

7. Skills assessments in education

Samuel Greiff, University of Luxembourg

Jan Dörendahl, University of Luxembourg

This chapter describes skills typically included in large-scale educational assessments and discusses how such assessments can be used for measuring the capabilities of artificial intelligence (AI). It focuses on two major skill domains covered in educational assessments: core domain skills such as mathematics, reading and science literacy, and transversal skills such as problem solving, collaboration and creativity. The chapter provides an overview of the skills in each domain, as well as their theoretical underpinning and measurement. In addition, it examines the role of these skill domains in occupational settings by drawing links to taxonomies of skill requirements in the workplace. The chapter concludes with recommendations regarding the use of education tests for scaling AI capabilities.

Introduction

Educational systems must provide environments that foster and facilitate the highly diverse skill sets needed for the digital world. These skill sets can generally be represented by three different skill domains. First, students require literacy in core domains such as mathematics, reading and science (OECD, 2017^[1]). Second, they need skills in transversal domains that span situations and contexts such as problem solving, collaboration and creativity (OECD, 2013^[2]; 2017^[3]; 2019^[4]). Finally, they need basic cognitive skills such as general mental ability, fluid reasoning or working memory (McGrew, 2009^[5]); these are often considered fundamental for acquiring more complex skill sets in the other two domains.

In response to this three-part challenge, global student assessment initiatives, such as the Programme for International Student Assessment (PISA), have included skills from the transversal domain in addition to domain-specific knowledge (OECD, 2019^[6]). Transversal skills are now part of many educational large-scale assessments and considered important markers of educational achievement (OECD, 2013^[2]; 2017^[1]; 2019^[6]). However, basic cognitive skills have been included to a lesser extent. On average, evidence suggests that countries improve with respect to the core domain skills, but basic cognitive skills are malleable to a lesser extent within education. This, in turn, has lessened the interest of practitioners and policy makers in basic cognitive skills.

For all the importance of developing skills in the core domains, transversal skills and basic cognitive skills, educational systems face another challenge. Given the emergence of artificial intelligence (AI) and robotics, which skills will become obsolete for humans, both as a requirement of the workforce and as an educational goal?

To approach this question, a taxonomy of skills is needed to assess and scale AI-related capabilities. Ideally, this taxonomy will relate to skill frameworks and the tasks associated with them, for instance, from international large-scale assessments. It also needs to distinguish skills from core domains (OECD, 2017^[1]), transversal skills (OECD, 2013^[2]; 2017^[3]; 2019^[4]) and basic cognitive skills (McGrew, 2009^[5]).

This chapter focuses on core domains and transversal skills, leaving basic cognitive skills for Chapter 3. First, it provides a brief overview of skills from core domains and the transversal domain. It also looks at specific skills typically assessed in education and provides a brief background on the underlying theories. Second, it presents assessment instruments to measure these skills with a focus on innovative and technology-based instruments. Drawing on these two points, it then assesses the extent to which such tests could assess the capabilities of AI.

Educationally relevant skills

Core domains and transversal skills have different theoretical backgrounds, partly due to disparate research traditions. Research on reading literacy (skill set: core domains), for example, originates in the educational and learning sciences. Conversely, research on collaborative problem solving (skill set: transversal domain) is largely rooted in educational large-scale assessments and social psychology.

Although the two skill domains and the nature of the associated skills vary in complexity, they might be equally important for success in life. Both domains relate to recognising, interacting with and solving real-world situations. While core domains and transversal skills are interdependent (OECD, 2014^[7]), this chapter considers them separately for ease of interpretation and readability.

Assessment of the two skill sets: General concepts

There is broad consensus that the two skill sets – core domain and transversal skills – are important across several outcomes. Thus, there is a strong need for measurement and assessment to keep track of them

and their development. This could include, for instance, international comparisons of educational systems or tracking individual student progress across grade levels.

Several skill sets continue to be the focus of international large-scale assessments, as well as of comprehensive research efforts. Some examples of specific measurements appear below. However, there are many ways of assessing the two skill sets, including classical paper-pencil assessments and highly innovative computer-simulations.

This section focuses on innovative item types and the potential of such items for assessing AI skills. Innovative item types often provide additional information such as behavioural patterns of students when working with dynamically changing problem-solving items such as MicroDYN (see Greiff and Funke (2009^[8]); skill set of transversal skills). Similarly, navigating complex texts in an online environment such as digital reading items (skill set of core domain skills) can also reveal behavioural patterns.

Specifically, the chapter considers the field of international large-scale assessments:

- PISA (OECD, 2013^[2]; 2017^[1]; 2019^[9])
- Programme for the International Assessment of Adult Competencies (PIAAC); (PIAAC Expert Group in Problem Solving in Technology-Rich Environments, 2009^[10]); (PIAAC Literacy Expert Group, 2009^[11])
- National Assessment of Educational Progress (NAEP) (National Assessment and Governing Board, 2019^[12]; 2019^[13]; 2019^[14])
- Trends in International Mathematics and Science Study (TIMSS) (Mullis and Martin, 2017^[15])
- Graduate Record Examination (GRE) and the SAT (formerly known as Scholastic Assessment Test).¹

For each of the two skill domains, this chapter describes typical skills for the respective overarching set of skills; provides an overview and examples of assessments that include these skills; and summarises the theoretical backgrounds and sub-processes for the respective skills. After these subsections, the chapter presents possible dimensions of a skill taxonomy in relation to scaling AI capabilities.

