

OPTIMIZING ASSESSMENT FOR ALL  
JUNE 2020

# OPTIMIZING ASSESSMENT FOR ALL

## Developing 21st Century Skills-Embedded Curriculum Tasks



BROOKINGS



## Authors

Esther Care is a senior 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.

Aynur Gul Sahin was a Fulbright scholar completing a post degree placement at the Brookings Institution.

Optimizing Assessment for All (OAA) is a project of the Brookings Institution. The aim of OAA is to support countries to improve the assessment, teaching, and learning of 21st century skills by 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 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.

In addition, Brookings gratefully acknowledges the support provided by Porticus.

Brookings recognizes that the value it provides is in its absolute commitment to quality, independence, and impact. Activities supported by its donors reflect this commitment and the analysis and recommendations are not determined or influenced by any donation.

Cover photo credit: Erdmiin Dalai Complex School, Dundgovi, Mongolia



## TABLE OF CONTENTS

**04 INTRODUCTION**

**06 DEVELOPMENT OF ASSESSMENTS**

06 DRAFT TASKS

07 THINK ALOUD

08 Revise

08 PANEL

09 Revise

10 SCORING

13 PILOT

14 Cultural differences

14 Translation issues

14 Task and item quality

15 Quality of scoring rubrics

**15 DISCUSSION**

17 REFERENCES

18 APPENDICES

19 Appendix A: Critical thinking - Shop earnings

25 Appendix B: Collaboration - Separating mixtures

31 Appendix C: Collaboration (and problem solving) - Serrekunda

34 Appendix D: Problem solving - Lemon tree



## INTRODUCTION

This fourth Optimizing Assessment for All (OAA) report describes the process of test and item development undertaken by Cambodia, Mongolia, and Nepal in OAA Asia and by the Democratic Republic of Congo, The Gambia, and Zambia in OAA Africa. As the fourth in the OAA series, the report synthesizes the two slightly different approaches taken in the two regions.

OAA itself was designed to explore how to assess 21st century skills (21CS) while also situating this exploration in a philosophy that uses assessment as a constructive tool for teaching and for exploring how the integration of 21CS in classrooms can deepen learning. This report outlines the main assessment task development and refinement processes, in addition to a focus on scoring protocols, and is complemented by a set of appendixes that provide examples of the iterative processes through which OAA assessment tasks and their items were increasingly refined. The report is written to provide guidance that can be used or adapted for further development of 21CS items.

United Nations Sustainable Development Goal 4 (SDG 4) is to “ensure inclusive and equitable quality education and promote lifelong learning opportunities for all.” To be successful in today’s world, students need to have a broad set of knowledge, skills, habits, and traits that go beyond what has traditionally been taught in the classroom, such as rote memorization-based skills. These broad sets of competencies—such as critical thinking, collaboration, problem solving, technology and information skills, and communication, among others—are often referred to as 21st century skills (21CS).

Despite broad agreement that schools should adopt an agenda for teaching 21CS, there is still a great deal of debate about 21CS—from which skills are most important to how such skills should be taught and assessed. Given a lack of clear consensus on the most effective ways of teaching and assessing these skills, educators are challenged in adapting the learning and teaching paradigm of 21CS to the classroom.

OAA is a project designed to improve the teaching and learning of 21CS in classroom settings through development of expertise in the assessment of these skills. The three main goals of the project include:

- Supporting regional and national stakeholders in measuring learning across 21CS
- Demonstrating how to align the assessment of skills with their inclusion in the written and taught curriculum
- Strengthening education systems’ capacity to integrate 21CS into their teaching and learning.

For more information about the OAA project, see the first report in this series ([Care & Kim, 2020](#)).

The purpose of this report is to help educators adapt and develop new ways of assessing 21CS in their classrooms. The report provides information about the process followed in the OAA project and includes examples of assessment strategies for capturing 21CS skills, specifically problem solving, collaboration, and critical thinking.



The report includes assessment tasks developed by the three focus countries in Asia (Cambodia, Mongolia, and Nepal) and the three in Africa (the Democratic Republic of Congo, The Gambia, and Zambia), providing a step-by-step guide to the development and adaptation of 21CS assessment tasks.

As nations worldwide look to integrate a breadth of skills perspective into their curricula—and into teaching and learning—the need to specify learning outcomes and performance standards for these skills has become apparent. Assessment is key not only to these specifications but also to teaching of the skills. Twenty-first century skills are transferable skills, recognized as difficult to measure. For educators who have focused for many years on mathematics, languages, sciences, and social sciences, the integration of both cognitive and social skills into their teaching and assessment constitutes a challenge. The approach described here draws together known types of assessment in the classroom together with the process required for tackling these hard-to-measure competencies.

The task development processes, undertaken by national teams across Asia and Africa, focused on assessment of collaborative, critical thinking, and problem solving competencies of students enrolled across the final years of primary school and early years of secondary school. None of the countries that engaged in the work yet have comprehensive definitions and descriptions of the skills. Hence the development of assessments was preceded by formulation of definitions and descriptions that provided the structure against which assessment tasks could be created. Nor have the countries yet formally specified performance outcomes for these competencies. This situation reflects the reality for most countries—where, because 21CS are so recently integrated into formal education, learning progressions that describe the skills are not yet developed. Accordingly, setting the difficulty level of the assessment tasks was achieved through the national teams and their teacher participants estimating likely level of student competency. The accuracy of these estimates can be seen in analysis of the results from the pilot assessment programs of the tasks in both the OAA Asia and OAA Africa regions. These are included in two preceding reports in this series ([Care, Vista, & Kim, 2020](#); [Kim & Care, 2020](#)).

Figure 1. Task development steps





## DEVELOPMENT OF ASSESSMENTS

The OAA project undertook assessment development slightly differently across its two regions of Asia and Africa. However, for the purposes of this report, Figure 1 depicts these differences synthesized into one combined set of steps: (1) draft tasks; (2) conduct think aloud sessions, and revise; (3) panel the tasks, and revise; and (4) pilot the tasks.

### DRAFT TASKS

This step includes four activities:

- Identify, define, and deconstruct the targeted skills
- Identify existing assessment tasks or develop new ones that can 'capture' 21CS
- Adapt or modify these tasks to explicitly target 21CS
- Identify which skills, strands, and subskills the task and its items are targeting.

**Identify, define, and deconstruct skills.** When drafting an assessment task, the first step is to identify the skill of interest. Its definition and description are needed. Many 21CS are complex skills; they have several contributing subskills. Therefore, the descriptions need to clarify these subskills so they can be reflected in the assessment tasks.

The test and scale development literature refers to these subskills and their components in several ways. In this report, complex skills are 'deconstructed' from the overarching level of 'skill' into strands, subskills, and sub-subskills. This makes it possible to draft tasks with items that target different parts of the complex skill.

**Identify, develop, or adapt assessment tasks to target 21CS.** Next, rather than starting from scratch to develop completely new items that target 21CS, existing assessment tools can be reviewed to examine whether they could be modified to assess 21CS, taking into account the specific strands and subskills of the targeted skill to be measured. When modifying the assessments, a few points should be considered in terms of quality. Specifically, a high-quality assessment has:

- A clear intention
- Language understood by most students
- A simple and authentic context;
- One correct answer or a set of acceptable responses
- Good psychometric properties.

**Identify the skills, strands, and subskills each task should target.**

Finally, once the assessment tasks have been drafted, the specific skills, strands, and subskills are identified for each task. The assessments described in this report consist of what is referred to as a 'task', each consisting of several items and based in one of several subject areas: mathematics, science, or social sciences. So the curricular subject matter provides the substance or knowledge upon which the students can exercise their skills.

For a robust, large-scale assessment program, a blueprint of the skills with their strands and subskills, and of the curricular content, would be drawn up to frame the development of the assessments.



In the case of the OAA program, with its exploratory approach to task development, the blueprint was less stringent. It identified which skills to target across which subject areas by the selected grade levels. The number of tasks and items which targeted specific strands and subskills was counted after final sets of assessment tasks were agreed upon.

The OAA project acknowledged that different countries will define and describe 21CS in different ways. A vast research literature has accumulated over the past few decades (e.g., Care, Griffin & Wilson, 2018; Jones, Bailey, Brush, Nelson & Barnes, 2016), but it is by no means definitive. How 21CS are perceived within different cultures and nations is acknowledged to be central to how teaching and learning are delivered, and these perceptions will frame how the skills are embedded in curricula, how they are taught, and how they are assessed. Testament to these differences are the slightly different component structures for skills identified by OAA Asia and OAA Africa. See, for example, the different structures for collaboration across Appendixes B and C. These naturally have implications for task creation and design.

To seed the drafting of tasks, two approaches were adopted. One was to take already existing assessments and adapt these to integrate 21CS; the other was to audit curricula for topics common across participating countries and select topics from these from which to create assessment tasks. Regardless of the approach adopted, the following task development process was followed. The appendix to this report provides worked examples of task adaptation to integrate 21CS.

## THINK ALOUD

Once a task has been drafted, there is a need to check whether the intended skills, with their strands and subskills, are actually being prompted by the task. 'Think aloud' (also known as cognitive laboratory) is a method of studying how a student responds to a task; whether the task appears to elicit the intended knowledge, skills, or competencies; and whether it is appropriate in terms of factors such as gender, culture, and expected performance range.

