Educational Research and Innovation

Is Education Losing the Race with Technology?

AI'S PROGRESS IN MATHS AND READING

OECD

Educational Research and Innovation

Is Education Losing the Race with Technology?

AI'S PROGRESS IN MATHS AND READING

OECD
BETTER POLICIES FOR BETTER LIVES

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 the Member countries of the OECD.

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.

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Note by the Republic of Türkiye
The information in this document with reference to “Cyprus” relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Türkiye recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Türkiye shall preserve its position concerning the “Cyprus issue”.

Note by all the European Union Member States of the OECD and the European Union
The Republic of Cyprus is recognised by all members of the United Nations with the exception of Türkiye. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

Please cite this publication as:
OECD (2023), *Is Education Losing the Race with Technology?: AI's Progress in Maths and Reading*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/73105f99-en.

ISBN 978-92-64-45137-7 (print)
ISBN 978-92-64-92037-8 (pdf)
ISBN 978-92-64-76515-3 (HTML)
ISBN 978-92-64-63967-6 (epub)

Educational Research and Innovation
ISSN 2076-9660 (print)
ISSN 2076-9679 (online)

Photo credits: Cover © LightField Studios/Shutterstock.com; © VH-studio/Shutterstock.com; © Rido/Shutterstock.com; © KlingSup/Shutterstock.com.

Corrigenda to publications may be found on line at: www.oecd.org/about/publishing/corrigenda.htm.

© OECD 2023

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

Foreword

Artificial intelligence (AI) has been rapidly advancing in recent years, bringing significant changes to various sectors and industries. As we continue to witness the evolution of AI, it is increasingly important to understand what this technology can and cannot do. The OECD recognises the need to systematically assess the capabilities of AI in relation to human skills, particularly in skill domains of key importance for employment and education. Such an assessment can help policy makers and educators better anticipate the impact of technological change on the workforce and prepare individuals for the demands of the future.

The present report follows up on an earlier pilot study and assesses the abilities of state-of-the-art AI in reading and mathematics, two key domains of human competence that are essential for success in many areas of work and life. The report shows how AI capabilities in these domains evolve over time and how they compare to the reading and mathematics skills of humans. The assessment of AI is based on evaluations from computer scientists.

The study is part of a comprehensive project for assessing AI capabilities and their implications for work and education, led by the OECD Centre for Educational Research and Innovation (CERI). The AI and the Future of Skills (AIFS) project aims to develop measures of AI capabilities that are understandable, comprehensive, repeatable and policy-relevant. By using various sources of information on AI, including expert evaluations, the project aims at providing policy makers with the knowledge they need to shape future-oriented education- and labour-market policies.

The report shows significant improvements in AI's capabilities in reading since 2016, which reflect advancements in natural language processing (NLP) in recent years. AI's capabilities to solve mathematical tasks have not improved at the same rate. However, experts predict that increasing investments in AI research and development will lead to significant advancements of AI in both reading and mathematics in the coming years.

The report also demonstrates the potential for AI to outperform large portions of the population in reading and maths. This has important implications for employment and education, as workers are likely to face increasing competition from machines in these skill domains in the future. It also highlights the need to strengthen the foundation skills of the workforce and prepare it to work together with AI in key domains.

By providing an example of how AI capabilities improve with respect to two key cognitive skills of humans, this study emphasises the importance of periodically and systematically monitoring the evolution of AI capabilities and comparing them to human skills. This will be useful to policy makers, educators, and researchers who are seeking to understand the implications of technological advancements for the future of work and education.

Acknowledgements

This study was carried out by the OECD's Artificial Intelligence and Future of Skills project team – Stuart Elliott (Project lead), Nóra Révai, Margarita Kalamova, Mila Staneva, Abel Baret and Aurelija Masiulytė. The report was drafted and prepared for publication by Mila Staneva. Aurelija Masiulytė and Abel Baret contributed to the formatting.

This publication would not have been possible without the invaluable contributions of the renowned computer science experts who are supporting the project.

Firstly, we would like to express our gratitude to the experts who participated in the assessment (in alphabetical order): Chandra Bhagavatula, Anthony G. Cohn, Pradeep Dasigi, Ernest Davis, Kenneth D. Forbus, Arthur C. Graesser, Yvette Graham, Daniel Hendrycks, José Hernández-Orallo, Jerry R. Hobbs, Aviv Keren, Rik Koncel-Kedziorski, Vasile Rus, Jim Spohrer and Michael Witbrock.

Secondly, we would like to thank Lucy Cheke, Charles Fadel, Michael Handel, Patrick Kyllonen, Frank Levy and Michael Schoenstein for their participation in the discussions and their fruitful feedback.

Additionally, we wish to thank our colleagues in the Centre for Educational Research and Innovation (CERI). Tia Loukkola, Head of CERI, provided oversight, direction and valuable advice during the process. Francois Keslair and Marco Paccagnella from the Directorate for Education and Skills (EDU) made important contributions to the analysis. Colleagues within the EDU communications team and the Public Affairs and Communications Directorate contributed to both formatting and the preparation of the publication.

Our thanks are extended to Mark Foss, who made substantive and structural editing to the publication, ensuring for coherent, comprehensible reading.

We are grateful for the encouragement and support of the CERI Governing Board in the development of the project.

This publication contributes to the OECD's Artificial Intelligence in Work, Innovation, Productivity and Skills (AI-WIPS) programme, which provides policy makers with new evidence and analysis to keep abreast of the fast-evolving changes in AI capabilities and diffusion, and their implications for the world of work. The programme aims to help ensure that adoption of AI in the world of work is effective, beneficial to all, people- centred and accepted by the population at large. AI-WIPS is supported by the German Federal Ministry of Labour and Social Affairs (BMAS) and will complement the work of the German AI Observatory in the Ministry's Policy Lab Digital, Work & Society. For more information, visit https://oecd.ai/work-innovation-productivity-skills and https://denkfabrik-bmas.de/.

Table of contents

Foreword 3

Acknowledgements 4

Executive summary 8

1 Setting the stage: Approaches to assessing AI's impact 11
- Previous studies that measure AI capabilities and impact 13
- Objective for the exploratory assessment of AI capabilities 15
- Plan of the report 17
- References 17
- Notes 19

2 Evolution of human skills versus AI capabilities 20
- Changes in skills supply 21
- Recent developments in AI capabilities 29
- The importance of measuring AI capabilities 35
- References 36
- Annex 2.A. Supplementary tables 39
- Notes 40

3 Methodology for assessing AI capabilities using the Survey of Adult Skills (PIAAC) 41
- Overview of the Survey of Adult Skills (PIAAC) 42
- Identifying a group of computer scientists 46
- Collecting expert judgement 47
- Developing the questionnaire 48
- Constructing aggregate measures of AI literacy and numeracy performance 49
- Challenges and lessons learned 49
- References 51

4 Experts' assessments of AI capabilities in literacy and numeracy 52
- Evaluation of AI capabilities in the domain of literacy 53
- Evaluation of AI capabilities in the domain of numeracy 62
- References 73
- Annex 4.A. Supplementary tables 74

5 Changes in AI capabilities in literacy and numeracy between 2016 and 2021 75
- Change in AI literacy capabilities over time 76

Change in AI numeracy capabilities over time 82
References 89
Annex 5.A. Supplementary figures 90
Notes 91

6 Implications of evolving AI capabilities for employment and education 92
Summary of results 93
Policy implications of evolving AI capabilities in literacy and numeracy 97
A new approach to assessing AI 105
References 107
Notes 108

FIGURES

Figure 2.1. Literacy proficiency levels of 15-65 year-olds, IALS and PIAAC 22
Figure 2.2. Literacy proficiency levels of working population, IALS and PIAAC 23
Figure 2.3. Long-term trends in average reading proficiency of 15-year-olds 24
Figure 2.4. Numeracy proficiency levels of 16-65 years-olds, ALL and PIAAC 26
Figure 2.5. Numeracy proficiency levels of the working population, ALL and PIAAC 27
Figure 2.6. Long-term trends in average mathematics proficiency of 15-year-olds 28
Figure 3.1. Literacy – Sample item 44
Figure 3.2. Numeracy - Sample item 45
Figure 4.1. AI literacy performance according to different computation methods 54
Figure 4.2. AI literacy performance by questions and difficulty levels 55
Figure 4.3. AI literacy performance by expert 56
Figure 4.4. AI literacy performance according to different rules for agreement 57
Figure 4.5. AI literacy performance using questions with high certainty 59
Figure 4.6. Literacy performance of AI and adults of different proficiency 60
Figure 4.7. Literacy performance of AI and average adults 60
Figure 4.8. AI numeracy performance according to different computation methods 63
Figure 4.9. AI numeracy performance by questions and difficulty levels 64
Figure 4.10. AI numeracy performance by expert 65
Figure 4.11. AI numeracy performance by expert group 66
Figure 4.12. AI numeracy performance using questions with high certainty 68
Figure 4.13. Numeracy performance of AI and adults of different proficiency 69
Figure 4.14. Numeracy performance of AI and average adults 69
Figure 5.1. Answer categories used in the 2016 and 2021 assessments 76
Figure 5.2. AI literacy performance in 2016 and 2021, by question difficulty 77
Figure 5.3. AI literacy performance in 2016 and 2021, counting Maybe as 50%-Yes 78
Figure 5.4. Average expert ratings in literacy in 2016 and 2021 79
Figure 5.5. AI literacy performance in 2016 and 2021 according to experts who participated in both assessments 80
Figure 5.6. Average majority size in rating literacy questions in 2016 and 2021 81
Figure 5.7. Share of literacy questions that receive three or more uncertain ratings in 2016 and 2021 81
Figure 5.8. Projected AI literacy performance for 2026, by question difficulty 82
Figure 5.9. AI numeracy performance in 2016 and 2021, by question difficulty 83
Figure 5.10. AI numeracy performance in 2016 and 2021, counting Maybe as 50%-Yes 84
Figure 5.11. Average expert ratings in numeracy in 2016 and 2021 85
Figure 5.12. AI numeracy performance in 2016 and 2021 according to experts who participated in both assessments 86
Figure 5.13. Average majority size in rating numeracy questions in 2016 and 2021 87
Figure 5.14. Share of numeracy questions that receive three or more uncertain ratings in 2016 and 2021 87
Figure 5.15. Projected AI numeracy performance for 2026, by question difficulty 88
Figure 6.1. Success rate of AI in PIAAC according to experts' assessments 95
Figure 6.2. Percentage of workers at different proficiency levels who use literacy and numeracy on a daily basis at work 98

Figure 6.3. Daily use of literacy and numeracy practices at work 99
Figure 6.4. Daily use of literacy and numeracy at work together with other skills 100
Figure 6.5. Proportion of workers with high literacy and numeracy proficiency 102
Figure 6.6. Literacy performance of AI and average adults by cognitive strategy required in PIAAC questions 103
Figure 6.7. Digital skills of adults 104
Figure 6.8. Proportion of workers with a well-balanced skill set 105

TABLES

Table 3.1. Computer scientists participating in the follow-up assessment of computer capabilities 46
Table 4.1. Experts' agreement on literacy questions 57
Table 4.2. Experts' uncertainty on literacy questions 58
Table 4.3. Experts' agreement on numeracy questions 67
Table 4.4. Experts' uncertainty on numeracy questions 67
Table 6.1. Summary of AI and adults' performance in PIAAC 93

Annex Table 2.A.1. List of online tables for Chapter 2 39
Annex Table 4.A.1. List of online tables for Chapter 4 74
Annex Table 5.A.1. List of online figures for Chapter 5 90

BOXES

Box 2.1. Example tasks from natural language processing benchmarks 31
Box 2.2. Example tasks from benchmarks on mathematical reasoning 34
Box 3.1. Example for literacy questions 44
Box 3.2. Example for numeracy questions 45
Box 6.1. ChatGPT as an example of AI literacy capabilities 96

Follow OECD Publications on:

https://twitter.com/OECD

https://www.facebook.com/theOECD

https://www.linkedin.com/company/organisation-eco-cooperation-development-organisation-cooperation-developpement-eco/

https://www.youtube.com/user/OECDiLibrary

https://www.oecd.org/newsletters/

This book has...

StatLinks
A service that delivers Excel® files from the printed page!

Look for the StatLink at the bottom of the tables or graphs in this book. To download the matching Excel® spreadsheet, just type the link into your Internet browser or click on the link from the digital version.

Executive summary

Advances in artificial intelligence (AI) are ushering in a large and rapid technological transformation. Understanding how the capabilities of AI relate to human skills and how they develop over time is crucial for understanding this ongoing process. Knowing what AI can do compared to humans can help predict which skills may become obsolete and which skills may become more significant in the years ahead. This knowledge base can help policy makers reshape education systems in ways that best prepare students for the future and provide opportunities to adult learners to renew their skills.

This report follows up an earlier pilot study, collecting expert evaluations of how well AI can do the literacy and numeracy tests of the OECD Survey of Adult Skills of the Programme for International Assessment of Adult Competencies (PIAAC). It shows how AI capabilities in these domains have evolved since the pilot assessment in 2016, up until mid-2022 (shortly before the release of ChatGPT). Assessing AI capabilities in literacy and numeracy is indicative of AI's potential impact on work and life since these skills are relevant in most social contexts and work situations.

