) by relating pieces of explicit and/or implicit information					x	x	x	x	x
		R2.2.4 Identify the sequence of events/actions/steps in a grade-level text				x	x	x	x	x	x
		R2.2.5 Identify, compare, or contrast points of view in a grade-level text					x	x	x	x	x
		R2.2.6 Identify, compare, or contrast evidence in a grade-level text to support or explain an idea, action, or statement				x	x	x	x	x	x
		R2.2.7 Draw a basic conclusion from a grade-level text by synthesizing information in the text						x	x	x	x
		R2.2.8 Apply information from a grade-level text to a new example or situation									x
	R2.3 Identify the main and secondary ideas in a grade-level text	R2.3.1 Identify the main idea in a grade-level text when it is not explicitly stated			x	x	x	x			
		R2.3.2 Distinguish between a prominent main idea and secondary ideas in a grade-level text						x	x	x	x

Performance standards: This is related to **HOW MUCH** content do learners need to know and be able to do. For each content standard, Table 5(a) students are classified and their know-how into three levels:

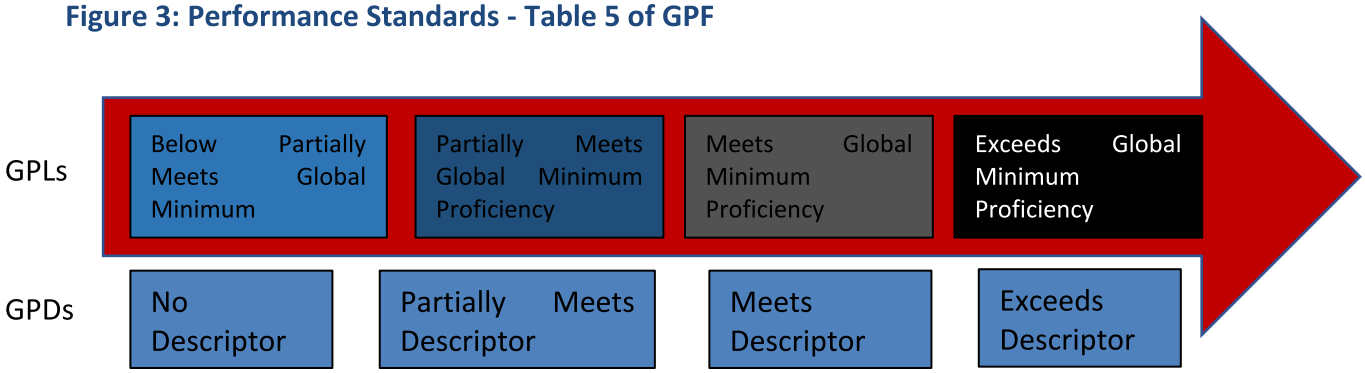
a) Partially meets

b) Meets

c) Exceeds

These levels are called GPL or Global Proficiency Levels. Each GPL has a descriptor - i.e., how much students should know and be able to do. These descriptions are called Global Proficiency Descriptors. For example, a learner who meets global minimum proficiency in the construct of 'interpret information' (Refer Picture 1), should be able to identify the main theme of a grade-level passage.

Figure 3: Performance Standards - Table 5 of GPF



Picture 2: Excerpt for Reading Descriptors from Performance Standards - Table 5 of GPF

GRADE 3: READING – DESCRIPTORS FOR THE THREE HIGHEST GLOBAL MINIMUM PROFICIENCY LEVELS (EXCERPT FROM THE GPF)		
Partially Meets Global Minimum Proficiency	Meets Global Minimum Proficiency	Exceeds Global Minimum Proficiency
R: READING COMPREHENSION		
R2: INTERPRET INFORMATION and R3: REFLECT ON INFORMATION		
R2.2: Make inferences in a grade-level text		
R2.2.1_P Make simple inferences in a grade 3-level text by relating two pieces of explicit information in consecutive sentences when there is no competing information. This will generally be in response to a "why" or "how" question.	R2.2.1_M Make simple inferences in a grade 3-level text by relating two pieces of explicit information in consecutive sentences when there is limited competing information. This will generally be in response to a "why" or "how" question.	R2.2.1_E Make simple inferences in a grade 3-level text by relating two pieces of explicit information in one or more paragraphs when there is more distance between the pieces of information that need to be related and/or a lot of competing information. This will generally be in response to a "why" or "how" question.
R2.3: Identify the main and secondary ideas in a grade-level text		
N/A	**R2.3.1_M** Identify the general topic of a grade 3- level text when it is prominent but not explicitly stated	R2.3.1_E Identify the general topic of a grade 3-level text when it is less prominent and not explicitly stated.
R3.1: Identify the purpose and audience of a text		
N/A	N/A	N/A

The facilitators contextualized the explanations for the performance levels, whenever necessary, to ensure that the panelists had conceptual clarity. For instance, the panelists were guided to group their students according to the different levels of performance based on their performance in a classroom-based test. This enabled the panelists to relate their real-life scenarios to the technical guidance on content standards and performance standards.

Task 1: Alignment

After the initial orientation on GPF and policy linking, the first exercise for the panelists was to check the alignment of the ORF passage in the literacy assessment to the GPF following a three-step process. A sample passage for grade 3 level is presented in Box 2 to help the reader of this document understand the process.

Task 1.1: Individual and independent ratings for each of the assessment items was made by the panelists based on the content standards. They recorded their ratings on an alignment form provided by the facilitators

Step 1: Review of each item was done to identify the knowledge or skill(s) required to answer the item correctly. This included assessing the oral reading fluency of each word on the sample paragraph followed by assessment of knowledge and skills required for answering each of the five comprehension questions (Refer Box 2)

Step 2: Panelists identified how much of the required knowledge or skill(s) overlap with the content standards listed in the GPF (Table 3)

Step 3: Individual and independent judgments were made by the panelists to rate how well the items align with the GPF. The format of the alignment form is presented in annexures.

Box 2: Sample text for oral reading fluency and comprehension for Grade-3

Noga is the smallest girl in her class. She does not like being small.

Her mother tells her not to worry. "It's ok to be small," she says. But Noga does not think it is ok to be small.

One day, Noga is out walking. She hears a chirping sound coming from a small hole in a tree. She crawls into the hole and sees a baby bird. She gently picks up the bird.

Noga crawls out of the hole and gently places the bird on a branch of the tree. The bird chirps happily.

"How lucky that I was walking past here, and not some big kid," Noga thinks. She smiles and walks home. She keeps smiling all the way home.

Ref #	Items	Key
R2.2.1_P	1. Why does Noga pick up the baby bird? A. To hold the bird B. To rescue the bird C. Because she likes the bird D. Because she saw the bird	B. To rescue the bird
R2.3.1_M	2. What is the main idea of this text? A. Being small can be good B. Big kids do not like to help C. Helping animals is lucky D. Mothers can be wrong	A. Being small can be good
R3.1.1_E	3. What is the purpose of this text? A. To provide a clue B. To explain an idea C. To give instructions D. To tell a story	D. To tell a story

Panelists based their ratings on how well the assessment item aligned with the content standards at grade level i.e., at grade 3 and plus one grade level above and one grade level below. They made independent and individual judgments on the degree to which 1) the assessment items aligned with at least one domain (depth) and 2) the sub constructs were covered by at least one assessment item (breadth). The lead facilitators explained the three-point alignment scale to the panelists as follows:

- Complete Fit (C) signified that all the content required to answer the item correctly was contained in the knowledge or skill.

- If an item had a rating of Complete Fit (C) with a knowledge or skill, the panelists should not match it with another knowledge or skill

- Partial Fit (P) signified that part of the content required to answer the item correctly was contained in the knowledge or skill.
    - If an item had a rating of Partial Fit (P) with a knowledge or skill, the panelists should generally match it to one or two other knowledge or skill(s)
- No Fit (N) signified that no amount of the content required to answer the item correctly was contained in the knowledge or skill.
    - If an item had a rating of No Fit (N) with a knowledge or skill, the panelists should not match it to any knowledge or skill.