In this case, the think aloud process was developed to check whether tasks elicited cognitive and social processes from the students that were consistent with the intended skills, strands, and subskills. During this process, a student completes a task while orally reporting his or her mental processes by explaining his or her thinking and reasoning. The stream of consciousness is recorded and used to evaluate whether the targeted skills are in fact prompted by the tasks.

In addition to general observations as students are completing the task, the following questions can guide the think aloud sessions and identify any issues with the task and its items:

- Does the task or item appear to test the targeted skill?
- Does the task or item appear to test the targeted strand or subskill?
- Is the difficulty level specified in the task or item appropriate for the targeted students?
- Is the time allotted for the task or item appropriate?
- Do students know exactly what they are being asked to produce?
- Are there other kinds of answers that have not been catered for in the scoring guide?



To check whether tasks are equally accessible for all students, several students from across the ability range should participate in the process. Where the process is undertaken to check on collaboration skills, both homogeneous and heterogeneous groups (in terms of ability) should participate.

## Revise

Think aloud activities provide teachers and task developers with information about how students approach a task and any functional issues that need to be addressed in revising the task. Based on this empirical evidence, revisions are made in this task development step. The evidence typically highlights administration, content, and reporting issues.

**Administration:** Students' difficulties as they engage with tasks, but which they subsequently solve, are made explicit through this process. This helps the task developers to consider what additional scaffolds (or hints) might be required in task instructions to support less able students. These immediate difficulties encountered by students can be used to ensure that administration instructions are adequate.

**Content:** Tasks that are developed to assess 21CS use subject-matter information as a base. It is important that students be familiar with this information before they can use or manipulate it. The think aloud results help to clarify whether the concepts or information reflected in the tasks have been covered in the delivered curriculum.

**Reporting:** Students may respond and react to tasks in ways not envisaged by the task developers.

These responses provide not only valuable clues about how students interpret the tasks and task instructions but also insights into student capacities that had not been considered in the original scoring categories. Tasks that stimulate responses that can be scored at different levels of quality provide a rich source of information that can inform teaching strategies.

## PANEL

Once revisions are made based on the think aloud step, the revised test items go through another round of verification, called 'paneling'. Paneling is a collaborative process to review the substantive and technical aspects of assessment tasks. It serves as a quality control check in the development process by examining the materials' validity and contextual appropriateness (including issues of bias). Paneling focuses on the task and items as presented, to determine whether they sample the skills at the levels intended; whether they are likely to discriminate between students of different ability; and whether, in combination with other tasks, the target construct will be adequately sampled.

An independent review panel consists of subject experts, assessment experts, and language experts as well as teachers of the students for whom the task is designed. A major feature of a well-selected panel is that the skills and experiences panel members bring are complementary rather than duplicated. At least one representative of the task creator is part of the panel. This representative acts as a resource for the panel—that is, to provide explanations as required, not to defend or advocate for the materials. The task, and any associated material, should be presented such that it can stand on its own.





During the paneling sessions, the panel reviews all parts of the assessment material in detail, including the stimulus, the questions posed, the materials' layout and presentation, and the associated scoring guides and criteria. The purpose of this process is to ensure that administration, completion, and scoring of test items could be undertaken effectively and efficiently. The panel participants' roles and responsibilities are outlined below.

A panel convenor:

- Collects and distributes the assessment materials to be examined
- Ensures that security is maintained
- Leads and monitors the discussion
- Ensures that the purpose of the test development sets the parameters of the discussion
- Summarizes what needs to be done in revising materials.

Individual panel members:

- Work through the materials, from the perspectives of both test developer and test taker, by working through the tasks or items, answering the questions, and reviewing the scoring criteria
- Thoroughly critique the items
- Exercise critical thinking and problem solving
- Identify weaknesses in tasks, where these exist, and identify alternatives
- Write comments that are a "snapshot" of their own thinking uninfluenced by others.

The item writer, upon request:

- Articulates reasons for selecting the stimulus or content and for creating the format or structure of the items
- Articulates the specific skills or content that the item is targeting.

The panel's final process is to summarize whether a task or item:

- Is acceptable as it is
- Should be modified in ways that address the flaws or problems identified
- Is unacceptable.

## Revise

After the paneling concludes, the panel's responses regarding the quality of the items should be consolidated and used to make any necessary changes before finalizing the items. Revision of the assessment materials continues until the panel convenor is satisfied that all major issues have been addressed. This does not necessarily mean that consensus will be reached on every issue, but all viewpoints and opinions should be given careful consideration.

It is important to keep in mind that finalizing the materials at this stage is not actually the last step of test development. That step occurs through examination of the process and outputs from the pilot of assessments.



## SCORING

Scoring is an intrinsic part of the development of assessment materials—and a design component that must be considered both early in task creation and throughout the process. Initial scoring rules are typically amended after the think aloud activities and again after piloting of assessments, based on the evidence presented through student responses.



Where an item within a task provides for an incontestably correct or incorrect answer, scoring is simple. However, where this is not the case, scoring rubrics are provided. A scoring rubric defines what is expected in terms of a response and how it should be interpreted. A good scoring rubric describes the criteria to be used in evaluating a response, with performance descriptors that are comprehensive and mutually exclusive—enabling the scorer to identify different levels of response quality (Figure 2).

For constructed response items (that is, those requiring a free-form response from the student) rubrics are particularly important because levels of quality are typically more difficult to identify or quantify. As an example of clear, if minimalist, rubrics for such a situation, see a task (D8) developed by the Democratic Republic of Congo team to sample problem solving competencies in the social sciences (Figure 3).

Figure 2. Scoring elements






Figure 3. Problem solving task D8 in the social sciences with generic rubrics

<p>In some cities and rural areas, when heavy rain falls, there is flooding. Houses and streets are flooded, and there are significant consequences.</p>	<p><b>D8a IG-Oi-Des</b>            0 = No response            1 = Irrelevant response            2 = Provides a possible cause            3 = Provides two possible causes</p>
<p>a) What are the causes of floods in cities and rural areas? Identify at least two.</p>	<p><b>D8b IG-Oi-Des</b>            0 = No response            1 = Irrelevant response            2 = Provides one consequence            3 = Provides two consequences</p>
<p>b) What are the consequences of flooding? Identify at least two.</p> <p>c) Using one of the causes and consequences you have identified, what is a possible solution?</p>	<p><b>D8c PS-Ge-Hyp</b>            0 = No response            1 = Irrelevant response            2 = Provides a possible solution but is not related to one of the causes and consequences identified            3 = Provides a possible solution that is related to one of the causes or consequences identified            4 = Provides a possible solution that is related to one of the causes and consequences identified</p>

Key. IG = Information Gathering; Oi = Organise information; Des = Describe; PS = Planning a Solution; Ge = Generate ideas; Hyp = Hypothesize

The rubrics provide skeletal but adequate guidance for a scorer to identify whether a student has addressed the question in terms of the skills aspect. The rubrics do not cover other aspects of student competence such as quality of writing or even the quality of the ideas; they cover whether causes have been identified, whether associations have been made, and whether solutions have been provided. 


For scoring for large-scale data analysis, these rubrics may be adequate. At the classroom level, however, the teacher may wish to include additional scoring criteria that address the quality of the social science knowledge that the student brings to the problem or the student's quality of literacy. So, in this Figure 3 example, the rubrics provide adequate guidance to identify whether the required skills processes have been undertaken, through the product evidence; they do not, however, provide guidance about how to estimate the product quality. Note that the zero scoring category in the figure was used only by OAA Africa owing to their perspective that effort should be rewarded. That such effort is not directly associated with the target of the assessment is an issue at the psychometric level.

Another example, Figure 4, illustrates the provision of guidance to scorers by listing acceptable answers as well as providing the generic guidelines. This task from OAA Asia, and also piloted by OAA Africa, was designed to capture problem solving skills in a mathematics context. The figure includes just two items that are based on this task.

A critical issue made explicit by these student response examples is the interplay of the problem solving skills, the mathematics knowledge and skill, and the literacy capabilities. This interplay is challenging for the scorer because the level of performance in any one of these three areas affects the student's ability to display competence in other areas.



