OPTIMIZING ASSESSMENT FOR ALL
MARCH 2020

# OPTIMIZING ASSESSMENT FOR ALL

## Focus on Asia

BROOKINGS

Esther Care is a senior fellow in the Global Economy and Development Program at the Brookings Institution

Alvin Vista was a fellow in the Global Economy and Development Program at the Brookings Institution

Helyn Kim was a fellow in the Global Economy and Development Program at the Brookings Institution

Optimizing Assessment for All (OAA) is a project of the Center for Universal Education at the Brookings Institution. The aim of OAA is to support countries to improve the assessment, teaching, and learning of 21st century skills through increasing assessment literacy among regional and national education stakeholders; focusing on the constructive use of assessment in education; and developing new methods for assessing 21st century skills.

## Acknowledgements

The authors and the OAA Asia National Technical Teams deeply appreciate the significant contribution of Dr. Claire Scoular of the Australian Council for Educational Research. In addition, the support of UNESCO’s Network on Education Quality Monitoring in the Asia-Pacific (NEQMAP) has been invaluable for our work, and our thanks are extended to Mark Manns, Jun Morohasi and Maki Hayashikawa. The authors gratefully acknowledge Ung Chinna, Sarin Sar and Khou Hav of the Cambodia National Team; Amarjargal Adiyasuren and Munkhjargal Davaasuren of the Mongolia National Team; and Deviram Acharya and Shyam Acharya of the Nepal National Team for their substantial contributions to their respective country sections of the report. Additionally, the authors thank Aynur Gul Sahin for her creative and editorial support.

The Brookings Institution is a nonprofit organization devoted to independent research and policy solutions. Its mission is to conduct high-quality, independent research and, based on that research, to provide innovative, practical recommendations for policymakers and the public. The conclusions and recommendations of any Brookings publication are solely those of its author(s), and do not reflect the views of the Institution, its management, or its other scholars.

Brookings gratefully acknowledges the support provided by Porticus.

Brookings recognizes that the value it provides is in its commitment to quality, independence, and impact. Activities supported by its donors reflect this commitment.

Cover photo credit: Preah Norodom Primary School, Cambodia

# EXECUTIVE STATEMENT

The Optimizing Assessment for All (OAA) project at Brookings is about supporting efforts to use assessment constructively in education systems, specifically through developing assessments of 21st century skills (21CS).

21CS are now firmly ensconced as new learning goals in education systems worldwide, but their implementation in teaching and assessment practices is lagging.

We have taken decades to understand how to teach mainstream education subjects like math, history, science, and language. But with these new learning goals, which prioritize how to get answers, rather than just providing a correct response, we are facing new challenges.

We are particularly challenged in the case of assessment. Assessment has a bad name—it is used to label students as pass or fail. However, there is increasing use of what we call "formative assessment," using assessment in the classroom to inform a teacher about what to teach students next. This approach to assessment provides an opportunity. If we can identify useful approaches to assessment of 21CS in the classroom, then both the assessment tools themselves as well as how students engage with them can provide insights for teaching the skills.

OAA, in collaboration with participating countries from Asia and Africa, has helped identify 21CS valued by these countries, hypothesized what these skills might look like in classroom assessment tasks, and developed these tasks with teachers to ensure that they are usable and valuable in the classrooms. Notably, OAA has worked with established approaches to assessment that teachers already know and adjusted them to reflect new learning goals. Of course, the work goes far beyond assessment to implications for how we think about education and what we value in the classroom. What we value are the thinking and social processes that individuals use to explore and understand their environment.

More comprehensive information about the complete OAA approach can be found in “Optimizing Assessment for All: Framework for understanding project goals and scope,” while in this report we focus on the collaborative activities undertaken in Asia by Cambodia, Mongolia, and Nepal to create 21CS assessment tasks. The mechanics of the activities are described in detail in order to illustrate the methods used in the project and by the countries. For examples and guides for task creation, as well as information about scaling and implementing the OAA approach, see forthcoming reports in the series.

# INTRODUCTION

One of the main goals of OAA was for participating countries to develop classroom assessment tasks that can measure 21CS. The project adopted a collaborative approach to develop capacity in assessment design. The project was structured so that national teams had the opportunity to develop these assessment tasks together at the regional level, as well as individually at the national level. The objective was to ensure that the national teams were confident in the usability of the developed assessment tasks and in their ability to continue to develop tasks that were localized to their particular conditions, needs, and curriculum. The development process was undertaken through a series of workshops, each convened in one of the participating countries so that all national teams had the opportunity to understand the conditions under which each was working. Between workshops, in-country development work continued, both within the national teams, and with teachers from participating schools in each country. The process of task development is embedded in the following descriptions by the three national teams—covering the workshops and in-country activities—and culminating in a pilot of the assessments across the countries.

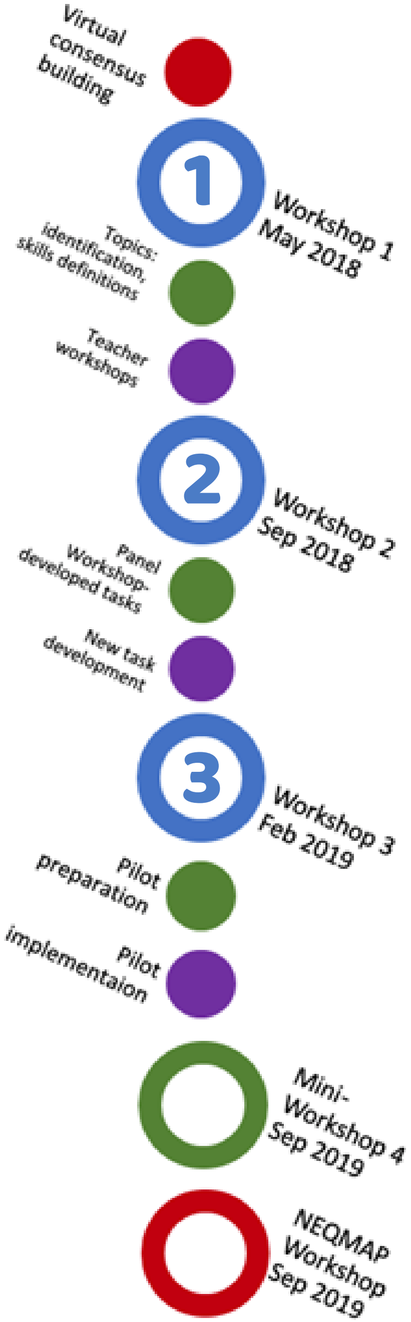
Every country is aware of children's need for 21CS. The United Nations Sustainable Development Goal 4.7 states that children should have transversal skills to prepare them as global citizens. The United Nations aims to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all." Education 2030 (UNESCO, 2017) identifies common competencies for all countries. Literature shows that most scholars agree on the need for skills like collaboration and team work, creativity and imagination, critical thinking, and problem solving (Partnership for 21st century skill, 2007; Trilling & Fadel, 2009). Brookings, in collaboration with Nepal, Cambodia, and Mongolia, have defined three of these 21CS for use in OAA. Assessing such complex skills is a challenging job. However, OAA has developed example tasks to sample these skills, tested them in the partner countries, and established a process that can be scaled to additional skills and grade levels (Nepal National Technical Team).

> ***"The Ministry plans for integration of transversal competencies in school curricula re-design drawing on the Ministry's learnings and experience from Brookings's OAA work, the Finnish Technical Assistance for Soft Skills project, and the British Council's Connecting Classrooms Project."***
>
> *Dr. Lekha Poudel, Director General Curriculum Development Centre, Ministry of Education, Science and Technology, Nepal*

# THE OAA PROCESS

The most intensive phase of the OAA Asia work took place over a 20-month period from 2018 to 2019. It included formally convened multi-country workshops, individual in-country workshops and convenings, regional meetings hosted by the Network on Education Quality Monitoring in the Asia-Pacific (NEQMAP), virtual communications, and maintenance of an online platform for document sharing and management. In this report, the experiences of the partner countries are described in the sequence in which they took place. The initiative included working together, consolidating the approach to assessment, and reviewing each other's progress and achievements throughout the process.

Figure 1. The OAA timeline



# 1 SETTING THE SCOPE: SKILLS, SUBJECTS, AND GRADES

The first task was to establish the scope of the work to ensure that the assessments would target the goals of interest. This required consideration of which 21CS to target, which grade levels, and on which curricular subjects to focus for the content which would carry the skills. The intention was to concentrate on 21CS that all three countries valued highly and to include both cognitive and social 21CS.

The first collaborative workshop for Asia was hosted in May 2018 by the Nepal National Technical Team, based in their Education Review Office of Nepal's Ministry of Education, Science, and Technology in Kathmandu, Nepal. The workshop was opened by Dr. Lekha Nath Poudel, then Joint Secretary of the Government of Nepal, Ministry of Education, Science and Technology. Experts participated from the Education Review Office, Curriculum Development Center, and Center for Education and Human Resource Development, as well as teachers from local Nepali schools who were contributing to the work.

The National Technical Teams from Cambodia and Mongolia, as well as the Brookings team with its technical members, came to the workshop having previously engaged in virtual meetings designed to establish shared understandings and likely selection of target 21CS, subjects, and grades. As part of each workshop, participants visited schools in each location in order to understand the classroom contexts of each country. In Nepal, schools in Kathmandu were visited.

The key goal for the first workshop was deciding on the definitions and descriptions of the three skills selected by the focus countries, as well as subjects and grades to target. An important feature of the first workshop was ensuring that the initiative, its philosophy, and approach made sense to the teachers.

*National and core teams hosted by Nepal's Education Review Office, then directed by Dr Lekha Poudel, for the first OAA Workshop in Kathmandu*

Twelve Nepali teachers from the schools participating in the project worked side by side with the local and international teams to define and describe the skills, and to consider the implications of these skills for classroom practice—both in terms of assessment and teaching.

An additional goal of the workshop was to develop an assessment framework for the classroom-based tasks. The framework was designed to consider the following elements:

- The purpose of an assessment framework and its development;
- The outcomes of a curriculum review of targeted grade levels across the three countries;
- The three focus skills and their definitions;
- The structural models for problem solving, critical thinking, and collaboration; and
- Examplar subject contexts for each of the skills.

The development of the assessment framework was therefore strongly informed by the preliminary curriculum reviews by each country, and comparisons across them.

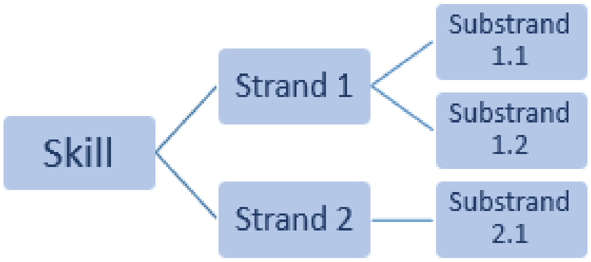
At the beginning of the workshop, each team provided a comprehensive overview of its country's education system, ensuring compatibility of the OAA work across all three countries. These compatibility issues came to the fore as the similarities and differences across the three country curricula for the target domain areas—mathematics, science, and social science—at the target grade levels explored and analyzed.

Although several members of the teams knew each other well through previous collaborative research (e.g., UNESCO, 2018), the detailed nature of task development aligned with curriculum required deeper understandings across countries. Ensuring that team members were familiar with the functions and forms of assessment in order to contextualize how the classroom-based assessment tasks should be designed was critical.

## The selected skills

The skills selected for assessment by the three focus countries were problem solving, critical thinking, and collaboration. All three countries had previously participated in Education Research Institutes Networks in the Asia-Pacific (ERI-Net) or NEQMAP studies of transversal competencies (Care & Luo, 2016; UNESCO, 2015, 2016a, 2016b), and were well placed to make decisions regarding the skills selection. Many 21CS are actually a mix of several skills. Figure 2 provides a structural model that demonstrates this.

Figure 2. Structure of skills



Of course, different skills have different numbers of contributing subskills, and sub-subskills. Labeling the structural levels of skill varies in different publications. Here we refer to the overarching construct as the skill, the next level as the strand, and the next as the substrand. As will be seen in the descriptions of these strands and substrands, some contribute to more than one 21CS. In other words, strands and substrands are not necessarily unique to a particular skill.