Task 1.2: Facilitators analyzed the ratings of all panelists and determined the degree of alignment of the test item with the GPF (none, minimal, additional, or strong alignment). The process of analysis and ratings is described below:

- Count complete and partial fit ratings to calculate the level of alignment of the test item and the GPF content standards (Table 3)
- For each panelist, data was calculated for both content depth (number of items has either partial or complete fits in the domains) and content breadth (number of items has either partial or complete fits in the sub constructs)
- Calculate the average of content depth and breadth separately across panelists to estimate the overall level of alignment between the assessment and the GPF
- Compare these averages against the global requirement.

Picture 3: A sample presentation of results after alignment

Level of Alignment		Grade 1	Grade 2	Grade 3	Grades 4-9
Minimally aligned	Domain/ Construct:	D (min 5 items)	D (min 5 items) AC (min 5 items)	RC (min 5 items)	RC (min 5 items)
	Subconstructs:	At least 50% of the D subconstructs	At least 50% of the D and AC subconstructs	At least 50% of the RC subconstructs	At least 50% of the RC subconstructs
Additionally aligned	Domain/ Construct:	D (min 5 items) AC or RC (min 5 items)	RC (min 5 items)	N/A	RC: B1 (min 5 items) RC: B2 (min 5 items)
	Subconstructs:	At least 50% of the D and AC or D and RC subconstructs	At least 50% of the RC subconstructs	N/A	At least 50% of the RC subconstructs
Strongly aligned	Domain/ Construct:	D (min 5 items) AC (min 5 items) RC (min 5 items)	RC: B1 (min 5 items) RC: B2 (min 5 items)	RC: B1 (min 5 items) RC: B2 (min 5 items)	RC: B1 (min 5 items) RC: B2 (min 5 items) RC: B3 (min 5 items)
	Subconstructs:	At least 50% of all subconstructs	At least 50% of the RC subconstructs	At least 50% of the RC subconstructs	At least 50% of the RC subconstructs

Key: D – Decoding; AC – Aural language comprehension; RC – Reading comprehension: B1 – Retrieve information; B2 – Interpret information; B3 – Reflect on information

Task 1.3: Facilitators and panelists discussed the implications of the alignment results. The discussions helped in answering the following questions:

What are the implications of items that align with the GPF content standards?

- Task 2: Matching these items with the GPLs and GPDs (performance standards – Table 5) will be possible since there is relevant content in the GPF
- Task 3: Rating these items to set benchmarks will be straightforward since it depends on the matching

What are the implications for items that do not align with the GPF content standards?

- Task 2: Matching these items with the GPLs and GPDs (performance standards – Table 5) will not be possible since there is no relevant content in the GPF
- Task 3: Rating these items to set benchmarks will be more difficult since there is no matching.

Panelists required considerable one-on-one support in familiarizing themselves with the technical concepts discussed in the workshop. The concept of a content standard being a "Complete Fit (C)" was confusing. The facilitators provided supportive mentoring in building concept clarity. The state coordinators also played an important role in helping their respective language groups understand the concepts better.

Mr. Chinmaya Holla, Facilitator, American Institutes of Research

Task 2: Matching

Building on the previous task of alignment to content standards, the facilitators supported the panelists to match the items with the GPLs and GPDs. The purpose was to: a) increase the panelists' knowledge of the items and GPF and b) improve the identification of the GPLs corresponding to the items, which would increase the accuracy and consistency of the item ratings in the next task of benchmarking. The panelists reviewed each word of the ORF passage and the comprehension questions in their respective language to identify the Level (performance standard) most appropriate for the item. Panelists had discussions in their respective language groups and focused on the following questions:

- What level of knowledge and skill is required to read a word or answer the comprehension question (referred to as 'item') correctly?
- What makes an item easy or difficult?
- What is the lowest level in the Descriptors that is most appropriate for the item?

The panelists noted the sub-construct and the level next to each of the items in the test booklet. If the item matched with more than one sub construct – which was usually the case with partial fit, the panelists wrote the additional subconstruct(s) and level(s) next to the item. The completion of this task was a prerequisite before setting the benchmark which was the next task to be done by the panelists.

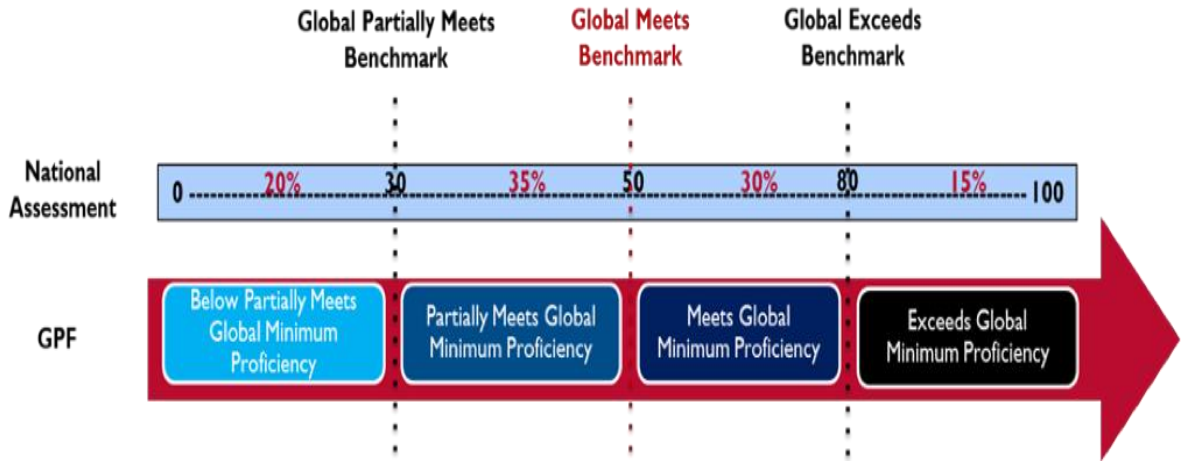
Task 3: Benchmarking

The next two days i.e. day 3 and 4, were dedicated for the panelists to undertake the benchmarking exercise, which was done in two rounds. The facilitators oriented the panelists on implementing the Yes-No variation of the Angoff method to set benchmarks that would align with global standards. The lead facilitators showed the panelists how the benchmarking method would link the data from the national assessment to the GPF. The four ratings, Just Partially Meets (JP), Just Meets (JM), Just Exceeds (JE), and Above Exceeds (AE), for categorizing learners were explained.

Picture 4 below illustrates the policy linking process through an example resulting in 3 benchmarks of partially meets = 30, meets = 50, and exceeds = 80 on a scale of 0 to 100. These 3 benchmarks created four GPLs with the following score ranges: below partially meets = 0-29, partially meets = 30-49, meets = 50-79, and exceeds = 80-100. The benchmarks and score ranges were applied to the assessment data to calculate the

percentages of learners in each GPL: below partially meets = 20%, partially meets = 35%, meets = 30%, and exceeds = 15%.

Picture 4: Example of link between assessment data and GPF



First, the panelists participated in a practice session on conducting item ratings using the features of the Angoff benchmarking method. For ORF, to establish benchmarks, panelists read each word/item individually and independently and then decided whether minimally proficient learners at each performance level (JP, JM, and JE) would be able read the word accurately or answer the question correctly. Each word in the paragraph was rated based on four expectations given below, i.e., chances of whether a minimally proficient learner would read the word accurately. The same process was also applied to reading comprehension questions.

