

4

Experts' assessments of AI capabilities in literacy and numeracy

This chapter describes the results of the follow-up assessment of computer capabilities with the Survey of Adult Skills (PIAAC). It first presents the results of the literacy assessment and then the results for numeracy. The chapter studies AI performance by question difficulty by exploring different ways of aggregating experts' ratings. It then shows the average evaluations of the individual experts and analyses disagreement and uncertainty among them. Subsequently, the chapter provides a comparison of artificial intelligence (AI) and adults' performance. Finally, the expert discussion of the rating exercise is summarised to illustrate challenges that experts faced in assessing AI with PIAAC.

This chapter describes the results of the follow-up assessment of computer capabilities with the Survey of Adult Skills (PIAAC). This assessment was carried out in 2021 by a group of 11 computer scientists using the approach described in Chapter 3. The participants rated the potential performance of current artificial intelligence (AI) with regard to each of the questions in the literacy and numeracy domains of PIAAC. In making these evaluations, experts considered a hypothetical development effort for adapting AI techniques to PIAAC that lasts no longer than one year and costs no more than USD 1 million.

Due to disagreement among experts with regard to AI capabilities in numeracy, four additional experts in mathematical reasoning of AI were invited to re-assess the numeracy test. This assessment followed a revised approach, where experts received more information on PIAAC in advance and were asked to provide more information on the technologies that can potentially carry out the test. The chapter first discusses the results of the literacy assessment and then the results for numeracy.

In general, the experts projected a pattern of performance for AI in the upper-middle part of the adult proficiency distribution on PIAAC. In literacy, the results suggest that current computer techniques can perform roughly like adults at proficiency Level 3 in the test. In numeracy, the results suggest that AI performance is closer to adult proficiency at Level 2 for easier questions, and to adult proficiency at Level 3 for harder questions. However, not all experts agree on the latter finding.

Evaluation of AI capabilities in the domain of literacy

Literacy in PIAAC is defined as “understanding, evaluating, using and engaging with written texts to participate in society, to achieve one’s goals, and to develop one’s knowledge and potential” (OECD, 2012_[1]). It is assessed with questions in different formats, including both print-based and digital texts, continuous prose and non-continuous document texts, as well as questions that mix several types of text or include multiple texts. These questions require the decoding of written words and sentences, as well as the comprehension, interpretation and evaluation of complex texts; they do not include writing. The questions are drawn from several contexts that will be familiar to most adults in developed countries, including work, personal life, society and community, and education and training.

Literacy questions are described in terms of six difficulty levels, ranging from below Level 1 to Level 5 (OECD, 2013_[2]). The easier test items involve short texts on familiar topics and questions with the same wording as the answer contained in the text. The harder test items involve longer and sometimes multiple texts on less familiar topics, questions that require some inference from the text and distracting information in the text that can lead to a wrong answer. In the following, below Level 1 and Level 1 are combined into one single question category as well as are Level 4 and Level 5. Seven of the 57 literacy questions in PIAAC are at Level 1 difficulty or below; 15 questions are at Level 2; 23 questions are at Level 3; and 12 questions are at Level 4 or above (see also Chapter 3 for an overview of PIAAC).

AI literacy ratings by question difficulty

Figure 4.1 shows the average share of literacy questions that AI can answer correctly at each difficulty level according to the majority of experts. For each question, experts provided a rating on a scale from “0% – No, AI cannot do it” to “100% – Yes, AI can do it”. This scale reflects both experts’ judgement of AI capabilities and their confidence in this judgement. Three types of aggregate measures are computed from these ratings:

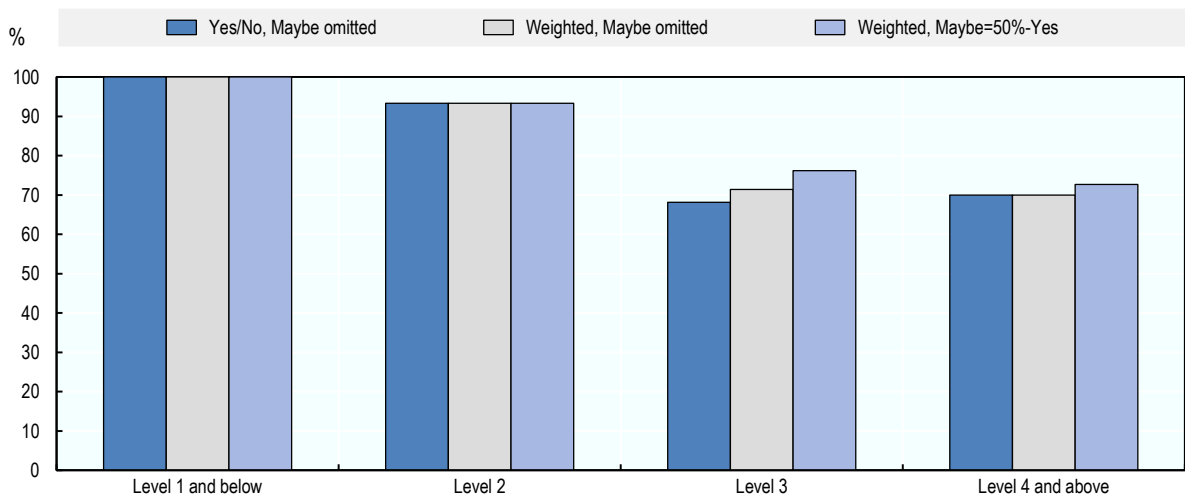
- Ratings of 0% and 25% are counted as No and answers of 75% and 100% are treated as Yes. PIAAC questions are then labelled as doable or not doable for AI according to the answer of more than half of the experts. Experts who gave Maybe- or Don’t know- answers are not considered. Finally, the share of questions that AI can answer correctly according to most experts is calculated for each difficulty level.

- The second version is similar to the first but it weighs ratings by experts' confidence. That is, ratings of 25% and 75% are given a smaller weight than confident ratings of 0% and 100%.
- A third version additionally includes Maybe-ratings as partial Yes-answers (Yes weighted by 0.5) to consider potentially differing interpretations of the Maybe- category.

All three aggregate measures provide similar results. AI is expected to solve all questions at Level 1 and below and 93% of the questions at Level 2, according to a simple majority vote. At Level 3 and Level 4 and above, AI is expected to answer around 70% of the questions correctly. This means that AI performance is highest at questions that are easier for adults and decreases as questions become more difficult for humans.

Figure 4.1. AI literacy performance according to different computation methods

Percentage share of literacy questions that AI can answer correctly according to the simple majority of experts