Core domain skills

The core domain skills [i.e. mathematic literacy, reading literacy, science literacy; OECD (2017^[1])] focus on knowledge and processes closely related to scholastic domains. Although the labels for these skills are not consistent across assessments, they are similar and show strong overlap (Table 7.1). While definitions and sub-processes of each skill might differ slightly across assessments (and even across different cycles within one assessment), their overlap is substantial. Table 7.1 summarises the definitions of the three skills and their sub-processes. Additionally, it provides an overview of several other large-scale assessments where these skills have been assessed.

Several frameworks exist for each of the skills, including those developed by scientific expert groups within PISA through expert opinions and from the scientific literature. The framework documents are constantly refined as the assessment cycles progress. This provides a theoretical foundation in defining the skills and fans out sub-processes. The frameworks also make suggestions and provide specific guidance on how the theoretical background can be translated into specific and actionable assessments that are ultimately run in the respective assessments such as PISA or PIAAC.

Table 7.1. A comparison of large-scale assessment frameworks

Skill	Definition	Sub-process	Examples of large-scale assessments assessing these skills (and labels used in the assessment)
Mathematical literacy	An individual's capacity to formulate, employ and interpret mathematics in a variety of contexts.	<ul style="list-style-type: none"> Formulate mathematics Employ mathematics Interpret mathematical results 	PISA (Mathematical literacy) PIAAC (Numeracy) NAEP (Mathematics) TIMSS (Mathematics) GRE (Quantitative fluid reasoning) SAT (Mathematics)
Reading literacy	The ability to make use of written texts, to achieve one's goals, to increase one's knowledge and potential, and to participate in society.	<ul style="list-style-type: none"> Access and retrieve Integrate and interpret Reflect and evaluate 	PISA (Reading literacy) PIAAC (Literacy) NAEP (Reading) GRE (Verbal fluid reasoning; analytical writing) SAT (English; Languages)
Science literacy	The ability to engage with science-related issues and with the ideas of science.	<ul style="list-style-type: none"> Explain phenomena scientifically Evaluate and design scientific enquiry Interpret data and evidence scientifically 	PISA (Science literacy) NAEP (Science) TIMSS (Science) SAT (Science)

Note: Definitions are partly direct quotes.

Source: OECD (2017_[1]).

Figure 7.1 provides an example of a set of items (labelled “unit” in the PISA context) that assesses mathematical literacy, as used in the PISA 2012 cycle. The unit consists of three items in a real-world context. The students need to solve them by interpreting and comparing the numbers in the table and performing calculations themselves. More technically, they use sub-processes indicated in Table 7.1 to interpret mathematical results for items 1 and 2, and then to employ them for item 3.

Figure 7.2 presents an example item assessing reading literacy in PISA 2018 (OECD, 2017_[3]). Test takers need to reflect on and evaluate three texts on space exploration. They then write a comment on its benefits afterwards (i.e. employing the “reflect and evaluate” sub-process; see Table 7.1).

Figure 7.3 displays an example item assessing science literacy in PISA 2012 (OECD, 2014_[16]). Test takers are asked to interpret scientific information and explain why they do not support the conclusion of a fellow student (i.e. employing the “explain phenomena scientifically” sub-process; see Table 7.1).

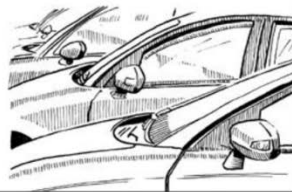
Figure 7.1. Sample item assessing mathematical literacy

WHICH CAR? – a unit from the PISA 2012 main survey

WHICH CAR?

Chris has just received her car driving licence and wants to buy her first car. This table below shows the details of four cars she finds at a local car dealer.

Model:	Alpha	Bolte	Castel	Dezal
Year	2003	2000	2001	1999
Advertised price (zeds)	4 800	4 450	4 250	3 990
Distance travelled (kilometres)	105 000	115 000	128 000	109 000
Engine capacity (litres)	1.79	1.796	1.82	1.783



WHICH CAR? – QUESTION 1

Chris wants a car that meets **all** of these conditions:

- The distance travelled is **not** higher than 120 000 kilometres.
- It was made in the year 2000 or a later year.
- The advertised price is **not** higher than 4 500 zeds.
- Which car meets Chris's conditions?

A. Alpha
B. Bolte
C. Castel
D. Dezal

WHICH CAR? – QUESTION 2

Which car's engine capacity is the smallest?

A. Alpha
B. Bolte
C. Castel
D. Dezal

WHICH CAR? – QUESTION 3

Chris will have to pay an extra 2.5% of the advertised cost of the car as taxes.

How much are the extra taxes for the Alpha?

Extra taxes in zeds:

Source: OECD (2014)^[16].

Figure 7.2. Sample item assessing reading literacy in PISA 2018

PISA 2018

Unit Title: Space Exploration

Question 5/5

Refer to the articles on the right. Type your answer to the questions in the space provided.

Think about how Scott Huffington wrote his article and the commenters responded. Based on this information, write a comment that explains two primary benefits of space exploration? Support your answer with details from the articles.