1. Probably not ("no");
2. Somewhat possible ("no");
3. Reasonably sure or ≥ 67% chance ("yes"); and
4. Absolutely positive ("yes")

The number of 'yes' responses by Level were summed and aggregated to yield an individual panelist's benchmark. The benchmarks from all panelists for the particular language were then averaged to determine the panel's benchmarks.

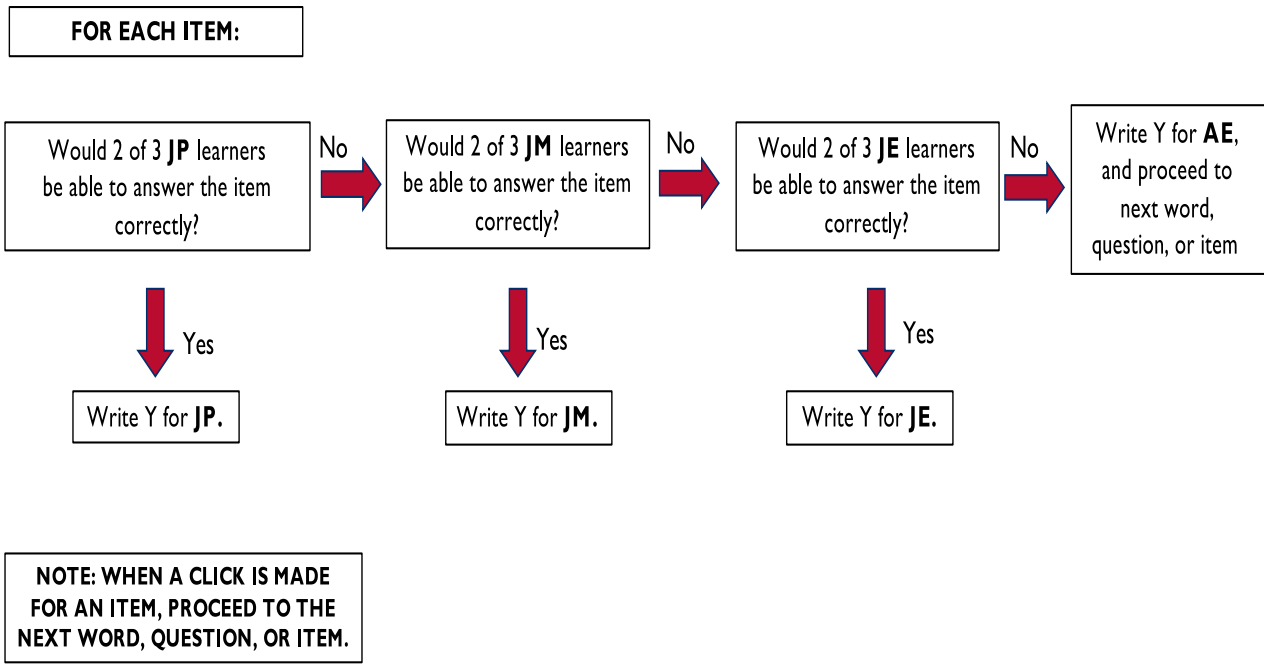
Picture 5: Features of the Angoff Method

Between Below Partially Meets and Partially Meets (**Partially Meets Benchmark**) → At or slightly above Partially Meets (Just Partially Meets or **JP Student**)

Between Partially Meets and Meets (**Meets Benchmark**) → At or slightly above Meets (Just Meets or **JM Student**)

Between Meets and Exceeds (**Exceeds Benchmark**) → At or slightly above Exceeds (Just Exceeds or **JE Student**)

Figure 4: Steps in Benchmarking Process



Panelists found two ideas challenging while benchmarking in the first round - 1) the idea that for each item, it was sufficient to mark a "Yes" for only the lowest performance level that would be able to answer that item correctly, and 2) the idea that they had to account for the time restriction that was implemented while the student was reading the passage. Multiple panelists marked a "Yes" in all the performance levels that they judged would answer it correctly. Panelists, sometimes, also marked a "Yes" for those performance levels that they had earlier been judged could not be read in in the passage in the 60 seconds time limit.

Mr. Chinmaya Holla, Facilitator, American Institutes of Research

After conducting the first round of ratings for each of the items, the lead facilitators did the following:

i. Compiled the ratings for each panelist to calculate their initial benchmarks
ii. Entered the panelists' benchmark data into the spreadsheets
iii. Calculated the initial benchmarks for the panels by averaging the benchmarks across the panelists, and
iv. Produced summaries of the benchmarks.

The fourth day of the workshop started with the facilitators sharing the first-round ratings with the panelists, before they undertook the second round of benchmarking based on a discussion on the results of the first round. The Round 1 summary was presented to the panelists, for each language including the following:

1. Initial benchmarks of each panelist, with each panelist only indicated by a code number in order to conceal individual identity of the panelists to ensure objectivity
2. Impact data with percentages of scores in the Levels based on the score distributions, and
3. Statistics showing the quality of the ratings, including inter-rater consistency and standard errors.

The facilitators discussed these initial ratings, pointing out the minimum and maximum scores given by panelists. The results presented from round 1 also included an analysis of how panelists had rated the level of difficulty of each word and the number of words that were read accurately. This helped the panelists to review their individual scores against the average score of the language group. After this review of the initial benchmarks and feedback, the panelists re-grouped into their language groups and

revisited their individual item ratings. It was left to the panelists to revisit their round 1 scores using the same steps as in the first round.

The guidance for the second round was that the panelists should a) focus on item content in relation to the descriptions of the knowledge and skills, b) consider what learners would be able to do given any issues related to the measurement error (for instance the factors in the learner's environment), and d) make adjustment to the ratings based on their individual judgments. After each panelist submitted their second round of responses, the data was further analyzed to arrive at the revised benchmark.

"Panelists had an opportunity to update their judgements from round 1 of benchmarking, based on analysis of the FLS data that the technical agency presented to them. This included summary statistics such as the percentage of students that correctly answered each item in a particular language as well as the percentages of students in that language categorized into performance standards based on the panelists' benchmarks in round 1.

Panelists in updated their benchmarks considerably between rounds 1 and 2. Conversations with the panelists, anecdotally, suggests that the information provided to them between rounds 1 and 2 had a role to play in them updating and perfecting their responses"

Mr. Chinmaya Holla, Workshop Facilitator, American Institutes of Research

After completing the second and final round of scoring, the process of benchmarking was completed. Panelists were informed that the data from the second round of scoring would inform the benchmark for ORF for their respective language. Since the data required further analysis the results from this round were not presented.

Panelists were requested to share their feedback on the workshop using a feedback form that was shared with them.

Conversations with panelists, during and after the workshop, indicated that they had found the workshop useful in building their skills. They also gave suggestions for helping them to participate better. For instance, panelists suggested that the workshop presentations should be interspersed with activities for them to practice a particular skill or discuss a certain concept within their language groups. Specifically, they requested for such an activity-based format on days where they were expected to work individually.

3.3. Post Workshop Activities

Additional Round 2 Analysis and Benchmark Approval

After the workshop, AIR conducted additional analysis to compute a range of ORF benchmarks and corresponding impact data through adjustments of standard errors, and then discussed it with NCERT. NCERT selected the ORF benchmarks that were deemed meaningful and realistic for each language and were consistent across languages.

Department of School Education & Literacy, Ministry of Education, approved the ORF benchmarks that NCERT had recommended. As part of the benchmarking process, this stage was considered a policy stage.

After ORF benchmarks were finalized, AIR estimated the number of reading comprehension questions the learners read at or above the benchmarks through descriptive statistics and simple regression analysis of ORF and reading comprehension scores. The reading comprehension scores were used as dependent variable and ORF scores as independent variable.

4. Benchmarking numeracy in India

The five day workshop for Benchmarking in Numeracy was held at the NCERT Campus, New Delhi, India from August 22-26, 2022 with participation from 66 panelists who came from India's 34 states and UTs (excluding Goa and Nagaland). The workshop was a collaborative effort between MOE, NCERT, AIR and UNICEF (Refer Table 1).

