3 Methodology for assessing AI capabilities using the Survey of Adult Skills (PIAAC)

This chapter describes the methodology of assessing computers' capabilities to solve the questions of the Survey of Adult Skills of the Programme for International Assessment of Adult Competencies (PIAAC). It first provides an overview of the PIAAC test, the skills it measures and the test questions used to measure them. The chapter then describes the methods used to select experts, to collect expert judgement, to develop the questionnaire and to construct aggregate measures of artificial intelligence (AI) capabilities in literacy and numeracy. The focus is on the methodological improvements on the assessment approach used in the pilot study. The chapter concludes with a summary of the methodological challenges encountered in the study and the attempts to solve them.

In 2016, the OECD asked a group of computer scientists to assess the capabilities of computers with regard to the core skills measured in the Survey of Adult Skills within the Programme for International Assessment of Adult Competencies (PIAAC) (Elliott, 2017[1]). The goal was to provide a way of anticipating how potential changes in technology could affect use of these skills in work and everyday life. The current follow-up study looks at how AI capabilities in literacy and numeracy have evolved since the last assessment. It explores new methods for collecting expert judgement on artificial intelligence (AI) skills to address some methodological challenges and refine existing measures.

This chapter describes the approach used to assess AI capabilities and the methodological improvements introduced in the course of the work. After an overview of PIAAC, the chapter outlines the techniques used to select experts, obtain judgements from them, obtain qualitative feedback on those judgements and produce aggregate ratings on AI capabilities. The last section discusses challenges in the study and steps taken to address them.

Overview of the Survey of Adult Skills (PIAAC)

The Survey of Adult Skills (PIAAC) examines the proficiency of adults aged 16-65 in literacy, numeracy and problem solving with computers. These skills are conceived as "key information-processing competencies" since they are necessary for fully integrating into work, education and social life, and are relevant to many social contexts and work situations (OECD, 2013[2]). In addition, the survey collects rich information on respondents' background and context, including participation in reading- and numeracy- related activities, the use of information and communication technologies at work and in everyday life, collaborating with others and organising one's time.

This study focuses on the numeracy and literacy assessments of PIAAC. Literacy and numeracy constitute a foundation upon which individuals can develop higher-order cognitive skills, such as analytic reasoning. In information-rich societies, these skills are essential for understanding specific domains of knowledge. Moreover, they are also needed for gaining access to information relevant for everyday life, such as reading medical prescriptions or handling money and budgets (OECD, 2012[3]). The following subsections provide more information on the approach to assessing these skills, describing the formats of test questions, as well as the contexts and cognitive strategies they address.

PIAAC is conducted every ten years. The First Cycle took place between 2011 and 2018. First results from the Second Cycle are expected in 2024. In the First Cycle, data from 39 countries and economies were gathered in three rounds. The first round surveyed around 166 000 adults in 24 countries (or regions within these countries) in 2011-12. These include Australia, Austria, Belgium (the data were collected in Flanders), Canada, Cyprus, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Ireland, Italy, Japan, Korea, the Netherlands, Norway, Poland, the Slovak Republic, Spain, Sweden, the United Kingdom (the data were collected in England and Northern Ireland) and the United States. The second round took place between 2014 and 2015 and covered Chile, Greece, Indonesia, Israel, Lithuania, New Zealand, Singapore, Slovenia and Türkiye. The third round was conducted in 2017 with Ecuador, Hungary, Kazakhstan, Mexico, Peru and the United States. Approximately 250 000 adults were surveyed in the First Cycle, with national samples ranging from about 4 000 to nearly 27 300 (OECD, 2019[4]).

In the process of scoring the assessment, a difficulty score is assigned to each task, based on the proportion of respondents who complete it successfully. These scores are represented on a 500-point scale for each of the three domains. Respondents are placed on the same 500-point scale, using the information about the number and difficulty of the questions they answer correctly. At each point on the scale, an individual with a proficiency score of that particular value has a 67% chance of successfully completing test items located at that point. This individual will also be able to complete more difficult items with a lower probability of success and easier items with a greater chance of success (OECD, 2013[5]).