Text 1
Text 2
Text 3

Is the Golden Era of Space Exploration Over?
by Scott Huffington • May 16, 201

Beginning with the launch of Sputnik in 1957 the focus of space exploration had one aim: be the first to go where no human had gone before. In 1961 Yuri Gagarin became the first man in space sparking an intense competition where astronauts and cosmonauts battled to break records, expand frontiers, and bring notoriety to their countries of origin. However, since July 22nd 1969 and Neil Armstrong's historic leap for mankind, space exploration has slowed.

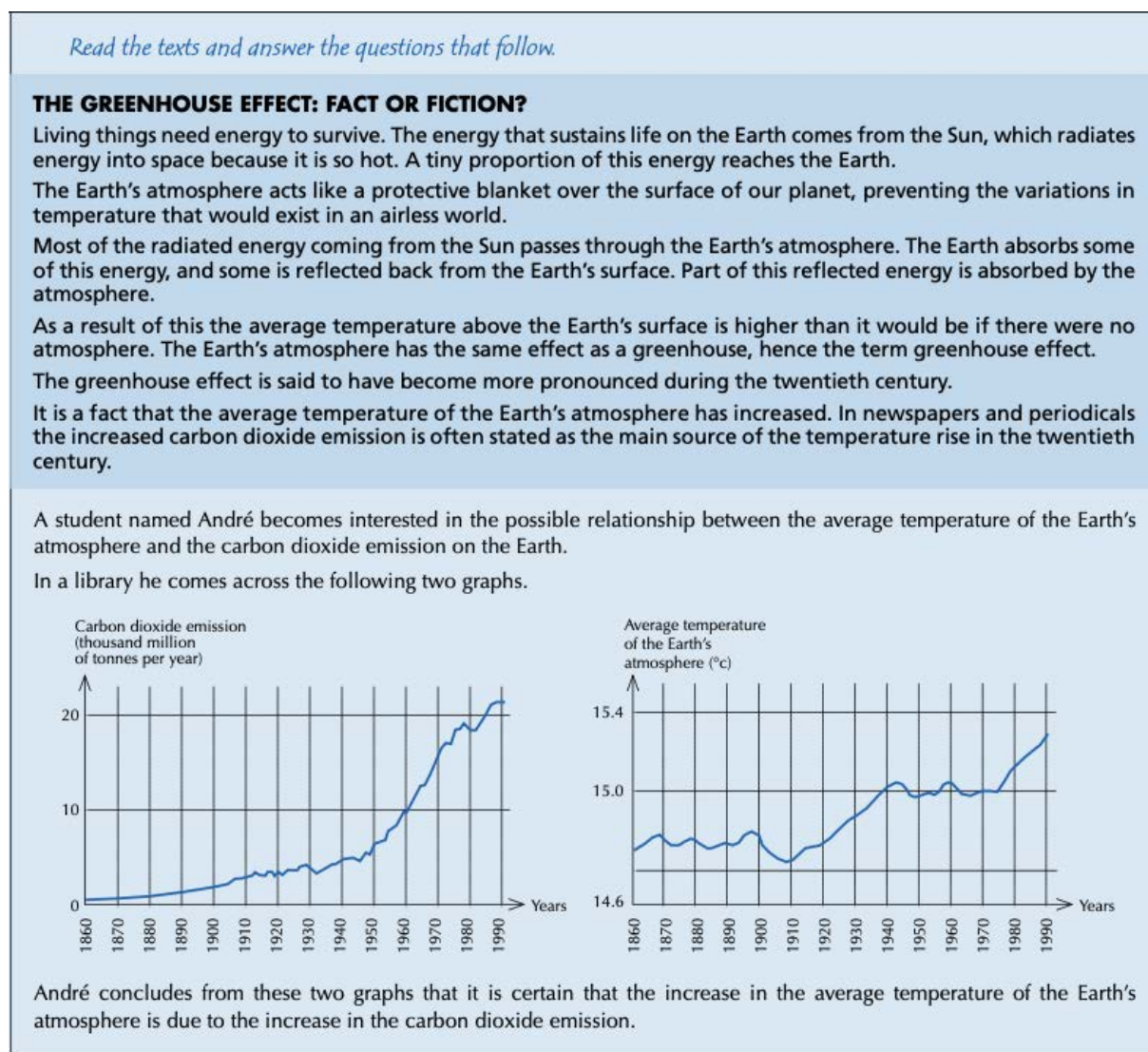
Since then, space programs have focused on creating a sustainable presence in low-Earth Orbit through the development and maintenance of space craft, space stations, and satellites. The Russian space station Mir and the US Skylab were the first space stations but proved too expensive to operate independently. We now have the International Space Station (ISS), an impressive international collaborative effort led by the United States, Russia, Canada, and Japan. Yet, the station was meant to be a stepping stone to bolder projects including a manned mission to Mars. Thirty years later, we are still maintaining the space station but we are no closer to achieving a manned mission to Mars.