Figure 4. Problem solving task A1 in mathematics with generic rubrics and examples to meet scoring criteria for two of six items

<p>A farmer has to fill a new fishpond with water.</p>  <p>a) What information does the farmer need to fill the fishpond? For example, you need to know where the water comes from. Write three other useful pieces of information.</p>	<p><b>A1 a) IG</b></p> <p>Score 0 = No response          Score 1 = Gives one or no relevant factors          Score 2 = Gives two different factors related to filling the fish pond          Score 3 = Gives three different factors related to filling the fish pond</p> <p><b>Examples of useful information:</b>  <i>How to get the water into the pond (from the source) e.g. hose, pipe, dig a channel etc</i>  <i>How large the pond is (or how much water it holds)</i>  <i>How much water is already in it (or how much more it needs)</i></p> <ul style="list-style-type: none"> <li>• <i>How long it will take to fill</i></li> <li>• <i>How much it will cost to fill</i></li> <li>• <i>Whether it needs to be kept full</i></li> <li>• <i>How much water will drain away through the walls and bottom of the pond</i></li> <li>• <i>How the fish will get into the pond (not directly related to filling it, but an important point)</i></li> </ul>
<p>b) A hose can pump 300 litres of water a minute. How many litres of water does this hose pump in half an hour?</p> <p>A. 10 L          B. 150 L          C. 9,000 L          D. 18,000 L</p> <p>c) The farmer finds another hose and uses it for 5 minutes. It pumped about 1,000 litres. Is this hose faster than the first hose (300 litres per minute)? Give a reason for your answer.</p> <p>d) The fishpond needs 36,000 litres of water to fill it. Using the hose that pumps 300 litres per minute, how many hours will it take to fill the pond?</p> <p>e) The farmer wants to put 1,000 fish in the pond, but he thinks that the pond might overflow. What should he do?</p>	<p><b>A1 d) IG-Oi-Ana/Des</b></p> <p>Score 0 = No response          Score 1 = Incorrect or insufficient response</p> <ul style="list-style-type: none"> <li>• <i>Yes, 1,000 is more than 300 (incorrect answer)</i></li> <li>• <i>No, it is not faster (correct answer, but no reason given)</i></li> </ul> <p>Score 2 = Explains why it is NOT faster, showing that it has a lower flow rate</p> <ul style="list-style-type: none"> <li>• <i>1,000 litres in 5 minutes is the same as 200 litres in 1 minute, which is less than the first hose</i></li> <li>• <i>The first hose can do 1,500 litres in 5 minutes, so it is faster</i></li> <li>• <i>No. <math>5 \times 300 = 1,500</math> is more than 1,000</i></li> </ul>

Key. IG = Information Gathering; Oi = Organise information; Ana/Des = Analyze/Describe



This reflects the reality of life, of course, where it is rare that demonstration of skill is not affected by other factors, and is a major justification for both teaching and assessment of 21CS within the context of subject-area studies.



In summary, questions such as the following can help in reviewing scoring rubrics:

- Are the performance descriptors clear?
  - Would students be able to self-assess by having the descriptors?
  - Would the descriptors give students enough information to know how to improve?
- Does the rubric reflect the range of levels at which students may perform the tasks?
- Are the criteria sufficiently defined to ensure that scoring is accurate, unbiased, and consistent?
- Does the rubric reflect process, product, or both?

## PILOT

Pilots of the assessment tasks were undertaken in both Asia and Africa. Most of the tasks administered within a region were developed within that region, although OAA Africa also administered three of the OAA Asia tasks. The numbers of students providing results varied: each OAA Asia country tested over 300 students, and the OAA Africa countries tested numbers varying from 120 to 209 to over 300. The student responses from these pilots were analyzed (with the quantitative results reported in the OAA Asia and OAA Africa reports within this series)([Care, Vista, & Kim, 2020](#); [Kim & Care, 2020](#)).

The pilot activity was designed to identify whether the tasks and their items functioned as expected. In other words, the function of the pilot is to check the quality of the assessments, not the performance of the students. Analysis of many students' responses to the assessment tasks provides information about whether the tasks sample across a range of abilities, whether they might be too difficult or too easy for students, and whether they target groups of the same competencies or skills. Where analyses show that students respond in unexpected ways to certain tasks or items, this provides the test developers with clues about whether the tasks might have been designed poorly, whether the scoring rubrics might have been unclear or applied randomly, or whether data entry and management might have been insufficiently standard.

Of particular interest for OAA, given its focus on assessing hard-to-measure constructs, was the adequacy of the scoring rubrics applied to student responses. This interest was due largely to the acknowledged lack of familiarity on the part of both teachers and students with the targeted skills and the modes of assessment employed. This lack of familiarity was a continuing theme that emerged from teacher comments across the six countries, and it was highlighted through both the engagement and occasional hesitation of students. For those teachers who participated in the scoring processes, therefore, the task was unlike their previous scoring experiences. The student response examples provided in Figure 5 were derived from the pilot assessments.

Where task- and item-level data were anomalous, the stimulus materials themselves and the scoring rubrics were reviewed. Possible reasons for anomalies include cultural differences; translation issues; and quality issues with tasks, items, or scoring rubrics.



## Cultural differences

The OAA Asia tasks were developed and aligned by the three countries, the design being that all countries would pilot all tasks. Therefore, in the development process, the countries had audited their curricula to ensure that students across each of the countries would all have been presented with the same subject-content learning opportunities. Despite this overall assurance of commonality, the final task versions were modified slightly to ensure that students in each country could access the tasks on similar terms—requiring use of different words for particular items, different currencies, and different naming conventions.

For OAA Africa, the design did not assume that all subject content would be similarly accessible across the three countries or that all countries would administer all the same tasks. In fact, however, all three countries did administer tasks developed both originally by their own teams as well as those developed by others. The three national teams had collaborated across all items throughout the process, so again there had already been some moderation of tasks that made them more accessible than otherwise would have been the case.

During OAA Africa's task development process, there was considerable discussion of country phenomena that were true for all countries—issues such as crowding, flooding, and disease, all of which formed subject matter for the tasks in the social sciences. How different countries might respond to such issues, and how their understandings of these might vary depending on the severity of the issues, could well affect student responses.

There is little doubt that attitudes and values could influence views on acceptable and unacceptable responses to some of these tasks and their items.

This highlights a further issue in assessment of these hard-to-measure constructs, where the competence demonstrated reflects not only knowledge and skill but also values or attitudes. Hence, where pilot response data show anomalies from just one country, one factor to consider is whether cultural issues may have affected student performance.

## Translation issues

Tasks were translated from the original language of development into each of the languages of instruction for OAA Asia: Khmer, Nepali, and Mongolian. Where a task was first drafted in one of these languages, it was translated first to English for moderation and then into languages of instruction.

For OAA Africa, the tasks were provided in English and French, with English being the language of instruction for The Gambia and Zambia. No back translations were undertaken, and there were likely some infelicities of expression in different versions. These may have influenced student understanding and interpretation of tasks and their items. Similar to the cultural issues, where pilot response data show anomalies from just one country, one factor to consider is whether language or translation issues may affect student performance.

## Task and item quality

Where tasks and their items do not appear to function in the way expected—in other words, if the most competent students respond inaccurately or inappropriately to the simplest items, or if the least able students respond accurately to items presumed to be the most difficult—this could signal problems with the tasks or items themselves. Most of the tasks drafted by OAA team included a relatively simple subject-oriented item as the first item.



The subsequent items rested more heavily on the skills being targeted, although still with the need for subject content to be accessible to the student. Hence, if those simplest items were not responded to correctly, it signaled either a flaw in the stimulus material for the task or in the wording of the item itself. The response data for a task can be analyzed to identify whether the stimulus material (that is, the whole task base) is problematic or whether there are problems isolated to particular items.

### Quality of scoring rubrics

The issues identified above about potential problems with tasks and items are difficult to distinguish from problems with the scoring rubrics. When the student response data are anomalous, each of the reasons above needs to be examined. Where anomalies are isolated to a particular item rather than many items associated with a task, then the process is to analyze the quality of the item and the scoring rubric attached to it. The large-scale pilot data point to the problem but cannot precisely identify that problem. And this is what takes us to the qualitative analysis of student responses, as shown in Figure 5.

For OAA Africa, responses such as those in Figure 5 were used to review the rubrics used for pilot scoring. These responses are from Congolese and Zambian students. Typically some 15 responses or so to a specific item were analyzed, with just four being shown in the figure here. Initially the analysts sort the responses from least to most proficient by referring back to the generic rubrics. The sorting provides a loose progression of very low to high competence. Where responses are not matched neatly by rubrics, those rubrics are then further refined. In Figure 5, as can be seen, all materials were provided in both English and French, with the national teams operating across both languages.

Figure 5. Student responses to task A1, Item(c)

(d) The farmer finds another hose and uses it for 5 minutes. It pumped about 1,000 litres. Is this hose faster than the first hose (300 litres per minute)? Give a reason for you answer.

*my reason is that the first one is faster than the second because it pumped 300 litres and second it pumped 3000*

d) L'agriculteur trouve un autre tuyau et l'utilise pendant 5 minutes. Il a pompé environ 1 000 litres. Ce tuyau est-il plus rapide que le premier (300 litres par minute)? Donnez une raison pour votre réponse.

*Ce premier est plus rapide* The first is fastest.

(d) The farmer finds another hose and uses it for 5 minutes. It pumped about 1,000 litres. Is this hose faster than the first hose (300 litres per minute)? Give a reason for you answer.

*Il a pompé en 5 minutes 1.000 litres  
300 litres en 2 minutes*

d) L'agriculteur trouve un autre tuyau et l'utilise pendant 5 minutes. Il a pompé environ 1 000 litres. Ce tuyau est-il plus rapide que le premier (300 litres par minute)? Donnez une raison pour votre réponse.

*Non, parce que le premier pompe 300 litres en 5 minutes.*

No, because the first pumps 300 liter in 5 minutes.

## DISCUSSION

To generate high-quality assessments that target 21CS, it is important to make sure that tasks and their items are capturing performances that rely on the skills targeted. Since assessment of the skills is embedded in subject-specific content, there is potential for students to have difficulties, either because they lack the relevant subject knowledge or because they cannot exercise the required skills.

This difficulty is not always easily addressed, but there are solutions: First, if the same skills are required across many tasks that rely on different subject areas, analysis of results over all assessments can identify patterns of achievement. Second, if tasks are created using subject content that is known to have been achieved by all students, then variable performance across students can be attributed to what is not known—the skill differential.