4.1. Preparatory Phase

Selection of panelists

The panelists included master teachers and pedagogy specialists from each state and territory with expertise in grade 3 numeracy. NCERT invited two panelists from each state and union territory for the benchmarking workshop. Selection of the panelists was based the below mentioned qualifications:

MASTER TEACHERS. At least five years of teaching experience at or adjacent to the relevant grade level, strong skills in mathematics/numeracy, native skills in the language of instruction and assessment, experience with students at different proficiency levels, knowledge of the instructional system and materials, and a teacher's college and/or university certification and licensing.

PEDAGOGY SPECIALISTS. At least five years of teaching experience at the primary school level, strong skills in mathematics/numeracy, native skills in the language of instruction and assessment, experience with learners at different proficiency levels, knowledge of the instructional system and materials, and teacher's college and/or university certification and licensing.

Following the guidelines of the policy-linking method, the composition of the panels had three characteristics

1. Each panel had at least 15 panelists.
2. Teachers comprised at least 70 percent of the panel.
3. Selection of teachers was representative, i.e., gender balanced from a variety of schools and geographical areas.

Additionally, one facilitator (with strong assessment and psychometric expertise) and data analysts (with statistics and data processing background) from NCERT, supported

the lead facilitators from AIR during the presentations and discussions with the panelists and in data entry respectively.

Pre-workshop exercise

Prior to coming for the benchmarking workshop, all panelists were required to administer the oral reading fluency assessment and the related comprehension questions, to at least nine children in their school. The primary objective of this was for panelists to develop an understanding of students reading fluency, as the experience would help them gain the knowledge and skills that were later applied in the workshop for setting benchmarks. For this, the panelists were provided a sample paragraph in the language known to them. Detailed instructions were shared with the panelists for administering the test item.

Materials used for the Workshop

1. FLS assessment booklet
2. Policy Linking Toolkit (PLT)
3. Global Proficiency Framework (GPF) – Numeracy for grade 3
4. Facilitation Slides
5. Rating Forms
6. Data Entry Templates
7. Workshop Evaluation Form

Agenda for the Workshop

The agenda for the 5-day policy linking workshop is presented below.

DAY	AGENDA
Day 1	Registration and opening speeches Task 1 Presentation: Assessments, GPF, and alignment Task 1 Activity: Align assessments and GPF
Day 2	Task 1 Presentation: Results from the alignment Task 2 Presentation: Assessments, GPF, and matching Task 2 Activity: Match assessments and GPF (part 1)

Day 3	Task 2 Activity: Match assessments and GPF (part 2) Task 3 Presentation: Global benchmarking
Day 4	Task 3 Presentation: Angoff benchmarking method Task 3 Activity: Conduct Round 1 rating
Day 5	Task 3 Presentation: Results from Round 1 rating Evaluation Task 3 Activity: Conduct Round 2 ratings Certificates and closing speeches

4.2. Implementation Phase

The workshop followed a standardized process with three key tasks:

1. Checking the alignment of the assessments and the GPF using the Frisbie (2003) method.
2. Matching the items with the GPLs and global proficiency descriptors (GPDs) in the GPF, and
3. Setting the benchmarks using the Angoff method (Angoff, 1971; Plake, Ferdous, & Buckendahl, 2005; Ferdous, 2019; Ferdous, Davis, & Kelly, 2019).

Each task is explained below for global benchmarking on numeracy.

Task 1: Checking the Alignment of the Assessments and the GPF (Days 1-2)

After receiving orientation on GPF, the master teachers and pedagogy specialists i.e. the panelists checked the alignment of the assessments with the content standards in the GPF Table 3 (Refer to Figure 2). The key domains covered in mathematics for Grade 3 are:

N = Number and operations

M = Measurement

G = Geometry

S = Statistics and probability

A = Algebra

They made individual and independent ratings of the fit (complete, partial, and no fit) between the items and content standards (Table 3 in the GPF). The facilitators compiled and analyzed the ratings to determine the degree to which the assessments align (or do not align) with the GPF.

Picture 6: Excerpt for Numeracy from Content Standards - Table 3 of GPF

Construct	Subconstruct	Knowledge or Skill	Grade								
			1	2	3	4	5	6	7	8	9
N1 Whole numbers	N1.1 Identify and count in whole numbers, and identify their relative magnitude	**N1.1.1** Count, read, and write whole numbers	x	x	x	x	x	x			
		N1.1.2 Compare and order whole numbers	x	x	x	x	x	x			
	N1.2 Represent whole numbers in equivalent ways	**N1.2.1** Determine or identify the equivalency between whole numbers represented as objects, pictures, and numerals	x	x	x						
		N1.2.2 Use place-value concepts		x	x	x	x	x			
	N1.3 Solve operations using whole numbers	**N1.3.1** Add and subtract whole numbers	x	x	x	x	x	x			
		N1.3.2 Find the double or half of a set of objects	x	x							
		N1.3.3 Multiply and divide whole numbers			x	x	x	x			
		N1.3.4 Demonstrate fluency with basic addition and subtraction facts			x						
		N1.3.5 Demonstrate fluency with basic multiplication and division facts				x					
		N1.3.6 Identify factors and multiples of whole numbers						x			

All the numeracy subtasks were included for setting the benchmarks to ensure a higher degree of alignment with GPF as well as provide strong evidence of what grade 3 students in India know and can do in relation to global minimum proficiency.

NCERT used Booklet #3 for the benchmarking workshop. Since the numeracy section appears first in Booklet #3, followed by the literacy section, it was easier for panelists to navigate subtasks and items. NCERT ensured that panelists receive the translated or adapted version of the booklet in their respective local languages. However, the translated or adapted version of the GPF document was not necessary.

The FLS Numeracy Booklet had 11 questions as follows:

Question	Domain	Learning Outcome	Description	No. of Items
Q1	Number & Operations	IL01	Read aloud whole numbers presented in a grid	24
Q2	Number & Operations	IL01	Compare pairs of whole numbers to identify the larger number	14
Q3	Number & Operations	IL02	Add and subtract whole numbers in a grid	8
Q4	Number & Operations	IL02	Solve word problems using addition and subtraction	6

Q5a	Number & Operations	IL03	Multiply whole numbers in a grid	4
Q5b	Number & Operations	IL03	Solve word problems using multiplication and division	4
Q6a	Measurement	IL04	Read time on a clock in hours and half-hours	3
Q6b	Measurement	IL04	Measure length in standard units (cm)	3
Q7	Number & Operations	IL05	Identify half, one-fourth, and three-fourths of a whole number	6
Q8	Algebra	IL06	Identify, extend, and communicate rules for simple patterns	8
Q9	Data Handling	IL07	Solve problems involving data displays (pictograph)	6
			Total Score Points (Items or Marks)	86

A demonstrative example of the alignment process with rating form is presented in the table below for Complete Fit, Partial Fit and No Fit.