For decades, the idea of human space exploration has widely been seen as the exclusive domain of government agencies like the Russian Federal Space Agency (RKA), the National Aeronautics and Space Administration (NASA) in the United States, and the European Space Agency (ESA) with 22 member countries. However, the rise of private companies making serious steps toward successful commercial space flights has many people questioning the relevance and necessity of government run and publicly funded space exploration programs. Add the highly publicized U.S. space shuttle disasters in 1986 and 2003 and the enthusiasm and commitment for space exploration has further eroded.

All of this leads me to conclude that the world has lost the focus and drive to explore new frontiers. I fear that the golden age of space exploration has passed, and we are rapidly progressing toward a decidedly

Source: OECD (2019)^[9].

Figure 7.3. Sample item assessing science literacy in PISA 2012



Another student, Jeanne, disagrees with André's conclusion. She compares the two graphs and says that some parts of the graphs do not support his conclusion.

Give an example of a part of the graphs that does not support André's conclusion. Explain your answer.

Source: OECD (2014_[16]).

Transversal skills

Transversal skills such as problem solving, collaboration, creative thinking, learning in a digital world and global competence are centred around two capacities (OECD, 2013_[2]; 2017_[3]; 2019_[4]). First, they allow people to navigate successfully in dynamically changing environments. Second, they allow them to act competently both on a cognitive and a non-cognitive level in today's world.

By definition, these skills are important in a variety of situations (i.e. domain-general) and involve adaptability and flexibility as defining parts (Griffin and Care, 2015^[17]). For instance, adaptability and flexibility are inherent parts of problem-solving activities that require dealing with unknown situations and successfully choosing the right actions to solve the problem. Indeed, adaptability and flexibility are at the core of adaptive problem solving that is planned for the PIAAC 2022 assessment (Greiff et al., 2017^[18]).

Large-scale assessments have repeatedly focused on transversal skills. PISA 2012, for example, focused on creative problem solving. Meanwhile, PIAAC 2012 looked at problem solving in technology-rich environments, and PIAAC 2022 will look at “adaptive problem solving” (PIAAC Expert Group in Problem Solving in Technology-Rich Environments, 2009^[10]; OECD, 2013^[2]; Greiff et al., 2017^[18]). PISA 2015 assessed collaborative problem solving (OECD, 2017^[3]) and PISA 2018 looked at global competence (OECD, 2019^[9]). PISA 2022 envisages assessment of other and equally diverse transversal skills such as creative thinking (OECD, 2019^[19]), while PISA 2025 plans to assess learning in a digital world.

As with core domain skills, conceptual frameworks exist or are in development for each of the transversal skills. Again, these have been derived through expert opinions and from the scientific literature. They provide a theoretical foundation both in defining the skills and their sub-processes, and in guiding the developments of appropriate assessment instruments.

Table 7.2 displays how frameworks from the completed cycles and drafts for the planned cycles define the above-mentioned transversal skills. It also displays definitions and sub-processes for the transversal skills of problem solving, collaboration, creative thinking, learning in a digital world and global competence on the basis of PISA 2012, 2015 and 2018 assessment frameworks, as well as for the draft of PISA 2022. Additionally, it provides an overview of the large-scale assessments where similar skills have been assessed or will be assessed in future cycles (comparable to Table 7.1 above).

Again, as with the core domain skills, the sub-processes of the transversal skills displayed in Table 7.2 largely define what items ultimately need to assess. However, unlike for core domain skills, innovative item formats are generally required for assessment of transversal skills. Both PISA and PIAAC typically use dynamic, scenario-based approaches where the assessment situation changes through actions of the test taker. In these scenarios, test takers are presented with a problem in a real-world setting, such as working out how to use a new air conditioner (see sample item presented in Figure 7.4).

Innovative item formats provide a more realistic simulation of real-world situations. They are particularly relevant for scaling AI as skills in general are measured against their real-world (and less their theoretical) relevance. In addition, the availability of such item types has implications both for face validity and for how well an assessment can represent the underlying theoretical concept.

To this end, innovative item formats come with several advantages. First, they provide dynamically changing and interactive environments that cannot be simulated using traditional paper-pencil formats. In this way, they allow for assessment of new constructs such as transversal skills [i.e. problem solving, collaboration, creative thinking, learning in a digital world and global competence; OECD (2013^[2])]. Second, these innovative items can be easily constructed in divergent ways and can include direct simulations of complex real-world situations. As such, they allow for an increasing construct coverage (Greiff and Funke, 2009^[8]). Third, they record the test takers’ behaviour, saving it into so-called log files (these come in addition to the final test score). This allows gathering of insights about the processes operating in solving the items (Greiff et al., 2016^[20]). Finally, as innovative item formats allow to simulate everyday situations, they offer increasing face validity and engagement, and increased ecological validity (Greiff and Funke, 2009^[8]).