To help interpret the results, the reporting scales for each domain are divided into a small number of proficiency levels. Six proficiency levels are defined for literacy and numeracy (Levels 1 through 5 plus below Level 1). With the exception of the lowest level (below Level 1), tasks located at a particular level can be successfully completed approximately 67% of the time by a person with a proficiency score in the middle of the range defining the level. In other words, a person with a score in the middle of Level 2 would score close to 67% in a test made up of items of Level 2 difficulty (OECD, 2013[5]).

The information on level and distribution of proficiency in the population is useful for policy makers and researchers concerned with issues such as the development of skills of the labour force or the efficacy of the education system. In addition, PIAAC data can help understand the relationship between key skills and economic and social outcomes, and the factors related to acquiring, maintaining and losing skills.

Assessing literacy in the Survey of Adult Skills (PIAAC)

The PIAAC literacy test measures adults' ability to understand, evaluate, use and engage with written texts in real-life situations. The tasks contain texts that adults typically encounter in work and personal life. Examples include job postings, webpages, newspaper articles and e-mails. These texts are presented in different formats – as print texts, digital texts, continuous texts, sentences formed into paragraphs or non- continuous texts, such as those appearing in charts, lists or maps. Items can also contain multiple texts that are independent from each other but linked for a particular purpose (OECD, 2012[3]; OECD, 2013[5]).

The literacy test requires readers to use three broad cognitive strategies when responding to a text:

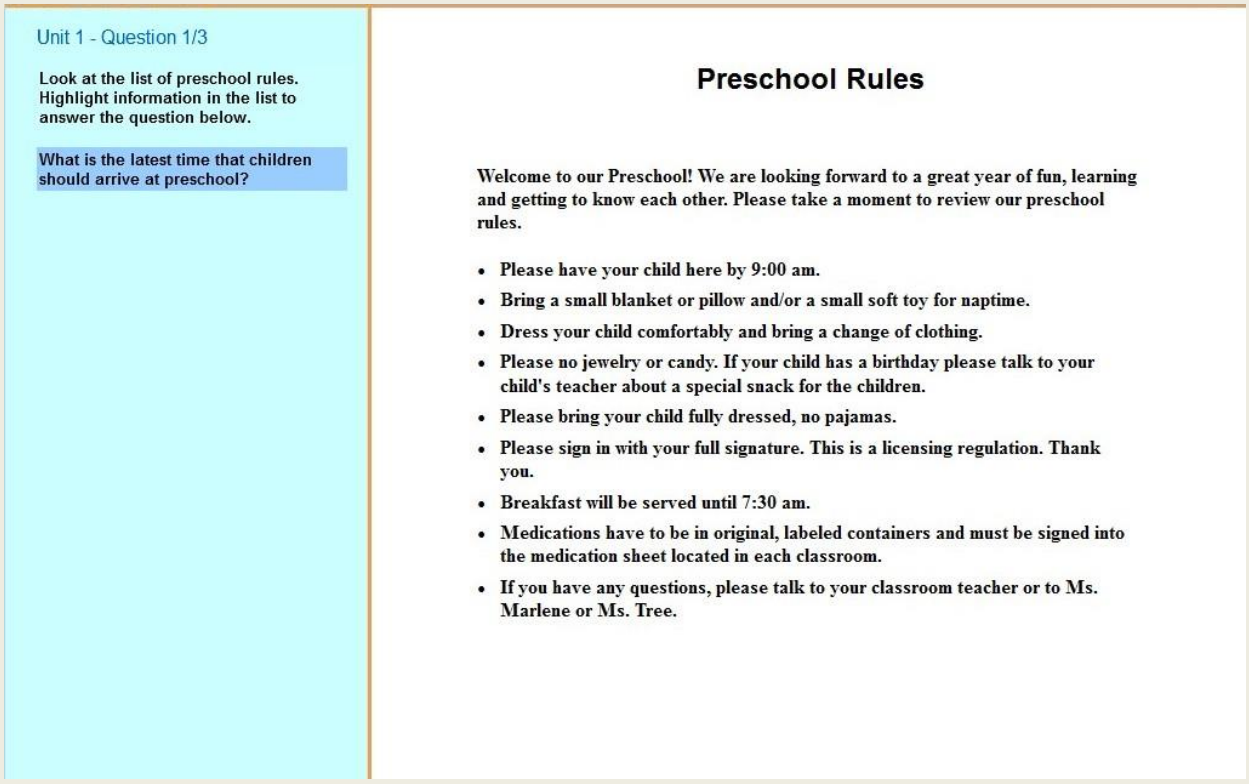
- *Access and identify*: tasks require the reader to locate items of information in a text. Sometimes this is relatively easy, as the required information is directly and plainly stated in the text. However, some tasks may require inferences and rhetorical understanding (e.g. identifying the reasons behind a policy by the local government).
- *Integrate and interpret*: tasks may require the reader to understand the relation(s) between different parts of a text, such as those of problem/solution or cause/effect. These relationships may be explicitly signalled (e.g. the text states that "the cause of X is Y") or may require the reader to make inferences.
- *Evaluate and reflect*: tasks may require readers to draw on knowledge or ideas external to the text, such as evaluating the relevance, credibility or argumentation of the text.