Ref #	Item	Key	Type of Fit
N1.1.1	1. What is two hundred and seventy-four written in standard form? A. 204 B. 247 C. 270 D. 274	D. 274	Domain: N Number and operations Construct: N1 Whole numbers Subconstruct: N1.1 Identify and count in whole numbers, and identify their relative magnitude Knowledge or skill: N1.1.1 Count, read, and write whole numbers (Grades 1-6) **Alignment: Complete fit**
N1.1.2 **N1.3.1**	2. What is the largest sum? A. 21 + 39 B. 22 + 37 C. 23 + 38 D. 24 + 36	C. 23 + 38	Domain: N Number and operations Construct: N1 Whole numbers Subconstruct: N1.1 Identify and count in whole numbers, and identify their relative magnitude Knowledge or skill: N1.1.2 Compare and order whole numbers (Grades 1-6) **Alignment: Partial fit** Domain: N Number and operations Construct: N1 Whole numbers Subconstruct: N1.3 Solve operations using whole numbers Knowledge or skill: N1.3.1 Add and subtract whole numbers (Grades 1-6) **Alignment: Partial fit**

N1.3.6	3. What are the factors of 6? A. 1, 2 B. 1, 2, 3 C. 1, 2, 3, 6 D. 1, 2, 3, 6, 12	C. 1, 2, 3, 6	Domain: Number and operations Construct: N1 Whole numbers Subconstruct: N1.3 Solve operations using whole numbers Knowledge or skill: N1.3.6 Identify factors and multiples of whole numbers (Grade 6) **Alignment: No fit**

Picture 7: Sample Rating Form

Panelist ID: _____

Policy Linking Workshop, New Delhi, India

Foundational Learning Study (FLS) – Grade 3 Numeracy

Alignment Rating Form

Item No.	Knowledge/Skill Reference No. (Table 3)	Fit Rating (C = Complete; P = Partial; N = No)	Item No.	Knowledge/Skill Reference No. (Table 3)	Fit Rating (C = Complete; P = Partial; N = No)
Q1	N1.1.1	C	Q16		
Q2	N1.1.2 N1.3.1	P	Q17		
Q3	N1.3.6	N	Q18		
Etc.			Etc.		

After collecting all the rating forms, facilitators compiled and analyzed the alignment ratings in the following manner:

- Count complete and partial fit ratings to calculate the level of alignment between the assessment and the GPF content standards (Table 3)
- For each panelist, calculated both content depth (number of items in the domains with either Complete or Partial fit) and content breadth (number of subconstructs covered with at least one item with either Complete or Partial fit)
- Calculated the average of content depth and content breadth across panelists to estimate the overall level of alignment between the assessment and the GPF
- Compared these averages for content depth and content breadth against the global requirements. The global community has developed the criteria for degrees of

alignment between mathematics/numeracy assessments and the GPF. There are three alignment levels (minimally, adequately, and strongly) based on two criteria (depth for domains and breadth for subconstructs):

Alignment Level	Criteria
Minimally Aligned	**Domain** (depth): Number and Operations (minimum five items) **Subconstructs** (breadth): Items covering at least 50 percent of the Number and Operations subconstructs
Adequately Aligned	**Domain** (depth): Number and Operations (minimum 5 items) and Measurement and Geometry (minimum 5 items) **Subconstructs** (breadth): Items covering at least 50 percent of the Number and Operations, Measurement, and Geometry subconstructs
Strongly Aligned	**Domain** (depth): Number and Operations (minimum five items) and Measurement and Geometry (minimum five items) and Statistics and Probability and Algebra (minimum five items) **Subconstructs** (breadth): Items covering at least 50 percent of all subconstructs

The implications of the alignment exercise on Task 2 (matching) and Task 3 (benchmarking) were then mapped over two scenarios.

- In the first scenario items aligned with the GPF, therefore matching these items with the GPLs and GPDs (Table 5) was possible in Task 2 since there is relevant content in the GPF. Rating these items to set benchmarks in Task 3 was also relatively straightforward since it depended on the matching.
- In scenario 2, items did not align for Tasks 2 and 3. Matching these items with the GPLs and GPDs (Table 5) was not possible since there was no relevant content in the GPF and rating these items to set benchmarks was more difficult since there was no matching.

This is further elucidated in sections below.

Task 2: Matching assessment items with GPLs & GPDs (Days 2-3)

In the matching exercise on Day 2, the panelists categorised the knowledge and skills with Global Minimum Proficiency Levels i.e. the minimum knowledge and skills required for a partially meets learners, meets learner and exceeds learner as defined in Table 5 of GPF. After orientation, the panelists were divided into sub-panels. They build on their knowledge gained in the alignment activity to reach a consensus on matching each assessment item/question (within their subtasks) with the performance standards (Table 5 in the GPF).

Picture 8: Excerpt for Numeracy Descriptors from Performance Standards - Table 5 of GPF

GRADE 3: MATHEMATICS – DESCRIPTORS FOR THE THREE HIGHEST GLOBAL MINIMUM PROFICIENCY LEVELS (EXCERPT FROM THE GPF)					
Partially Meets Global Minimum Proficiency		Meets Global Minimum Proficiency		Exceeds Global Minimum Proficiency	
N: NUMBER AND OPERATIONS					
N1: WHOLE NUMBERS					
N1.1: Identify and count in whole numbers, and identify their relative magnitude					
N1.1.1a_P	Count in whole numbers up to 100.	N1.1.1a_M	Count in whole numbers up to 1,000.	N1.1.1a_E	Count in whole numbers up to 10,000.
N1.1.1b_P	Read and write whole numbers up to 100 in words and numerals.	N1.1.1b_M	Read and write whole numbers up to 1,000 in words and numerals.	N1.1.1b_E	Read and write whole numbers up to 10,000 in words and numerals.
N1.1.2_P	Compare and order whole numbers up to 100.	N1.1.2_M	Compare and order whole numbers up to 1,000.	N1.1.2_E	Compare and order whole numbers up to 10,000.
N1.3: Solve operations using whole numbers					
N1.3.1_P	Add and subtract within 100 (i.e., where the sum or minuend does not surpass 100) and without regrouping	N1.3.1_M	N/A	N1.3.1_E	Add and subtract within 1,000 (i.e., where the sum or minuend does not surpass 1,000) and with and without regrouping

"Complete, consistent and clear understanding of the panelists on the knowledge and skills required to answer the assessment questions/item and where it matches with the Table 5 of the GPF i.e. whether a partially meets student or a meets student or a exceeds student will be able to answer the question is required.

The 86 test questions/items in the numeracy assessment are spread over different knowledge skills as wells as GPLs. A question may at times require two skills, say addition and comparison are required to answer the question. So then the panelists will have to look for both separately in Table 5 and then decide what level of student can answer it."

Dr. Abdullah Ferdous, Lead Facilitator, American Institutes of Research

"Drawing the line between knowledge and skill of a below partially meets student and a partially meets student is often difficult and subjective. The boundaries are difficult to set for each category till we don't clearly understand the knowledge and skill required for each question"

Anectodal evidence from a panelist

In the end of the matching exercise, the panelists presented their results to all panelists, who wrote this information (reference codes) on their assessment tools and copies of the GPFs.

Task 3: Setting the Benchmarks (Days 3-5)

The Angoff benchmarking method as described in the policy linking toolkit (PLT) was introduced to the panelists. The method has been in practice for over 50 years and is one of the most popular and well-accepted methods (globally) for setting benchmarks on assessments. It was selected by the global community and approved by the UN for benchmarking with policy linking.

The Angoff method adjusts the minimum proficiency levels to borderline levels. Under the method, based on teaching experience, the panelists were asked to conceptualize three students who would fall on the border of being classified as just partially meets, just meets and just exceeds. These were minimally proficient students in the 3 GPLs (Refer Picture 7).

Round 1 Benchmarking

To establish benchmarks, each panelist read each question individually and independently and decided whether minimally proficient students at each performance level (JP, JM, and JE) would be able answer the question correctly.

However, benchmarking has some steps that are different from matching. The key difference is that matching is based on "should" for general expectations while benchmarking is based on "would" for realistic expectations.

- Should refers to judgments based only on the content and performance standards
- Would is based on those judgments plus assessment conditions, e.g., difficulty of an item for a student, testing conditions, student anxiety, and random errors.