Table 7.2. Definitions and sub-processes for five key skills in large-scale assessments

Skill	Definition	Sub-process	Examples of large-scale assessments assessing/planning to assess these skills (and labels used in the assessment)
Problem solving	The ability to engage in cognitive processing to understand and resolve problem situations where a solution is not immediately obvious.	<ul style="list-style-type: none"> Explore and understand the problem. Represent and formulate a mental model of the problem. Plan and execute strategies to solve the problem. Monitor and reflect the progress made in solving the problem. 	PISA (Creative problem solving, 2012 cycle) PIAAC (Cycle 1 in 2012: Problem solving in technology-rich environments; Cycle 2 in 2022: Adaptive problem solving)
Collaboration	The ability of an individual to engage effectively in a process whereby two or more agents attempt to solve a problem.	<ul style="list-style-type: none"> Establish and maintain a shared understanding of the problem. Take appropriate actions to jointly solve the problem. Establish and maintain team organisation. 	PISA (2015 cycle: Collaborative problem solving)
Creative thinking	The ability to generate, evaluate and improve ideas directed towards original and effective solutions, advances in knowledge and impactful expressions of imagination.	<ul style="list-style-type: none"> Creative expression (written and visual). Problem solving (scientific and social). 	PISA (2022 cycle: Creative thinking)
Learning in a digital world	Forthcoming.	Forthcoming	PISA (2025 cycle: Learning in a digital world)
Global competence	A combination of skills, knowledge, values and attitudes facilitating individuals to act and interact respectfully, successfully and in sustainable manner on a local, global and intercultural level.	<ul style="list-style-type: none"> Examine local, global and intercultural issues. Understand and appreciate different perspectives and worldviews. Interact successfully and respectfully with others. Take responsible action towards sustainability and collective well-being. 	PISA (2018 cycle: Global competence)

Note: Definitions are partly direct quotes.

Source: OECD (2013^[2]; 2014^[7]; 2019^[9]; 2019^[19]); PIAAC Expert Group in Problem Solving in Technology-Rich Environments (2009^[10]).

Figure 7.4 provides an example item from PISA 2012 assessment of creative problem solving (OECD, 2014^[7]). In a simulated microworld, test takers need to figure out how to use a new air conditioner without further instructions. To this end, they must first explore how the three input variables (i.e. top control, central control and bottom control) influence the outcome variables of temperature and humidity (i.e. sub-process “explore and understand the problem”; see Table 7.2).

In a mental model (see bottom part of Figure 7.4), the test takers then represent how input and output variables are connected based on their explorations. This engages the “represent and formulate a mental model of the problem” sub-process; see Table 7.2). Based on this mental model, test takers derive strategies for solving the problem (i.e. reaching certain target values for temperature and humidity) and subsequently execute them (i.e. plan and execute strategies). At the same time, they monitor their progress (i.e. “monitor and reflect the progress” sub-process; see Table 7.2).

Figure 7.4. Sample item assessing problem solving in PISA 2012

CLIMATE CONTROL: Stimulus information

CLIMATE CONTROL

You have no instructions for your new air conditioner. You need to work out how to use it.

You can change the top, central and bottom controls on the left by using the sliders (▬). The initial setting for each control is indicated by ▲.

By clicking APPLY, you will see any changes in the temperature and humidity of the room in the temperature and humidity graphs. The box to the left of each graph shows the current level of temperature or humidity.

CLIMATE CONTROL: Item 2

Question 2: CLIMATE CONTROL CP025Q02

The correct relationship between the three controls, Temperature and Humidity is shown on the right.

Use the controls to set the temperature and humidity to the target levels. **Do this in a maximum of four steps.** The target levels are shown by the red bands across the Temperature and Humidity graphs. The range of values for each target level is 18-20 and is shown to the left of each red band. **You can only click APPLY four times and there is no RESET button.**

Source: OECD (2014_[7]).

Figure 7.5 displays an example item assessing collaboration in PISA 2015 (OECD, 2017_[3]). The item requires test takers to work in teams to gather information about a fictional country named Xandar. To this end, the test takers need to interact with computer agents in a chat to establish and maintain a shared understanding of the problem. They subsequently plan and execute strategies to solve the problem together with team members (see sub-processes in Table 7.2).

Figure 7.5. Sample item assessing collaboration in PISA 2015

The screenshot shows the PISA 2018 interface. On the left, there's a chat window titled 'Who's in the Chat' with participants 'YOU', 'Alice', and 'Zach'. The chat history shows Alice saying 'Hi. I'm not sure about the best way to do this.' and Zach replying 'Let's just get going.' Below this, a prompt says 'You are continuing the chat. Click on a choice below. Then click on Send.' There are four radio button options for 'You:': 'I wonder if some of the other teams have started yet.', 'I hope the questions are easy.', 'Maybe we should talk about strategy first.', and 'Alice, you can see what to do once we get started.' A 'Send' button is at the bottom. On the right, there's a 'Scorecard' table with columns for 'Geography', 'People', and 'Economy'. Each column has four empty rows for scores. Below the table are three buttons labeled 'Geography', 'People', and 'Economy'.

Scorecard		
Geography	People	Economy

Geography People Economy

Source: (OECD, 2017^[3]).

Role of the two skill sets in occupational settings

This section analyses how core domain and transversal skills relate to explicit skill models in the workforce. It focuses on commonalities and overlap between skills in education and in the workforce. To this end, it introduces two skill models from the workforce (ISCO-08 and O*NET). It also connects the two skill domains with the two skill models to provide an overview of which skills will be important for which types of jobs.

Given the focus of this chapter on scaling AI capabilities, it focuses only on skills found in education and relevant for the workforce. For skill requirements on the job, comprehensive skill models from the workforce, such as ISCO-08 (ILO, 2012^[21]) and O*NET (National Center for O*NET Development, 2020^[22]), have been derived. In ISCO-08, four skill levels with increasing complexity are distinguished (see Table 7.3).