These solutions notwithstanding, for teacher development of assessment tasks for classroom use, large-scale results analyses are not available. This means that local test developers and teachers are best advised to monitor the levels of subject matter that are assumed to be familiar and well understood by their students. As a result, there is less space to misinterpret why students are performing at different levels on tasks that capture demonstration of both skills and curriculum knowledge.

Twenty-first century skills include a wide range of competencies, covering both cognitive and social aspects of our performance and behavior. These skills vary in the degree to which they can be integrated into formal education and classroom practice. Some of the skills are typically demonstrated through media that are not compatible with easy capture—unlike skills like literacy, which are relatively easy to capture through the written word. OAA took traditional assessment practices and adapted these to the demands of a small number of skills. There are limitations to this approach, especially for some of the social 21CS. The collaboration tasks created in OAA were more difficult to administer than the problem solving and critical thinking tasks, and caused some confusion for students, in part because of greater familiarity with traditional competitive practices that have characterized assessment as distinct from the collaboration embedded in 21CS assessment tasks. Capture of whether processes were implemented, capture of whether all group members contributed, and capture of product are all viable in the classroom. The nuance of human dynamics that contributes so much to social skills, however, is less viable—particularly in large classes.

How comprehensively can we measure 21CS through approaches such as those outlined here? And why do we want to measure them?

What this report describes is a beginning. It is possible to define and describe 21CS, and it is possible to create subject-specific tasks that elicit and promote these skills. Measuring the skills provides teachers with information about student competencies that can then be used to nurture and develop these further. What needs to be kept in mind is that we practice and display these skills every day, in non-technological as well as technology-embedded environments. The approach described in this report may not appear exciting to a reader who anticipates that new learning goals will be matched by technologically complex modalities, but the daily human-to-human processes are as necessary to capture and assess in order to improve as are those that are digital.

The worked examples provided in the appendixes to the report illustrate approaches that can and need to be built upon. The approach assumes that skills frameworks have been developed to support creation of assessments. Beyond using frameworks such as those described in the appendixes, test developers need to develop knowledge, and then guidelines, about how each 21CS of interest progresses from simple to more complex. This will enable creation of better assessments that target wider ranges of performance and enable teachers to better estimate student performance.





## REFERENCES

- Care, E., Vista, A., & Kim, H. (2020). *Optimizing Assessment for All: Focus on Asia*. Washington D.C.: The Brookings Institution.  
<https://www.brookings.edu/research/optimizing-assessment-for-all-focus-on-asia/>
- Care, E., & Kim, H. (2020). *Optimizing Assessment for All: Framework for understanding project goals and scope*. Washington\_D.C.: The Brookings Institution.  
<https://www.brookings.edu/research/optimizing-assessment-for-all-framework-for-understanding-project-goals-and-scope/>
- Care, E., Griffin, P., & Wilson, M. (Eds.) (2018). *Assessment and Teaching of 21st Century Skills: Research and Applications*. Cham: Springer.
- Jones, S., Bailey, R., Brush, K., Nelson, B., & Barnes, S. (2016). *What is the same and what is different? Making sense of the "non-cognitive domain: Helping educators translate research into practice*. Easel Lab: Harvard.  
[https://easel.gse.harvard.edu/files/gse-easel-lab/files/words\\_matter\\_paper.pdf](https://easel.gse.harvard.edu/files/gse-easel-lab/files/words_matter_paper.pdf).
- Kim, H., & Care, E. (2020). *Optimizing Assessment for All: Classroom-based assessments of 21st century skills in the Democratic Republic of Congo, The Gambia, and Zambia*. Washington D.C.: The Brookings Institution.  
<https://www.brookings.edu/research/optimizing-assessment-for-all-classroom-based-assessments-of-21st-century-skills-in-the-democratic-republic-of-congo-the-gambia-and-zambia/>



## APPENDICES

The purpose of these appendixes is to provide guidance for developing assessments of 21st century skills. Items from the OAA project are used as examples to explain each step of the task and item development process as described in this report. One or two examples are provided for each skill: critical thinking, problem solving, and collaboration. The adaptation of the tasks to integrate 21st century skills rests on the adoption of a conceptual framework or structure of the targeted skill. These conceptual frameworks are drawn either from the OAA Asia or OAA Africa studies. The skill structures used as frameworks for the worked examples are at the beginning of each appendix.

Appendix A to Appendix E provide worked examples of task revisioning. Each example takes an original task that was designed for classroom use within the context of subject studies - such as mathematics, social science, or science. The example is refined to extend its capacity to the capture 21st century skills. One complex skillset is targeted for each task, although in some cases, the task also includes aspects of a second skillset.

Actual changes that are documented to the tasks are accompanied by annotations that make clear what the changes are, and why they were made. Changes in task versions associated with each of the steps are shown in red font with scoring changes or skills information shown in red italics.

For some of the examples, the iterations in task design and content are organized within the processes of (1) drafting tasks, (2) the think aloud step, and (3) paneling. Further refinement of tasks would also typically be made after the piloting of assessment tasks, where required. These are not included in this publication.

- The majority of tasks were initially conceived in a language other than English; the iterations were made in a combination of English and languages of the participating countries. In some cases this led to some infelicities of expression in English - addressing these has not been a focus of the work reported here.
- Some of the changes made to tasks consist of providing more scaffolding to test-takers. These changes have taken the shape of shorter sentences of instruction, more specificity, and providing response spaces or structures for students. In principle such changes do not provide the reader with any new insights, but because these changes did occur in the processes, some are reported here.
- The changes can be classified across: greater clarity and formatting; amendments to stimulus to stimulate the targeted skillset; amendments to individual items within a task to target specific subskills of the targeted skillset; development of rubrics to provide teachers and students with deeper understanding of the cognitive and social processes that the tasks are intended to stimulate.

APPENDIX	SKILL	NAME	SUBJECT AREA
A	Critical thinking	Shop earning	Mathematics
B	Collaboration	Separate mixes	Science
C	Problem solving	Rice yield	Social sciences
D	Collaboration	Serrekunda	Social sciences
E	Problem solving	Lemon tree	Science



## Appendix A: Critical thinking - Shop earnings

### DRAFT TASKS

Critical thinking was defined to describe its structure and its contributing strands and subskills. The process was informed by a summary of the research literature available on the topic. For the purposes of the OAA project, critical thinking was defined as analyzing information objectively and make a reasoned judgment through the synthesis of information using logical reasoning. The OAA Asia structure identified four strands, linked with behavioral indicators for the associated subskills (Table A1).

Table A1. OAA Asia structure for critical thinking

Strands	Subskills
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

Figure A1

**Stimulus:** The table below shows the earnings of a shop for a week.

Day	Total earning
Monday	\$370
Tuesday	\$410
Wednesday	\$325
Thursday	\$360
Friday	\$450
Saturday	\$320
Sunday	-

**Q1:** How much did the shop earn on Monday?

- A. \$410
- B. \$320
- C. \$370
- D. \$450

Answer: C

**Q2:** Which day has the most earnings?

Answer: Friday

Once critical thinking was defined, an existing task which had the potential to target critical thinking skills was identified. Figure A1 shows this classroom mathematics task. In its current form, the task does not require students to engage 21CS to respond. For example, to answer Question 1 (Q1), students need to look at the table in the stimulus and find the total earning amount listed next to Monday. The answer is provided in the table, and students need only some information literacy skills to access it. The same is true for Q2.



Figure A2

**Stimulus:** A shop has a number of goods for sale including juice, cake and cookie. The table below show a number sold for a week (a juice cost \$3, one cake is \$2, cookie is \$1).

Day	Juice	Cake	Cookie	Total earning
Monday	100	10	50	\$370
Tuesday	95	15	95	\$410
Wednesday	80	20	45	\$325
Thursday	80	20	45	\$360
Friday	100	25	100	\$450
Saturday	80	15	50	\$320
Sunday	90	40	52	-

**Q1:** How much did the shop earn on Monday from cake?

- A. \$10
- B. \$20
- C. \$100
- D. \$200

*[Strand: Information management]*

*[Subskill: Collate information]*

*Answer: B*

**Q2:** On Monday what percent of juice is sold of a week?

- A. 6.25%
- B. 16%
- C. 62.5%
- D. 100%

*[Strand: Information management]*

*[Subskill: Analyse information, synthesises information]*

*Answer: B*

**Q3:** On Sunday, how much does the shop earn in total?

*[Strand: Information management]*

*[Subskill: analyze information]*

*Answer: 402*

**Q4:** If the shop wants to close one day a week, which day is the best choice for the owner?

*[Strand: Judgement]*

*[Subskill: Evaluate the source, compare and contrast]*

*Answer is Saturday*

**Q5:** Wednesday is discount for Cookie of 20% and Sunday for Cake discount of 20%. You want to buy one cake and one cookie. Which day do you want to go?

*[Strand: Judgement]*

*[Subskill: Compare and contrast]*

*Answer: Sunday*

Figure A2 shows how the task can be modified to target critical thinking skills.



First, the stimulus has been changed to include categories of sales (Juice, Cake, Cookie) in the table and provide detail which makes the information more complex. Second, the existing questions have been amended so that they require the identified subskills to answer. Third, new questions have been added. Finally, each question now includes information about which specific strands and subskills are targeted. This can be useful for teachers in identifying which processes students are having difficulty with, in order to provide the necessary support.