"Benchmarking diversifies the thinking process to see from the eyes of the student and understand their perspective. There is no set standard for benchmarking. Benchmarks are REVEALED. And they are revealed when individual benchmarks of the panelists and group averages show consistency of thought and pattern while making judgements/decisions in the benchmark setting process. It's just that they always have to ask the question in 'would' and not 'should', coz 'should' implies curriculum expectations and not performance expectations"

Dr. Abdullah Ferdous, Lead Facilitator, American Institutes of Research

The panelists answered each question – "would a learner in the JPM/JM/JE level be able to answer this question correctly?" based on four realistic expectations as given below:

1. Probably not = No
2. Reasonably sure or 2 out of 3 (67%) = Yes
3. Somewhat possible = No
4. Absolutely positive = Yes

Question by Panelist: What if I identify one JP who can answer a question/item and my partner decides that two JPs can answer the same question. The decision making on rating will be stuck and cant put a 'yes' in JP?

Answer by Facilitator: Table 5 is the foundation of this method and need to be kept in mind at all times of decision making while giving ratings. This will reduce the variation in identifying the JP, JM and JE learners amongst panelists as then the panelists will have identified similar knowledge and skills for each GPL.

Another factor is to understand the face difficulty i.e. the cognitive understanding on how easy or difficult a question will be for a JP, JM or JE learner. For instance 3+5 is easy but 39+28 may be difficult. So panelist need to reiterate and ask themselves if 2 out 3 i.e. can 67% of the learners in the GPL category can answer this question or not. Only then it is a YES.

Question by Panelist: My classroom has only one learner whom I can conceptualise at JE, so how do I apply the question that 2 out 3 learners can answer?

Answer by Facilitator: The concept of conceptualisation is not limited to the current classroom. It is rather based on the teachers years of experience. The teacher can think of potential JE students from previous years and form a generalisation.

Observations from Q&A session

When making the ratings, the panelists proceeded from the lowest level (JP) to the highest level (JE or AE). So if the answer is Yes (Y) in JP, then the rating column for JM and JE was left blank as it was assumed that what a JP learner can do, the same can invariably be done/performed by a JM or a JE student. If the answer was No (N) in JP, then the panelists ascertained 'who can do it? – JM or JE?'. If JM was marked Y, then column in JE was left blank and so on and so forth (Refer to Picture 10). There was always only one YES to a question.

The benchmark as set by each panelist was calculated as follows:

- Sum of YES in JP column = Partially Meets Benchmark
- JP Benchmark + Sum of YES in JM column = Meets Benchmark
- JM Benchmark + Sum of YES in JE column = Exceeds Benchmark

Picture 9: Sample rating form for benchmarking filled by each panelist

	ROUND 1				ROUND 2			
Item #	Would two of the three **Just Partially Meets (JP)** students answer this item correctly? (**Reasonably Sure**)	Would two of the three **Just Meets (JM)** students answer this item correctly? (**Reasonably Sure**)	Would two of the three **Just Exceeds (JE)** students answer this item correctly? (**Reasonably Sure**)	If two of three **JE** students would not answer this item correctly, then **Above Just Exceeds** (**AE**) would	Would two of the three **Just Partially Meets (JP)** students answer this item correctly? (**Reasonably Sure**)	Would two of the three **Just Meets (JM)** students answer this item correctly? (**Reasonably Sure**)	Would two of the three **Just Exceeds (JE)** students answer this item correctly? (**Reasonably Sure**)	If two of three **JE** students would not answer this item correctly, then **Above Just Exceeds** (**AE**) would
Q1-01	Yes							
Q1-02	No	Yes						
Q1-03	No							
Q1-04								
Q1-05								
Q1-06								
Q1-07								

Picture 10: Example of calculation of benchmarks

Item	Panelist 1				Panelist 2				Panelist 3				Panel (Average)
	JP	JM	JE	AE	JP	JM	JE	AE	JP	JM	JE	AE	
1		Y					Y			Y			
2	Y					Y				Y			
3			Y				Y				Y		
4			Y		Y						Y		
5		Y			Y							Y	
6				Y		Y						Y	
SUM	**1**	**2**	**2**	**1**	**2**	**2**	**2**	**0**	**0**	**2**	**2**	**2**	
Partially Meets	**1**				**2**				**0**				**(1+2+0)/3 = 1**
Meets		**1+2 = 3**				**2+2 = 4**				**0+2 = 2**			**(3+4+2)/3 = 3**
Exceeds			**1+2+2 = 5**				**2+2+2 = 6**				**0+2+2 = 4**		**(5+6+4)/3 = 5**

Picture 11: Example of how the benchmarking results are tabulated and presented globally

Global Proficiency Level	Global Benchmark	Score Range (40 points maximum)	Percentage of Students
Below Partially Meets	N/A	0	%
Partially Meets	1	1-2	%
Meets	3	3-4	%
Exceeds	5	5-6	%
Total			100%

Table 2: Steps of Benchmarking for Round 1

Test	Step 1	Step 2	Step 3
Questions	Conceptualize minimally proficient students	Understand the knowledge or skill required to answer the question correctly Consider testing conditions	Provide ratings for each question on the test

Each panelist submitted their rating forms to the facilitators for data entry and processing at the end of Round 1. The facilitators and data entry specialist compiled the ratings of all panelists under JP, JM and JE and calculated an average in each level to find a group benchmark in all three levels (Refer Picture 9, 10, 11).

Round 2 Benchmarking

The Round 1 ratings were presented on Day 4 for further contemplation and revision in Round 2 with the help of additional information and insights. The additional information in Round 2 included the following:

1. **Location statistics:** Facilitators showed a compiled table and a graph of ratings of all panelists to help draw comparisons and understand the differences. This graph is called location statistics.

   The panelists identified the lowest rating and highest rating in JP, JM and JE. For instance, say the lowest rating given in JP was 29 and highest rating was 49. So the variation between 29 and 49 highlights the level of matching with Table 5 i.e. the knowledge and skill of a JP learner as marked by the panelists. This helped the panelists to make a self-assessment/reassessment of their matching vis a vis other panelists' perceptions on the knowledge and skills required for the GPL . Thereby

the consistency in understanding of knowledge and skills between panelists on the GPLs was reviewed/adjusted/improved in Round 2. This was the key reason that the data analysis of individual scores/ratings of Round 1 was important as the graph of location statistics helped understand the dispersion between GPLs and between different panelists across GPLs. Standard deviation was often reduced from Round 1 to Round 2.

Facilitators were also able to better assess the extent of (a) lack of understanding of Table 5 and GPF (b) lack of understanding of difficulty of questions amongst panelists and thereby plug gaps before the final benchmarking in Round 2.

Question by Panelist: Does Round 2 implies that my freedom to do ratings will now be restrained and I will have to abide and move with what the group perceives?

Answer by Facilitator: It is not about "loss of freedom" but rather about reassessing and reviewing how each of the panelists matched knowledge and skills with GPLs.

Question by Panelist: My school has its own set of problems. The students we get are mostly from fishing community and migrating for more than half of the year in search of work, so we do not have many JE's to conceptualise. These students have their own set of learning difficulties as well.

Answer by Facilitator: The overall conceptualisation needs to be widened to reduce the error that might emanate from local and contextual factors. In order to improve the ratings in Round 2, the panelists need to conceptualise a more wider group of students in JP, JM and JE so that the assessment is more pliable generally rather than. Having contextual observations

Observations from Q&A session

1. **Data from FLS Study:** The facilitators shared the actual results from the FLS on numeracy on each of the 86 questions and how the grade 3 learners actually responded to the questions. The Round 1 benchmarking scores, when applied to actual results from FLS gave a clear purview of realistic the benchmarks were set. This helped the panelists to quickly grasp the fact that their benchmarks may need improvement. This sharing of actual data provided evidence to trigger a cycle of reassessment in Round 2 based on factual evidence.
2. **Empirical Item Difficulty Scores:** In Round 1, the panelists were asked to make subjective judgements on an items difficulty level. In Round 2, this was fine-tuned

by provision of a pre-calculated list of item difficulty scores. The scores were calculated based on the number of students who answered the item correctly in FLS. So if the item difficulty score was nearer to one, it implied that the question was easy. For instance, a score of 0.98 implied a JP score. If an item had a difficulty score of 0.50 that it showed that the item would either be in JE or above exceeds category. Provision of these scores reduced judgement errors and replaced it with empirical evidence for improved benchmarking in Round 2.