Literacy tasks have six difficulty levels (OECD, 2012[3]; OECD, 2013[5]). Easy tasks (below Level 1 and at Level 1) require knowledge and skills in recognising basic vocabulary and reading short texts. Tasks typically require the respondent to locate a single piece of information within a brief text. In intermediate- level tasks (Levels 2 and 3), understanding text and rhetorical structures becomes more central, especially navigating complex digital texts. Texts are often dense or lengthy. They may require the respondent to construct meaning across larger chunks of text or perform multi-step operations to identify and formulate responses. Hard tasks (Levels 4 and 5) require complex inferences and application of background knowledge. Texts are complex and lengthy and often contain competing information that is seemingly as prominent as correct information. Many tasks require interpreting subtle evidence-based claims or persuasive discourse relationships.

Box 3.1. Example for literacy questions

The example literacy item presented below has a difficulty level of 3. It uses print-based materials (as opposed to digital stimuli such as simulated websites). It requires respondents to access and identify the correct information in the text.

Figure 3.1. Literacy – Sample item



Source: OECD (2012[3]), *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*, http://dx.doi.org/10.1787/9789264128859-en.

Assessing numeracy in the Survey of Adult Skills (PIAAC)

The PIAAC numeracy test measures the ability to access, use, interpret and communicate mathematical information and ideas to manage the mathematical demands of everyday life (OECD, 2012[3]; OECD, 2013[5]). The tasks are designed to resemble real situations from work and personal life, such as managing budgets and project resources, and interpreting quantitative information presented in the media. The mathematical information can be presented in many ways, including images, symbolic notations, formulae, diagrams, graphs, tables and maps. Mathematical information can be further expressed in textual form (e.g. “the crime rate increased by half”).

Tasks can require different cognitive strategies:

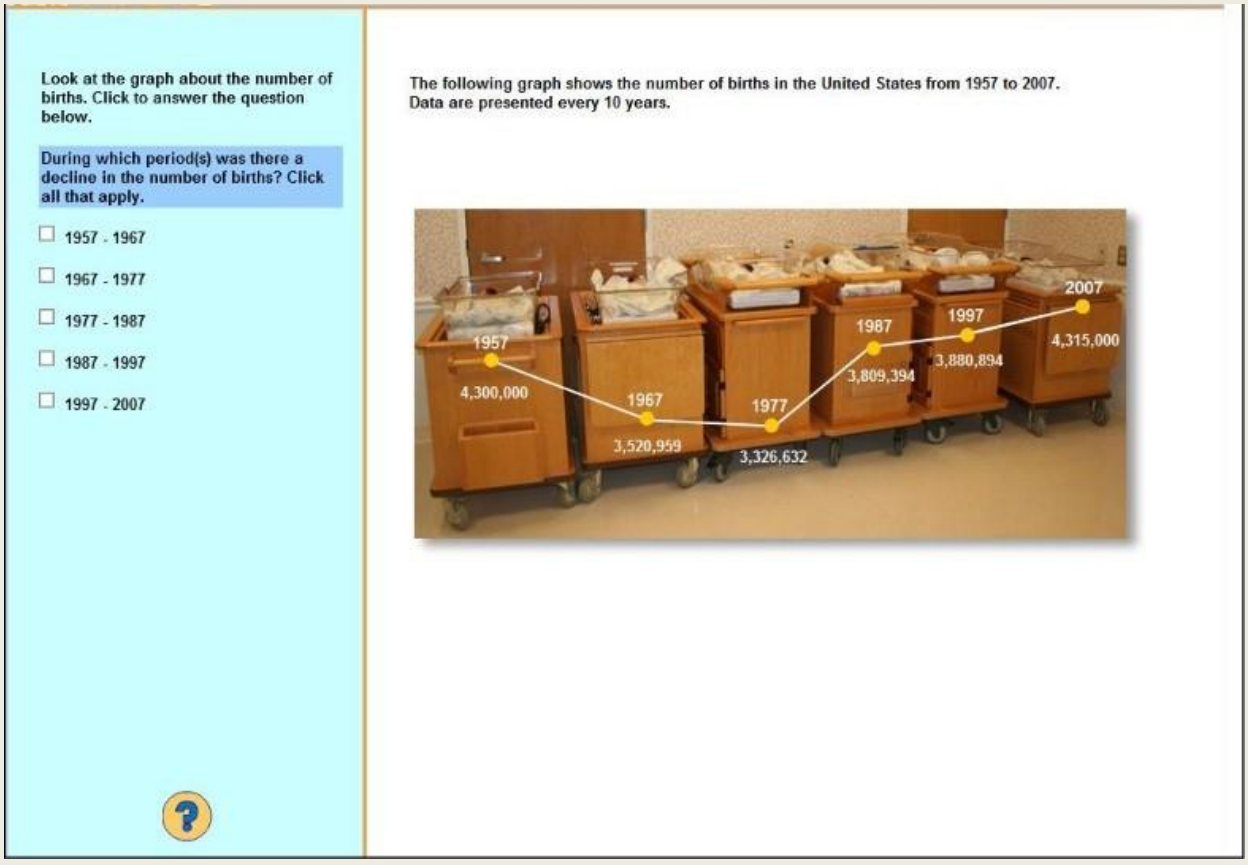
- *Identify, locate, or access mathematical information* that is present in the task and relevant to their purpose or goal.

- *Use mathematical knowledge*, i.e. apply known methods, rules or information, such as counting, ordering, sorting, estimating, using various measuring devices or using (or developing) a formula.
- *Interpret* the meaning and implications of mathematical information, e.g. regarding trends, changes or differences described in a graph or in a text.
- *Evaluate/analyse* the quality of the solution against some criteria or contextual demands (e.g. compare information regarding the costs of competing courses of action).

Box 3.2. Example for numeracy questions

This sample item is of difficulty level 3. It involves the cognitive strategies *Interpret* and *Evaluate*. Respondents are asked to click on one or more of the time periods provided in the left pane on the screen.

Figure 3.2. Numeracy - Sample item



Source: OECD (2012[3]), *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*, http://dx.doi.org/10.1787/9789264128859-en.

Tasks vary across six levels of difficulty (OECD, 2013[5]). Easy tasks (below Level 1 and Level 1) require respondents to carry out simple, one-step processes. Examples are counting, understanding simple percentages, or recognising common graphical representations. The mathematical content is easy to locate. Tasks at medium difficulty levels (Levels 2 and 3) require the application of two or more steps or

processes. This can involve calculation with decimal numbers, percentages and fractions, or the interpretation and basic analysis of data and statistics in texts, tables and graphs. The mathematical information is less explicit and can include distractors. Hard tasks (Levels 4 and 5) require understanding and integrating multiple types of mathematical information, such as statistics and chance, spatial relationships and change. The mathematical information is presented in complex and abstract ways or is embedded in longer texts.