Definitions and descriptions of the skills contributed by the countries as well as research literature contributed by the Brookings team were critiqued by groups as they analyzed the structures and components. The 12 Nepali teachers attending the workshop were tasked with teasing out the practical implications of the nature of the skills for classroom practice. Workshop output included a first draft of skill definitions and structures. The final skills definitions and corresponding strands and substrands, as defined by the OAA teams in Asia, are presented in Table 1. In the same way that a subject curriculum is broken down into topics and sub-topics, a 21CS is broken down into its component parts.

OAA described problem solving as a "process that involves conceptualizing a problem, considering options and strategies to implement a plan to reach a solution, and evaluating the implementation." Critical thinking was defined as "making judgements by analyzing arguments through the synthesis of information and use of logical reasoning." Collaboration was defined as "a process of working together, communicating to negotiate different perspectives, and making decisions to reach a common goal.

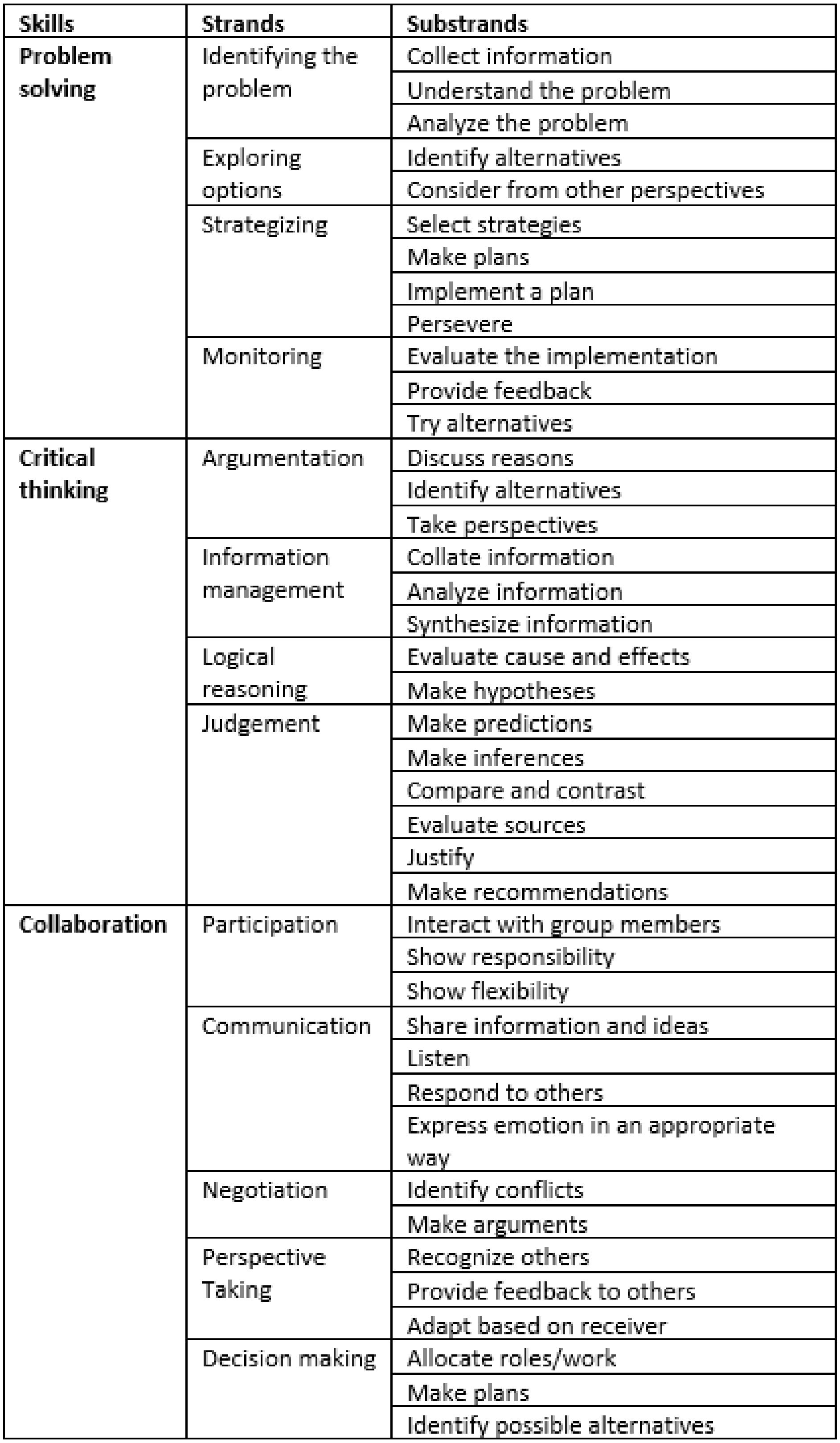
The core team, which included representatives from each country, and the Brookings and partner technical experts, designed an assessment framework with the following characteristics:

- The tasks to be developed were to be compatible with both formative and summative functions, and administrable by classroom teachers;
- Tasks were to be aligned with common Grade 5 curricula content;
- All three skills (problem solving, critical thinking and collaboration) were to be included;
- The three learning domains (mathematics, science and social science) were to be included; and
- The assessment blueprint would be designed such that testing of the hypothesis of transferability across domains could be explored.

*National teams sorting out the skills, strands, and substrands in Kathmandu—working across devices, across languages*

Following initial acceptance of the framework and the draft skills definitions, the strands and substrands were analyzed for their assessment potential, In other words, was it likely that assessment tasks could be developed that would both stimulate and provide the opportunity for evidence of the skills to be visible? Examples of assessment task types were discussed to reach understanding about the range of formats viable for 21CS assessment.

Table 1. 21CS definitions and structures



| Skills | Strands | Substrands |
|---|---|---|
| **Problem solving** | Identifying the problem | Collect information |
| | | Understand the problem |
| | | Analyze the problem |
| | Exploring options | Identify alternatives |
| | | Consider from other perspectives |
| | Strategizing | Select strategies |
| | | Make plans |
| | | Implement a plan |
| | | Persevere |
| | Monitoring | Evaluate the implementation |
| | | Provide feedback |
| | | Try alternatives |
| **Critical thinking** | Argumentation | Discuss reasons |
| | | Identify alternatives |
| | | Take perspectives |
| | Information management | Collate information |
| | | Analyze information |
| | | Synthesize information |
| | Logical reasoning | Evaluate cause and effects |
| | | Make hypotheses |
| | Judgement | Make predictions |
| | | Make inferences |
| | | Compare and contrast |
| | | Evaluate sources |
| | | Justify |
| | | Make recommendations |
| **Collaboration** | Participation | Interact with group members |
| | | Show responsibility |
| | | Show flexibility |
| | Communication | Share information and ideas |
| | | Listen |
| | | Respond to others |
| | | Express emotion in an appropriate way |
| | Negotiation | Identify conflicts |
| | | Make arguments |
| | Perspective Taking | Recognize others |
| | | Provide feedback to others |
| | | Adapt based on receiver |
| | Decision making | Allocate roles/work |
| | | Make plans |
| | | Identify possible alternatives |

## Curricular areas and topics

Three curricular areas, or subjects, were selected to provide the "content" to which the three skills would be applied in the assessment tasks. These were mathematics, science, and social science. Based on the curricula from each participating country, these areas were organized across topics. During the test development process, each assessment task was targeted such that the competency in a skill would be assessed in the context of a topic. Accordingly, each country's curriculum was compared to enable selection of topics common to all. An example of curriculum mapping across the countries to identify common topics is shown in Table 2.

The goal achieved was common skills and topics across the three countries, so that the assessment tasks developed could be used by all countries.

## Target group

Since the assessments of skills were to be embedded in curriculum at Grades 5-6 levels, the target groups were students in these grades, alongside their teachers who engaged in the project.

Nepal selected Grade 6 because it is the first grade of lower secondary school and the curriculum had relatively recently been reviewed.

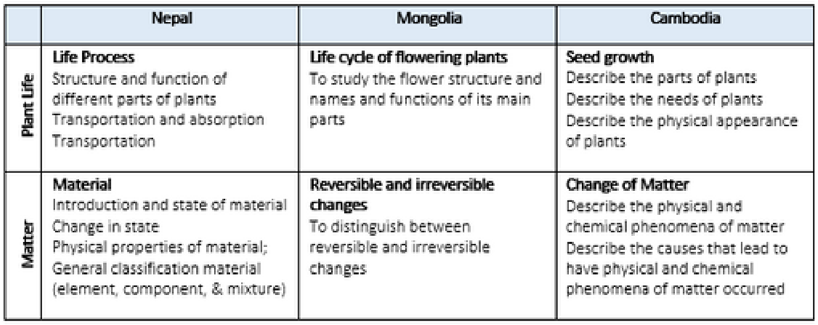
*First "all country" OAA meeting in Kathmandu with curriculum comparison and assessment development work ahead*

For Mongolia, Grade 5 was selected because it is the last year of primary education. Cambodia also focused on Grade 5, in order to align with the Southeast Asia Primary Learning Metrics (SEA-PLM), a new regional assessment, which target students in Grade 5. The slightly different curricular topics across grade levels of the three countries also played a role in the grade selection.

## An explicit focus on teachers

Teachers were a critical component of the OAA project because they support learning in the classroom. Accordingly, it was essential that their expertise was considered in knowledge building around the skills and their definitions. Teachers' understanding of the classroom context and their students was key to considering how the skills were to be integrated within assessment tasks.

Table 2. Common science topics and sub-topics within the “Plant Life and Matter” strands



| | Nepal | Mongolia | Cambodia |
|---|---|---|---|
| Plant Life | **Life Process**<br>Structure and function of different parts of plants<br>Transportation and absorption<br>Transportation | **Life cycle of flowering plants**<br>To study the flower structure and names and functions of its main parts | **Seed growth**<br>Describe the parts of plants<br>Describe the needs of plants<br>Describe the physical appearance of plants |
| Matter | **Material**<br>Introduction and state of material<br>Change in state<br>Physical properties of material;<br>General classification material (element, component, & mixture) | **Reversible and irreversible changes**<br>To distinguish between reversible and irreversible changes | **Change of Matter**<br>Describe the physical and chemical phenomena of matter<br>Describe the causes that lead to have physical and chemical phenomena of matter occurred |

*Note. Drawn directly from country curricula*

One teacher believed that 21CS are exercised mainly in the upper grade levels; lower grades are taught facts or knowledge only. These perspectives are embedded in the teachers' belief that older students have higher thinking capacities, and so can apply knowledge and discuss more complex issues including social problems. The lower grades are seen as platforms for building foundational knowledge.

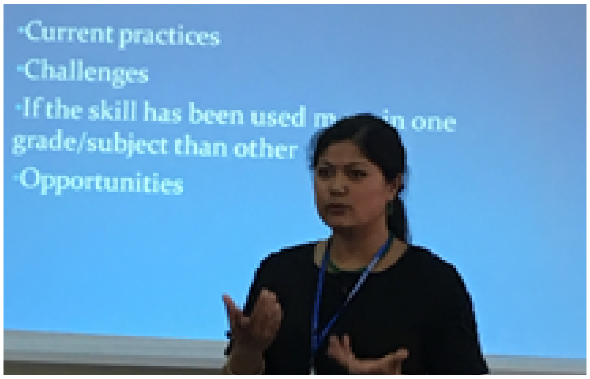
In terms of how 21CS might be applied, teachers saw critical thinking as most likely to be used in social sciences. For example, students could draw on their own knowledge of values, social problems, national heroes, or neighboring countries independent of the teacher. Based on the broad-reaching view that mathematics is difficult, the teachers viewed use of 21CS in mathematics as unlikely, as they relied on set procedures and formulae for teaching and learning in this subject.

Since the OAA project’s interest was in use of assessment in the context of teaching and learning, it was essential that tasks developed provide useful information to teachers, as well as provide lessons on assessment of 21CS applicable on a larger scale. Accordingly, teacher insights were drawn upon in the workshop to determine challenges not only on skills assessment but also on teaching. As the teams worked on the definitions and descriptions of the skills, it became clear that targeting skills rather than knowledge would have implications for teaching practice, and the teachers highlighted the challenges in teaching and assessment of 21CS.