3. **Inter-rater consistency between the rounds:** Based on the data analysis from Round 1, the facilitators calculated the inter-rater consistency between panelists i.e. the degree of agreement between different panelists. This helped in knowing if the panel was close to retrieving the truth or not? If the value of inter-rate index was less than one it implied that all panelists were closer in their perceptions/judgements and thereby were maintaining consistency in benchmark ratings and viceversa.

Facilitators calculated a total consistency figure for JP, JM and JE. For instance, the inter-rater consistency for Round 1 was 0.87. Although the value was less than one and acceptable but with correction of standard errors this rating could be further improved in Round 2.

After the panelists were equipped with this additional information, Round 2 of benchmarking began with independent work done by each panelists on re-setting their benchmarks from Round 1. It was common to see a trend of increase in percentage of students categorized in below partially meets and partially meets after Round 2 and a reduction of percentage of students in meets and exceeds category from Round 1 to Round 2.

"Motivation of panelists is a key factor which impacts the benchmarking process. It is important that each panelist understands the crucial role that they play as the selected representative from their state. They must realize that their judgements, their work, their ratings, their decisions on benchmarking will impact the children and the various stakeholders who contribute to their education in multiple ways and at multiple levels. It will set a precedence for their country. Because of their decisions today, a child will be classified as a JP or a JM or a JE student. Wrong ratings will lead to poor policy choices. So such is the weight of each panelists' motivation level"

Dr. Abdullah Ferdous, Lead Facilitator, American Institutes of Research

4.3. Post Workshop Activities

Additional Round 2 Analysis and Benchmark Approval

After the 5-day workshop was completed, AIR calculated the final benchmarks and conducted additional analysis (with error adjustments) along with statistics to check the reliability of the panelists' ratings. This was followed by a presentation to NCERT, MOE, and UNICEF with the recommended benchmarks, score ranges, and percentages of students in the 4 GPLs. An example is shown below for a numeracy assessment with 70 total score points and a scale of 0 to 70.

Global Minimum Proficiency Levels	Below Partially Meets GMP	Partially Meets GMP	Meets GMP	Exceeds GMP
Benchmarks	--	25	40	55
Score Ranges	0 - 24	25-39	40-54	55-70
Percentages (National)	25%	30%	25%	20%

5. Emerging observations and way forward

Based on informal conversations with participants, suggestions and feedback was gathered on future workshops.

5.1. Controllable factors which can be addressed to improve benchmarking

Accurate translations of the Global Proficiency Framework document and the assessment passage for ORF and reading comprehension to 20 different languages was challenging but imperative to an effective benchmarking process. The precision of translations impacted (i) the overall understanding of GPF by panelists, (ii) execution of the reading assessments as this led to lack of standardization of the assessment properties like number of words, passage length, etc (iii) matching of knowledge and skills with standards. Thus a system of multi-layer review of translations before finalization needs to set to reduce errors emanating from subjective understanding of the translators.

The lead facilitators observed that the **selection of state coordinators** play a role in an effective benchmarking process. Strong leaders were able to provide stringer supportive supervision to their respective language groups. Knowledge of local languages helped them to act as effective translators during the workshop. Good grip on subject matter and strong comprehension skills helped in bridging knowledge gaps with the panelists and eased the process of learning.

Selection of teachers is the key to an efficient and effective benchmarking process. A standard process needs to be developed for teacher selection as the benchmarking process gets engrained in India as a practice. Some of the aspects that need to be considered during the process of teacher selection are:

- Certain degree of minimum proficiency/familiarity and comfort with language of instruction used at the benchmark setting workshop. This helps in easier immersion of the panelist and reduces the need for translations.
- Level of comprehension of the teachers: This is required as the GPF document and benchmarking process makes use of several jargons which require understanding

to participate in the process and assimilate new learning. For instance, teachers often times were not familiar or could not comprehend the use of words like – explicit/implicit, inference etc.

- Teacher capacity in understanding the model of assessing learning outcomes for learners. The facilitators observed that there were instances when teachers could not identify the grade appropriate knowledge and skills of a learner. This posed a challenge in understanding Table 3 on content standards. Future in-service teachers trainings also needs to address this gap in capacity.

Pre-workshop orientation of State Coordinators is required so that they can streamline the process of selection of teachers in accordance to the knowledge and skill requirements. Simultaneously this will also help coordinators to orient the selected teachers before the benchmarking workshop on the content. State coordinators suggested that a blended-mode can be adopted where the orientation can be conducted online through virtual mode followed by on-site benchmarking.

Language groups with multi-state involvement required team-building so that all panelists could develop a sense of trust each on other's judgments and consensus could be built as it was key to arrive at benchmarks that would be acceptable across states. Many times, the same language has sub-dialects which are used in different states, where the spelling may vary slightly or the usage of the word, its tone, pronunciation may differ. For instance, Bengali is the medium of instruction in Assam, Tripura and West Bengal, yet there are some variations in its use. In such situations the panelists need to work as a cohesive team to achieve the common goal of setting benchmarks for Bengali that would be relevant and acceptable to all three states.

Since the process of benchmarking is highly interactive and based on long durations of group work, the **workshop venue and facilities play an important role** in the delivery of the exercise. Group work for different language-groups which runs simultaneously at the same time, calls for sufficient physical space such that effective discussion can be facilitated in a free and comfortable environment. The facilitators observed that venues which provided designated rooms for group-work were more effective in outputs. Room set ups in auditorium/theatre style, classroom style or boardroom styles are less enabling for groupwork. Banquet style or crescent round seating enabled better group work and interaction between participants and with the facilitator.

Since the workshop had high number of participants in a single room, the aids for training also played an important role. Large screen size for presentation was an enabler. Several

screens placed across the venue helped in keeping the participants engaged during training. Availability of additional tools like good audio systems, writing boards etc. also facilitate in effective imparting of training.

Add logistics issues as mentioned in participants evaluation form to corroborate the above.

Selection of passage for reading comprehension needs to be relevant across all languages. The facilitators received valuable feedback from teacher interactions during the benchmarking process that the sample passage often has words which may not be relevant in a particular language. For instance, 'swan' as a word is not familiar to students in Mizo language. Thus, selection of passage also needs to undergo multi-layer review with language experts such that the benchmarking process can be further strengthened.

5.2. External Factors which impacted Benchmarking

Given the mammoth task of including 20 languages in the benchmarking exercise for reading fluency and comprehension, the facilitators observed that the **learning outcomes differed as per the extent to which a particular language is developed or advanced.** For instance results differed across more developed languages like Hindi, English, Bengali than Garo, Khasi, Nepalese, Mizo etc. The latter languages are more dialectical in nature and do not have a well-developed science or grammar around their semantics.