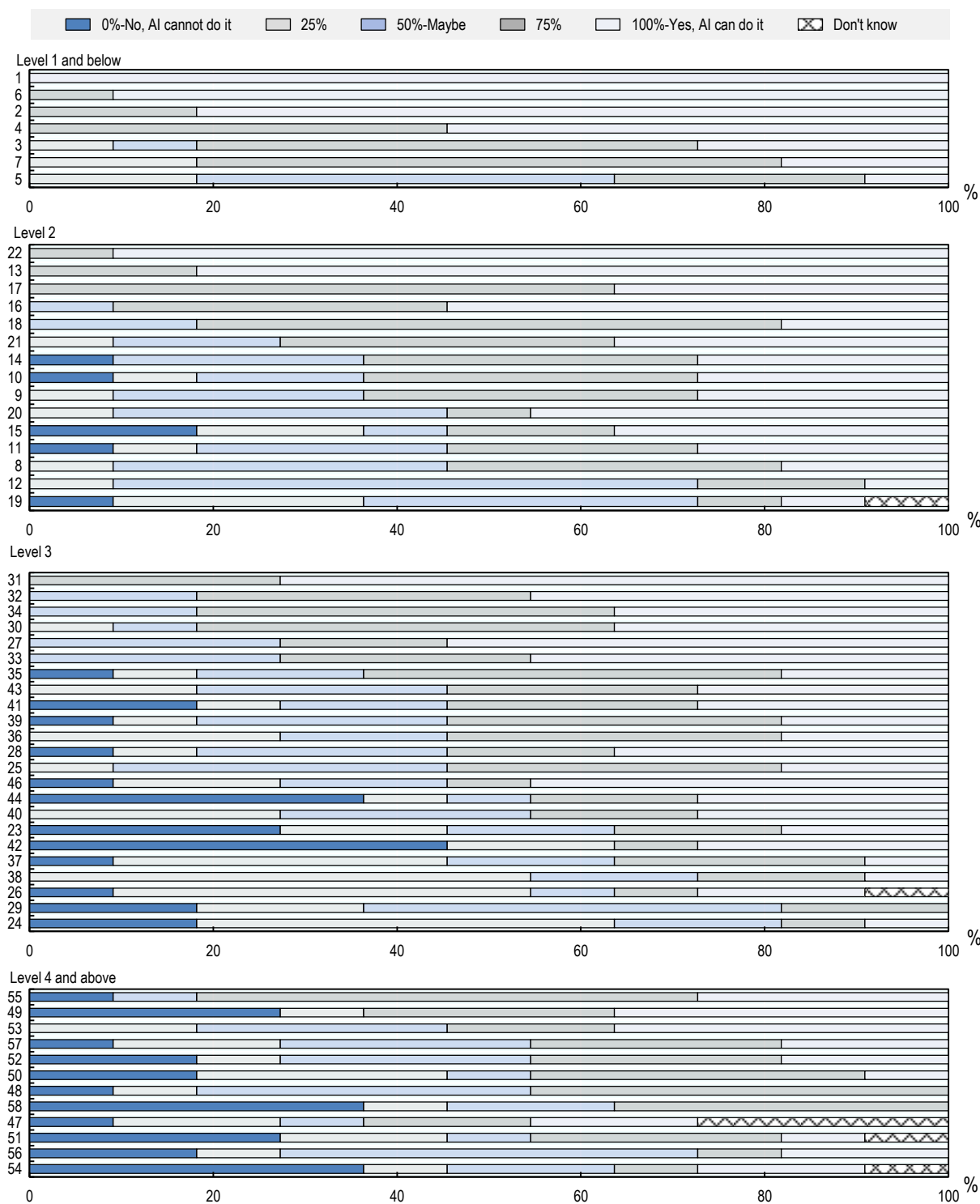


StatLink  <https://stat.link/lp57m8>

Figure 4.2 provides a more detailed picture of experts' ratings by looking at the distribution of ratings on each literacy question. It shows that questions at Level 1 and below and Level 2 receive only a few negative ratings. The evaluation of Level 1 questions is robust, as most experts rate AI performance high at these questions. At Level 2, there is more uncertainty in judgements, with bigger shares of experts providing a Maybe-answer to some questions. At Level 3 and Level 4 and above, the shares of negative ratings on individual questions increase. This indicates that experts expect AI performance to be lower at these levels. However, it also reflects disagreement among experts, as more questions at these levels receive roughly equal shares of opposing ratings. Possible reasons for disagreement are discussed below.

Figure 4.2. AI literacy performance by questions and difficulty levels

Distribution of expert ratings



StatLink  <https://stat.link/o51nt6>

AI literacy ratings by expert

The 11 computer scientists come from different subfields of AI research. Although they will most likely share the same knowledge on well-established techniques, each may have specific expertise when it comes to newer or less prominent approaches. This may affect experts' overall assessment of AI capabilities in literacy.

Figure 4.3. AI literacy performance by expert

Average ratings according to different computation rules

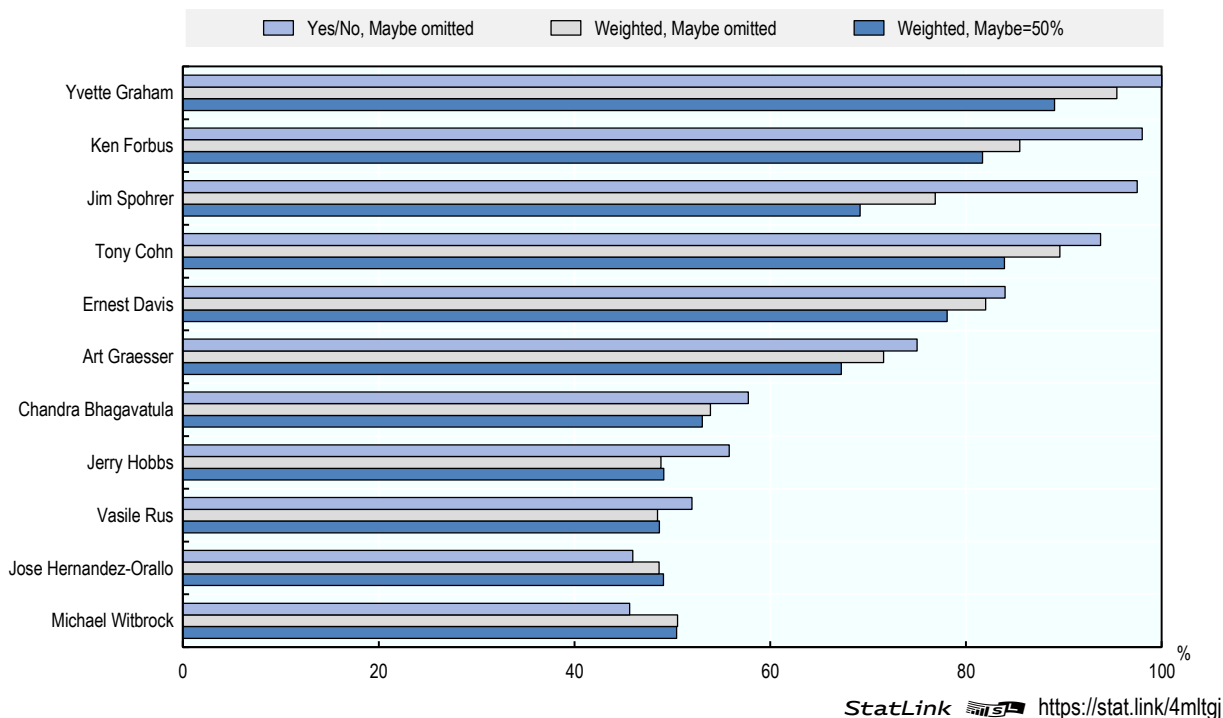


Figure 4.3 shows the average literacy ratings of experts. As in Figure 4.1, average ratings of experts are computed in three ways. First, all ratings are coded as either No (0%) or Yes (100%). In other words, less certain ratings of 25% are counted as 0% and 75%-ratings are counted as 100%. The average of an expert's ratings is then computed by omitting Maybe-answers. Second, averages are calculated by treating 25%- and 75%-ratings as such and by excluding Maybe-ratings. Third, the average of the original five categories is computed for each expert by treating the Maybe-category as 50%.

The results give a sense of experts' agreement on the overall performance of AI in literacy. They show that the average judgements of all experts are situated in the upper middle of the AI performance scale. The three different computations of experts' averages deliver similar results. However, averages where negative and positive answers are treated as No=0% and Yes=100%, respectively, show more variability. They range from 46% for the most pessimistic experts to 100% for those most optimistic about AI's potential performance in literacy. The weighted averages are generally closer to each other, ranging from 49-95% in the variant omitting the Maybe-category, and from 49-89% when Maybe-answers are included.

Disagreement among experts in literacy

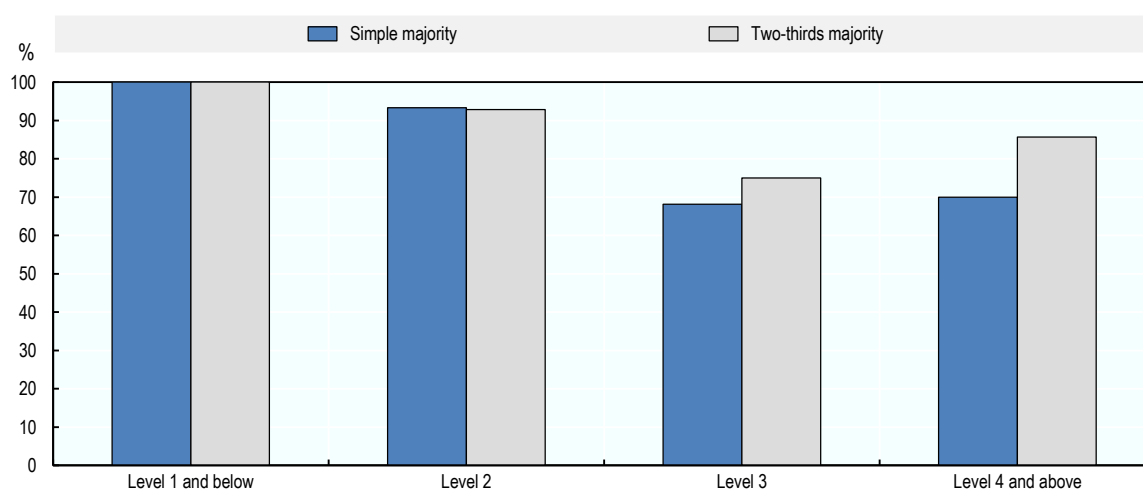
The analysis so far relied on a simple majority rule to determine whether experts rate AI as capable or incapable of answering PIAAC literacy questions correctly. However, as shown in Figure 4.2, some questions received similar shares of opposing ratings. This means that experts' agreement on these questions is low. This section presents a more rigorous approach, where two-thirds of experts must agree on whether AI can solve a PIAAC question.

Table 4.1. Experts' agreement on literacy questions

Question difficulty	N all items	Number of questions on which agreement is reached according to the following rule:					
		Simple majority			Two-thirds majority		
		Yes/No, Maybe omitted	Weighted, Maybe omitted	Weighted, Maybe=50%-Yes	Yes/No, Maybe omitted	Weighted, Maybe omitted	Weighted, Maybe=50%-Yes
Level 1 and below	7	7	7	7	7	6	7
Level 2	15	15	15	15	14	12	13
Level 3	23	22	21	21	20	10	16
Level 4 and above	12	10	10	11	7	2	6
All items	57	54	53	54	48	30	42

Figure 4.4. AI literacy performance according to different rules for agreement

Percentage share of literacy questions that AI can answer correctly according to a simple and two-thirds majority of experts; measures use Yes/No-ratings, Maybe omitted



StatLink  <https://stat.link/grcnml>

Table 4.1 shows the number of questions on which experts reach agreement according to different majority rules. Experts reach a simple majority on almost all questions. This means that more than half of those who provide answers other than “Maybe” and “Don’t know” determine whether AI can perform a PIAAC task. By contrast, two-thirds majority requires at least two-thirds of those with valid answers to be of the same opinion. This is the case on fewer questions. When ratings are viewed as either Yes or No, and Maybe-answers are excluded, only 48 of the 57 literacy items receive two-thirds agreement. When uncertain ratings of 25% and 75% are given smaller weight in the analysis, two-thirds majorities become even harder to reach. In the weighted variant, two-thirds agreement is reached on only 30 questions when

omitting Maybe-answers, and on 42 questions when including Maybe-answers as Yes-ratings weighted by 0.5.