**Teacher-identified challenges on teaching and assessment of 21CS**

- **Rote learning:** the practices of learning by memory and class level repetition discourages use of varied pedagogical strategies
- **Texts:** most texts follow the curriculum closely or stand in place of the curriculum; teacher focus is to "cover" these resources rather than ensure that students are understanding
- **Assessment culture:** assessments and especially examinations determine the teaching focus; correct answers based on recall are prioritized to achieve high marks on exams
- **Culture:** particularly in the lower grades, the norm is to have less interaction between students and the teacher, with the students adopting a listening role
- **Class size:** with up to 60 students in each class, and often cramped conditions, learner-centered approaches are not seen as viable
- **Lack of ability:** teachers believe it is important to teach 21CS but do not know how
- **Predetermined beliefs:** there is a widely held belief that students' backgrounds affect their ability to learn, and that pre-existing differences are immutable
- **Lack of resources:** there is a lack of resources in the form of teaching materials and pedagogical support

Overall, teachers believed strongly that 21CS are important to teach, but were concerned about the challenges. However, some teachers noted that participation in the OAA workshop, specifically in defining and decomposing the skills into strands and substrands, had alerted them that they were already teaching some of these skill strands in their classroom, but had not been conscious of it.


·Current practices
·Challenges
·If the skill has been used m in one grade/subject than other
·Opportunities

*Nepali teacher presenting on challenges to changing practices in the classroom*

The experience and learnings of the twelve Nepali teachers in the first workshop made clear that teachers needed more familiarization with the skills in order to contribute to the OAA project. Accordingly, sessions for teachers in Cambodia and Mongolia were arranged prior to the second workshop.

*Tea-time is "we" time—core and Nepal national team members learning about each other in Kathmandu at OAA's first workshop*

## 2 TASK DEVELOPMENT

The second collaborative workshop was hosted at the Mongolian National University of Education in Ulaanbaatar in September 2018. It was opened by Mrs. Gantsetseg, Ch., specialist for Policy and Development of Education, Department of General Education, and attended by the three NTTs, experts from the tertiary sector based in Ulaanbaatar, and teachers from participating Mongolian schools. The workshop was dedicated to the development of a first set of 11 assessment tasks.

The initial focus was on developing ideas for tasks. The ideas needed to be based on the curriculum and able to incorporate the skills. Following presentations and discussion of different task and item formats, participants divided into groups for brainstorming on task concepts. In order for tasks to provide optimal opportunity for the demand and exercise of substrands, task design required multiple steps or subtasks. To justify a student dedicating several minutes to understand the demands of a task, multiple data points are required. Accordingly, most tasks that were developed contain four to six items, each of which is centered around the same core stimulus material, but each of which may sample different substrands at different levels of difficulty.

## Task development at the workshop

The first day of task development at the Mongolian National University of Education was challenging. Participants had insufficient topic knowledge from the curriculum, and access to curricular resources was limited due to internet connectivity problems. However, these issues were overcome in the following days so that task drafting continued.

In the second half of the workshop, 10 Mongolian teachers joined the NTTs for task paneling, and three lecturers from the Mongolian National University of Education helped with translation of tasks that had been originally written in English. During these days, the NTTs and teachers worked with paneling checklists which guided the process of evaluating whether the draft tasks were appropriate for the purpose. The targeting of multiple substrands in some of the tasks raised queries about task design and effectiveness. This identification of concerns about assessment approaches that were unfamiliar to the countries remained a theme throughout the project. The teachers made a huge contribution to the process through their knowledge of the curricular competences of their students. However, teachers were less able to contribute actively when asked how to improve tasks. This was in large part due to their novice understanding of the skills, as well as the novelty of the form of assessment.

After the NTTs returned to their countries, they replicated the task paneling. This engaged review by local assessment developers and review with local teachers.

### Task formats

Several task formats were adopted for task construction,

**Multiple-choice items** provide students with three or four response options, of which only one is correct. Multiple choice items are widely used due to their capacity to generate validity and reliability indices, and to measure a construct in a relatively short testing time. However, they do not provide students with the opportunity to explain their choices or provide supporting statements. Although some individuals believe that multiple-choice items can only be used for assessment of non-complex learning, well-constructed items can assess both the simple and complex.

**Constructed-response items** require students to provide written responses, rather than select a response from a set of options. Because this format allows students to provide explanations, and support an answer with reasons or evidence, this format is well-suited for identifying the processes involved with 21CS. The disadvantage of this format is the need for some judgement by scorers, although well-crafted rubrics will minimize this error.

**Worksheet tasks** require students to complete a series of activities and log their actions and responses to items within the tasks. Some tasks may require students to find their own sources of information and report what they found, other tasks may require organization of information, and some tasks may require working with others and recording the details. In the OAA project, several of the collaboration tasks took this format.

## Paneling

### What is paneling?

Paneling is an essential part of test development. It helps provide quality assurance and establish validity of the tests (measuring what we think we are measuring). Paneling can expose accidental errors made during item development. It is a thorough and rigorous process that reduces waste in the trialing of items.

The paneling process relies on a checklist of issues to consider as the panel of experts evaluates each test item. Issues include:

- Does the item target the construct?
- Is the item fair to the intended test takers?
- Is the item well phrased and clear?
- Does the item respond to the likely capabilities and knowledge of the test takers?

A panel includes the item writer and two or three experts in the skill or domain being assessed. Experts include subject experts, assessment experts, language experts, or teachers who have sufficient knowledge to examine and revise the test items. There are conventions governing how the panel operates, so that the process remains objective.

The outcome of paneling is identifying whether an item:

- is acceptable for use;
- can be modified for inclusion in the item pool; or
- is unacceptable, and should be discarded.

### Paneling in Cambodia

Based on the knowledge built in the Mongolia workshop, the Cambodian NTT organized an in-country workshop to panel the workshop-developed items. Participants were Grade 5 teachers and directors from the four target schools, and representatives from relevant departments including the Primary Education Department, Curriculum Development Department, and Teacher Training Department. Participants were introduced to the procedures to be undertaken, the targeted subjects and key skills, and the skill assessment frameworks, and then divided into five groups to discuss and comment on 11 tasks.

The process of paneling items was a great opportunity for teachers and other participants to learn about the many elements employed in each task to measure the 21st century skills. It provided them with an inkling of how 21st century skills might look like in terms of student activity. The teachers could see how a single item could be developed to assess specific skills and how the items were targeted to their students' abilities. In addition, teachers had the opportunity to learn about quality of an item by exploring the criteria suggested. This helped them to reflect on the more routine classroom-based assessments, which they prepare and give to students every month.

There were a number of specific issues encountered during the paneling. First, the paneling record template did not provide for the variety of responses that panelists identified. This meant that for future work, some fine tuning of templates needs to be undertaken.

*OAA Teams learning about schools in Kathmandu and grateful for their commitment (with thanks to Shree Manohar Secondary School)*

Another issue was a substantive one: The items were originally drafted by the NTTs from Cambodia, Mongolia, and Nepal but some did not make immediate sense to the Cambodian reviewers. For instance, a question asked, "What is the best way to irrigate a large area of plantation with limited amount of water and cover the whole area?" The answer was "spray with pressured water," a response from the Mongolian context. Of course, identifying such issues was precisely the point of the in-country review. Finally, it was difficult for reviewers to comment on or set time estimates for completion of some tasks, since they did not have experience with these types of questions for student assessment.

*Members of the Cambodian, Nepali, and Mongolian OAA teams at work*

## Paneling in Nepal

Nepalese subject experts from the Curriculum Development Centre, National Centre for Educational Development, and Education Review Office (ERO) participated in a two-day in-country workshop for paneling of tasks in November 2018. The 12 teachers from participating schools were actively involved in the workshop.

Items developed in the Mongolia workshop were translated into Nepali language and paneled for evaluation. This led to much greater teacher understanding about the nature of 21CS and the approach to assessment, including familiarization with the characteristics of good items. The teachers appreciated the use of model items in mathematics, social studies, and science, and expressed enthusiasm in using the skills in teaching activities in their schools.

The sessions were conducted in Nepali language, making it easy to understand and interact. At the "summing up" session, the Director General of ERO, Krishna Prasad Kapri, noted how the technical processes had been an opportunity to learn and work at international standards. The notion of using curriculum-based standardized items to enhance the capacity of teachers was seen as innovative and full of potential.The sessions were conducted in Nepali language, making it easy to understand and interact.

## Cognitive laboratories, or "think alouds"

After the in-country panel sessions, all NTTs conducted "think aloud" sessions with small groups of students. The sessions enabled observation of the skills and substrands that students use when engaging in assessment tasks. The activity provided valuable information about whether students' target skills and substrands are actually drawn out by a task, as well as information about the tasks themselves. Country outputs from both the panel and "think-aloud" activities across the three NTTs were consolidated and synthesized by the core team.

**What is a "think aloud?"**

An important part of the process of developing assessments, or test items, is to ascertain that what we believe an item is measuring or sampling is matched by the reality. This is true of all assessment development, but particularly important when we are less familiar with the domain (the area of knowledge or competency) that we are targeting. With a skill like problem solving, for example, we want to know whether the problem solver is systematically and comprehensively identifying all the relevant features of the problem, understanding the relationships between them, hypothesizing and checking solutions, and so on. "Think alouds," sometimes referred to in the literature as cognitive laboratories, help us determine this.

The NTTs have been adapting knowledge-based (subject-based) test items to sample skills and check whether students actually activate those skills to solve the items.

The items are presented to the students who explain what they are thinking or doing as they work on the items.The resulting stream of consciousness from the students can be analyzed against the specific skills and substrands that an item is supposed to target. Detailed transcripts, video, or audio tapes are generated to check against a pre-developed set of questions about the items and the student responses.

What matters is whether the items are working, not how well the students are doing, but the variety of responses to the items across students of different abilities and backgrounds is important too. That way, not only do we know whether the items are stimulating student activation of the skills, but we also obtain more information about the different directions students can take items, and so can finetune the scoring methods.

One of the fascinating aspects of doing these think alouds is seeing how students can respond in so many ways to the same stimulus materials. It makes us aware of the vast opportunities that 21CS assessments can provide for students to take different perspectives, to experiment, and to imagine and create.

Facilitating think alouds is a skilled process. It requires understanding what the target competencies of the test items are; importantly, it requires patience to let the student maintain ownership of thinking. Through prompting we learn what the student is thinking rather than limit or guide the thinking. It is therefore essential that the facilitator does not provide interpretive comment but just prompt for the student to keep talking, to keep thinking.

### Cambodia’s think aloud

Prior to the think aloud exercise, the Cambodian NTT visited participating schools and worked with the teachers on how to engage in the think aloud activities. The twelve teachers' own experiences were material in their ability to engage in the process. They also needed to be familiar with the skills and substrands being tested, the difficulty level of items, and students' responses. Students were asked to talk aloud constantly while they were working individually or with their partners on a collaborative task. The test administrator sat near the students, but not in their personal space, in order to prompt with utterances such as "keep talking" or "what's happening." In total, 24 students and 12 teachers engaged in the think aloud sessions. Each student was assigned four tasks.

The think aloud activity is a unique method for teachers and the NTT to analyze the students' capabilities and the assessment. It provided an insight into the skills and substrands that students used when engaging in assessment tasks. It also shed light on how to study students' capabilities, providing valuable information about what tasks were beyond the abilities of students. Most teachers, accustomed to traditional tests in the classroom, were very curious to see the processes and concerned about their students' ability to complete the tasks. Among some of the issues, the think aloud exercise was not familiar to either teachers or students. Some teachers would forget to remind students to constantly talk aloud. Most students appeared reluctant and felt uneasy in doing the tasks.

*Convening in Phnom Penh to develop understanding of 21CS with Cambodian teachers*

This was in part due to the need to speak aloud and so students would sometimes keep silent while doing the tasks. Such focus on the individual in the formal education sector is unusual and is at odds with conventions of student-teacher interactions. Respect for teachers in Cambodia is signified by students acting modestly and unlikely to initiate discussion with teachers or elders, so having someone sitting nearby and listening to their thoughts felt awkward. In addition, it was difficult for the students to multi-task with talking aloud and writing at the same time.

As a result of these cultural issues, students may not have completed the tasks to the best of their ability. Moreover, for most students, the purpose of each task question, both for individual and group work, was not always clear. They were accustomed to much more scaffolding structure of their work; and so needed more detail to be provided. An associated issue lies in the level and expression of language used in the assessments: The targeting to students in Grade 5 or 6 needed reconsideration.