Identifying a group of computer scientists

The pilot study relied on the expertise of 11 computer scientists from various fields identified as key to the assessment, including natural language processing, reasoning, commonsense knowledge, computer vision, machine learning and integrated systems (Elliott, 2017[1]). These experts were recommended by social scientists working on the implications of AI for the economy or by other computer scientists. Six of these computer experts also participated in the follow-up study in 2021. Five new experts were recruited for the follow-up study, mostly based on recommendations from the initial expert group.

The assessment results obtained from the 11 experts in 2021 revealed big disagreements in the evaluation of AI capabilities in the numeracy domain (see below). Therefore, four additional experts with an explicit research focus on mathematical reasoning of AI were invited to participate in the follow-up. They re- assessed only the numeracy test of PIAAC. These experts were selected on the basis of their publication list and/or their participation in relevant conferences in the field.

Table 3.1. Computer scientists participating in the follow-up assessment of computer capabilities

Computer scientists	Expertise
Chandra Bhagavatula, Senior Research Scientist, Allen Institute for AI (AI2)	Commonsense reasoning, natural language generation, intersection of commonsense and vision
Anthony G. Cohn, Professor of Automated Reasoning, School of Computing, University of Leeds	Artificial intelligence, knowledge representation and reasoning, data and sensor fusion, cognitive vision, spatial representation and reasoning, geographical information science, robotics
Pradeep Dasigi*, Research Scientist, Allen Institute for AI (AI2)	Natural language understanding, question answering, reading comprehension, executable semantic parsing
Ernest Davis, Professor of Computer Science, Courant Institute, New York University	Representation of commonsense knowledge
Kenneth D. Forbus, Walter P. Murphy Professor of Computer Science and Professor of Education, Northwestern University	Qualitative reasoning, analogical reasoning and learning, spatial reasoning, sketch understanding, natural language understanding, cognitive architecture, reasoning system design, intelligent educational software, and the use of AI in interactive entertainment
Arthur C. Graesser, Emeritus Professor, Department of Psychology, University of Memphis	Cognitive and learning sciences, discourse processing, artificial intelligence and computational linguistics, text comprehension, emotions, problem solving, human and computer tutoring, design of educational software, human-computer interaction
Yvette Graham, Assistant Professor in Artificial Intelligence, School of Computer Science and Statistics, Trinity College Dublin	Natural language processing, dialogue systems, machine translation, information retrieval
Daniel Hendrycks*, Director, Center for AI Safety	Artificial intelligence, machine-learning safety, quantitative reasoning of AI
José Hernández-Orallo, Professor, Valencian Research Institute for Artificial Intelligence, Valencian Graduate School and Research Network of AI, Universitat Politècnica de València	Evaluation and measurement of intelligent systems in general and machine learning in particular

Computer scientists	Expertise
Jerry R. Hobbs, Emeritus Professor, Fellow and Chief Scientist for Natural Language Processing, Information Sciences Institute, University of Southern California	Computational linguistics, discourse analysis, artificial intelligence, parsing, syntax, semantic interpretation, information extraction, knowledge representation, encoding common sense knowledge
Aviv Keren*, Senior Applied Scientist, Anyword	Artificial Intelligence, philosophy of mathematics, mathematical cognition, mathematical logic, natural language processing
Rik Koncel-Kedziorski*, AI Research Scientist, Kensho Technologies	Artificial intelligence, natural language processing, question answering, general methods for representing meaning in natural language processing systems
Vasile Rus, Professor, Department of Computer Science and Institute for Intelligent Systems, University of Memphis	Natural language processing, natural language-based knowledge representations, semantic similarity, question answering, intelligent tutoring systems
Jim Spohrer, Retired Director, Global University Programs and Cognitive Systems Group, IBM	Artificial intelligence, cognitive systems for holistic service systems
Michael Witbrock, Professor, School of Computer Science, University of Auckland	Artificial intelligence, AI for social good, AI entrepreneurialism, natural language understanding, machine reasoning, knowledge representation, deep learning

Note: * Completed an assessment of AI with the PIAAC numeracy test in September 2022.