The study is part of a comprehensive ongoing project for assessing computer capabilities and their implications for work and education. The AI and the Future of Skills (AIFS) project at OECD's Centre for Educational Research and Innovation (CERI) uses several information sources to develop measures of AI capabilities that are understandable, comprehensive, repeatable and policy relevant.

Methodology

Both the pilot and this follow-up asked computer scientists to rate AI's capacity to answer the questions on PIAAC's literacy and numeracy tests. AI's likely performance on the tests was determined by looking at the majority expert opinion on each question. The use of standardised education tests enables the comparison to human capabilities, allows tracking AI progress across time and provides understandable AI measures. However, experts did not always agree in their evaluations. The study aimed at improving the methodology for eliciting expert knowledge on AI with standardised tests to address this challenge.

Key findings

Experts expect AI to perform well on both the literacy and numeracy tests of PIAAC.

- According to experts, AI can answer around 80% of the PIAAC literacy questions. It can solve most of the easy questions, which typically involve locating information in short texts and identifying basic vocabulary. It can also master many of the harder questions, which require navigating across larger chunks of text to formulate responses. This evaluation rests on high consensus among experts.
- According to experts, AI can solve around two-thirds of the PIAAC numeracy test. However, there is disagreement behind this result. Some experts imagined narrow AI solutions for separate

numeracy questions. Others considered general systems that can reason mathematically and process all kinds of numeracy questions similar to those in PIAAC. This led to diverging evaluations, with the latter experts giving lower ratings than the former.

AI capabilities in literacy have increased substantially since 2016.

- A comparison to the pilot assessment reveals considerable improvement in AI's literacy capabilities since 2016. The expected success rate of AI in the literacy test has increased by 25 percentage points since then. This reflects the technological breakthroughs in natural language processing (NLP) in the period, related to the introduction of pre-trained language models, such as GPT.
- The discussion with experts suggested that numeracy capabilities of AI are unlikely to have changed much between 2016 and 2021. While formal mathematics underlying numeracy problems are easily automatable, extracting formal models from tasks that require general knowledge and are expressed in language and in images has received less research attention.

According to experts, AI will be able to solve the entire literacy and numeracy tests by 2026.

- Given recent technological advancements and the heavy investment and research in NLP, experts judged that AI's literacy capabilities will continue to develop.
- More recently, large language models have been fine-tuned and applied for mathematical problems. The field has produced important benchmark tests as well as systems that perform well on these tests. These trends led experts to expect that AI will advance considerably in numeracy over the next few years.

AI can potentially outperform large shares of the population in literacy and numeracy.

- PIAAC assesses respondents' proficiency in literacy and numeracy on several levels – from low (Level 1 and below) to high (Levels 4-5). Following the evaluation of experts, AI's potential performance in literacy is close to that of adults with proficiency at Level 3. Across the OECD countries in PIAAC, on average, 90% of adults are at or below Level 3 in literacy and only 10% perform better than Level 3.
- The AI numeracy performance assessed by experts is close to that of adults at proficiency Level 2 on the easier and intermediate PIAAC questions, and similar to that of Level 3 adults on the harder questions. Across OECD countries with data, on average, 57% of adults are at or below Level 2 in numeracy, and 88% are at Level 3 or below that level.

Conclusions

- Despite its limitations, this study suggests that advancing AI capabilities with respect to literacy and numeracy may have important implications for employment and education. Most workers use these skills every day at work. At the same time, these skills have not improved in most countries in the last decades. By contrast, AI capabilities in literacy and numeracy are developing quickly.
- Across countries in PIAAC, on average, 59% of the workforce uses literacy skills daily at a proficiency comparable to or below that of computers. Between 27% and 44% of workers daily perform numeracy tasks at work, having numeracy proficiency below or at the level of AI. AI could affect the literacy- and numeracy-related tasks of these workers.
- Even the best-ranking countries to date cannot supply more than a quarter of their workforce with the literacy and numeracy skills needed to outperform AI. In this context, the focus of education

may need to shift more towards teaching students to use AI systems to perform literacy and numeracy tasks more effectively.

1 Setting the stage: Approaches to assessing AI's impact

This chapter introduces the study and situates it in a broader context of related research. The study assesses artificial intelligence (AI) capabilities by collecting expert judgements on whether AI can carry out tests from the OECD's Survey of Adult Skills of the Programme for International Assessment of Adult Competencies (PIAAC). It follows up an earlier study from 2016 to track changes in AI capabilities with regard to PIAAC over time. The chapter first provides an overview of past studies that assess computer capabilities and their impact on the economy. Against this background, it presents the objectives of this study and discusses potential strengths and weaknesses of the methodological approach the study uses to assess AI. The chapter concludes with an outline of the structure of this report.

New technologies can profoundly change the way people live and work. In the past, the steam engine, electricity and the computer transformed societies by accelerating productivity and growth and by shifting employment from agriculture to manufacturing and later to services. Today, advances in artificial intelligence (AI) and robotics are ushering in a larger and more rapid transformation. Compared to past technologies, AI and robotics can match or surpass humans in a larger number of tasks, especially those including image and speech recognition, predictions and pattern identification. This process is evolving faster than previous waves of technological progress due to steady improvements in computational power, storage capacity and algorithms.

Understanding how the capabilities of AI and robotics relate to human skills and how they develop over time is crucial for understanding the ongoing technological transformation. Knowing what AI can do compared to humans can help predict which work tasks may be automated, which skills may become obsolete and which skills may become more significant in the years ahead. This knowledge base can help develop effective labour-market policies to tackle the challenges of technological change. Moreover, it can enable policy makers to reshape education systems in ways that best prepare today's students for the future.

In 2016, OECD carried out a study that assessed AI capabilities with respect to core human skills (Elliott, 2017[1]). This pilot study used the OECD's Survey of Adult Skills, which is part of the Programme for International Assessment of Adult Competencies (PIAAC), as a tool to assess whether AI can carry out education tests designed for adults. The results showed that AI capabilities in literacy, numeracy and problem solving in technology-rich environments,[1] as assessed by experts, resemble the performance of adults at Level 2. In OECD countries and economies, on average, more than half of the adult population perform at Level 2 or below in these domains on the PIAAC test and so would not be able to "outperform" AI (OECD, 2019[2]). This shows that many people could potentially be affected by evolving computer capabilities at their work.[2]

The present report follows up the pilot study, collecting expert judgements on whether computers can carry out the PIAAC literacy and numeracy tests. It shows how AI capabilities in these domains have evolved since the previous assessment. Another goal is to improve the assessment framework for eliciting expert knowledge on AI using standardised tests. The study is part of a more comprehensive ongoing project for assessing the capabilities of computers and their implications for work and education. The AI and the Future of Skills (AIFS) project at OECD's Centre for Educational Research and Innovation (CERI) aims at developing measures of AI capabilities that are understandable, comprehensive, repeatable and policy relevant.[3] For this purpose, the project uses various sources of information on AI, including expert evaluations.

PIAAC assesses the proficiency of adults aged 16-65 in three general cognitive skills – literacy, numeracy and problem solving in technology-rich environments. These skills are key determinants of individuals' ability to participate effectively in the labour market, education and training, and social and civil life. Higher literacy proficiency, for example, is linked to higher wages, more participation in volunteer activities, higher levels of social trust, better employability and health (OECD, 2013[3]). Therefore, countries have large incentives to invest in the formation of these skills. They are associated with economic returns in the form of higher productivity and enhanced capacity for innovation. They are also linked to significant social returns such as social cohesion and civic engagement, and political and social trust.

Experts' assessments of AI performance on the PIAAC literacy and numeracy tests provide useful information for policy making. Assessing AI capabilities in these domains is indicative of AI's potential impact on work and life since literacy and numeracy are relevant in most social contexts and work situations. In addition, using human tests for the assessment makes it possible to compare AI and human capabilities and to draw conclusions on AI's capacity to reproduce human skills.

This chapter draws upon extensive research in social sciences, economics and computer science to provide an overview of studies that assess computer capabilities and their impact on the economy. It then

introduces the current study and its objectives. The chapter concludes with an outline of the structure of this report.

Previous studies that measure AI capabilities and impact

Most of the work on AI and robotics that is prominent in the policy discourse stems from economics and the social sciences. This literature typically focuses on AI's potential to replace workers in the workplace and assesses its capabilities with regard to job tasks. Other strains of research from computer science and psychology analyse AI from the perspective of skills and abilities. They measure which computer capabilities are available, how they evolve over time and how they relate to human skills.

The task-based approach

Many studies in the economic literature start their analysis by looking at occupations and their task content. They analyse whether occupational tasks are susceptible to automation, typically by drawing on the judgement of computer experts. The goal is to quantify the extent to which machines can carry out occupations. This information is then linked to labour-market data to study the impact of occupations' automatability on employment and wages. This section highlights key studies in this area.

The task-based approach originated in the seminal study of Autor, Levy and Murnane (2003[4]). This study assumes that machines can replace workers only in tasks that follow exact, routine procedures as these tasks can be easily codified. By contrast, non-routine tasks, such as those involving problem solving or social interaction, were not apt for automation because they are less explicable. The model predicts that declining prices of technology would affect the labour demand in these task domains differently. The demand for workers performing routine tasks would decrease as employers increasingly replace them with cheap machines. At the same time, more high-skilled workers will be needed for non-routine tasks emerging from the use of technology at the workplace, such as developing and operating machines.

Many studies have extended the approach of Autor, Levy and Murnane (2003[4]) to account for more recent technological advancements. The most widely cited study, Frey and Osborne (2017[5]), identifies three types of work tasks that are still hard to automate: perception and manipulation tasks, such as navigating in unstructured environments; creative intelligence tasks, such as composing music; and social intelligence tasks, such as negotiating and persuading. The authors study how these "bottleneck" tasks relate to how experts rate the automatability of 70 occupations. They use the estimated relationships to predict the probability of automation of more than 600 further occupations. The analysis relies on the Occupational Network (O*NET) database of the US Department of Labor – an occupation taxonomy that systematically links occupations to work tasks (National Center for O*NET Development, n.d.[6]). By mapping the measure of occupations' automatability to US labour-market data, the study estimates that 47% of US employment is at high risk of automation.

Two studies supported by the OECD – Arntz, Gregory and Zierahn (2016[7]) and Nedelkoska and Quintini (2018[8]) – improve upon the methodology of Frey and Osborne's framework. The studies include more countries in their analyses and use more fine-grained data on "bottleneck" tasks. Moreover, they estimate the measure of automatability at the level of jobs instead of occupations. This accounts for the fact that jobs within the same occupation may differ in their propensity for automation. The studies find a much smaller share of jobs prone to automation compared to Frey and Osborne (2017[5]): 9% on average across 21 OECD countries and economies as shown by Arntz et al. (2016[7]) and 14% across 32 countries and economies as shown by Nedelkoska and Quintini (2018[8]).

Assessing automation through the content of patents

Some studies draw on information from patents to measure the applicability of AI and robotics in the workplace. The study of Webb (2020[9]) scans patent descriptions for keywords such as “neural networks”, “deep learning” and “robot” to identify patents of AI and robotic technologies. It then studies the overlap between the text descriptions of such patents and the task descriptions of occupations available in O*NET. In this way, the study quantifies the exposure of occupations to these technologies. The results show that, while jobs occupied by low-skilled workers and low-wage jobs are most exposed to robotic technologies, it is the jobs of those with college degrees that are most exposed to AI. In addition, increases in occupations’ susceptibility for robotic technologies are linked to declines in employment and wages.

Squicciarini and Staccioli (2022[10]) adopted a similar approach to Webb (2020[9]). They identify patents of labour-saving robotic technologies using text-mining techniques and measure their textual proximity to occupation descriptions in ISCO08 – a standardised classification of occupations (ILO, 2012[11]). In this way, they estimate exposure of occupations to robotics. The study finds that low-skilled and blue-collar jobs, and also analytic professions, are the occupations most exposed to robotic technologies. However, there is no evidence of labour displacement, as employment shares in these occupations remain constant over time.

AI measures relying on benchmarks

In AI research, benchmarks are used to evaluate machines’ progress with regard to specific tasks and domains. A benchmark is a test dataset, on which systems perform a task or a set of tasks, and performance is rated with a standard numerical metric. This provides a common testbed for comparing different systems. Several studies connect the information from benchmarks to information on occupations to assess how evolving AI capabilities can impact the workplace.

Examples for popular benchmarks include ImageNet, a large publicly available dataset used to test systems’ ability to correctly classify images (Deng et al., 2009[12]). In the language domain, the General Language Understanding Evaluation (GLUE) benchmark tests systems on a multitude of tasks. These include predicting the sentiment of single sentences and detecting semantic similarity between the sentences in sentence pairs (Wang et al., 2018[13]). In reinforcement learning, the Arcade Learning Environment tests the ability of AI agents to maximise their performance on a defined task by testing various strategies to solving the tasks and identifying the most effective solutions (Bellemare et al., 2013[14]).