### Mongolia's think aloud

In Ulaanbaatar, Mongolia, the think aloud exercises were conducted in two schools during normal school hours. The NTT members engaged in the activity themselves, with each member taking a different skill.

It was clear that students were not at all familiar with the style of assessment tasks. For example, with problem solving tasks, students asked if they should continue to use the first stimulus for all items. The concept of completing a set of items all within the same task was new to them. For the NTT, it was reasonably simple to be able to estimate the student capability levels from their responses. For some tasks, such as the "Agriculture – Irrigation system," all students found this difficult and guessed the answer. This of course was useful to the NTT in that it helped to identify the range of task and item difficulty best suited to the grade level for future reference.

*Linguists and teachers from the Mongolian National University of Education working* on terminology and translations

For collaboration skills, again students and even teachers were unfamiliar with the approach. This lack of familiarity pertained both to the style of working together through a task, as well as to the format of the tasks themselves. It was necessary for the NTT to go beyond the test instructions to explain the tasks.

For the critical thinking tasks, 18 students in each school were involved. Students were allowed to talk to each other about the task, which allowed insight into their thinking. Although the task developers had looked at the curriculum and chosen the topics, at the time of the exercise, the concept of "percent" had not yet been introduced in mathematics and students could not do the item. Again, this was a useful learning for the team, making very clear the interdependence of content and skill in the tasks.

The think alouds were an important activity for teachers. It highlighted how students are thinking, and how teaching and assessments need to be aligned. The Mongolia NTT has continued discussions about how teachers might be trained in the technique to learn more about how students are learning.

### Nepal’s think aloud

In Nepal, teachers and the NTT had participated in the workshop for item translation and contextualization and so were well prepared for the think aloud exercise. The main objective of this exercise was to explore whether the tools were appropriate for the students' cognitive level, and about the quality of items, clarity of wording, and the degree to which the tasks actually targeted 21CS. The teachers themselves administered the tasks to selected students in Grade 6, while the Nepal NTT closely monitored the process. Between 3-6 students per task participated in the activity. Students were asked to read the items aloud, discuss, and then for the collaboration tasks to find solutions together. Students of high, medium, and low ability were selected for the activity. As was also the case in Mongolia, the NTT and teachers needed to support the students so that they could understand what was required of them in these previously un-encountered type of assessments.

The activity itself provided a great deal more information about the tasks and their items. One finding involved the time required for completing the tasks. There had been over-estimates of time needed for critical thinking and problem solving, but under-estimates for collaboration. Another set of findings centered around contextualization became clear, as illustrated by a task based on genetic modifications of food crops.

*Nepal's Education Review Office educators, and teachers from participating schools in Kathmandu exploring the nature of the targeted skills to assess*

## In-country additional task development

Each country then developed additional tasks using the regionally developed tasks as a guide. Outputs from the in-country paneling and cognitive laboratory activities were consolidated and synthesized to produce item templates for use in this item development.

Using the process that had been modeled in the second workshop, the NTTs engaged in development of additional tasks to supplement the three-countries' developed tasks. These were revised through an iterative process, including paneling across countries, and using online platforms for management of the activity, until the final set of tasks was ready for selection into the pilot phase.

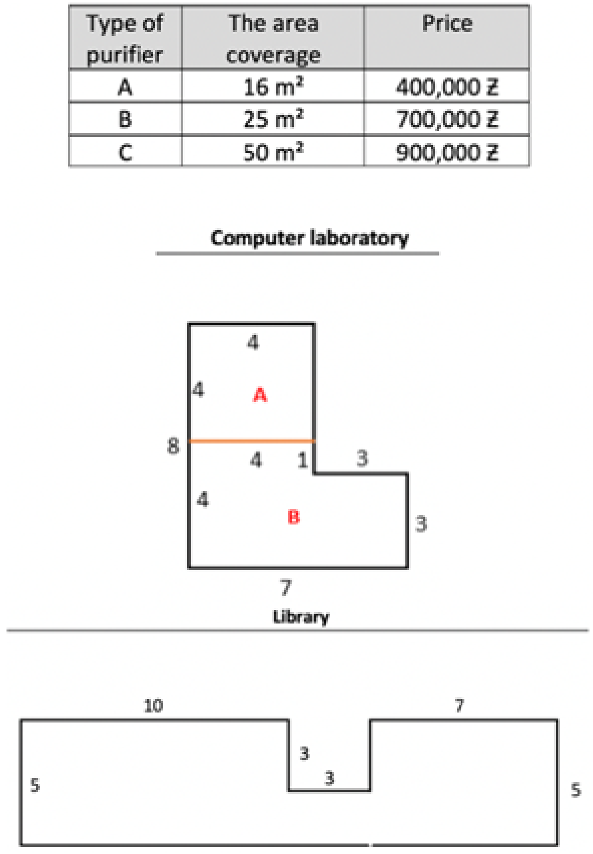
## Mongolia’s additional task development

Collaboration between the Mongolian NTT and teachers provided a perfect combination for the process of generating additional items. The national experts were experienced in developing "new format type," tasks while the teachers knew how Grade 5 students thought. Initially, the NTT asked teachers to identify a topic and develop some task concepts, but the teachers were not as creative as expected. They also were not familiar with the multiple-item structuring of the tasks. Accordingly, the NTT then selected potential topics, drafted some ideas for items, and then developed these together with the teachers.

Teachers’ contributions in adjusting the items in light of how Grade 5 students think, and finalizing the numbers, words, and phrases were enormous. As an example of the interdependence between curriculum and assessment expertise was the case of the "air purifier" task (see Figure 3). The NTT chose air pollution as a real-world problem topic that needs to be solved. The NTT drafted the first version of the task which covered the topic "volume of a cube or a parallelepiped." The national experts then researched an electronic shop website and found that air purifiers cover certain areas, rather than volume. Therefore, the task needed to be adjusted to an "area" problem. Accordingly, an item that required finding the area of an irregular rectangle was drafted. To finalize the denotation for this aspect of the task, teachers’ contribution was crucial. The development process illustrates the need for both curriculum and assessment expertise for integration of 21CS.

**Air Purifier Task:** Components of an assessment task in mathematics that require students to apply problem solving skills

Figure 3. A mathematics problem solving task



| Type of purifier | The area coverage | Price |
|---|---|---|
| A | 16 m² | 400,000 ₮ |
| B | 25 m² | 700,000 ₮ |
| C | 50 m² | 900,000 ₮ |

By the end of the processes, there were 25 nationally developed tasks, plus the original regionally developed 11. The whole set was reviewed by the NTTs during the third regional workshop as preparation for piloting at the class level.

# 3 TASK SELECTION AND PREPARATION FOR PILOT

All three countries shared their new tasks in the third collaborative workshop in Cambodia. Convened in Phnom Penh, and opened by His Excellency Im Koch, Secretary of State, Ministry of Education, Youth and Sport, the workshop took place in the Conference Hall of the Ministry of Education, Youth and Sport in February 2019. The primary goal of the workshop was to review all tasks, both regionally and nationally developed, and select a subset for piloting. The secondary goal was to ensure agreement on the numbers of students who would take part in the pilot, and on the pilot sampling design.

It was important that all three countries administer the same tasks for the pilot in order to establish cross-cultural relevance and to generate sufficient data for psychometric analysis. Developing new tasks that reflected common content across the three countries had proven challenging, in part since the teachers were familiar only with their own country curriculum and textbooks. However, topics such as global warming, disposal of waste, and use of urban spaces are relevant across all countries and provided some common themes for task creation.

For example, the Mongolia team developed a task based on public transportation that included specific destination names and real time and distances of travel. Localizing the task was difficult due to unique characteristics of the original locale, and would have required amendments such that the task itself would no longer be parallel across countries. Another example of non-selected task concerned local organizations and their roles. Even though the task was well developed, it was "too local" due to major differences in organizational structures across the countries.

The workshop review took into account whether the tasks could reasonably elicit the targeted skills, and the generalizability of the tasks for use across the countries.

*OAA national teams at workshop 3 with His Excellency Im Koch, Secretary of State, Ministry of Education, Youth and Sport, convening in Phnom Penh to finalize preparations for pilot of the assessments*

The process resulted in the selection of 18 classroom assessment tasks—six per skill (collaboration, problem solving, and critical thinking)—for the pilot. As part of the workshop process, each NTT presented their preferred tasks for discussion. The final decisions were based on analysis of three factors. These included:

- Views of NTTs around the appropriateness of the tasks in terms of their curricula and the appropriateness for Grade 5 and 6 students;
- Sufficient tasks across skills by subjects to develop indicators of the skills and strands; and
- Selection of regionally-developed and nationally-developed tasks.

### Skills coverage

It became clear that the subjects and topics selected for task development enabled targeting of some substrands to a greater degree than others. For example, there was a paucity of items for the monitoring strand within the problem solving skill, which demonstrated the potential inadequacy of the item pool for generation of scores for this strand. The next step for countries, beyond the OAA initiative, would therefore be to generate more items using item structures that worked effectively in the pilot.

### Scoring

An essential component of the task review was to evaluate the scoring of items. This consisted of ensuring that rubrics, where used, differentiated clearly between levels of quality of response. Where the differences between scoring categories were small or might lead to error, these were amended or combined.

### Skills targeting

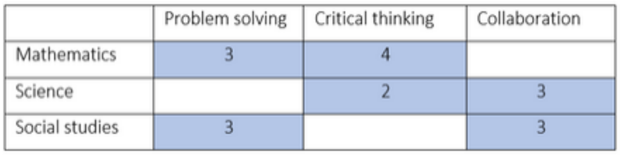
The review required re-visiting the nature of the skills to ensure that problem solving, critical thinking, and collaboration were being targeted. Problem solving was a particular concern; some country-developed tasks reflected normal mathematics assessment tasks rather than problems presented in a mathematics context. This issue of differentiating true problem solving processes from routine solutions is complex, both in terms of explanation and nuance.

## Pilot design

Assessment tasks were developed across combinations of subjects by skills. Important factors in the design were inclusion of each skill across at least two of the identified subject areas; and at least two skills within each subject. The design ensured that each skill was measured as it manifests in at least two domains of learning, while students would also have the opportunity to demonstrate two skills in each of the selected domains. This approach was designed to facilitate analysis of student performance across application of a skill to different subjects—keeping in mind the transversal character of 21CS.

The final number of tasks that were used for the pilot is summarized in Table 3, which shows the cross-representation of skills by subjects, and the number of tasks within each category. Each task consisted of between three and six sub-tasks or items, and was estimated to take between 15-40 minutes to complete.

Table 3. Design matrix for the pilot tasks

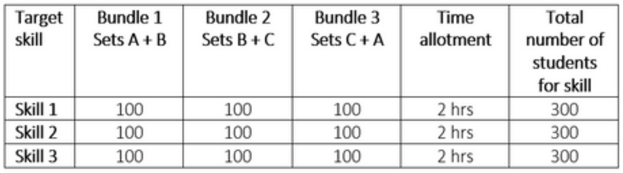


| | Problem solving | Critical thinking | Collaboration |
|---|---|---|---|
| Mathematics | 3 | 4 | |
| Science | | 2 | 3 |
| Social studies | 3 | | 3 |

## Bundles and sample

The final process for the NTTs involved the "bundling" of tasks so that decisions on the number of students needed in each country for a sufficient sample could be achieved (see Table 4).The embedding of two sets of tasks within each bundle ensured sufficient numbers of students completing common tasks to enable calibration across the full set of tasks for each skill, while limiting the number of hours that students would need to commit to the exercise.

Table 4. Bundles by skills by students



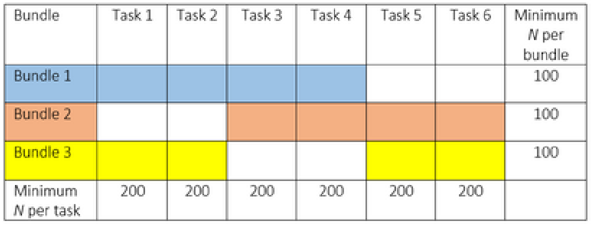
| Target skill | Bundle 1 Sets A + B | Bundle 2 Sets B + C | Bundle 3 Sets C + A | Time allotment | Total number of students for skill |
|---|---|---|---|---|---|
| Skill 1 | 100 | 100 | 100 | 2 hrs | 300 |
| Skill 2 | 100 | 100 | 100 | 2 hrs | 300 |
| Skill 3 | 100 | 100 | 100 | 2 hrs | 300 |

## Target number of responses

According to Linacre (1994), a minimum sample size of between 108 and 243 (depending on test targeting, with a higher minimum required for poorly targeted tests) is needed for item calibrations that are stable within ± 0.5 logits at 99% confidence level. Being in the pilot phase and therefore having no information related to test targeting, 200 was decided upon as the minimum sample size.