6. Annexures

6.1. Panelist Registration Form

Ministry of Education
Government of India

AIR
Advancing Evidence. Improving Lives.

unicef
for every child

Numeracy Benchmark Setting Workshop

NCERT, Delhi

August 22 – 26, 2022

Panelist Registration Form

Name: _____

Designation: _____

Postal address: _____

Email (if any): _____

Mobile Number (including area code): _____

Gender: 1) Female 2) Male

Social group: _____

Highest education level: 1) Primary 2) Secondary 3) High Secondary

4) Bachelor 5) Masters 6) Ph.D. 7) Other _____

Years of experience: _____

Experience teaching learners with disabilities: 1) Yes 2) No

Mother tongue: _____

Language(s) use for classroom instruction (for teachers only): _____

6.2. Sample of Angoff Rating Form for ORF

Ministry of Education
Government of India

AIR
Advancing Evidence. Improving Lives.

unicef
for every child

Language (please tick): 1) English 2) Konkani 3) Marathi **Panelist Code:**

Oral Reading Fluency with Comprehension Benchmark Setting Workshop
Regional Institutes of Education, Bhopal

Angoff Rating Form: Oral Reading Fluency

Total numbers of words in the passage		ROUND 1				ROUND 2			
		Number of words **Just Partially Meets (JP)** students would attempt to read in a minute	Number of words **Just Meets (JM)** students would attempt to read in a minute	Number of words **Just Exceeds (JE)** students would attempt to read in a minute	If two of three **JE** students would not read the word correctly, then Above Just Exceeds (**AE**) would read it correctly	Number of words **Just Partially Meets (JP)** students would attempt to read in a minute	Number of words **Just Meets (JM)** students would attempt to read in a minute	Number of words **Just Exceeds (JE)** students would attempt to read in a minute	If two of three **JE** students would not read the word correctly, then Above Just Exceeds (**AE**) would read it correctly
Word #	Word	Would two of the three **JP** students read this word correctly? (Reasonably Sure)	Would two of the three **JM** students read this word correctly? (Reasonably Sure)	Would two of the three **JE** students read this word correctly? (Reasonably Sure)		Would two of the three **JP** students read this word correctly? (Reasonably Sure)	Would two of the three **JM** students read this word correctly? (Reasonably Sure)	Would two of the three **JE** students read this word correctly? (Reasonably Sure)	
1	Renu								
2	and								
3	Shefali								
4	were								
5	good								
6	friends.								
7	They								
8	were								
9	neighbors								
10	too.								
11	One								
12	day,								
13	Shefali's								
14	father								

6.3. Sample of the Angoff rating form for reading comprehension

Ministry of Education
Government of India

AIR
Advancing Evidence. Improving Lives.

unicef
for every child

Language (please tick): 1) English 2) Konkani 3) Marathi **Panelist Code:**

Oral Reading Fluency with Comprehension Benchmark Setting Workshop
Regional Institutes of Education, Bhopal

Angoff Rating Form: Reading Comprehension

Total numbers of words in the passage		ROUND 1				ROUND 2			
		Number of words **Just Partially Meets (JP)** students would attempt to read in a minute	Number of words **Just Meets (JM)** students would attempt to read in a minute	Number of words **Just Exceeds (JE)** students would attempt to read in a minute	If two of three **JE** students would not read the word correctly, then Above Just Exceeds (**AE**) would read it correctly	Number of words **Just Partially Meets (JP)** students would attempt to read in a minute	Number of words **Just Meets (JM)** students would attempt to read in a minute	Number of words **Just Exceeds (JE)** students would attempt to read in a minute	If two of three **JE** students would not read the word correctly, then Above Just Exceeds (**AE**) would read it correctly
Item #	Question	Would two of the three **JP** students read this word correctly? (Reasonably Sure)	Would two of the three **JM** students read this word correctly? (Reasonably Sure)	Would two of the three **JE** students read this word correctly? (Reasonably Sure)		Would two of the three **JP** students read this word correctly? (Reasonably Sure)	Would two of the three **JM** students read this word correctly? (Reasonably Sure)	Would two of the three **JE** students read this word correctly? (Reasonably Sure)	
RC_1	Who were good friends? [Renu and Shefali]								
RC_2	Where did Shefali"s father get a new job? [Shefali"s father got a new job in Delhi]								
RC_3	What did Renu"s father do? [Renu"s father was a farmer]								
RC_4	How can you say that Renu and Shefali were good friends? [They became sad as they had to live at different places away from each other/ Shefali started visiting Renu in her summer vacations to spend time with her]								
RC_5	Did Shefali join a new school in Delhi? [Yes]								

6.4. Panelist Workshop Evaluation Form

Oral Reading Fluency with Comprehension Benchmark Setting Workshop
Panelist Evaluation Form

Policy Linking for Large Scale Assessment
The purpose of this evaluation is to learn your reactions and perceptions of the various components of the policy linking workshop. Please answer each question honestly and accurately; it is very important that we have your reactions to the activities of the workshop.

Please do not put your name on the Evaluation form, as we want your responses to be anonymous. Thank you for your time in completing this evaluation.

Part 1: Training on Global Proficiency Descriptors
You have been trained on the Global Performance Descriptors (GPDs). Please read the following statements carefully and place a tick in each category to indicate the degree to which you agree with each statement.

GPD training	Strongly disagree	Disagree	Agree	Strongly agree
I understand the purpose of the GPDs				
The GPDs were clear and easy to understand				
The discussion of the GPDs helped me understand what is expect of students in reading fluency with comprehension at the end of **3rd** grade				
The practical exercise using the GPDs was useful to improve my understanding				
There was an equal opportunity for everyone to contribute their ideas and opinions and to ask questions				
The amount of time spent on the GPD training was sufficient				

Do you have any additional comments on the GPD training?

Part II: Training on the assessment and policy linking method
You have been trained on the assessment on which we are undertaking the policy linking and the policy linking methodology. Please read the following statements carefully and place a tick in each category to indicate the degree to which you agree with each statement.

Assessment training	Strongly disagree	Disagree	Agree	Strongly agree
I understand the purpose of the assessment				
I understand the constructs assessed in the assessment				
I had a clear understanding of the level of difficultly of each of the assessment items (i.e., words and comprehension questions)				
The amount of time spent on the assessment training was sufficient				

Do you have any additional comments on the assessment training?

Policy linking training	Strongly disagree	Disagree	Agree	Strongly agree
I understand the process I need to follow to complete the policy linking exercise				
The discussion of the procedure was sufficient to allow me to feel confident in making decisions				
The practice exercise helped me to understand what I need to do				
There was an equal opportunity for everyone to contribute their ideas and opinions and to ask questions				
The amount of time spent on the policy linking method training was sufficient				

Do you have any additional comments on the policy linking training?

Part III: Round 1 evaluation

During Round 1, you were asked to predict whether borderline Does not Meet Minimum Proficiency, Partially Meets Minimum Proficiency, Meets Minimum Proficiency, and Exceeds Minimum Proficiency students would be able to answer the questions correctly.

Policy linking rating	Strongly disagree	Disagree	Agree	Strongly agree
I am confident about the performance predictions I made during round 1				
I was able to follow the instructions and complete the round 1 form accurately				
I was given sufficient time to complete the round 1 performance predictions				

Do you have any additional comments on round 1?

Part IV: Round 2 evaluation

During Round 2, you were given actual performance information and data about the impact of using the Round 1 results. You then were asked to give revised performance predictions. Please select the best answer below.

Policy linking rating	Strongly disagree	Disagree	Agree	Strongly agree
I am confident about the performance predictions I made during round 2				
My performance predictions were influenced by the information showing the ratings of other panelists				
My performance predictions were influenced by the reality information (i.e., item difficulty) showing the actual performance of learners on the assessment				
My performance predictions were influenced by the impact information (i.e., score frequency distribution, student classifications) showing the outcomes for the sample of students				
I was given sufficient time to complete the round 2 performance predictions				

Do you have any additional comments on round 2?

Part V: Overall Evaluation

How comfortable are you with your final performance predictions?

Very uncomfortable	Somewhat uncomfortable	Fairly comfortable	Very comfortable

If you ticked uncomfortable option, please explain why.

Overall, how would you rate the success of the Policy Linking Workshop?

a. Totally Successful
b. Successful
c. Unsuccessful
d. Totally Unsuccessful

How would you rate the organization of the Workshop?

a. Totally Successful
b. Successful
c. Unsuccessful
d. Totally Unsuccessful

Please provide any comments you feel would be helpful to us in planning future policy linking workshops.

Thank you for your involvement in the Workshop.