Figure 4.4 shows the aggregate measures for literacy based only on questions with two-thirds majority, and compares them to the measures that follow a simple majority vote. The focus is on aggregate measures using only Yes-answers (75% or 100%) and No-answers (0% or 25%) and excluding Maybe- ratings. Both agreement rules lead to similar expected AI performance at each level of question difficulty. Only at Level 4 and above is there a bigger difference between measures. In that case, the conservative measure indicates that AI can answer 86% of questions as opposed to 70% obtained from a simple majority vote. However, this difference should be interpreted with caution. Measures at Level 4 and above rely on very few questions – seven when using a two-thirds majority rule, and ten when using a simple majority vote.

Uncertainty of experts in literacy

Some experts may be unaware of AI's ability to tackle certain PIAAC questions. They may also have trouble understanding the requirements of a question for AI. A big share of experts providing an uncertain answer on a PIAAC question or not providing an answer at all may reflect a general ambiguity in the field about the required AI capabilities. It could also indicate a lack of clarity on how to use the question for evaluating AI. Questions with much uncertainty are, thus, less reliable measures of AI.

Table 4.2. Experts' uncertainty on literacy questions

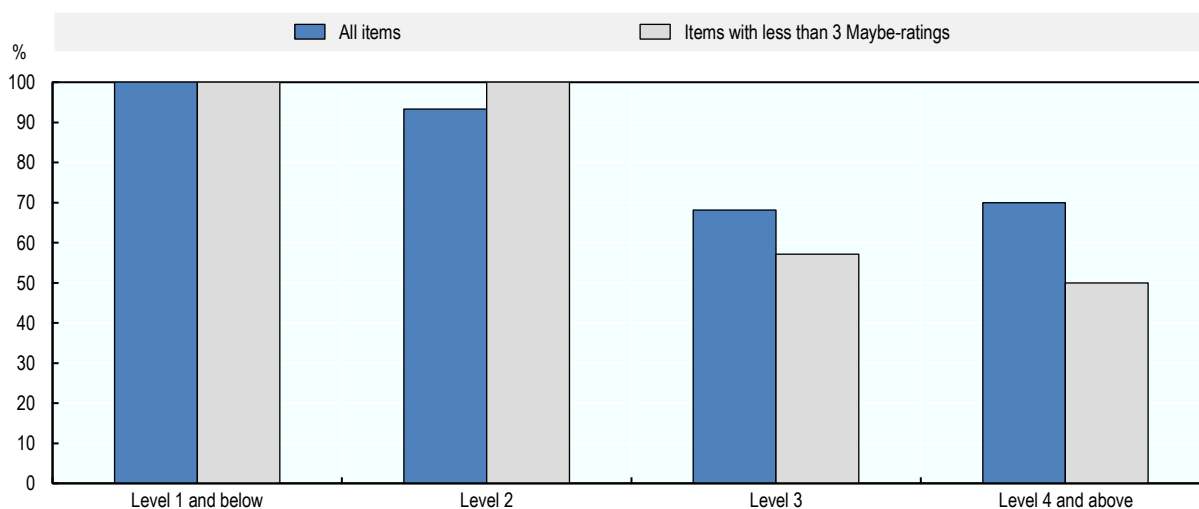
	N all items	Number of questions with Maybe- or Don't know-ratings:					Share of uncertain ratings
		No Maybe/NA	1 Maybe/NA	2 Maybe/NA	3 Maybe/NA	4+ Maybe/NA	
Level 1 and below	7	5	1	0	0	1	8%
Level 2	15	3	2	3	3	4	22%
Level 3	23	2	2	11	6	2	20%
Level 4 and above	12	1	2	2	4	3	23%
All items	57	11	7	16	13	10	20%


Table 4.2 provides an overview of the number of Maybe- and Don't know-ratings from experts. It shows that only 11 questions do not receive uncertain ratings and 10 receive 4 or more Maybe- or Don't know-answers. The last column shows the share of Maybe- and Don't know-answers from all possible answers. This gives an overview of the overall uncertainty at different difficulty levels. In total, 20% of all ratings in the literacy assessment are Maybe- or Don't know-answers. The share of uncertain answers at Levels 2 and above ranges from 20% to 23% and is lowest at Level 1 and below (8%).

Figure 4.5 shows an AI literacy performance measure computed only with questions that receive fewer than three uncertain answers. The measure is based on a simple majority vote, where 0%- and 25%-ratings are counted as No (0%); 75%- and 100%-ratings are counted as Yes (100%); and Maybe-ratings are omitted. The figure compares this measure to the one based on all questions where simple majority is reached. It shows that results remain roughly the same after excluding questions with high uncertainty, though there is a decrease in expected AI performance for the more difficult questions.

Figure 4.5. AI literacy performance using questions with high certainty

Percentage share of literacy questions that AI can answer correctly according to the simple majority of experts; measures use Yes/No-ratings, Maybe omitted



StatLink  <https://stat.link/3s20td>

Comparing the computer literacy ratings to human scores

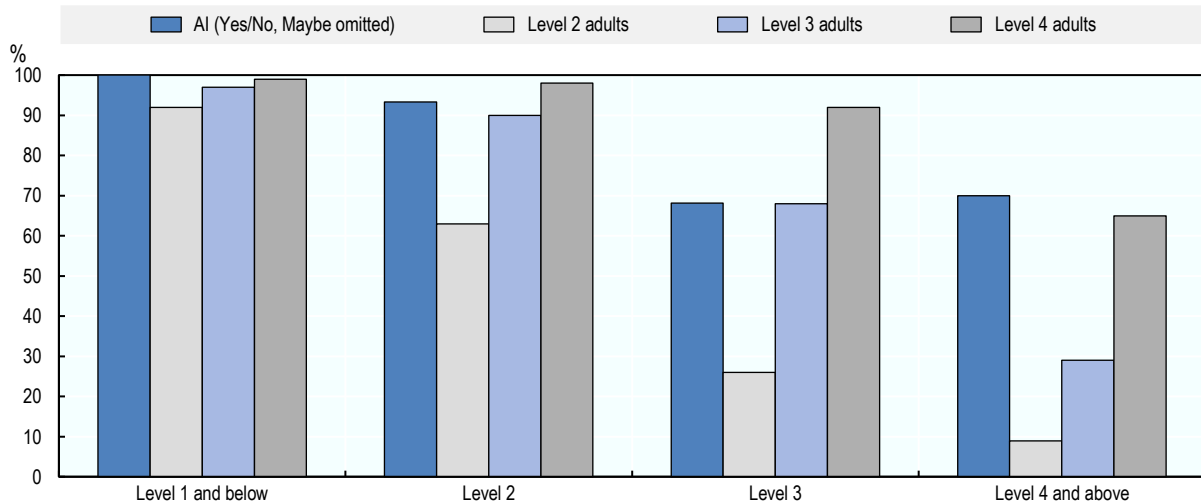
The scoring process for the Survey of Adult Skills uses item response theory to calculate difficulty scores for each question and proficiency scores for each adult. The scores for both questions and people are placed on the same 500-point scale (OECD, 2013_[2]). Each adult who takes the test is placed at the level where they answer two-thirds of questions successfully. As a result, an adult with a literacy proficiency of Level 2 can successfully answer Level 2 questions about two-thirds of the time. Generally, people are more likely to be successful in questions easier than their level and less likely to answer correctly questions harder than their level. For example, an average adult at the mid-point of Level 2 can answer 92% of Level 1 questions and only 26% of Level 3 questions (OECD, 2013, p. 70_[2]).

Figure 4.6 compares AI literacy performance with the expected performance of adults at three different levels of literacy proficiency. The AI performance measures rely on the simple majority among Yes-votes (ratings of 100% or 75%) and No-votes (ratings of 0% and 25%) of experts. The results show the scores of AI are close to those of adults at Level 3 proficiency at the first three levels of question difficulty. At Level 4 and above, AI's expected share of correctly answered literacy questions is closer to that of Level 4 adults. However, this latter result should be interpreted with caution. As shown in the preceding sections, there are only a few questions at Level 4 and above. They show somewhat higher degrees of disagreement and uncertainty among experts than questions of lower difficulty.

Figure 4.7 compares AI ratings and adults' average performance in the PIAAC literacy test. An average-performing adult in literacy is expected to complete successfully 90% of the questions at Level 1 and below; 68% of Level 2 questions; 43% of Level 3 questions; and 20% of the questions at Level 4 and above. Compared to these scores, AI is expected to solve a bigger share of questions at each level of difficulty, according to most computer experts.