Given the final set of 18 tasks (six tasks per skill), the bundling arrangement is described in Table 5. Because each skill is reflected across the same number of tasks, the arrangement was the same for all skills. As indicated in Table 4, this allocation means that each task is completed by a minimum of 200 students, and the overlap across bundles ensures that there are always two common tasks between any pair of bundles. The same students were allocated tasks across all three skills, and the bundles across these skills were administered on different days to minimize test fatigue.

Table 5. Bundling arrangement for each skill



| Bundle | Task 1 | Task 2 | Task 3 | Task 4 | Task 5 | Task 6 | Minimum *N* per bundle |
|---|---|---|---|---|---|---|---|
| Bundle 1 | 🟦 | 🟦 | 🟦 | 🟦 | | | 100 |
| Bundle 2 | | | 🟧 | 🟧 | 🟧 | 🟧 | 100 |
| Bundle 3 | 🟨 | 🟨 | | | 🟨 | 🟨 | 100 |
| Minimum *N* per task | 200 | 200 | 200 | 200 | 200 | 200 | |

## Translation and adaptation of the tasks

Following the third collaborative workshop, the NTTs prepared in-country for the pilots. Preparation included translation and adaptation of the tasks; training of Test Administrators; and preparation of test materials and arrangements with schools. The NTTs followed detailed guidelines on translation, and in some cases adaptation, to ensure that tasks would perform similarly across languages for the pilot.

### Cambodia

Translation and adaptation were undertaken by the Cambodian NTT. The team members were each allocated to one of three groups to work on the subject areas of mathematics, science, and social studies. Then, each group exchanged draft translations for reviewing. Finally, the translated materials were reviewed by teachers on content, word choices, and language structures. There is no doubt that language was an issue in the project, not only in terms of language used in the assessment tasks, but also the language used to describe the project and the skills. Understanding the project and skills was essential to ensure data collection at the grassroots level made sense to schools and teachers.

Adaptations were required to reflect the local context. For example, the term "cookie" used in a task was replaced by "sweet cake," a term with which Cambodian students were familiar. Other adaptations included names of persons, such as “Zana” and “Sono” being changed to “Bopha” and “Nary."

### Mongolia

The Mongolian NTT translated the tasks since they were already familiar with them and could draw on translations from the think aloud activities. For finalization, the NTT drew upon the Education Evaluation Center specialists for proofreading, and so some terms and grammar were corrected.

The translation process raised some issues concerning localization of items. The translations needed to be true to the original versions in order to maintain equivalence across countries. Issues encountered included the use of specific currencies, such as the Zed, and use of terminology such as “water convolvulus,” as opposed to the more familiar “cabbage.” The NTT had previously hypothesized that lack of familiarity with some of these terms might affect students’ performance (but this was later seen to be unfounded). For other terms in common use, the NTT checked with the pilot school teachers concerning whether Grade 5 students would be familiar with the words (for example the word "poster").

*Students in Cambodia engaging in the pilot of collaboration tasks*

*Students in Mongolia engaging in OAA assessment tasks*

### Nepal

The Nepal NTT used two strategies for task translation. First, language experts translated the tasks and then teachers reviewed for common usage. Second, subject teachers of social studies, mathematics, and science, NTT members from the ERO, and the participating non-government organization Samunnat Nepal representative also worked on translation, adaptation and finalization of the scoring rubrics. The goal was to translate as close to the original as possible, while maintaining the difficulty of the task associated with both content and skill, rather than associated with language. More simple words were substituted for the original words where possible. For example, “strategy” was replaced with “way,” and “challenges” with “difficulties.” During the translation process, some language issues were identified both in the test items and the scoring rubrics. These included matters of currency and science terms. Similarly, the names of persons were localized: “Samir” replaced “Sambat,” and “Ram” replaced “Zana.” Female variants of verbs were used where female characters were referenced. Measurement unit terms were changed from the generic "Zed" to "Rs." Some standard terms like 'long jump' were replaced in Nepali to “nªhd\k,” reflecting the English pronunciation, but in Nepali language. The term "hose," unknown in Nepali, was replaced by “flexible pipe.” The term “water convolvulus” was replaced by the term “water spinach” (पानी पालुङ्गो).

Through the translation process, additional issues came to light. The conventions for scoring, where instructions provided for clear discrimination between quality of responses with consequent allocation of codes, were not well understood by teachers. Traditionally teachers had been accustomed to awarding half marks, and moving toward a practice where different levels of quality responses were subjected to criteria in rubrics was not well understood.

Notwithstanding the paneling and subsequent revisions, teachers thought that the items would be difficult for students. Lack of familiarity with the style of items made it essential to make the instructions for how to approach the tasks more explicit. In addition, some tasks that students were required to undertake, such as “planning,” were not familiar. Although the teachers saw the items as appropriately reflective of the actual skills, they were concerned about the language facility required to engage with the assessment tasks.

# PILOT

The purpose of the pilot was two-fold. First it provided information on the feasibility of administration of these types of tasks at the class level across the three countries. Second, it provided student response data to enable analysis of the quality of the tasks and their items.

## Usability issues

Assessments in many classrooms rely on students working individually and answering either closed response or short constructed open response test items. Accordingly, an area of exploration relevant to the OAA project is how both teachers and students respond to differently presented tasks—tasks that require application of understanding and skills as opposed to items that target content knowledge, multi-step tasks, and multi-format tasks. Another area of interest is how teachers and students respond to tasks that value collaboration, rather than individual effort only. Since many of the tasks were unfamiliar in format and form, usability issues included how well the instructions would be understood by the students, the time it would take to complete each task in a classroom environment, and how manageable small groups might be in a formal collaborative assessment scenario.

*Students at the Mahendra Gram Secondary School in Kathmandu engaging in collaborative assessment tasks*

**Nepal's pilot**

Nepal initially selected four schools for the program. From these schools, twelve teachers participated in activities, including the first three-country collaborative workshop, as well as paneling, item development, and item adaptation. Teachers became familiar with the skills of critical thinking, problem solving, and collaboration.

For the pilot, another four schools were drawn upon to achieve the required number of students for each task. In working with these new schools, the NTT faced new challenges orienting them to the nature of the tasks. For example, where teachers saw item content that reflected unfamiliar knowledge, such as the map of Cambodia, they queried how the item could be appropriate. Similarly, the philosophy of engaging in collaborative tasks was novel, both for the students and the teachers. This also related to the logistic challenges of implementing and engaging in collaboration, particularly with given timeframes. Teachers were not accustomed to managing the seating arrangements in the class or having to manage activities simultaneously across groups of students. In Nepali classrooms, assessment practices do not cover group work. Therefore, it was a strange experience for the students, who have only experienced being assessed individually, and essentially in competition with each other.

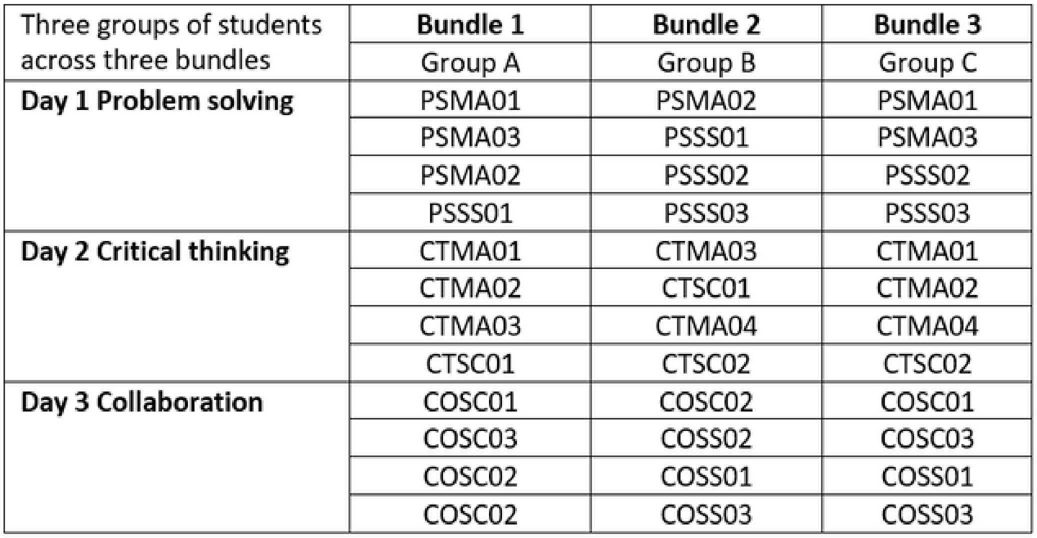
Students are familiar with textbook content and requirements. The demands of the critical thinking and problem solving tasks were very different from their textbook experiences, and they were not accustomed to thinking beyond what was presented. Notwithstanding instructions from the teacher to think beyond one's specific subject-based knowledge, very few students tried to solve, or think through, the questions. in the longer term, the teachers need to learn how to introduce questions differently from the textbook, provide time to students to reflect on their learning, and to practice collaboration in school.

There were difficulties in implementation of the collaborative tasks. One source of confusion emanated from the use of two different modalities in the collaborative task. Parts of the task required students to work independently and parts together. This was difficult for students and teachers to understand and manage. In addition, some collaboration tasks needed three students in one group, while for another four students were needed.

Beyond the school and classroom logistics, the NTT itself learned a great deal through the pilot process. The concept of bundling the tasks to collect sufficient data for the analyses was new to many and needed to be better understood to communicate to the schools' personnel (Table 6).

Notwithstanding this, the pilot was implemented successfully over three days. The target was reached with 306 students completing the tests, and consistent with the plan: a minimum 200 students sat for each task for each of the three bundles.

Table 6. Identification of specific items in the bundles



| Three groups of students across three bundles | **Bundle 1** | **Bundle 2** | **Bundle 3** |
|---|---|---|---|
| | Group A | Group B | Group C |
| **Day 1 Problem solving** | PSMA01 | PSMA02 | PSMA01 |
| | PSMA03 | PSSS01 | PSMA03 |
| | PSMA02 | PSSS02 | PSSS02 |
| | PSSS01 | PSSS03 | PSSS03 |
| **Day 2 Critical thinking** | CTMA01 | CTMA03 | CTMA01 |
| | CTMA02 | CTSC01 | CTMA02 |
| | CTMA03 | CTMA04 | CTMA04 |
| | CTSC01 | CTSC02 | CTSC02 |
| **Day 3 Collaboration** | COSC01 | COSC02 | COSC01 |
| | COSC03 | COSS02 | COSC03 |
| | COSC02 | COSS01 | COSS01 |
| | COSC02 | COSS03 | COSS03 |

*Key. PSMA01=Problem Solving, Math task #1; CTSC01=Critical Thinking, Science task #1; COSS01=Collaboration, Social Science task #1, etc.*

### Mongolia’s pilot

Twelve classes across four schools in Mongolia participated in the pilot: These included urban government, urban private, provincial government, and “soum” (or village) government schools. The class sizes varied from 24 students to 55, while the schools themselves were medium to large and representative of average to high academic performance and socio-economic status. The pilot was undertaken in the few weeks prior to the start of summer vacation (in the second half of May) in Mongolia. However, schools also have national examinations in late May and so logistics were difficult. The NTT monitored the sessions, and classroom teachers worked as test administrators in classes other than their own.

For the problem solving assessment, the time outlined in the field operation guide was strictly followed for the first school. The students completed the actual tasks in less time than estimated—about 30-45 minutes. For the next school, the amount of time different students needed covered a large range, and resulted in some students talking to each other, looking at others’ papers, and even changing their answers. Although the expected behaviors were explained to students, it was difficult for Grade 5 students to sit silent, and for large classes of 52-55 students, it was extremely hard to manage the classroom. Students in rural schools and the second urban school with small class sizes were very different.