Figure A2 Q1 is similar to Figure A1 Q1, but the slight changes in inserting "from cake" to the question, as well as the complexity added to the stimulus, requires the student to engage information collation subskills as well as information management. Q1 still includes mathematics content knowledge, but in addition the student needs to identify what is relevant, in order to generate an answer which is not explicit in the table.

Q2 and Q3 are similar to Q1 in that they require manipulation of the information found in the stimulus.

Q4 and Q5 include the Judgement strand by asking students to make choices based on comparisons.



**Q6:** Explain your answer in Q5.

[Strand: Judgement]  
[Subskill: Justify]

**Q7:** A student has \$10. Name two ways you could spend your money. You want to have at least one item from each category.

[Strand: Argumentation, information management]  
[Subskill: Identifying alternatives, analyze and synthesize information]

Q6 asks students to explain and justify their choice and so brings metacognitive skills into play.



Q7 requires students to engage in the multiple subskills listed.

## THINK ALOUD

Because there were significant modifications to the original item (Figure A1), teachers conducted think aloud activities with a few more students to examine whether the modified items elicit the critical thinking strands and subskills that were identified, as well as re-checking the appropriateness of the tasks for the targeted grade level and subject areas.

In general, feedback from the think aloud activities confirmed that the tasks were targeting the intended strands and subskills. Some items, especially Q2 and Q5, were too difficult for students across all ability levels. In some cases, the students were unable to answer the questions correctly because a particular topic area (e.g., unit of percentage or analysis of table data) was not yet taught. Students were also unfamiliar with some question formats, such as Q6, where they were asked to provide explanations. Finally, students did not understand what was being asked of them for Q7.

Based on the empirical evidence from the think aloud activities, the task was revised.

**Figure A3**

**Stimulus:** A shop has a number of goods for sale including juice, cake and cookie. The table below show a number sold for a week (a juice cost \$3, one cake is \$2, cookie is \$1).

Day	Juice	Cake	Cookie	Total earning
Monday	100	10	50	\$370
Tuesday	95	15	95	\$410
Wednesday	80	20	45	\$325
Thursday	80	20	45	\$360
Friday	100	25	100	\$450
Saturday	80	15	50	\$320
Sunday	90	40	52	-

In this Figure A3 version, the allocated time and the level of complexity (simple, average, and difficult) for each test item were added as a general guide for teachers. Information about possible answers was added or expanded for the constructed response questions to account for the multiple ways of responding (e.g., Q3 and Q4).





- Q1:** How much did the shop earn on Monday from cake?
- A. \$10
  - B. \$20
  - C. \$100
  - D. \$200

[Strand: Information management]  
 [Subskill: Collate information]  
 [Complexity: Simple]  
 [Time: 10mn]  
 Answer: B

- Q2:** On Monday what percent of juice is sold of a week?
- A. 6.25%
  - B. 16%
  - C. 62.5%
  - D. 100%

[Strand: Information management]  
 [Subskill: Analyse information, synthesises information]  
 [Complexity: average]  
 [Time: 15mn]  
 Answer: B

- Q3:** On Sunday, how much does the shop earn in total?

[Strand: Information management]  
 [Subskill: analyze information]  
 [Complexity: average]  
 [Time: 10mn]  
 The possible answer is:

- 402
- $(90 \times 3) + (40 \times 2) + (52 \times 1) = 402$
- Four hundred and two

- Q4:** If the shop wants to close one day a week, which day is the best choice for the owner?

[Strand: Judgement]  
 [Subskill: Evaluate the source, compare and contrast]  
 [Complexity: average]  
 [Time: 05 mn]  
 The possible answer is:

- Saturday
- \$320
- Saturday, \$320

- Q5a:** Wednesday is discount for Cookie of 20% and Sunday for Cake discount of 20%. You want to buy one cake and one cookie. Which day do you want to go?

[Strand: Judgement]  
 [Subskill: Compare and contrast]  
 [Complexity: difficult]  
 [Time: 15mn]  
 Answer: Sunday

- Q5b:** Explain your answer in Q5.

[Strand: Judgement]  
 [Subskill: Justify]  
 [Complexity: difficult]  
 [Time: 10 mn]  
 Possible answer:

- Purchase them on Sunday; spends less money than on Wednesday because the cake is more expensive than cookie
- Wednesday ( $\$2 + \$0.8 = \$2.8$ ) > Sunday ( $\$1.6 + \$1 = \$2.6$ ) so, purchasing on Sunday is spent less
- Students choose Sunday because of spending less money Students do not choose Wednesday because of spending more money

- Q6:** A student has \$10. Name two ways you could spend your money. You want to have at least one item from each category.

[Strand: Argumentation, information management]  
 [Subskill: Identifying alternatives, analyze and synthesize information]  
 [Complexity: difficult]  
 [Time: 10 mn]  
 Possible answers:

- $1j + 1ca + 5co$
- $1j + 2ca + 3co$
- $1j + 3ca + 1co$
- $2j + 1ca + 2co$

In Figure A2, Q5 and Q6 were separate test items. However, because of item dependency (i.e., students need to respond to Q5 correctly in order to respond to Q6), these two test items were amalgamated to be parts of a single item (Q5a and Q5b).





The revised task was then reviewed in a paneling session. During the paneling session, panel members (e.g., subject experts, assessment experts, and teachers) discussed each question and provided feedback. Based on the feedback from paneling session, the task went through a final round of revision.

Figure A4

**Stimulus:** A shop has a number of goods for sale including juice, cake and cookie. The table below shows the number sold for a week (a juice cost Z 3, one cake is Z 2, cookie is Z 1).

Day	Juice	Cake	Cookie	Total earning
Monday	100	10	50	Z370
Tuesday	95	15	95	Z410
Wednesday	80	20	45	Z325
Thursday	80	20	45	Z360
Friday	100	25	100	Z450
Saturday	80	15	50	Z320
Sunday	90	40	52	-

**Q1:** How much did the shop earn on Monday from cake?

- A. Z10
- B. Z20
- C. Z145
- D. Z375

[Strand: Information management]

[Subskill: Collate information]

[Complexity: Simple]

Answer: B

[Time: 10mn]

**Q2:** On Monday what percent of juice is sold of a week?

- A. 6.25%
- B. 16%
- C. 62.5%
- D. 100%

[Strand: Information management]

[Subskill: Analyse information, synthesises information]

[Complexity: average]

[Time: 15mn]

Answer: B

**Q3:** On Sunday, how much does the shop earn in total?

[Strand: Information management]

[Subskill: analyze information]

[Complexity: average]

[Time: 10mn]

The possible answer is:

- 402
- $(90 \times 3) + (40 \times 2) + (52 \times 1) = 402$
- Four hundred and two

In Figure A4, each country tailored aspects of the task. For example, the currency unit in the stimulus was revised with local currency to make the task more relevant to local daily life.

The response options C and D for Q1 were modified to reflect typical errors. The reasoning behind the options:

- A. Incorrectly using the number of cakes sold instead of earnings
- B. Correct answer
- C. Incorrectly sums number of cakes sold for the week
- D. Incorrectly uses the total earnings for Monday

Q2 and Q3 were removed. According to the paneling sessions, these items were not seen as appropriate for the grade level, nor reflective of the targeted construct.



**Q4:** If the shop wants to close one day a week, which day provides the most choice for the owner? **Explain why you choose that day.**

[Strand: Judgement]

[Subskill: Evaluate the source, compare and contrast]

[Complexity: average]

[Time: 05 mn]

The possible answer is:

- Saturday
- \$320
- Saturday, \$320

**Scoring for Q4:**

Score = 1 if explanation is incorrect or no explanation is given.

Score = 2 if answer and explanation are both correct.

**Q5a)** On Wednesdays, there is a half-price discount for Cookies. On Sundays, there is a half-price discount for Cakes. You want to buy one cake and one cookie. Which day do you want to go? Explain your answer.

[Strand: Judgement]

[Subskill: Compare and contrast]

[Complexity: difficult]

[Time: 15mn]

The answer is Sunday

**Q5b)** Explain your answer in Q3a).

[Strand: Judgement]

[Subskill: Justify]

[Complexity: difficult]

Possible answer:

- Purchase them on Sunday; it requires spending less money than that on Wednesday because the cake is more expensive than cookie.
- Wednesday ( $\$2 + \$0.50 = \$2.50$ ) will cost more than Sunday ( $\$1.00 + \$1.00 = \$2.00$ ) so, by purchasing on Sunday you will spend less
- Students choose Sunday because it will result in spending less money Students do not want to choose Wednesday because they will spend more money

**Scoring for Q5:**

Score 0 = No answer provided, or chosen day is incorrect.

Score 1 = Correct day (Sunday), but no explanation/reason is provided, or explanation is invalid.

Score 2 = Correct day (Sunday), and valid explanation/reason is provided.

**Q6:** You want to buy at least one juice, at least one cake and at least one cookie. You want to spend exactly \$10. **For example, 1 juice + 1 cake + 5 cookies = \$10 total.** Give 2 more ways you could spend your money.