Table 7.3. Skill levels and their description in the workforce skill model ISCO-08

Skill level	Description	Example tasks requiring the skills
1	Skills for performing simple and routine physical or manual tasks.	Cleaning; carrying materials; performing earthworks.
2	Skills for interacting with machines and information.	Operating, maintaining and repairing machines and electronic devices; manipulating, ordering and storing information.
3	Skills for performing complex technical and practical tasks that require extensive factual, technical and specialised knowledge.	Resource calculations for projects; technical support for professionals; ensuring compliance with regulations and schedules.
4	Skills involving complex problem solving, decision making and creativity.	Understanding and communication of complex information; research and diagnose; designing buildings and machines.

Note: Cell content is partly direct quotes.

Source: ILO (2012^[21]).

In the O*NET taxonomy (National Center for O*NET Development, 2020^[22]), skills are not arranged in levels of complexity but rather combined in six different groups (see Table 7.4).

Table 7.4. Skill groups and their description in the workforce skill model O*NET

Skill group	Description	Example skills
1	Basic skills	Mathematics; reading comprehension; writing; science.
2	Complex problem solving	Complex problem solving.
3	Resource management	Management of financial, material, human and time resources.
4	Social skills	Co-ordination; negotiation; persuasion.
5	System skills	Judgement and decision making; systems analysis; systems evaluation.
6	Technical skills	Equipment maintenance; operation and control; repairing.

Note: Cell content is partly direct quotes.

Source: National Center for O*NET Development (2020^[22]).

Table 7.5 summarises the connection of the two skill sets relevant for education (and their components) with the ISCO-08 levels and O*NET groups. In sum, core domain skills are relevant in almost any type of job according to ISCO-08 and O*NET (i.e. no isomorphic mapping). In contrast, transversal skills are mainly required in non-routine, cognitive occupations that are associated with higher ISCO-08 skill levels and the O*NET groups of complex problem solving, social skills and resource management skills.

For the core domain skills, mathematical and reading literacy are important skills in almost any occupation and can therefore be linked to all ISCO-08 skill levels. However, their involvement may vary across skill levels. For instance, ISCO-08 Skill Level 1 requires only minimal mathematical literacy and reading literacy whereas Skill Level 4 requires extensive proficiency in these skills. In contrast, science literacy is only required at the higher ISCO-08 Skill Levels 3 and 4.

With respect to the O*NET taxonomy, some skills are directly allocated (labelled as mathematics, reading comprehension, writing and science) in the group of basic skills. However, mathematical and reading literacy are required for all other skills groups except for group social skills.

Transversal skills are essential for performing non-routine cognitive tasks, and thus not required on the lowest ISCO-08 skill level. In general, these skills are required in ISCO-08 Skill Levels 3 and 4. However, collaboration and global competence might already be useful on Skill Level 3. This level includes jobs such as bus driver or police officer that involve social interactions and require individuals to understand their role and responsibilities in greater groups and society as a whole. In contrast, creative thinking is only connected to ISCO-08 Skill Level 4.

Table 7.5. Integration of core domain skills and transversal skills into the workforce skill models ISCO-08 and O*NET

Skill domain	Skill	ISCO-08 Skill level	O*NET Skill group
Core domain skills	Mathematical literacy	1-4	Primarily basic skills, but also complex problem-solving skills, resource management skills, systems skills and technical skills
	Reading literacy	1-4	Primarily basic skills, but also complex problem-solving skills, resource management skills, systems skills and technical skills
	Science literacy	3-4	Basic skills
Transversal skills	Problem solving	3-4	Complex problem-solving skills
	Collaboration	2-4	Social skills; resource management skills
	Creative thinking	4	Basic skills
	Global competence	2-4	Social skills; resource management skills

How to scale the capabilities of artificial intelligence? Some recommendations

This chapter identified two sets of skills relevant for concurrent educational efforts: skills from the core domains and transversal skills. It has provided definitions for each of these skill sets, referenced theoretical frameworks and provided examples of items. It has also drawn connections to ISCO-08 and O*NET, two commonly used frameworks to describe demands on the job. Through these frameworks, it has shown that many skills relevant in education can be found, in one way or another, in taxonomies of work requirements.

From a content perspective, both skill sets (and the specific skills therein) are relevant for educational success and beyond. The same holds for the set of basic cognitive skills covered in Chapter 3. Thus, in principle, all of them can be used to assess and scale the capability of AI. This is comparable to the assessment of the capacity of a student or educational system on different levels and skills. In fact, it is for good reasons that international large-scale assessments measure different dimensions to gain an

adequate overview. The same approach should be considered when assessing AI capabilities (i.e. looking at different dimensions from all available skill sets).

In addition to a broad content coverage of different skill sets from the set of domain-specific, transversal and basic skills, educational tests to scale AI capabilities should consider these five recommendations.