Collecting expert judgement

The assessment was carried out with an online survey, followed by a group discussion. The participants received the PIAAC test materials for review one week before the start of the survey. They had two weeks to complete it. During this period, they could access, re-access and modify their answers via an individualised survey link. In total, there were 113 test questions to rate, 57 in the literacy domain and 56 in the numeracy domain.

A four-hour online group discussion took place ten days after the online assessment. Prior to the meeting, each expert received a handout showing her or his individual rating on each PIAAC question next to the group average. In the meeting, experts received additional detailed feedback on how the group rated AI's ability to take the PIAAC test. Experts discussed these results, focusing on test questions where there was strong disagreement in the evaluation of AI performance. In addition, the experts described difficulties in understanding and rating the questions and provided feedback on the evaluation approach. After the meeting, the experts had the opportunity to re-enter the survey and revise their answers.

This assessment approach follows the so-called Delphi method for collecting expert judgement. Delphi is a structured group technique for eliciting judgements of multiple experts that aims at improving judgement quality and increasing consensus (Okoli and Pawlowski, 2004[6]; European Food Safety Authority, 2014[7]). It consists of at least two rounds of collecting experts' ratings, with feedback provided after each round on how the group rated on average. The iteration of survey rounds continues until consensus among experts is reached. During each round, experts provide their ratings anonymously and independently from each other. This should reduce potential bias from social conformity or from dominant individuals who impose their opinions on the group. By contrast, the feedback provided after each round should enable social learning and the modification of prior judgements due to new information. This feedback should ultimately increase consensus between experts.

In contrast to a classical Delphi approach, this study allowed for more communication among experts. It provided experts with the mailing list of the group and encouraged them to share any questions, comments or suggestions regarding the survey during the rating process. Several experts made use of this option. After the first round, experts could meet virtually to discuss the survey results. In addition, a group chat during the online meeting enabled them to exchange ideas and materials.

Communication is important for the assessment. All the experts are generally aware of the state of the art in AI domains relevant for performing PIAAC questions. However, they cannot possibly know all AI

applications, recent research results or other details that may be relevant for the evaluation. Only one or a few experts may have knowledge on particular AI systems that can perform a task. In such a case, these experts should be able to communicate this information to the group at any point of the rating process.

Providing more room for interaction is an improvement on the pilot study. In 2016, the assessment was held over a two-day meeting, with materials provided to participants in advance (Elliott, 2017[1]). Given the time constraints, the exchanges on details of a specific technique were limited to mentioning a relevant research article and experts were unable to work towards a full consensus understanding of different computer capabilities. In the follow-up study, experts did not reach consensus on many matters. However, they could share their views with the group during the entire process of data collection.

Developing the questionnaire

The online survey contained the literacy and numeracy questions from PIAAC. For each question, experts were asked about their confidence in AI technology carrying out the task. The response options were "0% – No, AI cannot do it", "25%", "50% – Maybe", "75%", "100% – Yes, AI can do it" and "Don't know". This scale combines both experts' confidence and their rating of the capability of AI. For example, "0% No, AI cannot do it" means that experts are quite certain that AI cannot carry out the task, while 25% means that experts think that AI probably cannot do it.

The study gave experts detailed instructions that defined the parameters for evaluating the potential use of AI on the PIAAC test. There was no reason to expect systems tailored to the tasks in the test. Therefore, experts considered the process of adapting techniques to the context of PIAAC. Such an adaptation can involve training the system on a set of relevant examples or coding information about specific vocabularies, relationships or types of knowledge representation, such as charts and tables. Experts needed boundaries on the size of the hypothetical development effort required to develop a computer system using current techniques to answer test questions. As in the past assessment, two rough criteria were used for experts to consider in their judgements.

First, the instructions asked experts to think of "current" computer techniques, meaning any available techniques addressed sufficiently in the literature. This is important since the assessment is intended to reflect the application of current systems not the creation of entirely new ones.