[Strand: Argumentation, information management]

[Subskill: Identifying alternatives, analyze and synthesize information]

[Complexity: difficult]

[time: 10 mn]

**Scoring for Q6:**

Full credit = Score 2

Any two of:

1 juice + 2 cake + 3 cookies

1 juice + 3 cake + 1 cookie

2 juice + 1 cake + 2 cookies

Partial Credit = Score 1

Any two combinations where at least one EITHER adds up to \$10 OR includes one of each food but not both. For example:

- 1 juice + 1 cake + 1 cookie (includes one of each food)
- 2 juice + 2 cake (adds up to \$10)
- Other options are no score = Score 0

Q4 was slightly revised to include a sentence asking students to explain the reasons behind their given choice in order to more explicitly engage critical thinking skills. In addition, the scoring rubric was added to distinguish the different levels of performance based on responses provided.

Q5 was changed because teachers reported that students understood the question, but they got the wrong answer because of difficulty with the mathematical operations using percentages. In order to address this issue, the discount percentage was revised to half price to make the operations easier, and to disentangle mathematical knowledge from critical thinking skills. More details about the rubric and scoring were added in order to identify the level of student responses.

For Q6, according to the data from the paneling session, additional explanation was recommended because it was difficult for students to understand the expected response format. Thus, providing an example response as part of the question may help students to better understand what is expected. In addition to this, the scoring rubric was included to give partial credit for a correct conclusion even if student gave the wrong answer because of a mistake in the arithmetic.





## Appendix B: Collaboration - Separating mixtures

### DRAFT TASKS

Collaboration was defined to identify its strands and subskills. For OAA Asia, collaboration was defined as a process of working together and communicating to negotiate different perspectives and make decisions to reach a common goal. Collaboration occurs when meeting the goal requires more than what each individual is able to manage alone, so there is a need to pool resources with others. Associated with the main strands, the behavioral indicators (or subskills) were identified: interacting effectively with others and making meaningful conversations; knowing when it is appropriate to listen or to speak (social regulation); working effectively in diverse teams (e.g., conflict resolution; team management); introducing new ideas and sharing resources; assuming shared responsibility for team work; and perspective taking (Table B1).

Table B1. OAA Asia structure for collaboration

Strands	Subskills
Participation	Interact with group members
	Show responsibility
	Show flexibility
Communication	Share information and ideas
	Listen
	Respond to others
	Express own 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

Figure B1

**Stimulus:** Work as a group to find out how to separate of water and salt mixture during science class.

Once collaboration was defined, an existing task which had the potential to target collaboration skills was identified. Figure B1 shows an example task that was being used in a classroom. At first glance, this may look like a task that assesses collaboration because students are asked to work in a group. However, in order to answer this question, students do not need to collaborate—that is, participate, recognize others' perspectives, provide feedback to others, etc. It is possible to work out the answer without the help of others.



Figure B2

**Stimulus:** Make a solution of water and salt in a pot and make a mixture of water and sand in another pot. Then separate salt and sand from the mixture. Also, demonstrate in front of the class and write a report.

**Q1: Preparation and roles allocation**

Within each group, allocate roles, bring the materials and apparatus required for the experiment, and fill in Table 1.

Table 1

Students	Materials	Purpose

*[Strand: Decision making]  
[Subskill: Allocation of roles/work]  
[Level of difficulty: medium]*

**Q2: Making plan**

All students make a plan of the experiment and complete the first column of Table 2.

Table 2

Task 2	Task 3	
Student 1 plan	Student 2 feedback	Student 3 feedback
Student 2 plan	Student 1 feedback	Student 3 feedback
Student 3 plan	Student 1 feedback	Student 2 feedback

*[Strand: Decision making]  
[Subskill: Making a plan]  
[Level of difficulty: hard]*

**Q3: Feedback to other students on their plan**


All of you have to provide feedback to other members of your group in the second or third column of Table 2.

*[Strand: Perspective Taking]  
[Subskill: Feedback to others; recognize others; adapt based on receiver]  
[Level of difficulty: medium]*

**Q4: Present the best plan of the group to others.**

Each of the groups will select their best plan and share the information of the experiments. Each group will provide feedback to other group's plans.

*[Strands: Communication]  
[Subskills: Share information and ideas; listen; respond to others]  
[Level of difficulty: medium]*

Figure B2 shows how the task can be modified to target collaboration skills. 

The nature of the task has been changed from a direct question to a project- or performance-based task. Instructions are provided for the overall task, with several sub-tasks, (referred to as questions or items), that provide structure to the completion of the task.

Figure B2 version of the task is more complex. It samples multiple activities and collaboration strands and subskills. For example, Q1 is about preparing and allocating roles to carry out the main instructions. Q1 is structured using a table, where students each identify their unique roles. This makes it clear to the students and to the teachers which strands and subskills are being targeted. Second, additional items have been added that walk students through the process. Finally, rather than a generic collaboration question, each item now includes the specific strands and subskills that are being targeted, as well as the level of difficulty.



**THINK  
ALOUD**

Because there were significant modifications to the original task (Figure B1), teachers conducted think aloud activities with students to examine whether the modified task elicited the collaboration skills, strands, and subskills intended; and the appropriateness of the task for the targeted grade level and subject areas.

Most feedback from the think aloud sessions confirmed that the task targeted the intended strands and subskills. The levels of difficulty were found to be appropriate for all items. Feedback from teachers mentioned that allocation of time for each item should be included, to guide teachers. Furthermore, Q1 needed more explanation for students to understand what to do. In Q2, revision of instructions to help teachers explain the task was requested.

Based on the empirical evidence from the think aloud sessions, the task was revised (see Figure B3).

**Figure B3**

**Stimulus:** Make a solution of water and salt in a pot and make a mixture of water and sand in another pot. Then separate salt and sand from the mixture. Also, demonstrate in front of the class and write a report.

**Q1: Preparation and roles allocation**

**Work in groups of three.** Within each group, allocate roles to each member, bring the materials and apparatus required for the experiment, and fill in Table 1.

*Table 1*

Students	Materials	Purpose
Name of student 1		
Name of student 2		
Name of student 3		

*[Strand: Decision making]  
[Subskill: Allocation of roles/work]  
Level of difficulty: medium  
[Duration: 5 minutes]*

**Scoring rubrics:**

Category 0	Category 1	Category 2	Category 3
No reason provided	Reasons are unrelated to the task; they do not demonstrate any evidence of working towards a common understanding or goal	Reasons for role allocation are somewhat related to the task and demonstrate some evidence of attempting to work towards a common understanding or goal, but otherwise are off-task	Reasons for role allocation are related to the task and show evidence of attempting to work towards achieving a common understanding or goal

In the Figure B3 version, the language of the items is revised to clarify how to approach the items. For example, information about number of students in each group were added in Q1.

In addition, the duration for each test item was added. Knowing the duration of each task ensures that teachers can plan structure of the session, and provide enough time to students. Details about the scoring of each test item were added. These rubrics provide teachers with specific information about what is being assessed, as well as a grading scale consisting of descriptors including progress, quality and/or proficiency levels. Using a scoring rubric which includes observable student behaviors supports objective judgments by teachers.





**Q2: Making plan**

Each student makes a plan of the experiment and completes the first column of Table 2.

Table 2

Task 2	Task 3	
Student 1 plan	Student 2 feedback	Student 3 feedback
Student 2 plan	Student 1 feedback	Student 3 feedback
Student 3 plan	Student 1 feedback	Student 2 feedback

[Strand: Decision making, Subskill: Making a plan]  
 [Level of difficulty: hard]  
 [Duration: 10 minutes]

**Scoring rubrics:**

Category 0	Category 1	Category 2	Category 3
Positive point(s) are simple affirmations of the plan without a specific reason	Positive point(s) relate to the plan as a whole	At least one positive point presented refers to a specific outcome of the corresponding plan	Two positive points are presented both of which refer to specific outcomes of the corresponding plan
Negative point(s) simply state the idea is not good without any explanation	Negative points are broad and generic criticism of the solution that can apply to most ideas	At least one negative point presented that describes a real world constraint or limitation that relates to the corresponding solution	Two negative points are presented both of which describe real world constraints or limitations that relate to the corresponding solution

**Q3: Feedback to other students on their plan**

Each student must provide feedback to other members of your group in the second or third column of Table 2.

[Strand: Perspective Taking, Subskill: Feedback to others; recognize others; adapt based on receiver]  
 [Level of difficulty: medium]  
 [Duration: 15 minutes]

**Scoring rubrics:**

Category 0	Category 1	Category 2	Category 3
Positive point(s) are simple affirmations of the plan without a specific reason	Positive point(s) relate to the plan as a whole	At least one positive point presented refers to a specific outcome of the corresponding plan.	Two positive points are presented both of which refer to specific outcomes of the corresponding plan
Negative point(s) simply state the idea is not good without any explanation	Negative points are broad and generic criticism of the solution that can apply to most ideas	At least one negative point presented that describes a real world constraint or limitation that relates to the corresponding solution	Two negative points are presented both of which describe real world constraints or limitations that relate to the corresponding solution

**Q4: Present the best plan of the group to others.**

- Each of the groups will select their best plan and share the information of the experiments.
- Each group will provide feedback to other group's plans.

Feedback from other groups

[Strands: Communication; Subskills: Share information and ideas; listen; respond to others]  
 [Level of difficulty: medium]  
 [Duration: 20 minutes]

**Scoring rubrics:**

Category 0	Category 1	Category 2	Category 3
Positive point(s) are simple affirmations of the plan without a specific reason	Positive point(s) relate to the plan as a whole	At least one positive point presented refers to a specific outcome of the corresponding plan	Two positive points are presented both of which refer to specific outcomes of the corresponding plan
Negative point(s) simply state the idea is not good without any explanation	Negative points are broad and generic criticism of the solution that can apply to most ideas	At least one negative point presented that describes a real world constraint or limitation that relates to the corresponding solution	Two negative points are presented both of which describe real world constraints or limitations that relate to the corresponding solution

In Q4, the format of the task was amended to include a space for students to write down feedback from other groups to make it clear what is required to complete the task successfully.