Figure 4.6. Literacy performance of AI and adults of different proficiency

Share of literacy questions that AI can answer correctly according to the majority of experts compared to the probability of successfully completing items of adults at different proficiency levels

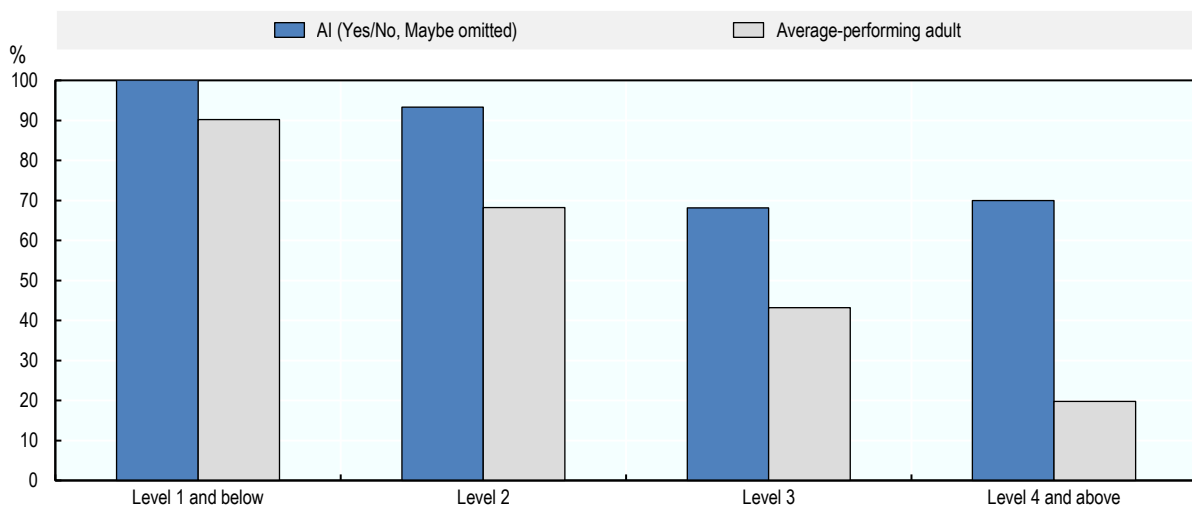


Source: OECD (2012^[3]; 2015^[4]; 2018^[5]), *Survey of Adult Skills (PIAAC) databases*, <http://www.oecd.org/skills/piaac/publicdataandanalysis/> (accessed on 23 January 2023).


StatLink  <https://stat.link/o9c3rg>

Figure 4.7. Literacy performance of AI and average adults

Share of literacy questions that AI can answer correctly according to the majority of experts compared to the probability of successfully completing items of average-performing adults



Source: OECD (2012^[3]; 2015^[4]; 2018^[5]), *Survey of Adult Skills (PIAAC) databases*, <http://www.oecd.org/skills/piaac/publicdataandanalysis/> (accessed on 23 January 2023).

StatLink  <https://stat.link/vd5jxz>

Discussion of the literacy assessment

The group discussion and the qualitative feedback gathered in the online survey centred on state-of-the-art natural language processing (NLP) technology, in general, and on question-answering systems, in particular. Experts often referred to large-scale pre-trained language models, such as GPT (Radford et al., 2018^[6]), or discussed specific solutions for solving single components of the tasks.

Overall, experts seemed at ease discussing the application of language processing systems on PIAAC. Some stated that the PIAAC literacy tasks are similar to those addressed by real-life applications of NLP. Others pointed to benchmark tests for evaluating NLP systems in AI research, such as the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016^[7]; Radford et al., 2018^[6]). They saw such tests as relevant for evaluating potential AI performance in PIAAC as they contain similar problems and tasks. However, some concerns about evaluating AI on PIAAC using expert judgement were raised as well.

Scope of tasks

A major difficulty of the rating exercise – in both literacy and numeracy – related to the range of tasks expected from a potential AI system. As described in Chapter 3, PIAAC tasks are presented in various formats, including texts, tables, graphics and images. Computer experts, on the other hand, are used to thinking of systems tailored for narrowly defined problems and trained on datasets with a definite set of tasks. The big variability of PIAAC questions thus raised uncertainty about the range of tasks on which a hypothetical system should be rated. Experts were explicitly instructed to think of one system for all tasks in a domain. However, some experts were inclined to view each task or set of similar tasks as a problem on its own and to judge current AI's capacity to solve this particular problem. By contrast, other experts assumed general systems designed to solve a wide range of tasks like those in PIAAC.

How experts interpreted the scope of PIAAC tasks affected how they viewed the AI capabilities required for solving the tasks and, ultimately, how they rated AI on PIAAC. One example for this relates to the degree of language interpretation experts assumed for systems. Some experts argued that certain literacy questions could be solved with only “shallow” language processing. Shallow processing involves pattern matching of various types, such as proposing a passage of text as an answer to a question based on its similarity to the question wording. These experts tended to rate AI on such questions higher, assuming that a simplistic approach would be good enough to spot the right answer in a text. However, other experts argued that for AI to be able to solve the entire literacy test, including similar tasks that are not part of the test, “deep” language processing would be necessary. Deep processing involves interpretation of the meaning of the language. The latter experts tended to rate AI literacy capabilities lower.

Question formats

Another example of diverging interpretations relates to questions using formats other than text. On several questions containing graphs, the group divided evenly between those who believed current techniques could answer the question and those who believed they could not. One such question was discussed in the workshop in more depth (Item #15 at Level 2). The item contains a short newspaper article on a financial topic, supplemented by two bar charts. The charts present a ranking of ten countries on two financial indicators, each of which is clearly stated in the chart title. The question asks respondents to indicate two countries with values falling in a specified range on one of the indicators. This requires respondents to identify the graph presenting the indicator in question, locate the bars that represent the values in the specified range, and see which countries these bars correspond to; it does not require reading the article.

Experts generally agreed that reading charts and processing images is still challenging for AI. However, experts argued that a system can be trained to solve the task with sufficient data containing similar charts.

This training would also meet the requirements in the rating instructions. These state that a hypothetical development effort to adjust current technology to PIAAC should take no longer than one year and cost no more than USD 1 million. Experts on the pessimistic side, on the other hand, argued that a general question-answering system for natural language arithmetic problems that can process graphs, images and other task formats does not exist yet. Moreover, developing such a system would require technological breakthroughs that would largely exceed the hypothetical investments stated in the rating instructions.

Response types

Other challenges in the rating exercise were discussed as well. One recurring topic was the variability of response types used in the questions. Some questions were multiple-choice, requiring the respondent to click a correct answer out of several possible alternatives. Other questions required typing the answer or highlighting it in a text. According to experts, computers may have considerable difficulties with some response types, such as clicking an answer.

Development conditions

Another discussion topic focused on the adequacy of the hypothetical advance preparation that experts were instructed to consider in their evaluations. As mentioned above, the hypothetical effort for adapting AI systems to PIAAC should require less than both one year and USD 1 million. The more optimistic experts noted that raising the budget threshold to more than USD 10 million would allow for developing systems to master the literacy test. However, the pessimists argued that budgetary limits are not the real challenge to developing systems for literacy. According to them, a general system for literacy tasks requires major technological advancements in NLP.

Overall, the discussion and written comments in the survey indicated considerable consensus among experts about the literacy capabilities of state-of-the-art NLP systems. Experts generally agreed that most PIAAC questions can be solved as isolated problems by systems trained on a sufficient volume of similar questions. However, these systems would be limited to PIAAC and have no practical implications. There was also general agreement that AI technology cannot yet master the entire PIAAC literacy test as well as a high-performing human. In other words, it could not understand the meaning of questions and process texts in different formats to answer these questions correctly.

However, experts differed in how they interpreted the requirements for the technology being evaluated. Some thought the technology should be narrow, solving only PIAAC questions. Others considered general systems, able to understand, evaluate and use written texts in various settings.

Evaluation of AI capabilities in the domain of numeracy

Numeracy in the Survey of Adult Skills is defined as the “ability to access, use, interpret and communicate mathematical information and ideas, in order to engage in and manage the mathematical demands of a range of situations in adult life” (OECD, 2012^[1]). The skill covers different mathematical operations, such as calculating; estimating proportions, percentages or rates of change; operating with spatial dimensions; using various measuring devices; discerning patterns, relationships and trends; and understanding statistical concepts related to probabilities or sampling. The mathematical information in the test is represented in a variety of formats, including objects and pictures, numbers and symbols, diagrams, maps, graphs, tables, texts and technology-based displays. The questions are drawn from the same familiar contexts used for the literacy test: work, personal life, society and community, and education and training.

Numeracy items are described in terms of six levels of difficulty, ranging from below Level 1 to Level 5 (OECD, 2013^[2]). For simplicity, below Level 1 and Level 1, as well as Level 4 and Level 5, are grouped

into single categories. Nine of the PIAAC numeracy items are at Level 1 or below, 21 items are at Level 2, 20 items have Level 3-difficulty, and only six items are at Level 4 and above.

The test items at the lowest difficulty levels involve single-step processes. Examples are counting, sorting, performing basic arithmetic operations with whole numbers or money, understanding simple percentages such as 50%, or recognising common graphical or spatial representations. The harder test items require the respondent to undertake multiple steps to solve the task and to use different types of mathematical content. For example, the respondent should analyse, apply more complex reasoning, draw inferences or evaluate solutions or choices. The mathematical information is presented in complex and abstract ways or is embedded in longer texts (see also Chapter 3 for an overview of PIAAC).