Recommendations

- **Select skills based on established relevance**

There is a limited number of skills towards which the capability of AI to reproduce human capabilities can be assessed and, subsequently, scaled. With this in mind, selection of skills should be driven by the available body of existing theory-driven empirical research. Moreover, it should include only skills for which there is a minimal level of agreement across researchers, practitioners and policy makers about their theoretical, empirical and curricular relevance. For instance, mathematics and science are clearly relevant skills that have been part of curricular teaching across the globe for decades. Conversely, transversal skills have accumulated less research and are not (yet) consistently included in school curricula. A hierarchy of skills that considers consensus on theoretical, empirical and practical levels, and curricular relevance in education and the workforce, will be beneficial when assessing and scaling AI-related capabilities.

- **Ensure enough high-quality items are available to measure the skill**

AI capabilities, as all other skills, should not be judged on specific items and their content but rather on underlying psychological constructs (so-called latent traits). To this end, assessments need sufficiently high numbers of items that all tap into the same construct in a reliable and valid manner. More specifically, this implies that empirical research has identified a set of items as reliable and valid indicators of a measurement. Moreover, many items are needed to allow for testing across different item contexts. This becomes particularly challenging when looking at scenario-based, computer-administered item types. Such item types are laborious to develop and so usually fewer are available, even though they are better representations of a real-world scenario. Thus, when choosing assessments to scale AI capabilities, constructs with large and empirically tested item banks should be preferred.

- **Use only skills linked to a measurement theory to scale AI-related capabilities**

Only skills for which items as the direct observable entities and the construct as the latent entity are explicitly linked through a measurement theory should be used to scale AI-related capabilities. Different measurement theories are available but most large-scale assessments use Item-Response-Theory, which links specific item responses to the underlying traits in a probabilistic way. The recommendation primarily relates to measurement theory but extends to the need for scoring rules and linking procedures that are clearly spelled out and theoretically justified. Linking of item banks from different studies is difficult and requires specific statistical procedures. It is, for instance, not possible to link items of different assessments empirically such as PISA or GRE, even if they theoretically claim to measure the same concept. Thus, skills for which measurement is rooted in an established psychological measurement theory and for which enough items are empirically linked to each other are preferred.

- **Ensure the underlying process towards the correct solution of item can be described**

Scaling AI capabilities is unlikely to stop at the evaluation of whether an AI algorithm can solve a particular item or where it stands on a dimensionally measured construct. It will be more informative to measure and describe at what point an AI algorithm fails in solving an item and its distance from a pre-specified goal (i.e. the solution). Scenario-based items usually provide information that goes beyond a correct/incorrect judgement of the response and bear the potential to explicate more fine-grained information. Of note, it would also be interesting to see whether AI can improve after receiving feedback (related to the concept

of formative assessment). To this end, preference should be given to assessments that provide information beyond the mere correctness of a response and provides data on the underlying solution process.

- **Use items to assess skills that AI can understand and perform**

Only tests that involve tasks AI can understand and that contain operators that, in principle, AI can perform should be used. For instance, it is not meaningful to confront an AI algorithm with a task that requires some physical intervention.

The five recommendations should be carefully weighed against each other when scaling AI capabilities for educationally relevant skills. This exercise allows choosing a taxonomy that predicts which skills AI can replace and make human input – at least to some extent – superfluous.

This is an interdisciplinary undertaking that requires substantial evaluation and great expertise. It should involve experts from relevant fields including education, cognitive science, computer science, AI and machine learning, and economy with both scientific and policy perspective. Together, they can fill such a taxonomy with the aim of making informed judgements on how the state of the art allows to predict the role of AI in education and the workforce within the next decades.

References

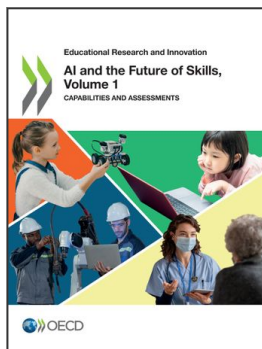
- Autor, D., F. Levy and R. Murnane (2003), “The skill content of recent technological change: An empirical exploration”, *The Quarterly Journal of Economics*, Vol. 118/4, pp. 1279-1333, <https://doi.org/10.1162/003355303322552801>. [23]
- Greiff, S. and J. Funke (2009), “Measuring complex problem solving: The MicroDYN approach”, in Scheuermann, F. and J. Björnsson (eds.), *The Transition to Computer-based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing*, OPOCE, Luxembourg. [8]
- Greiff, S. et al. (2016), “Understanding students’ performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files”, *Computers in Human Behavior*, Vol. 61, pp. 36-46, <https://doi.org/10.1016/j.chb.2016.02.095>. [20]
- Greiff, S. et al. (2017), “Adaptive problem solving: Moving towards a new assessment domain in the second cycle of PIAAC”, *OECD Education Working Papers*, No. 156, OECD Publishing, Paris, <https://dx.doi.org/10.1787/90fde2f4-en>. [18]
- Griffin, P. and E. Care (2015), *Assessment and Teaching of 21st century Skills: Methods and Approach (1st ed.)*, Springer, <https://doi.org/10.1007/978-94-017-9395-7>. [17]
- ILO (2012), *International Standard Classification of Occupations: ISCO-08*, International Labour Organization, Geneva. [21]
- International Baccalaureate Organization (n.d.), “Curriculum”, webpage, <https://www.ibo.org/programmes/diploma-programme/curriculum/> (accessed on 20 October 2021). [27]
- International Baccalaureate Organization (n.d.), “Why the IB is Different”, webpage, <https://www.ibo.org/benefits/why-the-ib-is-different/> (accessed on 20 October 2021). [28]