The study of Felten, Raj and Seamans (2019[15]) uses evaluation results from benchmarks to measure progress across major AI application domains, such as image and speech recognition. The authors asked gig workers on a crowdsourcing platform to rate how each AI application domain is linked to key abilities required in occupations from O*NET. By linking the AI domains to occupations, the authors assessed the extent to which occupations are exposed to AI. They assumed that occupations requiring abilities related to more rapidly advancing AI domains are more exposed to AI. The study finds that AI’s occupational impact is positively linked to wage growth, but not to employment.

Tolan et al. (2021[16]) use research output related to 328 AI benchmarks (e.g. research publications, news, blog entries) to measure the direction of AI progress (see also Martínez-Plumed et al. (2020[17])). They link these measures to tasks within occupations, obtained from labour force surveys, as well as O*NET. The link between AI progress and work tasks goes through an intermediate layer of key cognitive abilities. The latter are derived from work in psychology, animal cognition and AI, and include broad, basic capabilities, such as visual processing and navigation. Concretely, the study connects the three – AI benchmarks to cognitive abilities, abilities to work tasks, and AI benchmarks to work tasks via cognitive abilities – by drawing on expert judgement from various disciplines. The results suggest relatively high AI exposure for

high-income occupations, such as medical doctors, and low AI impact on low-income occupations, such as drivers or cleaners.

Using AI-related job postings as an indicator for use of AI in firms

The demand for AI experts in firms can serve as a proxy for the use of AI in the workplace. This assumes that firms deploying AI technology also need workers with AI-related skills to operate and maintain it. Studies following this approach obtain information on firms' skills needs from job postings.

Alekseeva et al. (2021[18]) scan job postings for AI-related skills. Both the job postings and a list of pre-defined AI skills are obtained from Burning Glass Technologies (BGT), a company that collects online vacancies daily and provides systematic information on their skill requirements. The study shows that firms with high demand for AI skills offer higher wages for both their AI and non-AI vacancies. According to the authors, this evidence supports the view that use of AI in the workplace raises the demand for complementary tasks that require advanced skills, such as project and people management tasks.

Instead of using pre-specified AI keywords, Babina and colleagues (2020[19]) estimate how frequently the skills contained in the BGT data co-occur with core AI concepts within vacancies, such as "artificial intelligence" and "machine learning". The idea is that skills often mentioned with core AI terms are relevant for AI. In this way, the authors assess the AI-relatedness of the skill requirements of job postings. They find that firms demanding AI-related skills grow faster in terms of sales, employment and market share within the industry.

Skills-based assessment: A new approach

The impact of AI on work can also be measured by comparing the capabilities of AI to the full range of human skills required in the workplace. This comparison directly addresses the question of whether AI can replace humans in their jobs. Moreover, it provides information on the impacts of AI that go beyond the scope of current occupations. For example, it can show how occupations should be rearranged in future to better reconcile AI and human skills, and how education should evolve in response.

An AI-human comparison can be achieved by assessing AI capabilities with standardised tests developed for humans. Computer science research has already used different types of human tests for AI evaluation, including IQ tests (e.g. Liu et al., (2019[20])), school exams in mathematics (Saxton et al., 2019[21]) and science (Clark et al., 2019[22]).

Objective for the exploratory assessment of AI capabilities

This study aims at assessing AI capabilities using expert judgement on whether AI can carry out the PIAAC test. It is part of a bigger effort at the OECD to assess AI. The AIFS project aims at developing measures of AI capabilities. These are intended to help policy makers and the public to understand AI's implications for education and work.

These measures should meet several criteria:

- They should provide an accepted framework to describe AI capabilities, which shows the most important strengths and limitations of AI and highlights when AI capabilities change substantially.
- As with any measures, they should be valid and reliable. In other words, they should both reflect the aspects of AI they claim to measure (validity) and provide consistent information (reliability).
- Measures should be understandable for non-experts, repeatable and comprehensive, meaning they should cover all key aspects of AI. They should also be relevant to policy, helping to draw out the implications of AI for education, work and the economy.

The AIFS project draws on various sources of information about AI capabilities to develop AI measures: direct assessment of AI capabilities through benchmarks and expert judgement (OECD, 2021[23]).

Direct assessments of AI capabilities are made through benchmarks, competitions and evaluation campaigns in the AI field to track progress and evaluate systems' performance. However, direct measures are typically only available for areas of current research and development, leaving many tasks and skills relevant for work uncovered. Moreover, direct measures are centred around state-of-the-art AI and do not assess performance on tasks that are too easy or too difficult for current systems.

Expert judgement can complement the assessment framework in areas in which information from direct measures is lacking. By filling these gaps, measures relying on expert judgement can contribute to a more comprehensive assessment of AI capabilities.

The project uses a battery of different tests to collect expert judgement on AI. As a complement to PIAAC, it uses the Programme for International Student Assessment (PISA) to measure key cognitive skills, while assessing occupation-specific skills with tests from vocational education and training. In addition, tests from the fields of animal cognition and child development will be used to assess basic low-level skills of all healthy adult humans, but which AI does not necessarily have (e.g. spatial and episodic memory) (OECD, 2021[23]).

The pilot study conducted in 2016 served as a stepping stone into the AIFS project (Elliott, 2017[1]). Expert judgement on whether AI can carry out human tests constitutes a valuable source of information for the study. Both the pilot and this follow-up intend to explore the assessment of AI capabilities with the Survey of Adult Skills using expert judgement. This new approach reveals a number of strengths:

- Rating based on specific test items enables a more precise estimate of computer capabilities. Test items provide experts judging computer capabilities with precise, contextualised and granular descriptions of the task. This allows computer experts to rate potential AI performance on the task without making additional assumptions about the task requirements. This implies greater reliability across raters and greater reproducibility.
- Using human tests makes it possible to compare computer to human capabilities. In particular, PIAAC enables fine-grained analyses of skill supply across different contexts, different age groups and occupations. This allows comparing AI to the average performance of particular groups of human workers. Moreover, the test offers a graduated progression from simple to complex tasks, allowing assessment and comparison of the level of proficiency of AI and humans.
- Standardised tests allow tracking AI progress across time. They enable the reproducibility of the assessment – both across experts and across different time points.
- Assessing AI against standardised tests provides understandable measures. Using PIAAC to describe AI capabilities provides information that is meaningful to educators and education researchers. Educators and education researchers are usually familiar with the types of skills assessed on tests like the Survey of Adult Skills. They are also familiar with the ways those skills are developed in education and potentially used at work and in daily life.

However, the assessment approach bears some challenges as well:

- Overfitting is a common danger, not only with respect to using human tests on AI but also with regard to any evaluation instrument. Overfitting means that an AI system can excel on a test without being able to perform other tasks that differ only slightly from the test. This happens because AI systems are generally "narrow" or trained to perform specific tasks.
- As another challenge, tests designed for humans typically take for granted skills that all humans (without severe disabilities) share, such as vision and common sense. Because such skills cannot be assumed for AI, human tests can have different implications for humans and machines. For

example, the simple task to count the objects in a picture tests humans' ability to count; for AI, it becomes a test for object recognition.

Plan of the report

This report presents the motivation, the methodological approach and the results of the assessment of AI capabilities with PIAAC. Chapter 2 provides background information on how human skills in literacy and numeracy have changed over time and how technologies processing language and solving mathematical tasks have evolved in the same period. By showing that computer capabilities develop much more rapidly than capabilities of humans in key domains, the chapter highlights the need for periodically assessing and comparing both. Chapter 3 describes the methodological approach of collecting expert judgements on whether AI can carry out the PIAAC test. Chapters 4 and 5 present the results. Chapter 4 presents the results of this follow-up study, while Chapter 5 compares these results to the results of the pilot study to track changes in the assessed AI capabilities in literacy and numeracy since 2016. Chapter 6 discusses the policy implications of evolving AI capabilities for education and work.

References

Alekseeva, L. et al. (2021), "The demand for AI skills in the labor market", *Labour Economics*, Vol. 71, p. 102002, https://doi.org/10.1016/j.labeco.2021.102002. [18]

Arntz, M., T. Gregory and U. Zierahn (2016), "The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis", *OECD Social, Employment and Migration Working Papers*, No. 189, OECD Publishing, Paris, https://doi.org/10.1787/5jlz9h56dvq7-en. [7]

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. [4]

Babina, T. et al. (2020), "Artificial Intelligence, Firm Growth, and Industry Concentration", *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.3651052. [19]

Bellemare, M. et al. (2013), "The Arcade Learning Environment: An Evaluation Platform for General Agents", *Journal of Artificial Intelligence Research*, Vol. 47, pp. 253-279, https://doi.org/10.1613/jair.3912. [14]

Clark, P. et al. (2019), "From 'F' to 'A' on the N.Y. Regents Science Exams: An Overview of the Aristo Project". [22]

Deng, J. et al. (2009), "ImageNet: A large-scale hierarchical image database", *2009 IEEE Conference on Computer Vision and Pattern Recognition*, https://doi.org/10.1109/cvpr.2009.5206848. [12]

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]

Felten, E., M. Raj and R. Seamans (2019), "The Variable Impact of Artificial Intelligence on Labor: The Role of Complementary Skills and Technologies", *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.3368605. [15]

Frey, C. and M. Osborne (2017), "The future of employment: How susceptible are jobs to computerisation?", *Technological Forecasting and Social Change*, Vol. 114, pp. 254-280, https://doi.org/10.1016/j.techfore.2016.08.019. [5]

ILO (2012), *International Standard Classification of Occupations. ISCO-08. Volume 1: Structure, group definitions and correspondence tables*, International Labour Organization. [11]

Liu, Y. et al. (2019), "How Well Do Machines Perform on IQ tests: a Comparison Study on a Large-Scale Dataset", *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, https://doi.org/10.24963/ijcai.2019/846. [20]

Martínez-Plumed, F. et al. (2020), "Does AI Qualify for the Job?", *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, https://doi.org/10.1145/3375627.3375831. [17]

National Center for O*NET Development (n.d.), *O*NET 27.2 Database*, https://www.onetcenter.org/database.html (accessed on 24 February 2023). [6]

Nedelkoska, L. and G. Quintini (2018), "Automation, skills use and training", *OECD Social, Employment and Migration Working Papers*, No. 202, OECD Publishing, Paris, https://doi.org/10.1787/2e2f4eea-en. [8]

OECD (2021), *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/5ee71f34-en. [23]

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. [25]

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. [2]

OECD (2013), *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*, OECD Publishing, Paris, https://doi.org/10.1787/9789264204256-en. [3]

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. [24]

Saxton, D. et al. (2019), "Analysing Mathematical Reasoning Abilities of Neural Models". [21]

Squicciarini, M. and J. Staccioli (2022), "Labour-saving technologies and employment levels: Are robots really making workers redundant?", *OECD Science, Technology and Industry Policy Papers*, No. 124, OECD Publishing, Paris, https://doi.org/10.1787/9ce86ca5-en. [10]

Tolan, S. et al. (2021), "Measuring the Occupational Impact of AI: Tasks, Cognitive Abilities and AI Benchmarks", *Journal of Artificial Intelligence Research*, Vol. 71, pp. 191-236, https://doi.org/10.1613/jair.1.12647. [16]

Wang, A. et al. (2018), "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". [13]

Webb, M. (2020), "The Impact of Artificial Intelligence on the Labor Market", *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.3482150. [9]

Notes

[1] PIAAC defines problem solving in technology-rich environments as the ability to use "digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks" (OECD, 2012[24]). The focus is not on "computer literacy", but rather on the cognitive skills required in the information age. Examples include locating and evaluating information on the Internet for quality and credibility, managing personal finances using spreadsheets, statistical packages, or operating a computer.

These skills are assessed only in the First Cycle of PIAAC (2011-17), which is the focus of this report. The Second Cycle, which is under way, assesses adaptive problem solving instead. This is the ability of problem solvers to handle dynamic and changing situations, and to adapt their initial solution to new information or circumstances (OECD, 2021[25]).

[2] Throughout this report, the term "computers" is used to refer generally to AI, robots and other types of information and communications technologies.

[3] See https://www.oecd.org/education/ceri/future-of-skills.htm (accessed on 21 February 2023).

2 Evolution of human skills versus AI capabilities

This chapter offers an overview of changes in human skills and computer capabilities in the domains of literacy and numeracy over time. It first analyses changes in the skill levels of adults aged 16 to 65, working adults and students aged 15 using data from the Programme for International Student Assessment (PISA), the Survey of Adult Skills (PIAAC), the International Adult Literacy Survey (IALS) and the Adult Literacy and Life Skills Survey (ALL). The chapter then describes recent trends in the fields of natural language processing and mathematical reasoning of artificial intelligence (AI). These technological developments are relevant for the potential performance of AI on the PIAAC test. By showing that technological progress develops much faster than human skills in key skill domains, the chapter highlights the need for periodically and systematically monitoring the evolution of AI capabilities and comparing them to human skills.