As described in Chapter 3, 11 experts evaluated AI on PIAAC's literacy and numeracy tests. Subsequently, four additional specialists in mathematical reasoning for AI were invited to assess AI in numeracy only. The following results present the ratings of all 15 experts who participated in the numeracy assessment.

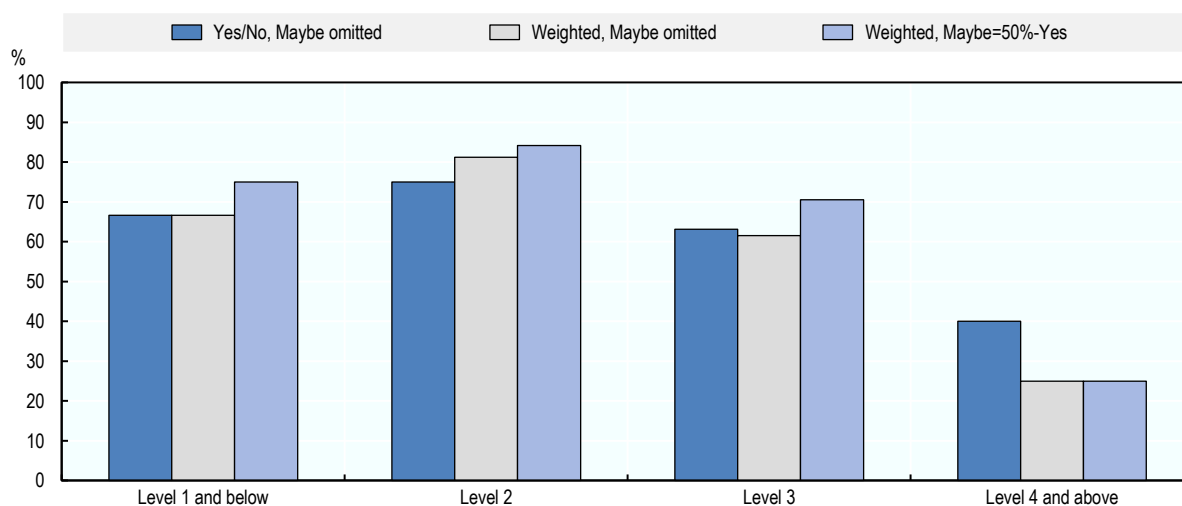
AI numeracy ratings by question difficulty

The aggregate measures of AI capabilities for the numeracy questions are illustrated in Figure 4.8. These measures are computed by counting the shares of Yes- and No-ratings on each question, assigning to questions the rating that receives the majority share of experts' votes, and then estimating the share of questions with a Yes-vote at each level of question difficulty. The measures thus show the share of questions that AI can answer correctly at each difficulty level, according to the majority of experts.

As in the literacy analysis, three versions of the aggregate measures are calculated, dependent on how Yes- and No-ratings are handled. The first version counts uncertain answers of 25%- and 75%-ratings as No (0%) and Yes (100%), respectively, and ignores Maybe-ratings. The second version considers experts' uncertainty, by giving 25%- and 75%-ratings a lower weight. That is, 25%-ratings are treated as 0.75-No and 75%-ratings are included as 0.75-Yes. The third version is similar to the second, except it includes Maybe-ratings as 0.5-Yes.

Figure 4.8. AI numeracy performance according to different computation methods

Percentage share of numeracy questions that AI can answer correctly according to the simple majority of experts




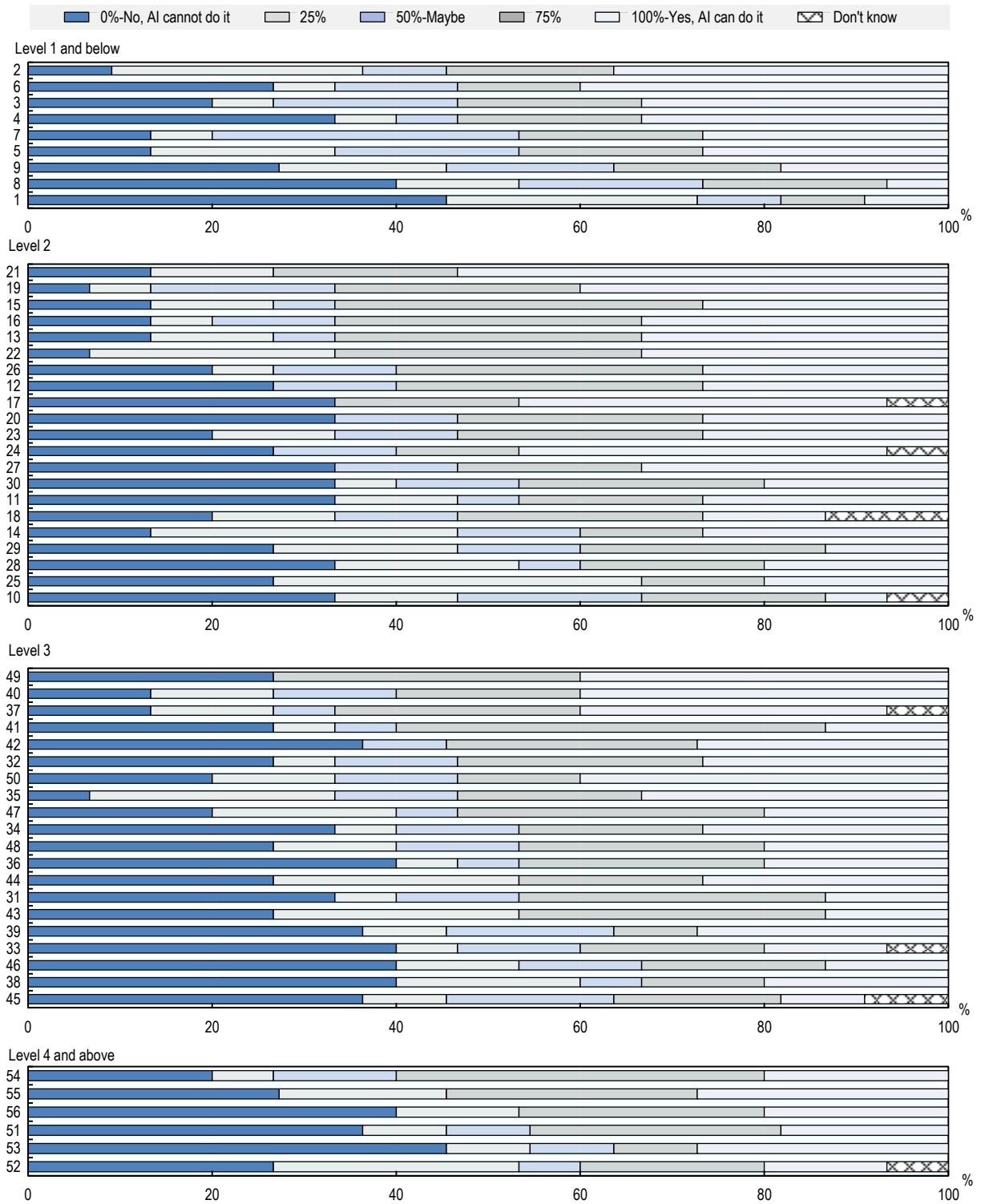
StatLink  <https://stat.link/xfea14>

Figure 4.9. AI numeracy performance by questions and difficulty levels

Distribution of expert ratings



StatLink  <https://stat.link/9txzbv>

Following the first version of the measure, AI can answer correctly 67% of the questions at Level 1 and below, 75% of the Level 2 questions, 63% of the Level 3 questions and 40% of the questions at Level 4 and above (see Figure 4.8). The second version of the measure produces similar results at the first three levels of question difficulty and a lower share of 25% of correctly answered questions at Level 4 and above. The third version, which treats Maybe-ratings as partial Yes, indicates higher AI performance than the other measures at Level 1 and below, Level 2 and Level 3, and a performance level at 25% at Level 4 and above. All three measures draw a pattern of performance for AI, which is different than the one for humans. That is, according to experts, AI is expected to perform better at questions of medium difficulty for humans and somewhat worse at questions that are easiest of humans.

Figure 4.9 shows the distribution of ratings at individual questions by difficulty of questions. It shows that all questions receive both certain negative and certain positive ratings. The shares of these opposing evaluations are often close to each other, indicating that only thin majorities decide on AI's capabilities in numeracy. At Level 1 and below, several questions receive a high share of uncertain ratings of about 20% and higher.

AI numeracy ratings by expert

The following analysis looks at the individual ratings of the 15 experts who assessed AI in the numeracy domain. It shows how ratings vary both between and within experts to provide insights into the congruence of experts' evaluations and into individual rating patterns.