Students in rural schools and the second urban school with small class sizes were very different. They obeyed teachers and could sit silent when they finished their work early. In typical Mongolian classrooms, two students share a desk, but a lesson learned was that students need to sit alone at their desk.

For the collaboration assessments, different challenges were encountered. Although class sizes are sometimes large with 52-55 students in one class, the actual physical space is not that big. This had implications for the collaboration tasks due to the difficulty in moving desks and chairs. From the perspective of the actual demands of the collaboration tasks, students were typically not able to finish the tasks within given times due to their lack of familiarity with the assessment style, as well as the tasks themselves. The pilot experience was salutary. It made clear the need for review of classroom management, familiarization of students, and comprehensive logistics protocols.

### Cambodia’s pilot

Cambodia used a cluster sampling approach to select participants. This allowed researchers to identify specific schools and test all students in those schools. Three classes from two schools in Phnom Penh and two classes from two schools in Kandal province were selected, providing a total of ten classes from the four target schools with about 380 students participating.

Each student was assigned tasks across all three skills but completed different bundles of tasks through use of a "rank function" in Microsoft Excel. This guaranteed that each class had access to the same bundles but that students within the classes were assigned different bundles. The approach reduced the possibility of students copying from each other. Test administration across the four schools was scheduled based on the class shifts (morning or afternoon). Over the three days of testing, the first day was given to critical thinking, the second to problem solving, and the third to collaboration. During the in-country workshops and meetings, the teachers recognized that most of the content used in the items and tasks reflected the students' daily life. However, as the actual pilot event loomed closer, many teachers were concerned about the capacity of their students to complete the tasks, especially for collaboration. These concerns were based on the lack of familiarity with the practices reflected in the tasks, and lack of experience with the different format of testing.

In the collaboration session, students were set to work in a group of three or four based on the topics required. The leftover student(s) were assigned to observe the group discussion. Since these students seemed to be working less than their peers, some discouraged students wished to be an observer for the next task. Also, the need to rearrange groups due to the task requirements presented both logistic, and later scoring, challenges.

The collaboration tasks raised many issues. The situation was tough for teachers, when they recognized that some of their outstanding students were experiencing difficulties, and sometimes being grouped with poorer performing students. It was also observed that some students were not happy to collaborate with others in their groups—they wished to be with their close friends. Moreover, there were influential members within groups, so especially a student who took a note taker role would contribute greater substance, often without consultation. Finally, there were groups that missed an opportunity to present, listen, and comment to the other groups, as required by the tasks, due to lack of task completion. These issues highlighted the very real challenges that will face teachers designing and encouraging collaborative skills in the classroom.

## Data cleaning, verification, and scoring

Each NTT was responsible for ensuring that scored student responses were accurately entered into databases. Instructions for identification of missing data provided guidance on how to deal with lack of responses on test items in order to ensure that all countries followed the same protocols. Scoring procedures were similarly outlined so that errors in coding of student responses would be minimized. Each country adapted the instructions and guidelines according to the structures of their teams and available personnel.

Each NTT appointed a data manager who was responsible for the overall process, including checking student lists and completed test forms. For scoring, the NTT had prepared a set of rubrics based on a booklet designed in Khmer language. Before coding, a technical training session focused on scoring procedures/rubrics and identified all possible scores for each item. Scorers worked on sample responses as a practice, checking agreement with their counterparts.

All countries used the OAA data entry Excel spreadsheet and followed instructions for entry of data. Use of Excel for data entry was a concern for the Cambodian NTT. Accustomed to working with secure forms in Windem, the team was concerned about mistakes in data entry, as well as loss of data due to unstable electricity supplies and out-of-date computer operating systems. However, use of Excel for OAA had been decided upon due to its ubiquity and sustainability of practices given that no additional, expensive software would need to be purchased by participating countries.

The instructions were explicit concerning use of separate tabs in spreadsheets, and maintenance of separate records for double scoring. For Cambodia, the data cleaning process for critical thinking and problem solving went smoothly. Similar to issues experienced by Mongolia and Nepal, the process for collaboration was problematic. The task design where different numbers of students worked on some tasks complicated the identification of group membership at scoring and data entry processing stages.

Nepal allocated eleven personnel across the full data scoring and data entry tasks that each focused on just one skill for one subject. The scoring took place over one working week, with two personnel over an additional week's timeframe for the data entry. The Nepal NTT cross-checked data entry, and a 5% sample was cross-verified. Double scoring was not undertaken.

For both Nepal and Mongolia personnel, data entry provided the opportunity for capacity development. They had not previously used such data entry templates, and it became clear during the process that the level of expertise required had been underestimated. This meant that the amount of time dedicated to the process was greater than had been anticipated.

## Scoring reliability

Since the task formats and forms were not familiar to those responsible for the scoring, of particular interest was how the scoring guides and associated rubrics worked. A double-scoring activity was undertaken for selected tasks, in order to provide data on inter-rater reliability. This also provided for partial investigation of average rater quality within each country. Each country approached this issue in slightly different ways.

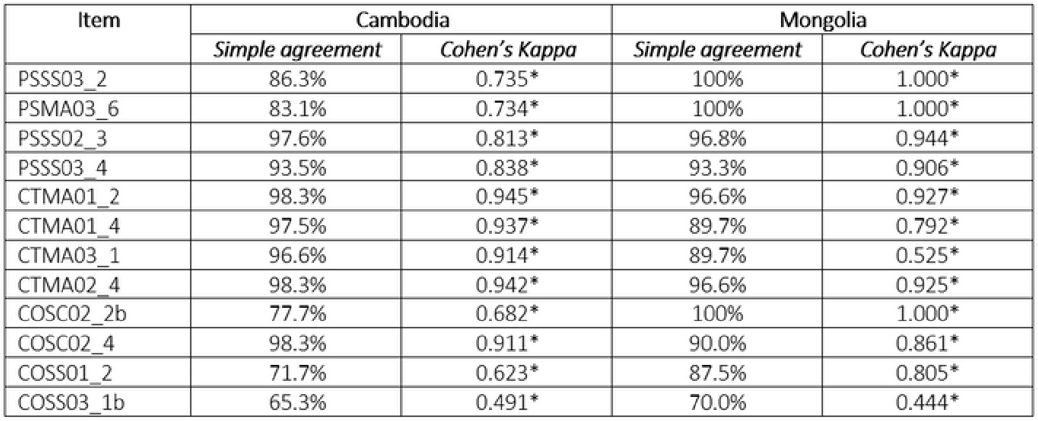
Cambodia's procedure for double scoring was conducted in two steps. The first scorer coded directly into the scoring worksheet, and the second scorer coded directly onto the student response booklet. The consistency of coding was first manually checked and then indices of simple agreement were used calculated through Excel. Where disagreement was found, both scorers came together to resolve the discrepancy. Where the two scorers were not able to reach a solution, the issue was brought to the attention of the data manager to make the final judgment.

As a reliability check for Mongolia’s scoring, the data from one school and one bundle were selected for double scoring. These data amounted to around 10 percent of the total sample for the selected tasks.

The Nepali NTT checked approximately 10 percent of the response sheets for double scoring. Due to lack of significant inconsistencies, a double scoring dataset was not created. The consistency of scoring was attributed to strict use of the scoring rubrics and guidelines.

Double-scored data by Cambodia and Mongolia were used to evaluate rater agreement on selected partial credit items. The double-scoring process was conducted by having two independent raters separately score the same set of randomly selected tasks using the same rubrics. The interrater analyses were conducted using the irr package (Gamer, et al., 2019).

Table 7. Inter-rater agreement results for selected items



| Item | Cambodia | | Mongolia | |
|---|---|---|---|---|
| | *Simple agreement* | *Cohen’s Kappa* | *Simple agreement* | *Cohen’s Kappa* |
| PSSS03_2 | 86.3% | 0.735* | 100% | 1.000* |
| PSMA03_6 | 83.1% | 0.734* | 100% | 1.000* |
| PSSS02_3 | 97.6% | 0.813* | 96.8% | 0.944* |
| PSSS03_4 | 93.5% | 0.838* | 93.3% | 0.906* |
| CTMA01_2 | 98.3% | 0.945* | 96.6% | 0.927* |
| CTMA01_4 | 97.5% | 0.937* | 89.7% | 0.792* |
| CTMA03_1 | 96.6% | 0.914* | 89.7% | 0.525* |
| CTMA02_4 | 98.3% | 0.942* | 96.6% | 0.925* |
| COSC02_2b | 77.7% | 0.682* | 100% | 1.000* |
| COSC02_4 | 98.3% | 0.911* | 90.0% | 0.861* |
| COSS01_2 | 71.7% | 0.623* | 87.5% | 0.805* |
| COSS03_1b | 65.3% | 0.491* | 70.0% | 0.444* |

*$p < .01$*

The results provided evidence on the quality of raters and allowed for partial investigation of what might affect rater quality. Two metrics were used to evaluate rater agreement:

1) **Simple agreement,** or the percent of scores that perfectly agree between the two raters (i.e., the number of cases where the scores given by both raters are exactly matching, divided by the total number of cases).

2) **Cohen's Kappa** (1960) for interrater agreement between two raters.

The results show very high agreement across the selected items for both Cambodian and Mongolian raters, with all Kappa values highly significant at a = .01. Although the agreement metrics show high agreement (Table 7), the results also highlight the relative challenges of scoring tasks that assess collaboration. For these tasks, it may be that more subjective judgment is used to score the responses, hence perfect agreement is relatively more difficult to achieve.

The inter-rater reliability check was useful in two ways. First it established the relative robustness of the scoring procedures. Second, it identified which tasks were subject to more error than others, so providing information about what types of tasks and items, and what types of scoring rubrics are most problematic. This information is invaluable for how countries might continue to approach future task development and scoring protocols.

The differences in agreement across Cambodia and Mongolia are assumed to be attributable to the slightly smaller number of scorers fielded by the Mongolia NTT. However, additional checking for possible error due to translation of the actual items, and the scoring rubrics across the two countries, is warranted. This possibility is fueled by noting the differences in reliabilities for two of the critical thinking math tasks and two of the problem solving tasks across the two countries.

The tasks that were subject to more error than others tended to be the collaboration tasks. The rubrics for these tasks may well have been less familiar to the scorers than those used for the cognitive tasks, as represented by problem solving and critical thinking.

## Quality of the assessment tasks

The majority of tasks and their items worked well. They mapped onto the relevant skill, and levels of competence across coding categories were discriminable. The collaboration skill was more problematic than either problem solving or critical thinking. To a large degree this issue was due to discriminability between proximate coding categories: In other words, differences between a code of '1' and a code of '2' did not capture differences in quality reliably. This issue suggests the need to review some scoring rubrics. Use of rubrics designed to estimate quality of performance rather than correct or incorrect responses was unfamiliar to the NTTs and their participating teachers. Also, there were differences in how some rubrics functioned across the three countries, suggesting that the issue could be either in translation or in scoring reliability for particular items. To explore this issue, analyses were undertaken by the countries, as well as at the aggregate level.

These analyses provided the information needed for each country to review their translations and review their coding records.

As suspected by some of the teachers, the overall level of difficulty of the tasks was beyond the capabilities of the Grade 5 and 6 students. Given that the tasks were based on the curriculum, it might be assumed that the difficulty would lie with the skills variance in the tasks rather than the curricular knowledge. However, not all students were at the same point across the three countries in their curricular learning and so equal progress against curricular goals cannot be assumed. Notwithstanding this, the unfamiliarity of the task design, as well as the skills variance, is presumed to have contributed to the overall difficulty levels. Figures 4 through 6 provide an illustration of the capabilities of the students in the context of the demands of the tasks.

In each of the figures, the distribution of students is on the left-hand side of the graph, and on the right, the item numbers and coding according to the partial credit models. The clustering of items at the top of each of the three graphs, demonstrates that these are beyond the ability of the students, while toward the bottom of the graphs, the lack of items indicates more items are needed to provide information about student capabilities at the less proficient levels. This is precisely the information that was being sought through the pilot procedure. Not only does the data reveal what items work well, they also tell us what an appropriate range of difficulty would be for students in Grades 5 and 6. Using this information, the NTTs can make decisions about developing more lower difficulty items for Grades 5 and 6, as well as the possibility of trying out the items with students in the higher grades.