The skill level of a population is key to a country's capacity for innovation, growth and competitiveness. Therefore, countries have large incentives to raise the supply of skills and optimise the available stock of skills. They use different policies to achieve this. Governments invest in education and try to improve the labour-market relevance of training programmes to develop the "right" skills in the future workforce – the skills that are needed by the economy and that help individuals thrive. Other policies aim at up-skilling and re-skilling the workforce, for example, by encouraging employers to offer more learning opportunities at the workplace; by strengthening lifelong learning; by activating the unemployed; or by training migrants to help them enter the labour market. However, all these policy efforts take time to contribute effectively to the formation of skills in the workforce.

By contrast, technological progress is moving fast and machines can reproduce more and more of the skills of human workers. Advances in big data, computational power, storage capacity and algorithmic techniques have driven big improvements especially in artificial intelligence (AI) and robotics capabilities over the past decade. AI is now faster, less biased and more accurate on a variety of tasks compared to humans. Language processing technology, for example, already exceeds human-level performance in speech recognition and in translation in a restricted domain. It has also made considerable progress on linguistic challenges that require logical reasoning or commonsense knowledge. In the field of vision, AI has surpassed humans in object detection, face recognition and many medical diagnostics tasks based on images. In robotics, systems are still constrained in unstructured environments. However, they have become more agile, mainly due to advances in machine learning and increased availability of sophisticated sensor systems (Zhang et al., 2022[1]).

This chapter provides background information on how human skills and computer capabilities with regard to numeracy and literacy evolve over time. By showing how much more rapidly the latter progress, the chapter highlights the need for periodically and systematically monitoring the evolution of AI capabilities and comparing them to human skills.

This chapter first analyses the skill level of adults aged 16 to 65 in the domains of literacy and numeracy and shows how it changes over time. The analysis draws on the Survey of Adult Skills (PIAAC), as well as on comparable data from two earlier skills assessments – the International Adult Literacy Survey (IALS) carried out in 1994-98 and the Adult Literacy and Life Skills Survey (ALL) carried out in 2003-07. Additional analyses focus on the reading and mathematical skills of students using data of the Programme for International Student Assessment (PISA) from 2000 to 2018. The chapter then provides an overview of recent technological developments in the fields of natural language processing (NLP) and quantitative reasoning of AI.

Changes in skills supply

Long-term developments in human skills are hard to assess. Economic studies have traditionally used average years of schooling and qualifications and diplomas attained as proxies for the supply of skills. From this perspective, skills supply should have increased across the OECD over the last decade because all member countries had an increase of the share of the adult population holding a tertiary degree (OECD, 2023[2]).

However, formal educational qualifications do not always fully capture the actual skills of individuals. For example, they do not account for skills and knowledge acquired after formal education. Nor do they capture loss of skills due to inactivity or ageing (OECD, 2012[3]). Skills assessments like PIAAC, by contrast, offer a direct measure of skills, although they are necessarily limited to a narrow set of skills.

In the following, the levels of literacy and numeracy skills of adults are presented using results from PIAAC. So far, data from only one survey wave are available. These data are combined with comparable data from IALS to address changes in literacy skills over time. Nineteen countries or economies participated in both

PIAAC and IALS, with results 13-18 years apart, depending on the country.[1] Changes in numeracy skills are analysed by comparing PIAAC results with those from ALL conducted in 2003 and then again between 2006 and 2008. This comparison is possible for only seven countries and has a shorter time frame of five to nine years.[2]

Changes in literacy skills

The comparability of the literacy data from PIAAC and IALS is limited because of changes in the assessment instruments between the two surveys. The literacy domain in PIAAC incorporates material assessed in two separate domains of prose and document literacy in IALS (OECD, 2016[4]). However, the data from IALS were re-analysed to create scores for a comparable joint literacy domain (OECD, 2013[5]). Over half of the literacy items used in PIAAC had also been used in IALS, and these linking items provided the basis for constructing comparable scales for the two surveys.

Skills are assessed on a 500-point scale, which is used to describe both the difficulty of individual test questions and the proficiency of individual adults who took the survey. For ease in understanding, the continuous scale is often described using six difficulty/proficiency levels – from below Level 1 to Level 5. Literacy questions at the lower difficulty levels (Level 1 and below) use short texts of a few sentences and ask about information that can be clearly identified in the text from the words used in the question. At the higher levels, the texts are longer and the questions may require interpreting or synthesising, as well as avoiding misleading information that may superficially appear to provide the answer. Individuals at each proficiency level can successfully complete two-thirds of the questions at that level. They also have higher chances of completing less difficult questions and lower chances of answering more difficult ones.

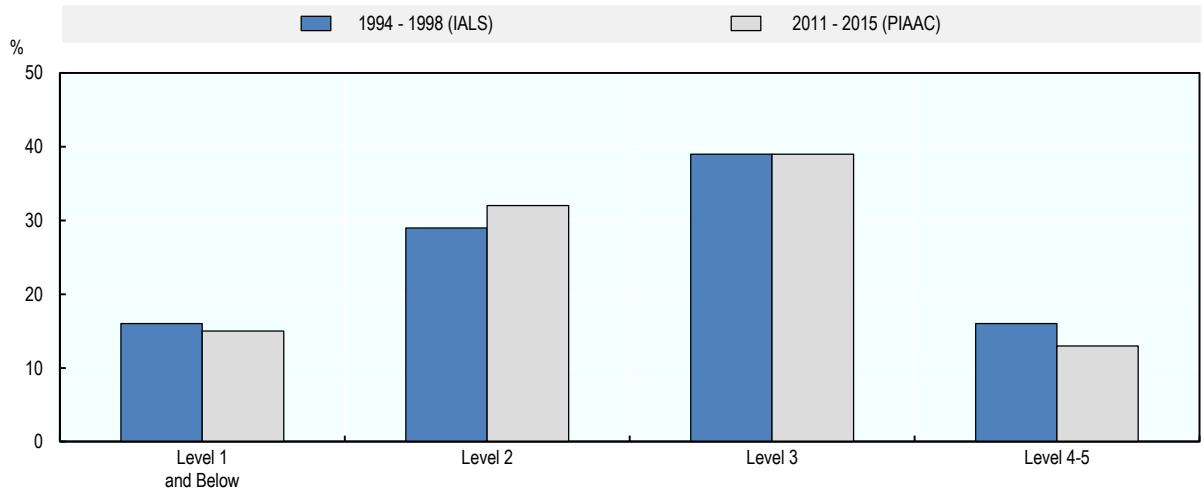
Figure 2.1. Literacy proficiency levels of 15-65 year-olds, IALS and PIAAC

Source: Adapted from Elliott, S. (2017[6]), *Computers and the Future of Skill Demand*, Figure 2.1, https://doi.org/10.1787/9789264284395-en.

Figure 2.1 shows the literacy proficiency results of adults aged 16 to 65, averaged across the 19 OECD countries and economies that participated in both IALS and PIAAC. Because of relatively small numbers of adults at the top and bottom of the scale, respondents with proficiency scores at Level 1 and below Level 1 are combined in a single category, as are those at Levels 4 and 5. In PIAAC, over two-thirds of adults have literacy proficiency at Levels 2 or 3. Over the gap of 13-18 years between IALS and PIAAC, skill levels in the adult population have shifted marginally, on average. The share of adults at Level 2

increased by four percentage points, while the shares in the bottom and top categories decreased by two percentage points each.

Figure 2.2. Literacy proficiency levels of working population, IALS and PIAAC



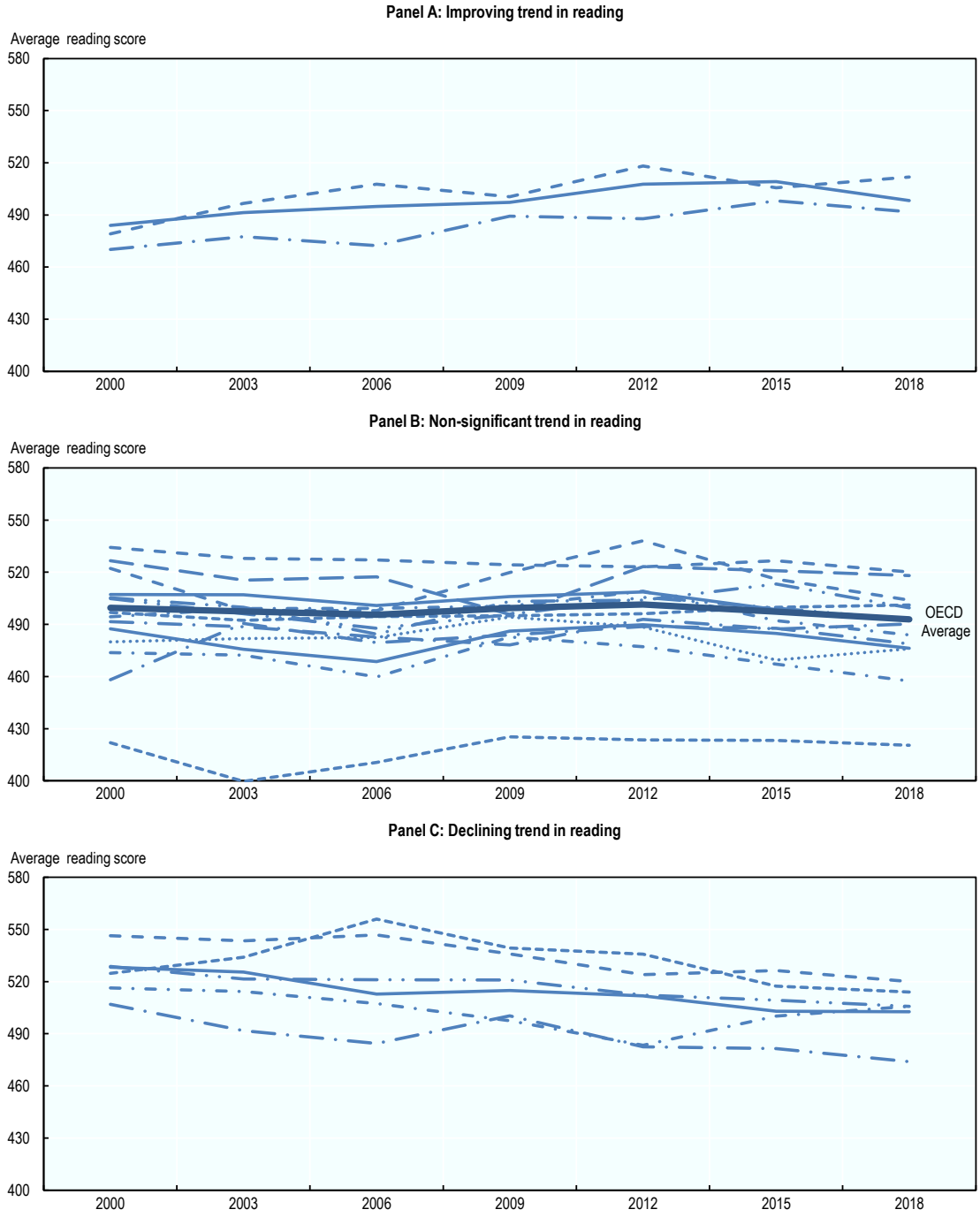
Source: Adapted from Elliott, S. (2017[6]), *Computers and the Future of Skill Demand*, Figure 2.2, https://doi.org/10.1787/9789264284395-en.

For comparison, Figure 2.2 shows the literacy proficiency of working adults only. The working population has similar skill levels as the full adult population. Nearly three-quarters (71%) of the employed have literacy skills at Levels 2 and 3 and 13% are proficient at the highest levels. Comparison with the IALS results shows that literacy skills of the working population have slightly decreased over time. The share of individuals at Levels 4 and 5 decreased by three percentage points, while the share of Level 2 individuals increased by four percentage points. A look at the individual countries reveals this pattern of decreasing literacy skills is more strongly pronounced in Canada, Denmark, Germany, Norway, Sweden and the United States (see Table A2.2 in Annex 2.A). Only Australia, Poland and Slovenia have shifted the distribution of literacy skills towards higher skill levels.

That literacy skills of adults do not improve over time in most countries, despite increases in educational attainment, is related to changes in the composition of the adult population (Paccagnella, 2016[7]). In the period between both surveys, all countries experienced increases in the average age of the population. In addition, in all countries, immigration has led to a higher proportion of foreign-born adults in the population. Both trends are linked to lower levels of literacy skills and counterbalance the literacy gains made by increased educational attainment.

The skills supply of a country depends not only on the skill level of the active population, but also on how well a country develops the skills of youth cohorts in preparing them to enter the workforce. PISA provides results on the knowledge and skills of young people in reading, mathematics and science. The assessment has taken place every three years since 2000, thus enabling the observation of long-term trends in students' skills. Each round focuses on one of the three subjects and provides basic results for the other two. The first full assessment of a subject sets the starting point for future trend comparisons in this subject. Since the very first round had reading as a major domain, trends in reading performance of students can be observed since 2000.