**PANEL**

The revised task was then reviewed in a paneling session to check and improve items if necessary. During the paneling session, panel members (e.g., subject experts, assessment experts, and teachers) discussed each question and provided feedback. Based on the feedback from paneling session, the task went through a final round of revision in Figure B4.

Figure B4

**Stimulus:** There is a mixture of salt and sand (or any non-dissolvable material that is similar to sand) in a pot. Your task is to separate the salt and sand from the mixture. Make a plan within your group that all of you can do together with each member of the group having a role. Describe the plan that your group would do in order to accomplish the task.

**Q1: Making plan**

Students work in groups of three. Each member of the group has to make a plan of the experiment that will accomplish the task and fill the first column of Table 1. Table 1. Fill in the first column only for this task.

Task 1	Task 2	
Student 1 plan	Student 2 feedback	Student 3 feedback
Student 2 plan	Student 1 feedback	Student 3 feedback
Student 3 plan	Student 1 feedback	Student 2 feedback

[Strand: Decision making  
[Subskill: Making a plan, difficulty]  
[Level of difficulty: hard]

**Scoring rubrics:**

Category 0	Category 1	Category 2	Category 3
No plan provided	The plan doesn't address all three requirements and doesn't relate to the role	The plan doesn't address all three requirements but what is presented relates to the role	The plan addresses all three requirements and has relates to the role

**Q2: Feedback to other students on their plan**

All of you have to provide feedback to other members of your group in the second or third column of Table 1. Provide one or more positive and negative feedback on the others' plans.

*Strand: Perspective Taking, Subskill: Feedback to others; recognize others; adapt based on receiver]*  
[Difficulty: medium]

**Scoring rubrics:**

Category 0	Category 1	Category 2	Category 3
Positive point(s) are simple affirmations of the plan without a specific reason	Positive point(s) relate to the plan as a whole	At least one positive point presented refers to a specific outcome of the corresponding plan	Two positive points are presented both of which refer to specific outcomes of the corresponding plan
Negative point(s) simply state the idea is not good without any explanation	Negative points are broad and generic criticism of the solution that can apply to most ideas	At least one negative point presented that describes a real world constraint or limitation that relates to the corresponding solution	Two negative points are presented both of which describe real world constraints or limitations that relate to the corresponding solution

To increase authenticity and provide an opportunity for groups to generate creative ideas, the stimulus was changed to separating the mixture of sand and salt, rather than water and salt.

In addition, the order of Q1, Q2 and Q3 was changed, so that the sequencing of the tasks made more sense. The table in Q1 was modified to allow for Q2 responses (i.e., feedback to other students).



### Q3: Preparation and roles allocation

Discuss the plans made by each member and decide on the best plan or combine the plans into a final plan.

To accomplish the best plan that was decided, allocate each member roles, describe the materials and apparatus required for the experiment, and fill in the table below.

Finally, share this final plan with the class for feedback from other groups. Example of roles (make sure that each member has a role and that the roles are not duplicated):

- Responsible for bringing the materials needed
- Responsible for separating the sand from the solution

Students	Role	Purpose of the role
Name of student 1		
Name of student 2		
Name of student 3		

[Strand: Decision making]

[Subskill: Allocation of roles/work]

[Difficulty: complex]

#### Scoring rubrics:

Category 0	Category 1	Category 2	Category 3
No reason provided	Reasons are unrelated to the task; they do not demonstrate any evidence of working towards a common understanding or goal  The roles are incomplete and will not accomplish the task	Reasons for role allocation are somewhat related to the task and demonstrate some evidence of attempting to work towards a common understanding or goal, but otherwise are off-task. Some of the roles are duplicated or overlapping	Reasons for role allocation are related to the task and show evidence of attempting to work towards achieving a common understanding or goal  None of the roles are duplicated or overlapping

### Q4: Present the best plan of the group to others

Each group will provide feedback to other group's plans.

Each group will select the best plan across all groups and provide a reason for their choice.

Feedback to other groups	Group 1	Group 2	Group 3	Group 4
<b>Positive points:</b>				
<b>Negative points:</b>				
<b>Chosen best plan across the groups and reasons</b>				

[Strands: Communication]

[Subskills: Share information and ideas; listen; respond to others]

[Difficulty: medium]

#### Scoring rubrics:

Category 0	Category 1	Category 2	Category 3
Positive point(s) are simple affirmations of the plan without a specific reason	Positive point(s) relate to the plan as a whole	At least one positive point presented refers to a specific outcome of the corresponding plan	Two positive points are presented both of which refer to specific outcomes of the corresponding plan
Negative point(s) simply state the idea is not good without any explanation	Negative points are broad and generic criticism of the solution that can apply to most ideas	At least one negative point presented that describes a real world constraint or limitation that relates to the corresponding solution	Two negative points are presented both of which describe real world constraints or limitations that relate to the corresponding solution

#### Group feedback

#### Choice of best plan

Category 0	Category 1	Category 2
No best choice or reason provided	The chosen plan is incomplete and will not accomplish the task or will only partly accomplish the task. For example, only separate the sand from the salt solution  Reasons are not based on science concepts	The chosen plan is complete and will accomplish separating sand from the salt solution, as well as separating salt from the water  Reasons are correctly based on science concepts

In Q3, details were added in the instructions to better guide the students in filling out the table. The scoring rubric was also revised.

The prompt for Q4 was expanded to make it clear to the students what they were supposed to do. In the previous version, students were asked to provide other groups with feedback in order to choose the best plan among the groups. The task is still asking for the same thing; however, the table asks specifically for both positive and negative feedback for each group. In addition, groups are now expected to collaborate to make the best plan and provide a reason together. This requires students to more explicitly engage their collaborative skills, including sharing ideas, listening, responding to others, and making decisions to come to a consensus.

The scoring rubric was revised in order to better align with the expectations of the task. Two separate scoring rubrics are included—one that assesses the feedback to groups and one that assesses the selection of the best plan and relevant reasons provided by the group.



## Appendix C: Collaboration (and problem solving) - Serrekunda

### DRAFT TASKS

Collaboration was defined to describe its structure and its contributing strands and subskills. The process was informed by a summary of the research literature available on the topic. The observable behavioral indicators for the subskills were identified, linked with four strands of the collaboration construct. The collaboration structure used for revisioning of tasks by OAA Africa is shown in Table C1. See Appendix D for the problem solving structure.

Table C1. OAA Africa structure for collaboration

Strands	Subskills
<b>Participation (P)</b>	Taking responsibility (Tr)
	Sharing (Sh)
	Turn taking (Tt)
	Engagement (En)
<b>Communication (C)</b>	Receptive (Re)
	Expressive (Ex)
<b>Negotiation (N)</b>	Compromising (Co)
	Perspective taking (Pt)
<b>Decision making (D)</b>	Analysis (An)
	Evaluation (Ev)
	Planning (Pl)

Figure C1

**Stimulus:** In your country, several families live together in over-crowded houses. Identify the causes and the consequences of this phenomenon, and the solutions you recommend to solve this population problem.

The task was originally conceptualized for problem solving.

Figure C2

**Stimulus:** Serrekunda is an over crowded town. Families are living in very small houses in larger numbers raising health concerns. In your group, work independently to determine the causes, consequences and then propose solutions to this problem. In your group discuss.

- Imagine that you are the mayor of Serrekunda, what measures will you put in place to address these health concerns?
- What challenges are you likely to face in addressing this issue?
- How do you plan to overcome these challenges?

*[Collaboration: Participation, communication, negotiation, and decision making]*

*[Problem solving: Gathering information, Planning a solution]*

Figure C2 shows that although the task retains its problem solving features, it is re-phrased to require working in a group.

Specific subskills of collaboration that the task targets, are identified. This iteration of the task was designed to make the problem more real through naming the town, and being specific that over-crowding was a community rather than just a family issue. To focus student responses, the mayoral role was introduced.



## THINK ALOUD

Because there were significant modifications to the original item (Figure C1), teachers conducted think aloud activities with students to examine whether the modified task elicited the skills strands and subskills intended, as well as appropriateness of the task for the targeted grade level and subject areas. Feedback from that process led to another round of revisions.

**Figure C3**

Serrekunda is an over-crowded town. **Very large families including grandparents, and their adult children with their wives and husbands and children, are all living in small two or three room houses.**

### Question Set 1

In groups of three students, discuss:

- a) Why might such large families be living together in small spaces?**
- b) What problems might occur due to large number of people living in small houses?**

### Question Set 2

The following questions should be allocated across the group, each question to one student:


- a) As a grandparent, what might you feel about the mayor proposing that extended families be housed in different houses? Give one positive and one negative possible response.**
- b) As a mother of four children, what might you feel about the mayor proposing that extended families be housed in different houses? Give one positive and one negative possible response.**
- c) As a child, what might you feel about the mayor proposing that extended families be housed in different houses? Give one positive and one negative possible response.**

### Question Set 3

Taking on the role of the mayor, how does the group plan to overcome the challenges identified by each family member?