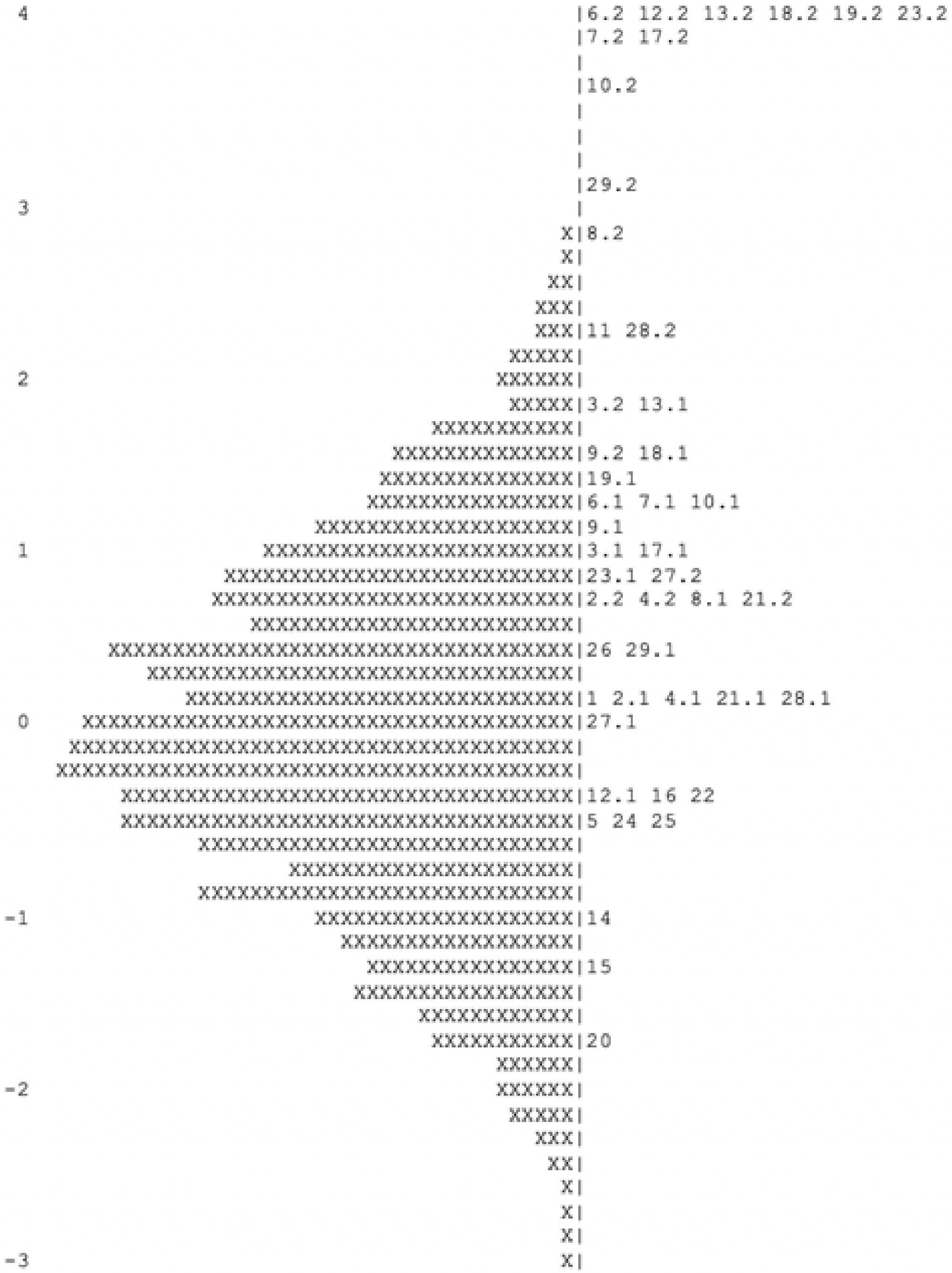
**How to interpret the item-person maps?**

The left side of the map shows a distribution of "Xs" that represent student abilities (each X = 1.5 students in this sample), ranging from low estimated person ability at the bottom of the scale to high ability at the top of the scale. Along the left-hand side of the figure is the "logit" scale—the numeric equivalents of the graphed locations.

The right side of the maps display items that represent score steps for items that can be responded to at different levels of quality. For example, in Figure 4, item #29.1 (at a logit of about 0.45) represents the lower level of quality response for item #29, while item #29.2 (at a logit of about 3.0) represents a higher level of quality response. Students represented by 'X' at the same horizontal level on the graph as the item have a 50% chance of getting the item correct, and of course a higher probability of getting items correct that are located further down the figure.

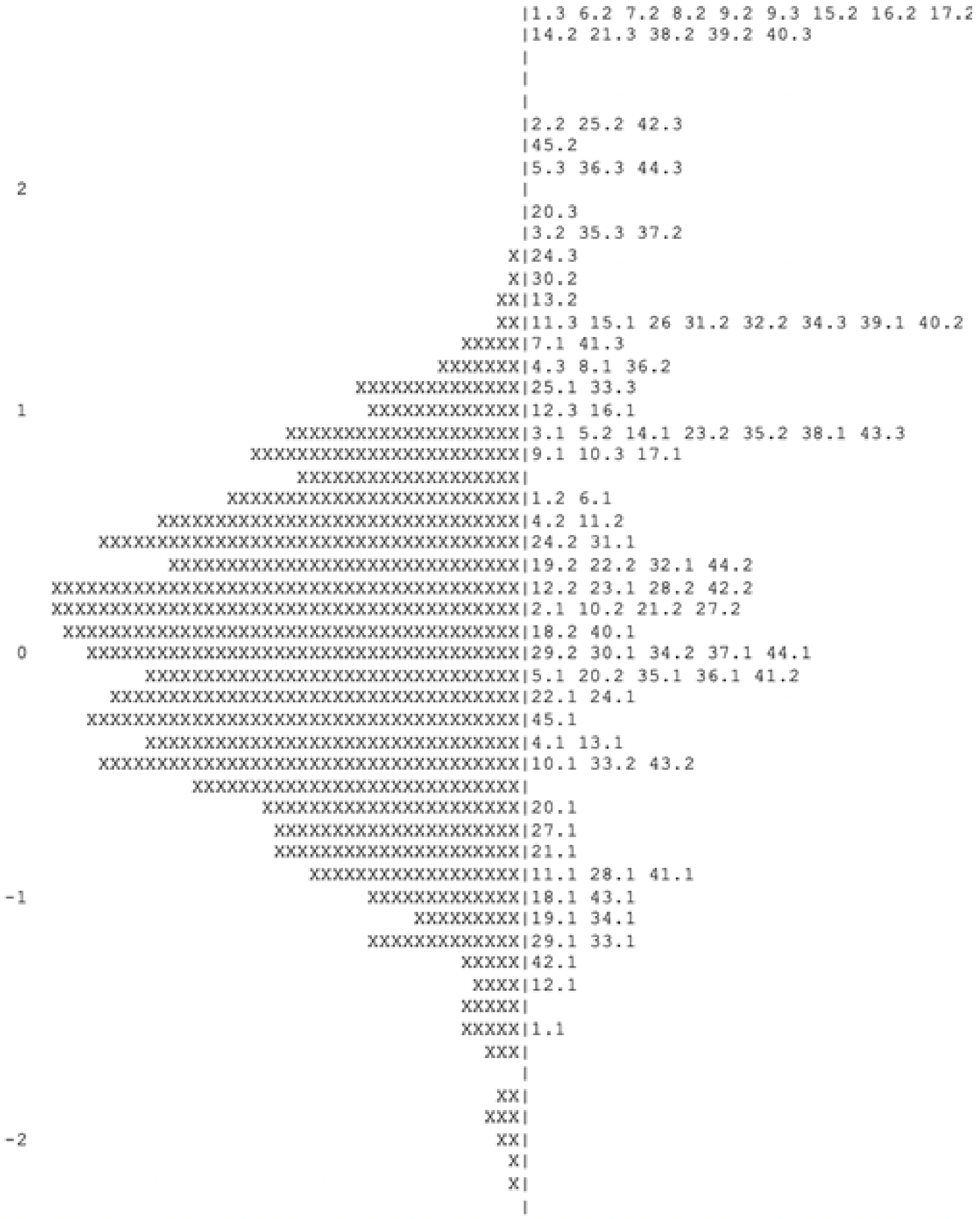
Therefore, score steps that require only low levels of proficiency are shown lower on the scale, and a progressively higher location on the scale is associated with higher difficulty.

Figure 4. Item-person map for critical thinking (each “X” represents 1.5 cases)



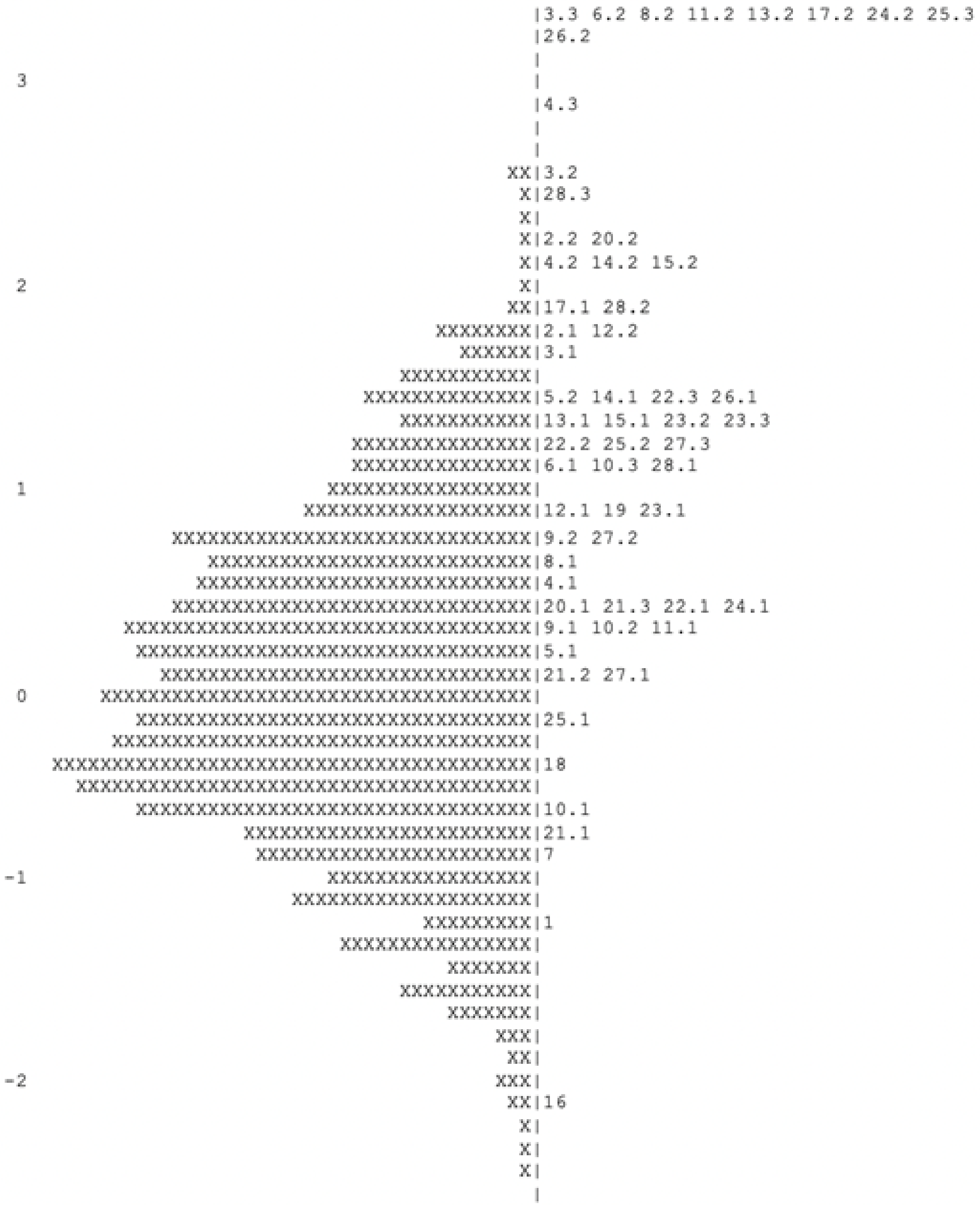
```
 4                                          |6.2 12.2 13.2 18.2 19.2 23.2
                                            |7.2 17.2
                                            |
                                            |10.2
                                            |
                                            |
                                            |
                                            |29.2
 3                                          |
                                           X|8.2
                                           X|
                                          XX|
                                         XXX|
                                         XXX|11 28.2
                                       XXXXX|
 2                                    XXXXXX|
                                       XXXXX|3.2 13.1
                                 XXXXXXXXXXX|
                              XXXXXXXXXXXXXX|9.2 18.1
                             XXXXXXXXXXXXXXX|19.1
                            XXXXXXXXXXXXXXXX|6.1 7.1 10.1
                        XXXXXXXXXXXXXXXXXXXX|9.1
 1                  XXXXXXXXXXXXXXXXXXXXXXXX|3.1 17.1
                 XXXXXXXXXXXXXXXXXXXXXXXXXXX|23.1 27.2
                XXXXXXXXXXXXXXXXXXXXXXXXXXXX|2.2 4.2 8.1 21.2
                   XXXXXXXXXXXXXXXXXXXXXXXXX|
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|26 29.1
           XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
              XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|1 2.1 4.1 21.1 28.1
 0    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|27.1
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
         XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|12.1 16 22
         XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|5 24 25
               XXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
                      XXXXXXXXXXXXXXXXXXXXXX|
               XXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
-1                      XXXXXXXXXXXXXXXXXXXX|14
                          XXXXXXXXXXXXXXXXXX|
                            XXXXXXXXXXXXXXXX|15
                           XXXXXXXXXXXXXXXXX|
                                XXXXXXXXXXXX|
                                 XXXXXXXXXXX|20
                                      XXXXXX|
-2                                    XXXXXX|
                                       XXXXX|
                                         XXX|
                                          XX|
                                           X|
                                           X|
                                           X|
-3                                         X|
```