Figure 2.3. Long-term trends in average reading proficiency of 15-year-olds



Note: Average reading scores of 23 OECD countries that took part in all PISA reading assessments since 2000. Countries' performance trends are classified as improving, not significantly changing or declining in accordance with the average three-year trend in mean performance. The average three-year trend is the average change, per three-year period, between the earliest available measurement in PISA and PISA 2018, calculated by a linear regression. Panel A presents countries with significantly positive three-year trends, Panel B presents countries with non-significant three-year trends, and Panel C shows countries with significantly negative trends.

Source: OECD (2019[8]), *PISA 2018 Results (Volume I): What Students Know and Can Do*, Table I.B1.10, https://doi.org/10.1787/5f07c754-en.

StatLink https://stat.link/b6uc98

PISA defines reading literacy as the capacity "to understand, use, evaluate, reflect on and engage with texts in order to achieve one's goals, develop one's knowledge and potential, and participate in society" (OECD, 2019, p. 14[9]). Reading performance is scored in relation to the variation of the results observed across all test participants in the first main assessment. That is, scores do not have a substantive meaning. Instead, they are scaled to fit a normal distribution with a mean of 500 score points and a standard deviation of 100 score points.

Figure 2.3 presents trends in countries' average reading scores since 2000. It focuses on 23 OECD countries that took part in all PISA reading assessments. Their performance trends are classified as improving, not significantly changing or declining in accordance with the average three-year change in mean performance between assessments. The average three-year trend in students' reading performance is significantly positive in only three countries – Germany, Poland and Portugal (see Table A2.3, Annex 2.A). In most participating countries, young people's reading skills have not changed significantly over time. Six countries – Australia, Finland, Iceland, Korea, New Zealand and Sweden – have declining trends in reading (see Table A2.3, Annex 2.A).

Changes in numeracy skills

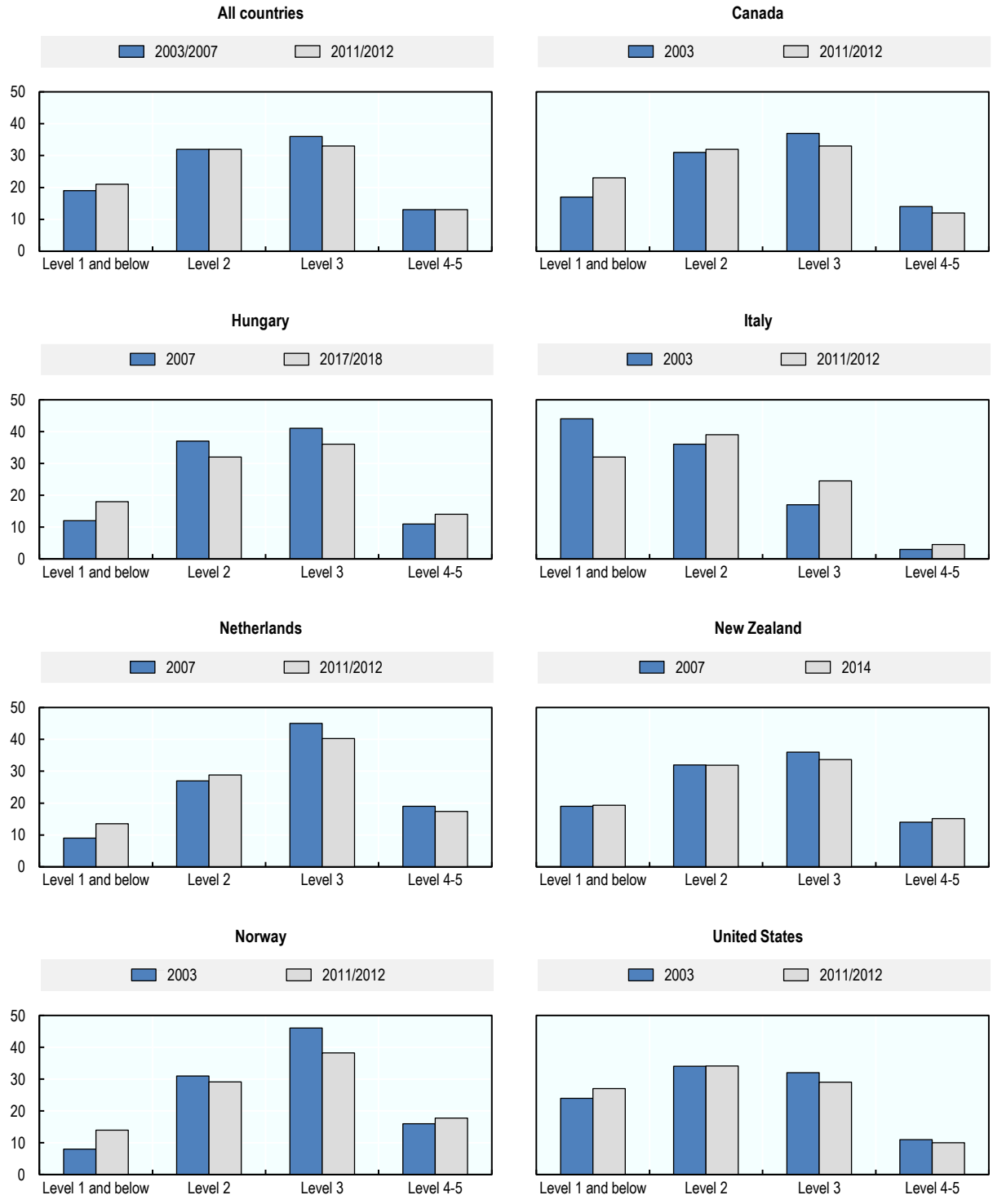
It is not possible to compare data on numeracy from PIAAC and IALS. Because the numeracy domain in PIAAC is substantially different from the quantitative literacy domain included in IALS, it is not possible to construct a comparable scale for the earlier survey.

The numeracy assessment of the PIAAC is similar to that used in ALL in terms of the constructs measured and the test content (OECD, 2013[10]; Paccagnella, 2016[7]). Most of the numeracy test items used in PIAAC were used in ALL. The numeracy results of ALL have also been re-estimated to fit the measurement scale used in PIAAC. Such comparable data on numeracy is available for seven OECD countries – Canada, Hungary, Italy, the Netherlands, New Zealand, Norway and the United States.

Numeracy in PIAAC is assessed on the same 500-point scale as literacy, which, in the following, is again broken down into four proficiency levels. Respondents at a given proficiency level can solve approximately two-thirds of the questions at that level. They are also more successful on less difficult questions and less successful on more difficult ones. Questions at lower difficulty levels require respondents to perform simple, one-step operations, such as counting or sorting. Questions at higher levels of difficulty, by contrast, typically require understanding and integrating several mathematical procedures, such as reading graphs, calculating rate of change and applying formulae.

Figure 2.4. presents the distribution of the 16-65 year-old population across the four numeracy proficiency levels in ALL and PIAAC. Similar to the findings for literacy, most of the population of the observed countries has numeracy skills at medium proficiency levels (Levels 2 and 3). Comparing PIAAC to ALL results shows that, across the seven countries, on average, skills have shifted slightly from higher to lower proficiency levels. Specifically, the proportion of adults at Level 3 has decreased by three percentage points, and the proportion of adults with the poorest numeracy skills has increased by two percentage points. This trend is more strongly pronounced in Canada, Hungary, the Netherlands, Norway, and, to some extent, in the United States. Among the observed countries, numeracy skills of adults have improved only in Italy over time. However, this increase started from a large share of poorly skilled individuals.

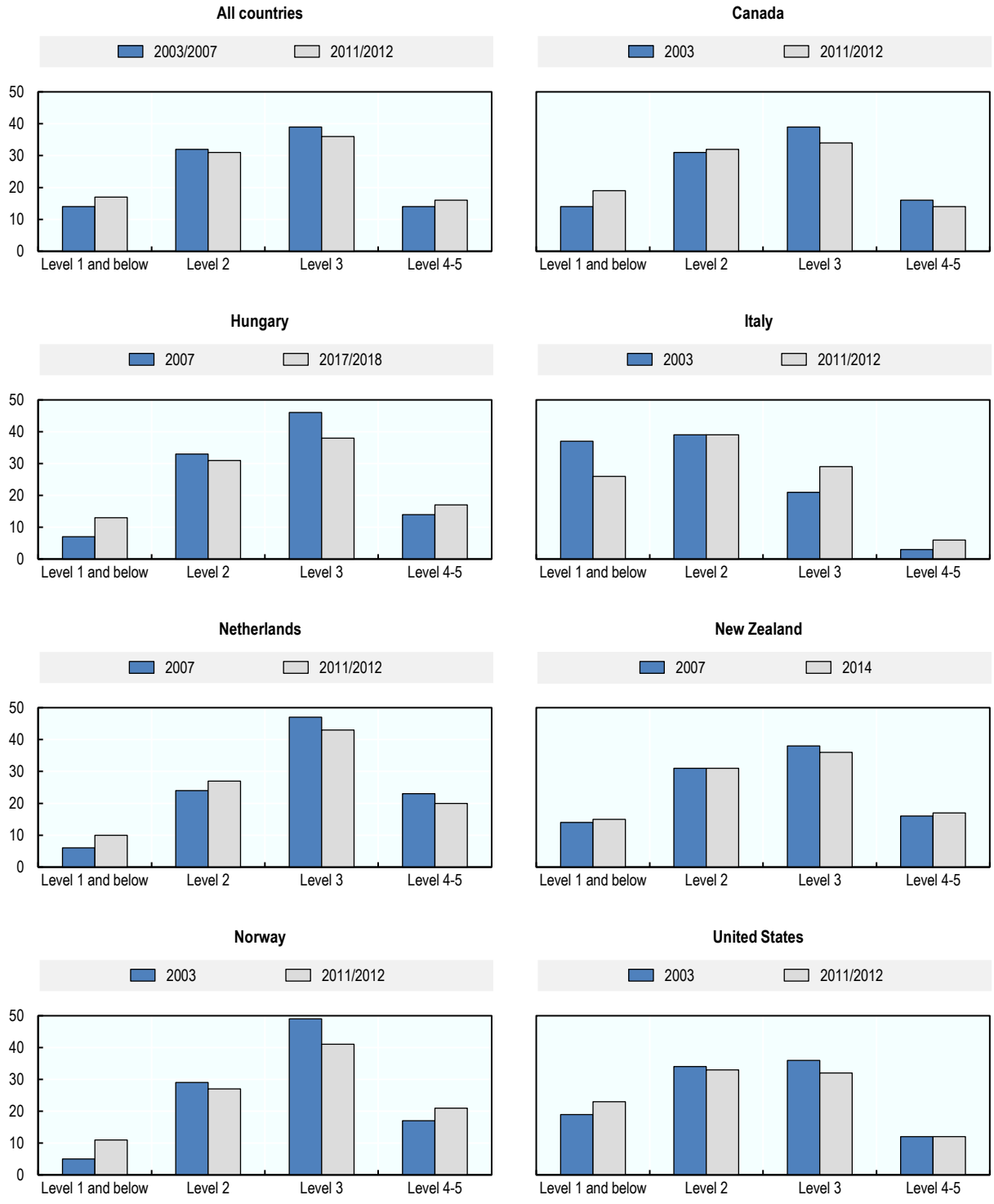
Figure 2.4. Numeracy proficiency levels of 16-65 years-olds, ALL and PIAAC



Source: US Department of Education, National Center for Education Statistics, Statistics Canada and OECD (2020[11]), *Adult Literacy and Life Skills Survey (ALL) 2003-2008 and PIAAC 2012-2017 Literacy, Numeracy, and Problem Solving TRE Assessments*, https://nces.ed.gov/surveys/piaac/ideuspiaac (accessed on 31 August 2022).