- Mainert, J. et al. (2019), "The incremental contribution of complex problem-solving skills to the prediction of job level, job complexity and salary", *Journal of Business Psychology*, Vol. 34, pp. 825-845, <https://doi.org/10.1007/s1>. [24]
- McGrew, K. (2009), "CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research", *Intelligence*, Vol. 37/1, pp. 1-10, <https://doi.org/10.1016/j.intell.2008.08.004>. [5]
- Mullis, I. and M. Martin (eds.) (2017), *TIMSS 2019 Assessment Frameworks*, TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement, Amsterdam. [15]
- National Assessment and Governing Board (2019), *Mathematics Framework for the 2019 National Assessment of Educational Progress*, National Assessment Governing Board, Washington, DC. [12]
- National Assessment and Governing Board (2019), *Reading Framework for the 2019 National Assessment of Educational Progress*, National Assessment Governing Board, Washington, DC. [13]
- National Assessment and Governing Board (2019), *Science Framework for the 2019 National Assessment of Educational Progress*, National Assessment Governing Board, Washington, DC. [14]
- National Center for O*NET Development (2020), "O*NET OnLine", webpage, <https://www.onetonline.org/> (accessed on 20 October 2021). [22]
- OECD (2019), *PISA 2018 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris, <https://dx.doi.org/10.1787/b25efab8-en>. [9]
- OECD (2019), *PISA 2018 Results (Volume I): What Students Know and Can Do*, PISA, OECD Publishing, Paris, <https://dx.doi.org/10.1787/5f07c754-en>. [6]
- OECD (2019), *PISA 2021 Creative Thinking Framework (Third Draft)*, PISA, OECD Publishing, Paris. [19]
- OECD (2019), *Skills Matter: Additional Results from the Survey of Adult Skills*, OECD Skills Studies, OECD Publishing, Paris, <https://dx.doi.org/10.1787/1f029d8f-en>. [4]
- OECD (2017), *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*, PISA, OECD Publishing, Paris, <https://dx.doi.org/10.1787/9789264281820-en>. [1]
- OECD (2017), *PISA 2015 Results (Volume V): Collaborative Problem Solving*, PISA, OECD Publishing, Paris, <https://dx.doi.org/10.1787/9789264285521-en>. [3]
- OECD (2014), *PISA 2012 Results: Creative Problem Solving (Volume V): Students' Skills in Tackling Real-Life Problems*, PISA, OECD Publishing, Paris, <https://dx.doi.org/10.1787/9789264208070-en>. [7]
- OECD (2014), *PISA 2012 Results: What Students Know and Can Do (Volume I, Revised edition, February 2014): Student Performance in Mathematics, Reading and Science*, PISA, OECD Publishing, Paris, <https://dx.doi.org/10.1787/9789264208780-en>. [16]

- OECD (2013), *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, PISA, OECD Publishing, Paris, <https://dx.doi.org/10.1787/9789264190511-en>. [2]
- Pellegrino, J. and M. Hilton (2013), *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*, National Academies Press, Washington, D.C. [25]
- PIAAC Expert Group in Problem Solving in Technology-Rich Environments (2009), "PIAAC Problem Solving in Technology-Rich Environments: A Conceptual Framework", *OECD Education Working Papers*, No. 36, OECD Publishing, Paris, <https://dx.doi.org/10.1787/220262483674>. [10]
- PIAAC Literacy Expert Group (2009), "PIAAC Literacy: A Conceptual Framework", *OECD Education Working Papers*, No. 34, OECD Publishing, Paris, <https://dx.doi.org/10.1787/220348414075>. [11]
- PIAAC Numeracy Expert Group (2009), "PIAAC Numeracy: A Conceptual Framework", *OECD Education Working Papers*, No. 35, OECD Publishing, Paris, <https://dx.doi.org/10.1787/220337421165>. [26]

Notes

¹ Another assessment program, the *International Baccalaureate* (International Baccalaureate Organization, n.d.^[28]), is not strictly a large-scale assessment. The IB aims at equipping students with subject-specific as well as subject-general skills through different programs. In the Diploma Programme, for example, students complete three core disciplines: theory of knowledge (i.e. reflecting on the concept of knowledge), extended essay (i.e. conducting and reporting a piece of research), creativity, activity, and service (International Baccalaureate Organization, n.d.^[27]). In addition, students choose one subject out of each of the following six subject groups: Studies in language and literature, Language acquisition, Individuals and societies, Sciences, Mathematics and the arts (International Baccalaureate Organization, n.d.^[27]).



From:

AI and the Future of Skills, Volume 1 Capabilities and Assessments

Access the complete publication at:

<https://doi.org/10.1787/5ee71f34-en>

Please cite this chapter as:

Greiff, Samuel and Jan Dörendahl (2021), "Skills assessments in education", in OECD, *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, Paris.

DOI: <https://doi.org/10.1787/68191ce9-en>

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Extracts from publications may be subject to additional disclaimers, which are set out in the complete version of the publication, available at the link provided.

The use of this work, whether digital or print, is governed by the Terms and Conditions to be found at <http://www.oecd.org/termsandconditions>.