Second, the instructions asked experts to consider a "reasonable advance preparation" to adapt current techniques to PIAAC. This was defined as USD 1 million over one year for a research team to build and refine a system to work with PIAAC questions using current techniques. In addition, the instructions asked experts to imagine development of two separate systems – one for solving all literacy items, the other for the numeracy test.

The follow-up study attempted to address some methodological challenges encountered in the pilot. Experts had pointed out that tests developed for humans generally omit capabilities that most people share but machines do not (Elliott, 2017[1]). In other words, computers may perform poorly due to a lack of capabilities taken for granted in humans rather than from a lack of the primary capabilities being assessed. This raises problems for interpreting computer performance on human tests. One task in PIAAC, for example, requires counting packaged bottles in an image. This question is clearly easy for most adults. The numerical reasoning aspect of the question is also easy for machines. However, the experts gave AI the lowest rating on this question because the packaging makes many of the bottles unrecognisable for machines. The question becomes a misleading measure of computer numeracy on its own because it requires additional object recognition capabilities.

There was a need to disentangle the literacy and numeracy skills being measured from the capabilities needed for a task but not subject to PIAAC. Some experts suggested two stages for the rating process: identifying different types of capabilities needed for each task, and then evaluating AI performance in each

area. However, such an exercise would require experts to agree on a set of categories to describe the different types of capabilities and determine the ones needed for each task.

Instead of adopting the "two-stage" solution, the survey included an additional open-ended question: "If you think that AI cannot carry out the entire task or you are uncertain about it, would you say that AI can carry out parts of the task? If so, which part(s)?" This question was intended to specify the elements of the task that are easy and hard for machines to perform. In this way, it would provide more precise information on computer performance on challenging PIAAC tasks.

In addition, the follow-up study attempted to collect more qualitative information on the rationales behind experts' ratings compared to the pilot study. To that end, an open-ended question followed each PIAAC question. It asked experts to explain their answers about AI performance on that question. At the end of the literacy and the numeracy parts of the test, experts could report any difficulties in understanding or answering the questions in the domain or leave any comments or suggestions.

Finally, the follow-up survey asked all experts to predict the capabilities of AI with respect to each PIAAC question in 2026. These projections were assessed to explore possibilities for tracking AI development over time. The pilot study had asked experts to predict technological improvements ten years in the future. By contrast, the follow-up study used a period of five years. Many grant applications require investigators to project the results of their own research over three to five years. Researchers, thus, have regular experience in estimating the degree of change that can occur over this shorter period.

Constructing aggregate measures of AI literacy and numeracy performance

The follow-up study considers both the extent of agreement and the extent of uncertainty among experts in aggregating their ratings into single measures for literacy and numeracy performance of AI.

First, it labels each PIAAC question as possible or impossible for AI to solve based on what most experts judged. It thus excludes questions on which experts could not reach majority agreement from the analysis. It then constructs the aggregate measures for AI performance in literacy and numeracy as the percentage share of PIAAC questions in a domain that AI can answer correctly according to the majority of computer experts. These measures are presented for different levels of question difficulty to provide a more detailed picture of potential AI performance on PIAAC.

Second, the study presents different versions of the measures to account for uncertainty among experts. One type of measures relies on ratings weighted by the confidence level that experts report. For example, with respect to the question of how confident experts are that AI can carry out the task, an answer of "75%" is considered as a 75%-Yes. This means it is given a smaller weight than a confident answer of 100%. In some versions of the measures, the Maybe-ratings are omitted because they do not provide a meaningful evaluation of AI. In other versions, the Maybe-ratings are counted as 50%-Yes; this reflects that some experts interpret this answer category as a not very certain Yes. Additional analyses are performed after excluding questions that receive many Maybe- and Don't know-answers to test whether experts' uncertainty influences overall ratings.

Challenges and lessons learned

The follow-up study started with collecting judgements from 11 AI experts. Disagreement among the experts was a major challenge in the assessment, especially around the potential performance of AI on numeracy questions. Two extreme groups emerged: four experts were pessimistic and another four were optimistic about AI's capabilities to perform the numeracy test.