StatLink https://stat.link/3hn6kj

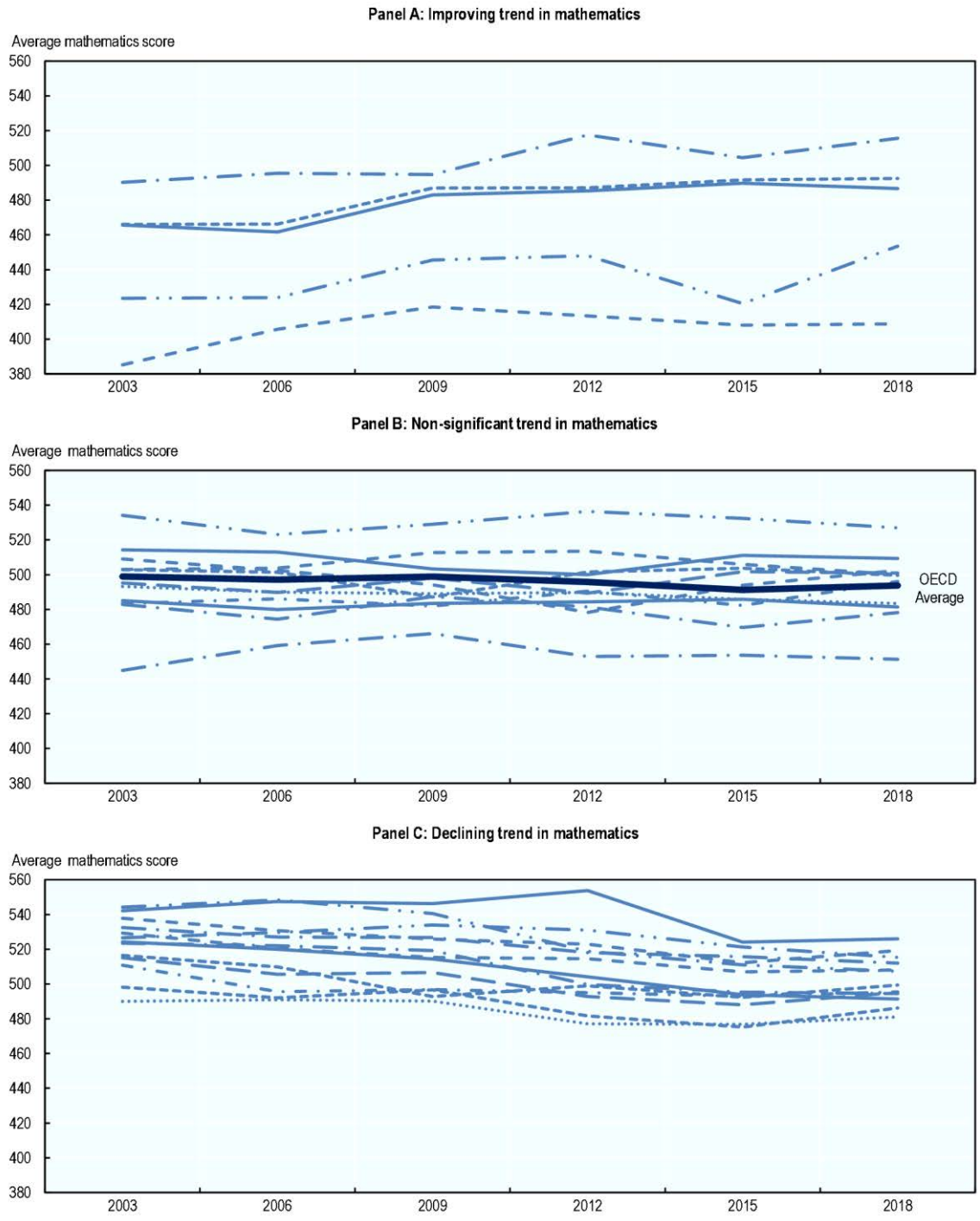
Figure 2.5. Numeracy proficiency levels of the working population, ALL and PIAAC



Source: US Department of Education, National Center for Education Statistics, Statistics Canada and OECD (2020[11]), *Adult Literacy and Life Skills Survey (ALL) 2003-2008 and PIAAC 2012-2017 Literacy, Numeracy, and Problem Solving TRE Assessments*, https://nces.ed.gov/surveys/piaac/ideuspiaac (accessed on 31 August 2022).

StatLink https://stat.link/k8ow7y

Figure 2.6. Long-term trends in average mathematics proficiency of 15-year-olds



Note: Average mathematics scores of 29 OECD countries that took part in all PISA mathematics assessments since 2003. Countries' performance trends are classified as improving, not significantly changing or declining in accordance with the average three-year trend in mean performance. The average three-year trend is the average change, per three-year period, between the earliest available measurement in PISA and PISA 2018, calculated by a linear regression. Panel A presents countries with significantly positive three-year trends, Panel B presents countries with non-significant three-year trends, and Panel C shows countries with significantly negative trends.
Source: OECD (2019[8]), *PISA 2018 Results (Volume I): What Students Know and Can Do*, Table I.B1.11, https://doi.org/10.1787/5f07c754-en.

StatLink https://stat.link/dioz4m

Figure 2.5 shows the numeracy proficiency of the working population and its change over time. Compared to the full adult population, the working population demonstrates higher numeracy skills, with smaller shares of workers having poor numeracy skills. Across all observed countries, on average, the shares of working adults with medium numeracy proficiency have slightly decreased over time. Meanwhile, the margins of the skills distribution – the shares of workers with the lowest and the higher numeracy proficiency – have increased. This pattern is observed for Hungary and Norway. In Canada and the Netherlands, the numeracy skills of working adults have shifted from higher (Levels 3-5) to lower (Level 1 and below and Level 2) proficiency level. Only Italy shows an improving trend in workers' numeracy skills.

PISA has provided information on the mathematics skills of 15-year-old students. Comparisons of mathematics performance across time are possible from 2003 on. Mathematics skills in PISA are defined as students' "capacity to formulate, employ and interpret mathematics in a variety of contexts" (OECD, 2019, p. 75[9]). This includes mathematical reasoning and the use of mathematical concepts and procedures to describe, explain and predict phenomena. Mathematics skills are assessed similarly to reading: scores are scaled to fit a normal distribution with a mean of 500 score points and a standard deviation of 100 score points. Students with the lowest scores can identify mathematical information that is clearly stated and perform routine mathematical procedures. Students with the highest proficiency can understand, use and conceptualise mathematical information of various types and apply advanced mathematical reasoning to solve complex problems (OECD, 2019[9]).

Figure 2.6 presents trends in average mathematical scores of 15-year-olds. The focus is on 29 OECD countries that participated in all mathematical assessments since 2003. Countries are grouped according to their average three-year score change into countries with significantly positive average change (Panel A), countries with a non-significant trend (Panel B) and countries with a significantly negative average three-year change (Panel C). The figure shows that only five of the observed countries experienced an improvement of young people's mathematics performance since 2003 (Italy, Mexico, Poland, Portugal and Türkiye, see Table A2.4, Annex 2.A). Eleven countries have non-significant trends, while mathematical average scores declined over time in 13 countries.

In sum, numeracy and literacy skills have not changed much over time, either for the adult or the young population. Only a few countries experienced improvements in foundation skills. This may result from various factors, including population ageing, immigration or changing skill proficiency of particular groups (Paccagnella, 2016[7]). However, the small to moderate changes in literacy and numeracy show that lifting up the supply of skills is challenging for governments.

Recent developments in AI capabilities

In contrast to human skills, AI capabilities develop fast. Over the past decade, a wave of technological progress has occurred in many AI fields, including vision, NLP, speech recognition, image understanding, reinforcement learning and robotics (Littman et al., 2022[12]). This has led to the proliferation of AI applications in various contexts, such as translation, games, medical diagnosis, stock trading, autonomous driving and science. Some observers have labelled this recent uptake in AI development and deployment "a golden decade of deep learning" (Dean, 2022[13]).

The following sections briefly summarise technological developments that are relevant for the evolution of computer capabilities in the domains of literacy and numeracy. Specifically, recent progress in the fields of NLP and quantitative reasoning are described and discussed.

Recent developments in natural language processing

NLP is a major domain in AI. It aims at building computer capabilities that allow AI systems to process and interpret spoken and written language to perform different linguistic tasks. These include extracting

information from large amounts of text data, correctly categorising and synthesising text content or communicating with humans.

The field consists of various sub-domains, each of which is centred around one major task or challenge. For example, Speech Recognition is a sub-domain that aims at reliably converting voice data into text, while Question-Answering deals with the automatic retrieval or generation of answers to questions posed in text or speech. Natural language technologies are typically developed within such narrow domains and focus on specific tasks. Their performance is evaluated accordingly – on domain-specific benchmarks that provide a standard for comparing different approaches in the domain. Benchmarks are test datasets, on which systems perform a task or a set of tasks (see example tasks in Box 2.1).

NLP has experienced a major surge in the last several years. In many domains, AI systems' performance has outpaced the tests developed to measure it (Zhang et al., 2022[1]). In Question-Answering, AI systems improved so rapidly that researchers launched a more challenging version of the Stanford Question Answering Dataset (SQuAD) only two years after the benchmark's initial release in 2016 (Rajpurkar et al., 2016[14]; Rajpurkar, Jia and Liang, 2018[15]). It took another year until systems reached human-level performance on SQuAD 2.0.[3] Similarly, AI exceeded human performance on the General Language Understanding Evaluation (GLUE) benchmark within a year, and on its successor SuperGLUE soon after.[4] Both benchmarks test systems on a number of distinct tasks, such as Question-Answering and commonsense reading comprehension (Wang et al., 2018[16]; Wang et al., 2019[17]).

Performance has also improved in Natural Language Inference, the task of "understanding" the relationship between sentences, e.g. whether two sentences contradict or entail each other. This is shown by benchmarks such as Stanford Natural Language Inference (SNLI) (Bowman et al., 2015[18]) and Abductive Natural Language Inference (aNLI) (Bhagavatula et al., 2019[19]). Considerable progress was also registered in text summarisation, translation and sentiment analysis (Zhang et al., 2022[1]).

This breakthrough in NLP was driven by the emergence of large pre-trained language models, such as Embeddings from Language Models (ELMo) by Peters et al. (2018[20]), Generative Pre-Trained Transformer (GPT) by Radford et al. (2018[21]) and Bidirectional Encoder Representations from Transformers (BERT) by Devlin et al. (2018[22]). These models are used to further develop specific NLP systems for particular tasks and domains. Specifically, they are trained once, on a large corpus of unlabelled text data to "learn" general language patterns and the semantic of words. The models can be then "fine-tuned" to downstream tasks, meaning they are adapted to a target task with additional training. This fine-tuning or extra training uses domain-specific training data to allow the general pre-trained models to learn the vocabulary, idioms and syntactic structures common in a new domain.

Box 2.1. Example tasks from natural language processing benchmarks

The General Language Understanding Evaluation dataset (SuperGLUE), Wang et al. (2019[17])

Text: Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.

Question: Is Barq's root beer a Pepsi product?

Answer: No

The Stanford Question Answering Dataset (SQuAD) 2.0, Rajpurkar, Jia and Liang (2018[15])

Text: Southern California, often abbreviated SoCal, is a geographic and cultural region that generally comprises California's southernmost 10 counties. The region is traditionally described as "eight counties", based on demographics and economic ties: Imperial, Los Angeles, Orange, Riverside, San Bernardino, San Diego, Santa Barbara, and Ventura. The more extensive 10-county definition, including Kern and San Luis Obispo counties, is also used based on historical political divisions. Southern California is a major economic centre for the state of California and the United States.

Question: What is a major importance of Southern California in relation to California and the United States?

Answer: economic centre

The Stanford Natural Language Inference (SNLI) Corpus, Bowman et al. (2015[18])

Text: A man inspects the uniform of a figure in some East Asian country.

Hypothesis: The man is sleeping.

Answer: contradiction

Choice of Plausible Alternatives (COPA), Roemmele, Adrian Bejan and S. Gordon (2011[23])

Premise: The man broke his toe. What was the CAUSE of this?

Alternative 1: He got a hole in his sock.

Alternative 2: He dropped a hammer on his foot.

Answer: Alternative 2

The Cloze Test by Teachers (CLOTH) benchmark, Xie et al. (2017[24])

Text: Nancy had just got a job as a secretary in a company. Monday was the first day she went to work, so she was very … and arrived early.

Question: A. depressed B. encouraged C. excited D. surprised

Answer: C

The introduction of these large-scale models has considerably pushed the state of the art of NLP forward. When ELMo was first introduced in 2018, it helped exceed system performance on various tasks in the domains of Question-Answering, textual entailment and sentiment analysis (Storks, Gao and Chai, 2019[25]). The release of GPT pushed forward AI performance on 12 benchmarks, including GLUE, SNLI and the Choice of Plausible Alternatives (COPA) benchmark (Roemmele, Adrian Bejan and S. Gordon, 2011[23]), which evaluates commonsense causal reasoning. Its later updated versions GPT-2 and GPT-3 further improved state-of-the-art results on numerous language modelling tasks. Similarly, BERT topped several benchmark performance rankings when first released, such as those of GLUE, SQuAD, COPA, Situations With Adversarial Generation (SWAG) (Zellers et al., 2018[26]) on commonsense reasoning and CLOze test by TeacHers (CLOTH) (Xie et al., 2017[24]), a collection of questions from middle and high school-level English language exams.

Remarkably, these models perform well on novel tasks without considerable additional training. When applied without any subsequent fine-tuning, GPT-3, for example, achieves strong performance on numerous language tasks. Indeed, in many cases, it outpaces state-of-the-art systems designed for the task (Brown et al., 2020[27]). Performance is even better in settings, where the model is provided with only one demonstration of the novel task or with few prior examples (typically 10-100).

As another remarkable feature, pre-trained language models can perform a huge variety of tasks, even without being explicitly trained for those tasks. In the above example, a GPT-3 without fine-tuning performed well on tasks as diverse as translation, Question-Answering, reading comprehension, reasoning or three-digit arithmetic. These qualities of language models are a gateway towards producing more general AI systems – systems that can adapt to new situations and solve problems from different domains without extensive additional training.

The success of pre-trained language models is mostly due to the use of self-supervised learning. This allows for training models on unprecedented amounts of training data. In self-supervised learning, neural network models are trained on texts, parts of which are made hidden. The task of the model is to predict the hidden words as function of the context in which they appear. In this way, the model “learns” grammar rules and semantics. This approach does not require a human to label training examples as true or false, thus enabling the use of more training data. Moreover, training occurs only once, which significantly reduces the costs and time to develop a system. Researchers can simply download a general pre-trained language model and fine-tune it for a specific task on a smaller amount of domain-specific data.