Figure 5. Item-person map for collaboration (each “X” represents 1.5 cases)



```
                                            |1.3 6.2 7.2 8.2 9.2 9.3 15.2 16.2 17.2
                                            |14.2 21.3 38.2 39.2 40.3
                                            |
                                            |
                                            |
                                            |2.2 25.2 42.3
                                            |45.2
                                            |5.3 36.3 44.3
 2                                          |
                                            |20.3
                                            |3.2 35.3 37.2
                                           X|24.3
                                           X|30.2
                                          XX|13.2
                                          XX|11.3 15.1 26 31.2 32.2 34.3 39.1 40.2
                                       XXXXX|7.1 41.3
                                     XXXXXXX|4.3 8.1 36.2
                              XXXXXXXXXXXXXX|25.1 33.3
 1                             XXXXXXXXXXXXX|12.3 16.1
                        XXXXXXXXXXXXXXXXXXXX|3.1 5.2 14.1 23.2 35.2 38.1 43.3
                     XXXXXXXXXXXXXXXXXXXXXXX|9.1 10.3 17.1
                         XXXXXXXXXXXXXXXXXXX|
                   XXXXXXXXXXXXXXXXXXXXXXXXX|1.2 6.1
             XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|4.2 11.2
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|24.2 31.1
              XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|19.2 22.2 32.1 44.2
    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|12.2 23.1 28.2 42.2
    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|2.1 10.2 21.2 27.2
     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|18.2 40.1
 0     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|29.2 30.1 34.2 37.1 44.1
            XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|5.1 20.2 35.1 36.1 41.2
         XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|22.1 24.1
       XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|45.1
            XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|4.1 13.1
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|10.1 33.2 43.2
                XXXXXXXXXXXXXXXXXXXXXXXXXXXX|
                      XXXXXXXXXXXXXXXXXXXXXX|20.1
                       XXXXXXXXXXXXXXXXXXXXX|27.1
                       XXXXXXXXXXXXXXXXXXXXX|21.1
                          XXXXXXXXXXXXXXXXXX|11.1 28.1 41.1
-1                             XXXXXXXXXXXXX|18.1 43.1
                                   XXXXXXXXX|19.1 34.1
                               XXXXXXXXXXXXX|29.1 33.1
                                       XXXXX|42.1
                                        XXXX|12.1
                                       XXXXX|
                                       XXXXX|1.1
                                         XXX|
                                            |
                                          XX|
                                         XXX|
-2                                        XX|
                                           X|
                                           X|
                                            |
```

Figure 6. Item-person map for problem solving (each “X” represents 1.5 cases)



```
                                               |3.3 6.2 8.2 11.2 13.2 17.2 24.2 25.3
                                               |26.2
                                               |
    3                                          |
                                               |4.3
                                               |
                                               |
                                             XX|3.2
                                              X|28.3
                                              X|
                                              X|2.2 20.2
                                              X|4.2 14.2 15.2
    2                                         X|
                                             XX|17.1 28.2
                                       XXXXXXXX|2.1 12.2
                                         XXXXXX|3.1
                                    XXXXXXXXXXX|
                                 XXXXXXXXXXXXXX|5.2 14.1 22.3 26.1
                                    XXXXXXXXXXX|13.1 15.1 23.2 23.3
                                XXXXXXXXXXXXXXX|22.2 25.2 27.3
                                XXXXXXXXXXXXXXX|6.1 10.3 28.1
    1                         XXXXXXXXXXXXXXXXX|
                            XXXXXXXXXXXXXXXXXXX|12.1 19 23.1
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|9.2 27.2
                    XXXXXXXXXXXXXXXXXXXXXXXXXXX|8.1
                   XXXXXXXXXXXXXXXXXXXXXXXXXXXX|4.1
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|20.1 21.3 22.1 24.1
             XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|9.1 10.2 11.1
              XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|5.1
                XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|21.2 27.1
    0      XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
              XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|25.1
            XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
       XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|18
         XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
              XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|10.1
                       XXXXXXXXXXXXXXXXXXXXXXXX|21.1
                        XXXXXXXXXXXXXXXXXXXXXXX|7
   -1                         XXXXXXXXXXXXXXXXX|
                           XXXXXXXXXXXXXXXXXXXX|
                                      XXXXXXXXX|1
                               XXXXXXXXXXXXXXXX|
                                        XXXXXXX|
                                    XXXXXXXXXXX|
                                        XXXXXXX|
                                            XXX|
                                             XX|
   -2                                       XXX|
                                             XX|16
                                              X|
                                              X|
                                              X|
                                               |
```

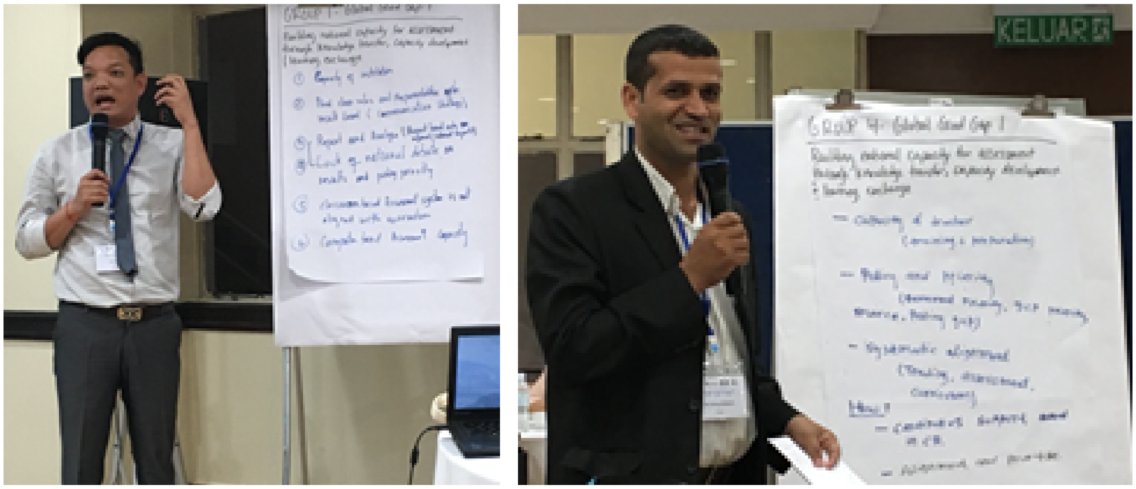
# CONCLUSION

From the country reports of the pilot processes, it becomes clear that task development is not the only or greatest challenge for the introduction of 21CS assessment. Aside from the matter of national team readiness and preparation for the pilot, the realities of assessing unfamiliar competencies in unfamiliar ways in the classroom raise both basic logistical and educational philosophy issues. In this report, the individual experiences of the three countries have been described by their teams to ensure transparency about the issues experienced, as well as their achievements. Teaching and assessment culture and how 21st century approaches to teaching and learning have impact beyond pedagogical norms remains an outstanding issue to address.

The complexity of the task development, the logistics of the pilot, and the coding and data entry phases comprise the main learnings of the project for the NTTs. Frequently assessment studies focus primarily on the tools or tests.

For the OAA project, the primary focus was on the process: How do you develop tasks that reflect current curricula and that integrate 21CS into teaching and learning practices? Of course, the quality of the tasks themselves is a major factor that will enable the countries to integrate 21CS into teaching and learning practices.

As a Nepali teacher stated in the first OAA workshop in Kathmandu, after working through descriptions and definitions of the skills, said "but this means we need to change the way we teach." It is clear that some re-thinking of the classroom culture is necessary for some teachers and education systems. Concerns expressed during the pilot about students cheating highlight some of the complexities associated with collaborative work, about not always prioritizing correct answers, and about building knowledge together—as opposed to competitive systems in which being ranked provides benefits to some and deprives others of opportunities.



*OAA national team representatives from Cambodia and Nepal sharing and leading in NEQMAP regional convening in Penang, Malaysia 2018*

The OAA initiative took just one step in the process of integration of 21CS. Acting as a lever, the development and introduction of assessment tasks in the classroom is a disruptive force. The challenge is how to deal with that force adaptively—and that is the next and current focus of the national teams. For descriptive information about the actual assessment tasks, and how Cambodian, Mongolian and Nepali student responses to these clarify aspects of skills that students find both easier and more difficult, see forthcoming report "Guide to OAA Assessments."

Introduction of assessment of competencies to be taught, prior to that teaching, may appear to be putting the cart before the horse. However, this approach has demonstrated to the NTTs that the way we assess has significant implications for the way we teach and can raise awareness of alternative approaches. If 21CS are to be integrated into school systems, the consequences for both teaching and assessment are considerable. Cambodia, Mongolia, and Nepal have seen this first-hand, and are therefore better prepared to scale teaching and assessment of 21CS in the classroom.

United Nations Educational, Scientific and Cultural Organization
Sustainable Development Goals
NEQMAP
Network on Education Quality Monitoring in the Asia-Pacific
GLOBAL PARTNERSHIP for EDUCATION
KICE
NEQMAP's 6th Annual Meeting
Opening the dialogue: Promoting inclusion and equity through quality assessment
13-14 DECEMBER 2018
BANGKOK, THAILAND

*NEQMAP supports OAA through promoting the* importance of regional relationships and common understandings

# REFERENCES

Care, E., & Luo, R. (2016) *Assessment of transversal competencies: Policy and practice in the Asia-Pacific region*. Bangkok, Thailand: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000246590.

Trilling, B., & Fadel, C. (2009). 21st century skills: Learning for life in our times. John Wiley & Sons.

UNESCO (2015). *Transversal competencies in education policy and practice (Phase I): Regional synthesis report.* Paris, France: UNESCO. http://unesdoc.unesco.org/images/0023/002319/231907E.pdf.

UNESCO (2016a). *School and teaching practices for twenty-first century challenges: Lessons from the Asia-Pacific region (Phase II):* Regional synthesis report. Paris, France: UNESCO. http://unesdoc.unesco.org/images/0024/002440/244022E.pdf.

UNESCO (2016b). *Preparing and supporting teachers in the Asia-Pacific to meet the challenges of twenty-first century learning (Phase III): Regional synthesis report.* Paris, France: UNESCO. http://unesdoc.unesco.org/images/0024/002468/246852E.pdf.

UNESCO (2018). A*ssessment of transversal competencies: Current tools in the Asian region.* Paris, France: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000368479.