Figure 4.10. AI numeracy performance by expert

Average ratings according to different computation rules

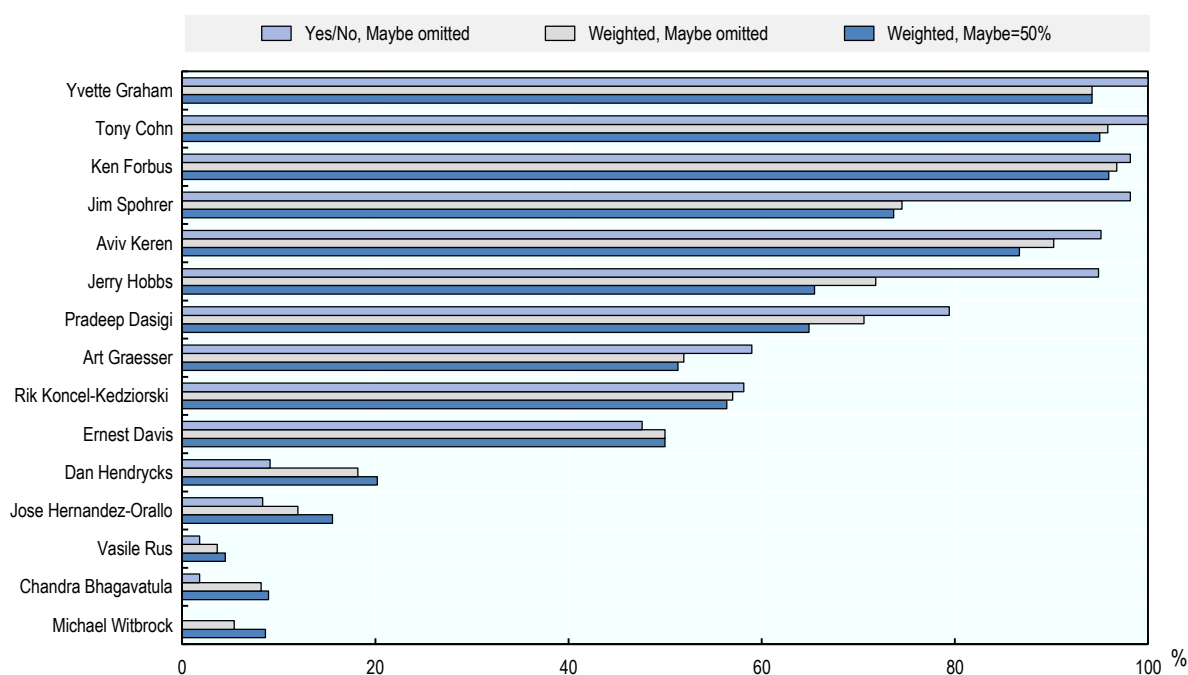
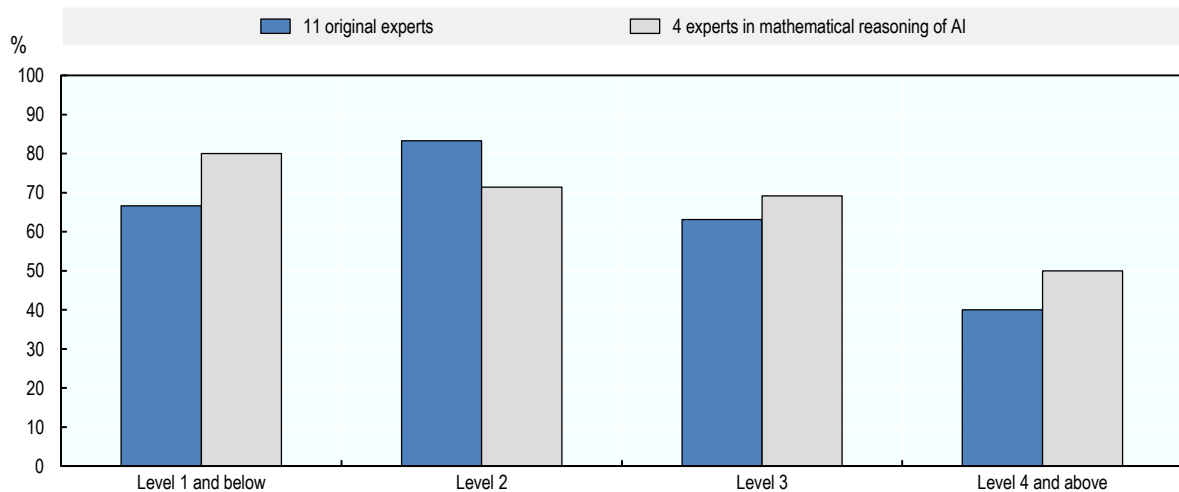


Figure 4.10 presents the averages of experts' ratings computed in three ways. First, it omits Maybe- answers and treats 25%- and 75%-ratings as 0% and 100%, respectively. Second, it only considers ratings of 0%, 25%, 75% and 100%. Third, it considers all ratings, including Maybe-ratings as 50%. The figure reveals a big variability in experts' opinions, with average ratings covering the entire scale of AI's capability in the numeracy test. Two extreme groups emerge: five experts with averages between 0-20%, depending on the type of measure, and four experts with averages between 80-100%.

Figure 4.11. AI numeracy performance by expert group

Comparison of ratings of core eleven experts with those of the four experts in mathematical reasoning of AI




StatLink  <https://stat.link/j4aysr>

Figure 4.11 compares results from the 11 original experts with those of the 4 experts in mathematical reasoning of AI who completed the assessment with a revised framework. The revisions included mainly providing more information and examples on PIAAC, as well as asking experts to describe an AI approach for addressing all questions in the domain at once.

Overall, the results from both assessments are similar, following a measure that relies on the simple majority between positive (75% and 100%) and negative (0% and 25%) ratings. At Level 1 and below and Level 3 and higher, the aggregate ratings from the first assessment are somewhat lower than those from the four experts in mathematical reasoning. At Level 2 of question difficulty, the results from the 11 original experts are 12 percentage points higher than the ratings from the subsequent re-assessment. These small differences indicate that neither the changes introduced in the assessment framework nor the changes in the focus of expertise substantially affect group ratings in numeracy.

Disagreement among experts in numeracy

Table 4.3 provides additional insights into experts' agreement. It shows the number of questions on which computer experts reach a simple or a two-thirds majority, following different computations of Yes- and No-votes. Experts reach a simple majority on 53 questions, when counting ratings of 75% and 100% as Yes-answers and ratings of 0% and 25% as No-answers. When ratings of 25% and 75% are given a smaller weight in the calculation of Yes- and No-votes, experts reach the 50%-threshold to majority on only 42 questions. In the case where Maybe-answers are additionally counted as a partial Yes-answer, a simple majority is reached on 48 questions. This indicates the weighted aggregate AI measures shown above rely on a considerably smaller number of numeracy questions.

Two-thirds majorities cannot be reached on most questions in the numeracy domain. Only 18 questions receive two-thirds agreement in the variant, which counts ratings as either Yes- or No-votes and omits Maybe-answers. In the weighted variant, two-thirds agreement is achieved on only three questions when omitting Maybe-answers, and on eight questions when including Maybe-answers as 50%-Yes. These few questions are clearly insufficient for evaluating AI's capabilities in numeracy.

Table 4.3. Experts' agreement on numeracy questions

Question difficulty	N all items	Number of questions on which agreement is reached according to the following rule:					
		Simple majority			Two-thirds majority		
		Yes/No, Maybe omitted	Weighted, Maybe omitted	Weighted, Maybe=50%	Yes/No, Maybe omitted	Weighted, Maybe omitted	Weighted, Maybe=50%
Level 1 and below	9	9	9	8	2	1	2
Level 2	21	20	16	19	11	2	5
Level 3	20	19	13	17	4	0	1
Level 4 and above	6	5	4	4	1	0	0
All items	56	53	42	48	18	3	8

Uncertainty among experts in numeracy

Table 4.4 provides an overview of the amount of uncertain evaluations in the numeracy assessment. Overall, uncertainty is lower than in the literacy assessment. Only 12% of answers are Maybe- or Don't know-answers compared to 20% on the literacy questions. In contrast to literacy, where there is more uncertainty in evaluating harder questions, uncertainty in numeracy is highest for questions at Level 1 and below and lowest for questions at Level 4 and above. That is, 17% of all ratings on the easiest numeracy questions are Maybe- or Don't know-answers compared to a share of 8% on questions at Level 4 and above. Only a few numeracy questions receive a high number of uncertain ratings – seven questions have three uncertain ratings, while three questions have four or more uncertain ratings.