The Transformer architecture is one of the most advanced and widely used self-supervised approaches. Both BERT and GPT are pre-trained on Transformers. Its crucial feature is a “self-attention” mechanism that allows for capturing long-range dependencies between words (e.g. words far apart in a sentence) (Littman et al., 2022[12]). In addition, Transformers and similar architectures allow for learning the meaning of words in context. For example, the word “rose” would be represented differently in the sentences “Roses are red” and “The sun rose”. This is a considerable advantage over earlier pre-trained word-embedding models like word2vec (Mikolov et al., 2013[28]). In such models, words are represented with the same vector, independent of the context, in which they are used.

While these new approaches considerably advanced the field of NLP, they are still far from producing AI systems that can process language as humans do. The reason is that NLP systems still lack a deep understanding of speech and text. This limits their capacity to perform more sophisticated language tasks that require commonsense knowledge and complex reasoning.

Recent developments in mathematical reasoning of AI

Research on automating mathematical reasoning has a long history with important achievements. These include the development of tools, such as Maple, Mathematica and Matlab, that can perform numerical and symbolic mathematical operations. Significant effort has also gone into automated theorem proving,

with major achievements, such as proving the four-colour theorem among others (Appel and Haken, 1977[29]). This section focuses narrowly on mathematical benchmarks that are comparable to the PIAAC numeracy test.

From the perspective of AI research, mathematical problems can be divided roughly into the following categories (Davis, 2023[30]):

- *Symbolic problems*: problems formulated in mathematical notation with minimal use of natural language. For example, "Solve $x^3 - 6x^2 + 11x - 6 = 0$".
- *Word problems*: problems stated in (more than minimal) natural language, possibly in combination with symbols.
    - Purely mathematical word problems: problems with minimal reference to non-mathematical concepts. E.g. "Find a prime number p such that p+36 is a square number."
    - Real-world word problems: problems whose solution requires using non-mathematical knowledge.
    - Commonsense word problems (CSW): problems involving significant use of commonsense knowledge and possibly common knowledge, but not encyclopedic or expert knowledge. Elementary CSWs require only elementary mathematics (Davis, 2023[30]).

Researchers have been trying to develop systems that solve mathematical word problems since the 1960s (Davis, 2023[30]). However, the dominance of machine-learning techniques over the past two decades has also affected AI research in mathematical reasoning. As a result, much recent work aims at generalising large-scale, pre-trained language models to mathematical problems and the quantitative aspects of natural language (Lewkowycz et al., 2022[31]; Saxton et al., 2019[32]). While acknowledging other types of AI research, this brief summary focuses on recent efforts, featuring some prominent benchmarks and the performance of deep/machine-learning systems on these (see Box 2.2).

Mathematical reasoning has received less attention in AI research than language modelling or vision because it has relatively less applicability and commercial use. However, some AI experts argue that mathematical reasoning poses an interesting challenge for AI (Saxton et al., 2019[32]). It requires learning, planning, inferring and exploiting laws, axioms and mathematical rules, among others. These capabilities can enable more powerful and sophisticated systems that can solve more complex, real-world problems.

Mathematics is generally considered hard for AI (Choi, 2021[33]). In 2019, researchers from DeepMind Technologies, an AI-focused company owned by Google, tested the performance of state-of-the-art NLP models in the domain of mathematics (Saxton et al., 2019[32]). For that purpose, they developed a test dataset of questions from the areas of algebra, arithmetic, calculus, comparisons and measurement, among others. In addition, they evaluated systems on publicly available mathematics exams for 16-year-old schoolchildren in Britain. The best-performing model in the study, the Transformer, achieved moderate results on the test dataset. It also failed the school maths exam, answering correctly only 14 of the 40 questions (O'Neill, 2019[34]).

To facilitate research in the field, researchers from University of California, Berkeley introduced MATH in 2021, a test dataset containing 12 500 challenging mathematics problems (Hendrycks et al., 2021[35]). The questions are in textual format and span different fields of mathematics. Large language models, pre-trained on mathematical content and presented with examples from MATH, achieved poor results on the benchmark at the time of release, with accuracy of 3-6.9%. However, MATH is also challenging for humans. A three-time gold medallist in the International Mathematical Olympiad attained 90% on the test, while a PhD student in computer science achieved 40%.

Box 2.2. Example tasks from benchmarks on mathematical reasoning

MATH Dataset, Hendrycks et al. (2021[35])

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colours ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is 1 + 6 = 7.

GSM8K, Cobbe et al. (2021[36])

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?

Solution: Beth bakes 4 2 dozen batches of cookies for a total of 4*2=8 dozen cookies. There are 12 cookies in a dozen, and she makes 8 dozen cookies for a total of 12*8=96 cookies. She splits the 96 cookies amongst 16 people so they each eat 96/16=6 cookies.

Saxton et al. (2019[32])

Question: Solve -42*r + 27*c = -1167 and 130*r + 4*c = 372 for r.

Answer: 4

MathQA, Amini et al. (2019[37])

Question: A train running at the speed of 48 km/hr crosses a pole in 9 seconds. What is the length of the train? a) 140, b) 130, c) 120, d) 170, e) 160

Answer: C

NumGLUE, Mishra et al. (2022[38])

Question: A man can lift one box in each of his hands. How many boxes can a group of 5 people hold in total?

Answer: 10

Similarly, in 2021, researchers from OpenAI, a prominent AI research laboratory, released GSM8K, a dataset of 8 500 diverse mathematics problems at grade school-level in textual form (Cobbe et al., 2021[36]). The questions require a sequence of several simple arithmetic operations to solve. They are generally easier than MATH test questions. A well-performing middle-school student is expected to correctly solve the entire test, for example. Nevertheless, competing AI methods achieved low to moderate results on the test when it was released.

Some developments illustrate how large language models have been applied in the field of mathematical reasoning. In 2022, Google introduced Minerva, a large language model pre-trained on general natural

language data and further fine-tuned on technical content (Lewkowycz et al., 2022[31]). Currently, the model tops the MATH benchmark (as of 21 February 2023).[5] In addition, Minerva achieved good results on questions from engineering, chemistry, physics, biology and computer science, stemming from the Massive Multitask Language Understanding dataset (Hendrycks et al., 2021[35]). The model also obtained a score of 57% on the National Maths Exam in Poland, which corresponds to the average human performance in 2021.

In the same year, Codex (Chen et al., 2021[39]), a system developed by Open AI, achieved high accuracy on subsets of the MATH dataset (Hendrycks et al., 2021[35]), as well as on questions from university-level mathematics courses (Drori et al., 2021[40]). The model is a neural network pre-trained on text and fine- tuned on publicly available code. While the system's achievement has been acknowledged by the community, its reported high performance has been questioned on numerous grounds (Davis, 2022[41]). Criticism includes that it is not the neural network that solves the problems but a mathematical tool that it invokes (a Python algebra package) and that the system may work based on correct answers recorded in the test corpus. In addition, as language models, both Minerva and Codex are limited to textual input, and cannot handle diagrams and graphs, which are often essential elements of mathematical problems.

More recently, in 2022, several datasets have been assembled in the collection LILA that combines 23 existing benchmarks and covers a variety of linguistic complexity and mathematical difficulty (Mishra et al., 2022[42]). Systems, such as Codex, GPT-3, Neo-P and Bashkara, perform with varying success in the different mathematical domains on LILA (Davis, 2023[30]).

Despite some successes, AI is still far from mastering mathematical problems. State-of-the-art results on the MATH or LILA datasets, for example, are still far below the benchmark's ceiling. There is also room for improvement with regard to performance on GSM8K. To date, no AI system can solve elementary commonsense word problems reliably (Davis, 2023[30]). Moreover, prominent benchmarks in the field focus strongly on quantitative problems stated in text – "math word problems" – leaving other mathematical tasks unaddressed. In particular, mathematical tasks that include visual content, such as figures, tables, diagrams or other images, have received less attention.

The importance of measuring AI capabilities

The analysis presented in this chapter shows that AI capabilities in core domains develop much faster over time than human skills. Over the last two decades, literacy and numeracy skills of adults, of adult workers and of youth have increased in only but a few countries. This highlights the fact that uplifting the supply of skills in an economy is not a trivial task for policy makers and education providers. At the same time, AI technology develops quickly, excelling its capabilities and acquiring new ones. The last five years have seen tremendous breakthroughs in NLP, leading to improved capabilities of AI in literacy. In the domain of numeracy, technological progress, although at a smaller scale, is under way.

These developments give rise to important questions for policy and education:

- Will new AI capabilities result in substantial numbers of people whose skills are below those of AI across important capabilities used at work?
- What education and training will be needed for most people to develop some work-related capabilities beyond those of AI and robotics?
- What human capabilities will be too difficult for AI and robotics to reproduce over the next few decades?

While this chapter focuses on the supply side of skills, much previous research studies how technological change affects the *demand* for human skills. The notion is often that technological change is task-biased, meaning that machines can substitute workers in some tasks better than in others (Autor, Levy and

Murnane, 2003[43]; Frey and Osborne, 2017[44]). This would result in decreasing demand for workers for tasks that are automatable. At the same time, demand for workers in tasks that relate to the deployment and monitoring of machines at the workplace would increase. Many studies try to understand which tasks machines can automate (see Chapter 1). This has important implications:

- Which occupations are at high risk of automation?
- Will AI have a greater effect on the demand for low-skill or high-skill workers? Younger or older workers? Workers with more or less education?
- How might new AI capabilities change the overall amount of education or the types of capabilities that people need for work?

A systematic assessment of AI and robotic capabilities that allows for comparisons with human skills can provide answers to the above questions. The following study demonstrates how the use of standardised skills assessments and expert knowledge can help track AI capabilities in core domains of human skills.

References

Amini, A. et al. (2019), "MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms". [37]

Appel, K. and W. Haken (1977), "The Solution of the Four-Color-Map Problem.", *Scientific American*, Vol. 237/4, pp. 108-121, http://www.jstor.org/stable/24953967. [29]

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. [43]

Bhagavatula, C. et al. (2019), "Abductive Commonsense Reasoning". [19]

Bowman, S. et al. (2015), "A large annotated corpus for learning natural language inference". [18]

Brown, T. et al. (2020), "Language Models are Few-Shot Learners". [27]

Chen, M. et al. (2021), "Evaluating Large Language Models Trained on Code". [39]

Choi, C. (2021), *7 revealing ways AIs fail. Neural networks can be disastrously brittle, forgetful, and surprisingly bad at math*, IEEE Spectrum for the Technology Insider, https://spectrum.ieee.org/ai-failures (accessed on 1 February 2023). [33]

Cobbe, K. et al. (2021), "Training Verifiers to Solve Math Word Problems". [36]

Davis, E. (2023), *Mathematics, word problems, common sense, and artificial intelligence*, https://arxiv.org/pdf/2301.09723.pdf (accessed on 28 February 2023). [30]

Davis, E. (2022), *Limits of an AI program for solving college math problems*, https://arxiv.org/pdf/2208.06906.pdf (accessed on 5 February 2023). [41]

Dean, J. (2022), "A Golden Decade of Deep Learning: Computing Systems & Applications", *Daedalus*, Vol. 151/2, pp. 58-74, https://doi.org/10.1162/daed_a_01900. [13]

Devlin, J. et al. (2018), "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". [22]

Drori, I. et al. (2021), "A Neural Network Solves, Explains, and Generates University Math Problems by Program Synthesis and Few-Shot Learning at Human Level", https://doi.org/10.1073/pnas.2123433119. [40]

Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/9789264284395-en. [6]

Frey, C. and M. Osborne (2017), "The future of employment: How susceptible are jobs to computerisation?", *Technological Forecasting and Social Change*, Vol. 114, pp. 254-280, https://doi.org/10.1016/j.techfore.2016.08.019. [44]

Hendrycks, D. et al. (2021), "Measuring Mathematical Problem Solving With the MATH Dataset". [35]

Lewkowycz, A. et al. (2022), "Solving Quantitative Reasoning Problems with Language Models". [31]

Littman, M. et al. (2022), "Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report". [12]

Mikolov, T. et al. (2013), "Efficient Estimation of Word Representations in Vector Space". [28]

Mishra, S. et al. (2022), "Lila: A Unified Benchmark for Mathematical Reasoning". [42]

Mishra, S. et al. (2022), "NumGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks". [38]

OECD (2023), *Adult education level* (indicator), https://doi.org/10.1787/36bce3fe-en (accessed on 1 February 2023). [2]

OECD (2019), *PISA 2018 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris, https://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://doi.org/10.1787/5f07c754-en. [8]

OECD (2016), *The Survey of Adult Skills: Reader's Companion, Second Edition*, OECD Skills Studies, OECD Publishing, Paris, https://doi.org/10.1787/9789264258075-en. [4]

OECD (2013), *Technical Report of the Survey of Adult Skills (PIAAC)*, https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf (accessed on 1 February 2023). [5]