#### *Question Set 1*

- a) Plausible responses include lack of buildings, poverty, culture***
- b) Plausible response includes extended family shared responsibilities***

More details were provided in the stimulus material at the beginning of the task. 

Due to the multiple embedded questions now in the task, it was re-structured across three question sets for ease of classroom administration.

To facilitate collaboration, more detailed instructions for working together were included.

As a result of the amendments, the overall strands and subskills being targeted, increased. Some plausible responses were provided as illustrations.





The revised task was then reviewed in a paneling session. During the paneling session, panel members (e.g., subject experts, assessment experts, and teachers) discussed each question and provided feedback. Based on this feedback, the task went through a final round of revision (Figure C4).

*Figure C4*

Serrekunda is an over-crowded town. Large families including grandparents, and their adult children with their wives and husbands and children, all live in small two or three room houses. **One issue that these families experience is that there is not enough coal for everyone to use as needed.**

- Different members of the family have different daily needs and uses for coal.
- The mother needs coal to make food
- The grandmother needs coal for heat during the day to stay warm and watch the baby grand-daughter.
- The young student needs coal for light to study.

Each of these activities require different daily amounts of coal.

- 7 units of coal to make food
- 8 units of coal for light
- 8 units of coal for heat

The family has only 15 units of coal total for each day.

In a group of three students, identify who will take on which family role (mother, grandmother, or young student).

- Individually, state and justify what is the best outcome for you in your nominated role.
- As a group, discuss and decide on the best outcome for all members of the family. Explain your answer.

**a) Ge>ConCom**

0 = No or irrelevant response

1 = Provides relevant outcome for self

**b) Ge>ConCom**

0 = No or irrelevant response

1 = Provides relevant outcome for group

**P>Sh**

0 = Less than 3 people share best outcome for self

1 = All three members share best outcome for self

**C>Ex**

0 = Did not make any expressive contributions/ irrelevant contribution

1 = Made at least 1 relevant expressive contribution

**Decision-making (D)**

0 = Did not agree on best outcome for all

1 = Did agree on best outcome for all

**Ju>Ex**

0 = No or irrelevant response

1 = States reason for why the outcome was chosen

2 = States reason and explains what the decision means for each family member

The revisions to the task reflected the structure of the "collaborative task template" (Kim & Care, 2020). This was a significant move from focus on a specific task, to re-structuring a task according to template structure. In turn, this means greater potential for development of multiple tasks that follow the same structure and therefore provide the opportunity to collect multiple indicators of the same strand/subskill across a set of tasks.

The first major change made was to make the problem scenario more specific. This provided a more robust context within which students could generate ideas, consider and compare options, justify and explain, participate and share, communicate and make decisions within a collaborative.

The second major change was to provide for simplified marking and scoring. As can be seen, the b) responses provide for scoring across multiple strands/subskills.



## Appendix D: Problem solving - Lemon tree

### DRAFT TASKS

Problem solving was defined to describe its structure and its contributing strands and subskills. The process was informed by a summary of the research literature available on the topic. The observable behavioral indicators for the subskills were identified, linked with four strands of the collaboration construct. The problem solving structure used for revisioning of tasks by OAA Africa is shown in Table D1.

Table D1. OAA Africa structure for problem solving

Strands	Subskills	Sub-subskills
<b>Information gathering (IG)</b>	Ask questions related to the problem (Aq) Organise information (Oi)	Classify (Cla) Analyse (verify, discriminate, compare (Ana) Describe (Des)
<b>Planning a solution (PS)</b>	Generate ideas, options, hypotheses (Ge)	Hypothesize (Hyp) Consider and compare options (ConCom)
	Developing plan (Dp)	Discriminate (Dis) Identify relationships (Rel) Predict (Pre)
<b>Managing information (MI)</b>	Follow a plan (Fp)	
	Compare outcomes with plan (Cf)	Compare evidence with predict (Com) Check logical flow (Clf)
	Justification of the process (Ju)	Explain (Exp)
	Synthesise (Sy)	Summarise (Sum)

Figure D1

**Stimulus:** In a science project, Chabota and his two friends planted lemon trees. Chabota planted his under a big tree. His friends planted theirs where logs of wood were burnt. He discovered that his plant was not growing as well as those of his friends. What would account for the differences in the growth of Chabota's lemon plant and those of his friends?

The task shown in Figure D1 was derived from a science topic, and adapted for problem solving.



Figure D2

In a science project, Chabota and his two friends planted lemon trees. Chabota planted his under a big tree. His friends planted theirs where logs of wood were burnt. He discovered that his plant was not growing as well as those of his friends.

- a) What would account for the differences in the growth of Chabota's lemon plant and those of his friends? **Identify at least two factors and explain how the factors account for the differences.**

In Figure D2, the initial question was provided with more structure in order to clarify what was expected of the student.





**b)** Chabota's friends found him preparing to cut the tree that was close to the lemon tree he had planted. If you were one of the friends of Chabota, what advice would you give him, and why?

- a)** IG>Aq>Ana/Des  
**b)** PS>Ge/Dp  
 MI>Exp

Another question was added into the task, in order to stimulate students to generate ideas about possible factors or causes.

The problem solving strands, subskills and sub-subskills relevant to each process were identified.

## THINK ALOUD

Because there were significant modifications to the original task (Figure D1), teachers conducted think aloud activities with students to examine whether the modified task elicited the collaboration skills, strands, and subskills intended; and the appropriateness of the task for the targeted grade level and subject areas. As a result of the activity, the task was substantially amended and expanded as in Figure D3.

**Figure D3**

In a science project, Chabota discovered that his lemon tree was not growing as well compared to those of his two friends. His friend's trees were in a place where logs of wood had been burnt while his was under a tree.

- a)** What would account for the differences in growth? Identify at least two factors and explain how the factors account for the differences.
- b)** Chabota's friends found him preparing to cut the tree that was close to his lemon tree. If you were one of the friends of Chabota, what advice would you give him, and why?
- c)** Suggest two ways in which Chabota could ensure that his tree grows as well as those of his friends.

- a)** IG>Aq>Ana/Des  
 0 = No response/unrelated  
 1 = Presentation of one factor  
 2 = Presents factors associated with the two environments
- b)** PS>Ge/Dp  
 MI>Exp  
 0 = No advice given/unrelated  
 1 = Advice given without justification  
 2 = Advice given with justification
- c)** PS>Dp/Pre  
 MI>Ju/Ex  
 0 = No suggestion/ unrelated  
 1 = One suggestion  
 2 = Two suggestions

The task was slightly re-worded for clarity, and the locations of the planting were not directly attributed to the three friends. This was intended to help the student respondents to focus on the current situation, rather than on who may have made mistakes in the initial locating of best planting areas.

Question c) was added, which requires the student respondent to consider the advice given, and how that would manifest in action - looking to developing and then following of a plan.

The problem solving strands, subskills and sub-subskills relevant to each process were identified for question c).

Scoring rubrics were added which focus on the processes and products associated with each of the questions. Note that these rubrics do not provide for quality of response, only for whether the requisite processes and products were enacted. Another level of rubrics could be added to discriminate levels of quality.



The revised task was then reviewed in a paneling session. During the paneling session, panel members (e.g., subject experts, assessment experts, and teachers) discussed each question and provided feedback. Based on the feedback from paneling session, the task went through a final round of revision as in Figure D4.

*Figure D4*

In a science project, Chabota discovered that his lemon tree was not growing as well compared to those of his two friends. His friend's trees were in a place where logs of wood had been burnt while his was under a tree.

- a) What would account for the differences in growth? Identify at least two factors and explain how the factors account for the differences.
- b) Chabota's friends found him preparing to cut the tree that was close to his lemon tree. If you were one of the friends of Chabota, what advice would you give him, and why?
- c) Suggest two ways in which Chabota would ensure that his tree grows as well as those of his friends.

**a-i) IG>Aq>Ana**

- 0 = No factors provided
- 1 = Factors provided but unrelated
- 2 = One relevant factor provided (e.g., only shade or nutrient)
- 3 = Two relevant factors provided

**a-ii) IG>Oi/Des**


- 0 = No Explanation provided
- 1 = Explanation is unrelated to the factors
- 2 = Explanation is provided for only one of the factors (e.g., *The shade from bigger tree blocks sunlight from reaching Chabota's tree; [while] his friends trees were in the open and received a lot of sunlight; Chabota's lemon tree has competition for nutrients from the roots of the big tree; [while] his friends lemon plants have less or no competition; Chabota's friends plants have Potash.*)
- 3 = Explanation is provided for both factors

**b-ii) MI>Exp**

- 0 = No justification provided
- 1 = Justification is not relevant
- 2 = Justification is related to the specific advice provided.

**c) PS>Dp/Pre**

- MI>Ju/Ex
- 0 = No suggestion/ unrelated
- 1 = One suggestion
- 2 = Two suggestions

The stimulus and questions are unchanged in version D4. 

Amendments were made to the marking rubrics, and to the scoring. These ensured that credit would be given for students responding to multiple parts of each item, as in a) and b).

Examples of acceptable explanations are provided as guidance for marking tasks and items such as these.

The amount of information included for the marking and scoring, while useful in test and item development, may exceed the needs of the classroom teacher in terms of identifying student proficiencies. In such cases, the actual marking and scoring can be reduced to the levels of the skill alone, the strands, or the subskills.