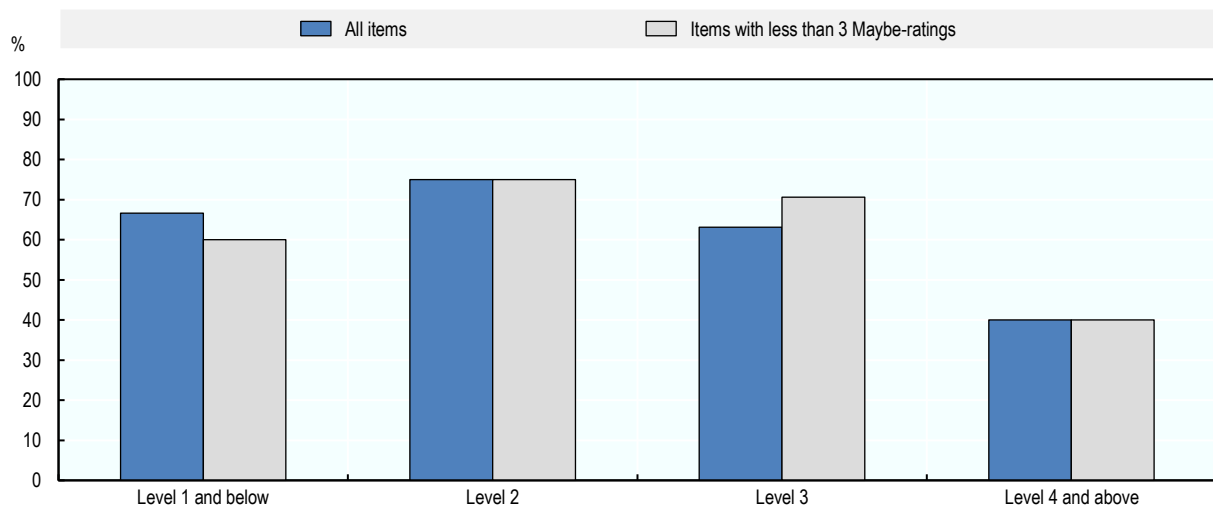
Table 4.4. Experts' uncertainty on numeracy questions


Question difficulty	N all items	Number of questions with Maybe- or Don't know-ratings:					Share of uncertain ratings
		No Maybe/NA	1 Maybe/NA	2 Maybe/NA	3 Maybe/NA	4+ Maybe/NA	
Level 1 and below	9	0	3	2	3	1	17%
Level 2	21	3	5	9	2	2	12%
Level 3	20	3	5	10	2	0	11%
Level 4 and above	6	2	2	2	0	0	8%
All items	56	8	15	23	7	3	12%

Figure 4.12 presents the aggregate AI numeracy measure computed after excluding the ten questions with three or more uncertain ratings. The measure uses a simple majority of Yes- versus No-votes (100%- and 75%-ratings versus 0%- and 25%-ratings) and excludes Maybe-ratings. It shows similar results to those of the measure using all questions with simple majority. The only differences are at Levels 1 and 3, where the measure built on questions with high certainty produces somewhat lower and higher AI scores, respectively.

Figure 4.12. AI numeracy performance using questions with high certainty

Percentage share of numeracy questions that AI can answer correctly according to the simple majority of experts; measures using Yes/No-ratings, Maybe omitted



StatLink  <https://stat.link/1styv3>

Comparing the computer numeracy ratings to human scores

As described in Chapter 3, question difficulty and performance in PIAAC are rated on the same 500-point scale. Respondents are evaluated depending on the number and difficulty of questions they answer correctly. For simplicity, the scale is summarised into six levels of question difficulty or respondents' proficiency. A respondent with a proficiency score at a given level has a 67% chance of successfully completing test questions at that level. This individual will also likely complete more difficult questions with a lower probability of success and answer easier questions with a greater chance of success.

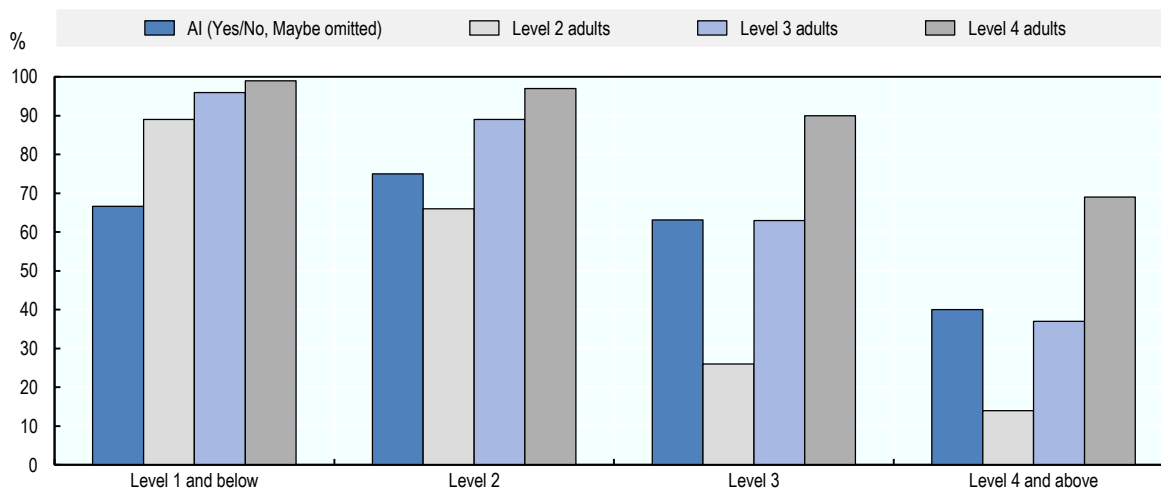
Figure 4.13 compares AI numeracy performance with the average performance of adults at proficiency levels 2, 3 and 4. The AI performance measure shows the share of questions that AI can answer correctly according to the simple majority among the 15 experts. It relies only on positive (75% and 100%) and negative (0% and 25%) ratings of experts, excluding Maybe-answers. The performance of adults can be interpreted similarly: the percentage share of questions that a respondent with a score at the middle of a given level of proficiency is expected to complete successfully.

The results show that AI numeracy performance varies less across the difficulty of questions than human performance does. That is, AI performance is similar across questions, whereas adults perform better at the easiest and worse at the hardest questions. At Level 1 and below, the performance gap between AI and humans is biggest, with AI being expected to solve 67% of questions and a Level 2 adult 89%. At Level 2 difficulty, AI's expected probability of success (75%) lies between that of Level 2 (66%) and Level 3 (89%) adults. At Levels 3 and 4 and above, AI performance matches that of Level 3 adults.

In addition, Figure 4.14 compares AI and average-performing adults in PIAAC. Compared to average human performance, AI numeracy performance is expected to be lower at Level 1 and below, similar at Level 2, and lower at Levels 3 and 4 and above.

Figure 4.13. Numeracy performance of AI and adults of different proficiency

Share of numeracy questions that AI can answer correctly according to the majority of experts compared to the probability of successfully completing items of adults at different proficiency levels

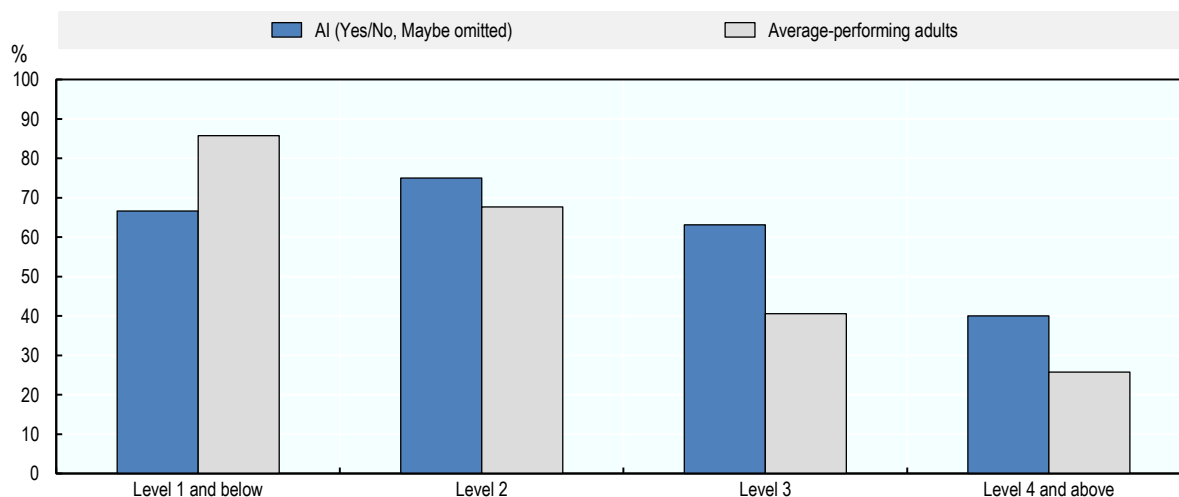


Source: OECD (2012^[3]; 2015^[4]; 2018^[5]), *Survey of Adult Skills (PIAAC) databases*, <http://www.oecd.org/skills/piaac/publicdataandanalysis/> (accessed on 23 January 2023).

StatLink  <https://stat.link/mnpktw>

Figure 4.14. Numeracy performance of AI and average adults

Share of numeracy questions that AI can answer correctly according to the majority of experts compared to the probability of successfully completing items of average-performing adults



Source: OECD (2012^[3]; 2015^[4]; 2018^[5]), *Survey of Adult Skills (PIAAC) databases*, <http://www.oecd.org/skills/piaac/publicdataandanalysis/> (accessed on 23 January 2023).

StatLink  <https://stat.link/ofb56d>

Overall, these results should be treated with caution. AI numeracy measures rely on only thin agreement among experts as to whether AI can perform the PIAAC numeracy tasks. The following section provides more insights into experts' agreement behind the quantitative AI measures in numeracy.

Discussion of the numeracy assessment

During the group discussion, the 11 computer experts elaborated on the difficulties they faced in the literacy and numeracy assessments. This provided first insights into the factors causing dissent and uncertainty in the numeracy domain. In a second workshop, some of these experts discussed how to improve the assessment framework to address these challenges. Subsequently, four additional specialists in mathematical reasoning of AI were invited to complete the numeracy assessment with a revised framework and to discuss the exercise in an online workshop. The following section describes the feedback received from experts in the three workshops, as well as steps taken to improve the assessment, following this feedback.