The qualitative information collected in the online survey and the group discussion provided some insights into experts' disagreement. While several points seemed to cause dissent, the major reason for disagreement related to how general the computer capabilities being assessed are supposed to be. Some experts considered general computer techniques that should be successful on a wide range of comparable questions. They tended to give lower ratings for AI capabilities since such general techniques are still limited. Other experts, by contrast, assumed techniques geared specifically to work on a single question and evaluated such "narrow" capabilities more positively. To reach agreement, the experts thus needed clarification on the generality of the AI capabilities being evaluated.

More examples of test questions is one way to clarify generality of AI capabilities. This can help experts picture the full range of problems that AI is supposed to solve in each of the domains. However, providing more examples is not possible since PIAAC has a limited set of questions. Therefore, several other steps were taken to revise the instructions for rating.

First, information from PIAAC's assessment framework was used to describe more precisely the literacy and numeracy skills subject to the evaluation (OECD, 2012[3]). The document defines these skills, describes the contexts and situations in which they are typically applied and characterises the tasks used to measure them. This information was synthesised and supplemented by nine example items of low, medium and high levels of difficulty. This should help experts better understand the domain, the tasks it involves and the capabilities required for carrying them out.

Second, the revised instructions ask experts to imagine and describe an AI system for each domain, based on the synthesised information from PIAAC's assessment framework and the examples provided. Experts are then asked to rate the potential success of their imagined system on each of the PIAAC items in the online survey. During the discussion, experts often argued about the technicalities of producing a system that can manage tasks as variable as those included in the numeracy test. However, there was no time for all experts to share their views. Asking experts in advance to describe a potential system for the test and making these descriptions accessible to all participants may eventually help experts to reach a consensus.

None of the initial 11 experts made use of the opportunity to revise their ratings of AI capabilities after the group discussion. The follow-up study invited four additional experts in mathematical reasoning of AI to re- assess the numeracy test to improve evaluation in the numeracy domain. These experts were identified based on publications and participation in relevant events in the field. They evaluated AI on each of the numeracy questions, following the revised instructions for rating, and met in an online meeting to discuss the survey results.

Nevertheless, the four experts delivered diverging evaluations of AI capabilities in numeracy. However, the discussion showed this is not the result of ambiguity regarding the rating exercise. Experts were clear on AI capabilities required for the numeracy test and how broad these should be. Instead, the instruction to consider some advance preparation for adapting AI systems to PIAAC seemed to make it difficult for experts to provide precise ratings. Some experts argued that, given the recent surge in AI research on mathematical reasoning, AI numeracy capabilities will improve within the period specified for preparing systems for PIAAC. Others focused on the current state of AI techniques. However, all experts agreed that AI is currently not at the stage of solving the numeracy test but will reach this stage soon.

References

Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/9789264284395-en. [1]

European Food Safety Authority (2014), "Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment", *EFSA Journal*, Vol. 12/6, https://doi.org/10.2903/j.efsa.2014.3734. [7]

OECD (2021), *The Assessment Frameworks for Cycle 2 of the Programme for the International Assessment of Adult Competencies*, OECD Skills Studies, OECD Publishing, Paris, https://doi.org/10.1787/4bc2342d-en. [8]

OECD (2019), *Skills Matter: Additional Results from the Survey of Adult Skills*, OECD Skills Studies, OECD Publishing, Paris, https://doi.org/10.1787/1f029d8f-en. [4]

OECD (2013), *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*, OECD Publishing, Paris, https://doi.org/10.1787/9789264204256-en. [2]

OECD (2013), *The Survey of Adult Skills: Reader's Companion*, OECD Publishing, Paris, https://doi.org/10.1787/9789264204027-en. [5]

OECD (2012), *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*, OECD Publishing, Paris, https://doi.org/10.1787/9789264128859-en. [3]

Okoli, C. and S. Pawlowski (2004), "The Delphi method as a research tool: an example, design considerations and applications", *Information & Management*, Vol. 42/1, pp. 15-29, https://doi.org/10.1016/j.im.2003.11.002. [6]