Challenges with numeracy questions

Generally, the 11 experts who first rated AI in numeracy described the exercise as less straightforward than the literacy assessment. They saw the numeracy questions as more distant from problems typically addressed by AI research. Compared to the literacy tasks, the numeracy tasks have received less attention in the field because of their limited practical applicability. According to the experts, these tasks do not pose a bigger challenge to AI technology than the literacy ones. However, the tasks will be harder for current systems to solve precisely because of the lack of interest and investment in solving them.

During the workshop, the 11 experts discussed the requirements of the numeracy test for AI. Overall, there was more ambiguity about the range of tasks that a hypothetical system is supposed to master than in the literacy assessment. This is because the numeracy questions are more diverse, including more graphs, images, tables and maps. This led some experts to view the numeracy questions as separate, narrow problems and to evaluate AI's capacity to solve them independently from each other. By contrast, other experts focused on the entire test, viewing it as a general challenge for AI to reason mathematically and to process multimodal inputs in various settings. How experts saw the scope of the numeracy test affected their evaluations. The ones who focused on narrow problems generally gave more positive ratings than those who focused on general challenges.

A discussion of one numeracy question with high disagreement in ratings exemplifies this divergence. The item (#20 at Level 2) shows a logbook that keeps track of the miles travelled by a salesperson on her work trips. The question asks respondents to estimate the reimbursement of travel expenses for the last trip. This requires applying a simple formula that multiplies the number of miles travelled with the amount paid per mile and adds the fixed amount paid per day for additional expenses. One group of experts argued that a general question-answering system can be fine-tuned to work with similar tables with sufficient training data. These experts gave higher ratings on this question. Another group of experts opposed to this that, while the single question may be solvable with sufficient fine-tuning, a much bigger effort would be needed to develop solutions for all numeracy problems and to integrate them into a single system. These experts gave lower ratings on the question because they doubted a system could solve this and all other questions in the numeracy test.

Development approaches to exemplify experts' evaluations

Much of the following discussion focused on how to develop an architecture that allows a single system to address the different question types. Three approaches received more attention.

The first approach, proposed by one of the optimists among the experts, combined dedicated systems for different question types using a classifier. Each of the dedicated systems would be trained individually on a huge amount of data that resembles a particular question type. The classifier would then read in the type of a particular PIAAC task and channel the task to the corresponding solution. According to the experts, at the current stage of technology, such specialised systems are possible, given sufficient training data. However, they offer only a narrow solution, which is limited to the PIAAC test and “brittle” to small changes in the tasks.

The second approach was proposed by an expert at the middle of the ratings distribution as an alternative to the machine-learning approaches that most experts described. It consists in engineering a set of components to address the different capabilities required for performing the test at a more general level. For example, the approach would combine separate components for language understanding, analogical reasoning, image processing and problem solving.

The third approach, suggested by some experts who gave lower ratings, is a multimodal system, trained on different types of tasks simultaneously. Learning different types of tasks jointly by processing different types of data increases the generality and reasoning capacity of a system. However, multitask, multimodal learning is still at a development stage, which explains the lower ratings of the experts who support this approach.

This discussion showed that encouraging experts to elaborate on a concrete approach can benefit the rating exercise. By stating more explicitly that a single system should tackle all types of problems in a domain at once, it gave experts a common ground for the evaluation. It also facilitated understanding and communication, which may help experts reach agreement in their evaluations. Therefore, the study added a survey question to the rating exercise that asked experts to briefly describe an AI system that could carry out all questions in a test domain.

Providing experts with more information on PIAAC

Experts offered other suggestions for revising the rating exercise, expressing the need for more information on PIAAC. This could help them determine the scope of problems to be addressed and the breadth of the hypothetical system to be evaluated. Experts were provided with information from PIAAC’s assessment framework (OECD, 2012_[1]). The materials include both conceptual information on the underlying skills targeted by the assessment, as well as practical information on the types and formats of the test questions. Nine test questions were added to this information to provide concrete examples for tasks to experts. These questions were selected to represent different difficulty levels and formats.

A second workshop was organised with some of the experts to discuss the proposed improvements. Experts received the materials on PIAAC and the task examples in advance. They were asked to describe a high-level approach for solving the tests using this information. In the workshop, experts discussed the usefulness and feasibility of the revised assessment framework. They agreed the additional information and examples helped them better understand the requirements of the tests for AI systems. In addition, experts proposed revising the instruction to consider a hypothetical investment of USD 1 million to adapt existing techniques to the test. Instead, the hypothetical effort should fit the size of a major commercial AI development project to better reflect reality in the field. Based on this feedback, the OECD team finalised the materials describing PIAAC and revised the instructions for rating.

The revised assessment framework was tested with four additional specialists in mathematical reasoning for AI. They were invited to complete the numeracy assessment only and to discuss the results in an online workshop. Despite the revisions, the assessment produced mixed results. One expert provided overly negative ratings, while another had mostly positive ratings. The evaluations of the other two experts were in the middle of the performance range.

Quantitative disagreement, qualitative agreement

The discussion showed that ambiguity in PIAAC's requirements is not responsible for the difference in numerical ratings. The four experts found the test description and the rating exercise clear. They did not consider the variability of tasks as a challenge to evaluating a single system. Instead, they discussed the fast pace at which AI research in mathematical reasoning has been developing over the past year. They also reflected on the likelihood of AI solving the numeracy test in the near future.

In between the first and second numeracy assessment – the period between December 2021 and September 2022 – the field has taken major steps. This includes the release of the MATH dataset, the leading benchmark for mathematical reasoning (Hendrycks et al., 2021^[8]); and the development of several systems such as Google's Minerva, Codex and Bashkara, which are all large language models fine-tuned for quantitative problems (Lewkowycz et al., 2022^[9]; Davis, 2023^[10]). In addition, prominent AI labs have been working on multimodal systems that can process both images and text. This was reflected differently in experts' evaluations.

The three experts with middle to high ratings argued that, given the recent advancements in the field, AI is close to solving the PIAAC numeracy test. Therefore, a hypothetical engineering effort in this direction would likely produce the desired outcomes in less than one year. By contrast, the expert with the lowest ratings focused on the current state of AI techniques, which are not yet able to solve the numeracy test. However, he agreed that AI will likely reach this stage within a year.

Overall, the changes introduced in the assessment framework, particularly the inclusion of more information and examples on PIAAC, have increased clarity and consensus about the AI capabilities targeted by the numeracy tests. The four experts who completed the numeracy assessment with the revised framework generally agreed that current systems are close to processing the different types of formats used in the test. To translate this qualitative agreement into coherent quantitative ratings, the time frames in the instructions for rating need to be shortened. This would enable more precise evaluations of the state of the art in AI technology.

References

- Davis, E. (2023), *Mathematics, word problems, common sense, and artificial intelligence*, [10]
<https://arxiv.org/pdf/2301.09723.pdf> (accessed on 28 February 2023).
- Hendrycks, D. et al. (2021), “Measuring Mathematical Problem Solving With the MATH Dataset”. [8]
- Lewkowycz, A. et al. (2022), “Solving Quantitative Reasoning Problems with Language Models”. [9]
- OECD (2018), *Survey of Adult Skills (PIAAC) database*, [5]
<http://www.oecd.org/skills/piaac/publicdataandanalysis/> (accessed on 23 January 2023).
- OECD (2015), *Survey of Adult Skills (PIAAC) database*, [4]
<http://www.oecd.org/skills/piaac/publicdataandanalysis/> (accessed on 23 January 2023).
- OECD (2013), *The Survey of Adult Skills: Reader’s Companion*, OECD Publishing, Paris, [2]
<https://doi.org/10.1787/9789264204027-en>.
- OECD (2012), *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*, OECD Publishing, Paris, [1]
<https://doi.org/10.1787/9789264128859-en>.
- OECD (2012), *Survey of Adult Skills (PIAAC) database*, [3]
<http://www.oecd.org/skills/piaac/publicdataandanalysis/> (accessed on 23 January 2023).
- Radford, A. et al. (2018), *Improving Language Understanding by Generative Pre-Training*, [6]
https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 1 February 2023).
- Rajpurkar, P. et al. (2016), “SQuAD: 100,000+ Questions for Machine Comprehension of Text”. [7]

Annex 4.A. Supplementary tables

Annex Table 4.A.1. List of online tables for Chapter 4

Table Number	Table Title
Table A4.1	Individual expert judgements on current computer capabilities for answering PIAAC literacy questions
Table A4.2	Individual expert judgements on current computer capabilities for answering PIAAC numeracy questions
Table A4.3	Individual expert judgements on computer capabilities in 2026 for answering PIAAC literacy questions
Table A4.4	Individual judgements of the 11 core experts on computer capabilities in 2026 for answering PIAAC numeracy questions
Table A4.5	Individual judgements of the 4 experts in mathematical reasoning of AI on computer capabilities in 2026 for answering PIAAC numeracy questions

StatLink  <https://stat.link/7bx9mt>



From:
Is Education Losing the Race with Technology?
AI's Progress in Maths and Reading

Access the complete publication at:
<https://doi.org/10.1787/73105f99-en>

Please cite this chapter as:

OECD (2023), "Experts' assessments of AI capabilities in literacy and numeracy", in *Is Education Losing the Race with Technology?: AI's Progress in Maths and Reading*, OECD Publishing, Paris.

DOI: <https://doi.org/10.1787/134fa8aa-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>.