OECD (2013), *The Survey of Adult Skills: Reader's Companion*, OECD Publishing, Paris, https://doi.org/10.1787/9789264204027-en. [10]

OECD (2012), *Better Skills, Better Jobs, Better Lives: A Strategic Approach to Skills Policies*, OECD Publishing, Paris, https://doi.org/10.1787/9789264177338-en. [3]

O'Neill, S. (2019), "Mathematical Reasoning Challenges Artificial Intelligence", *Engineering*, Vol. 5/5, pp. 817-818, https://doi.org/10.1016/j.eng.2019.08.009. [34]

Paccagnella, M. (2016), "Literacy and Numeracy Proficiency in IALS, ALL and PIAAC", *OECD Education Working Papers*, No. 142, OECD Publishing, Paris, https://doi.org/10.1787/5jlpq7qglx5g-en. [7]

Peters, M. et al. (2018), "Deep contextualized word representations". [20]

Radford, A. et al. (2018), *Improving Language Understanding by Generative Pre-Training*, https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 1 February 2023). [21]

Rajpurkar, P., R. Jia and P. Liang (2018), "Know What You Don't Know: Unanswerable Questions for SQuAD". [15]

Rajpurkar, P. et al. (2016), "SQuAD: 100,000+ Questions for Machine Comprehension of Text". [14]

Roemmele, M., C. Adrian Bejan and A. S. Gordon (2011), *Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning*, http://commonsensereasoning.org/2011/papers/Roemmele.pdf (accessed on 1 February 2023). [23]

Saxton, D. et al. (2019), "Analysing Mathematical Reasoning Abilities of Neural Models". [32]

SQuAD2.0 (2023), *The Stanford Question Answering Dataset. Leaderboard*, https://rajpurkar.github.io/SQuAD-explorer/ (accessed on 21 January 2023). [45]

Storks, S., Q. Gao and J. Chai (2019), "Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches". [25]

US Department of Education, National Center for Education Statistics, Statistics Canada and OECD (2020), *Program for the International Assessment of Adult Competencies (PIAAC), Adult Literacy and Life Skills Survey (ALL) 2003-2008 and PIAAC 2012-2017 Literacy, Numeracy, and Problem Solving TRE Assessments*, https://nces.ed.gov/surveys/piaac/ideuspiaac (accessed on 31 August 2022). [11]

Wang, A. et al. (2019), "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems". [17]

Wang, A. et al. (2018), "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". [16]

Xie, Q. et al. (2017), "Large-scale Cloze Test Dataset Created by Teachers". [24]

Zellers, R. et al. (2018), "SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference". [26]

Zhang, D. et al. (2022), "The AI Index 2022 Annual Report", https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf (accessed on 20 February 2023). [1]

Annex 2.A. Supplementary tables

Annex Table 2.A.1. List of online tables for Chapter 2

Table Number	Table Title
Table A2.1	Distribution of adult population by level of literacy, IALS and PIAAC
Table A2.2	Distribution of workers by level of literacy, IALS and PIAAC
Table A2.3	Mean reading PISA score since 2000 and average 3-year trend in reading performance, by country
Table A2.4	Mean mathematics PISA score since 2003 and average 3-year trend in mathematics performance, by country

StatLink https://stat.link/cl96uw

Notes

[1] The countries or economies participating in both PIAAC and IALS comprise Australia, Canada, Chile, the Czech Republic, Denmark, England (United Kingdom), Finland, Flanders (Belgium), Germany, Ireland, Italy, the Netherlands, New Zealand, Northern Ireland (United Kingdom), Norway, Poland, Slovenia, Sweden and the United States.

[2] The countries participating in both PIAAC and ALL comprise Canada, Hungary, Italy, the Netherlands, New Zealand, Norway and the United States.

[3] The leaderboard of SQuAD 2.0 can be found under: https://rajpurkar.github.io/SQuAD-explorer/ (accessed on 21 January 2023)

[4] The leaderboard of SuperGLUE can be found under: https://super.gluebenchmark.com/leaderboard (accessed on 21 January 2023)

[5] A ranking of systems' performance on MATH can be found under: www.paperswithcode.com/sota/math-word-problem-solving-on-math (accessed on 21 February 2023).

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:

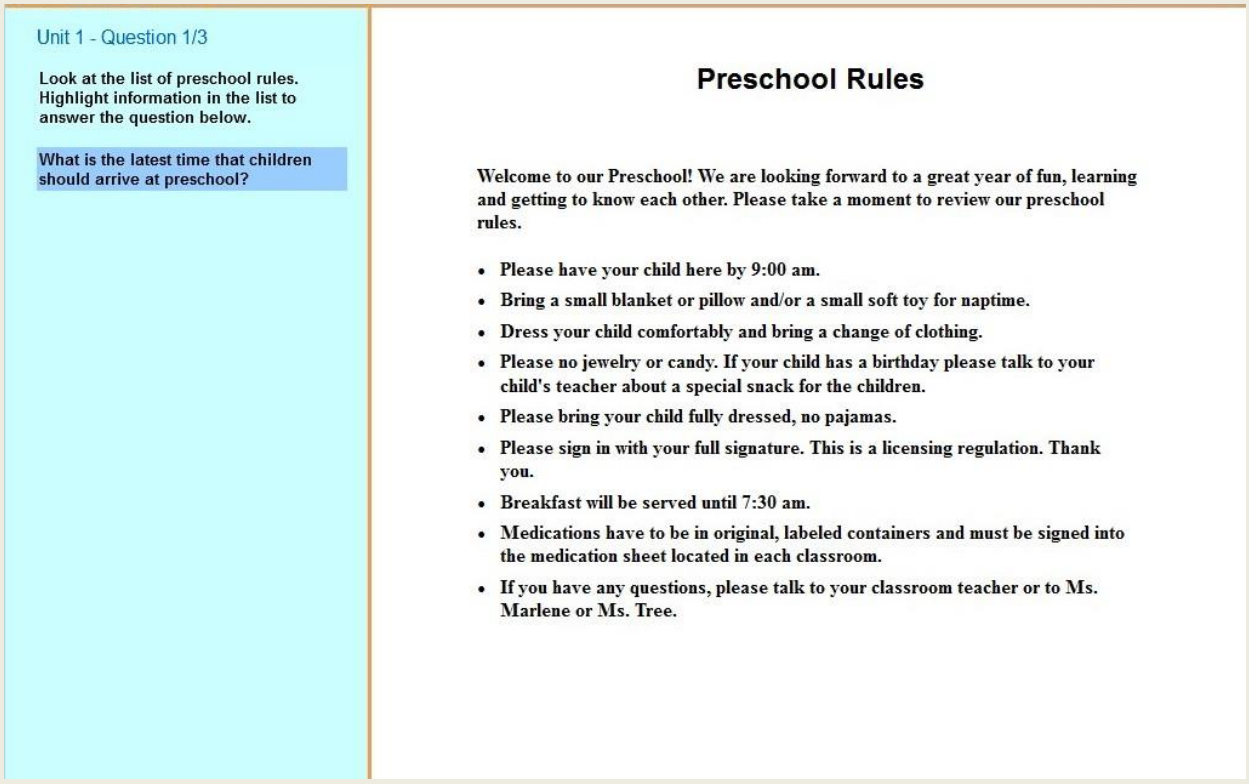
- *Access and identify*: tasks require the reader to locate items of information in a text. Sometimes this is relatively easy, as the required information is directly and plainly stated in the text. However, some tasks may require inferences and rhetorical understanding (e.g. identifying the reasons 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:

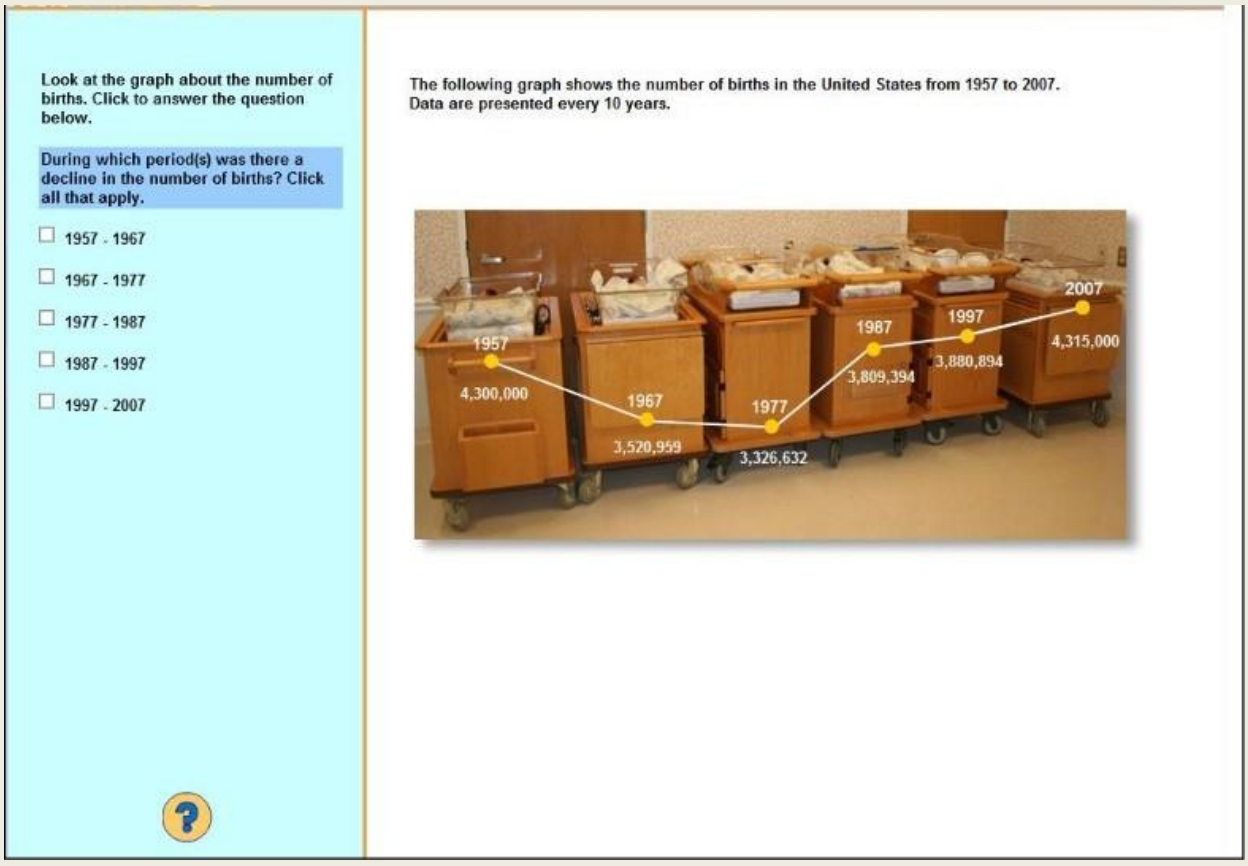
- *Identify, locate, or access mathematical information* that is present in the task and relevant to their purpose or goal.

- *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]

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:

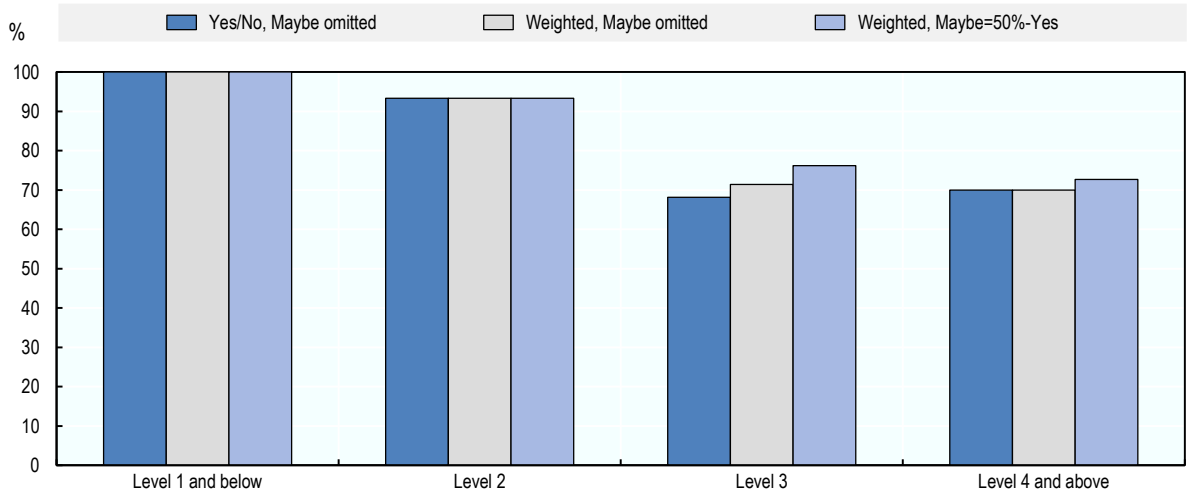
- Ratings of 0% and 25% are counted as No and answers of 75% and 100% are treated as Yes. PIAAC questions are then labelled as doable or not doable for AI according to the answer of more than half of the experts. Experts who gave Maybe- or Don't know- answers are not considered. Finally, the share of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

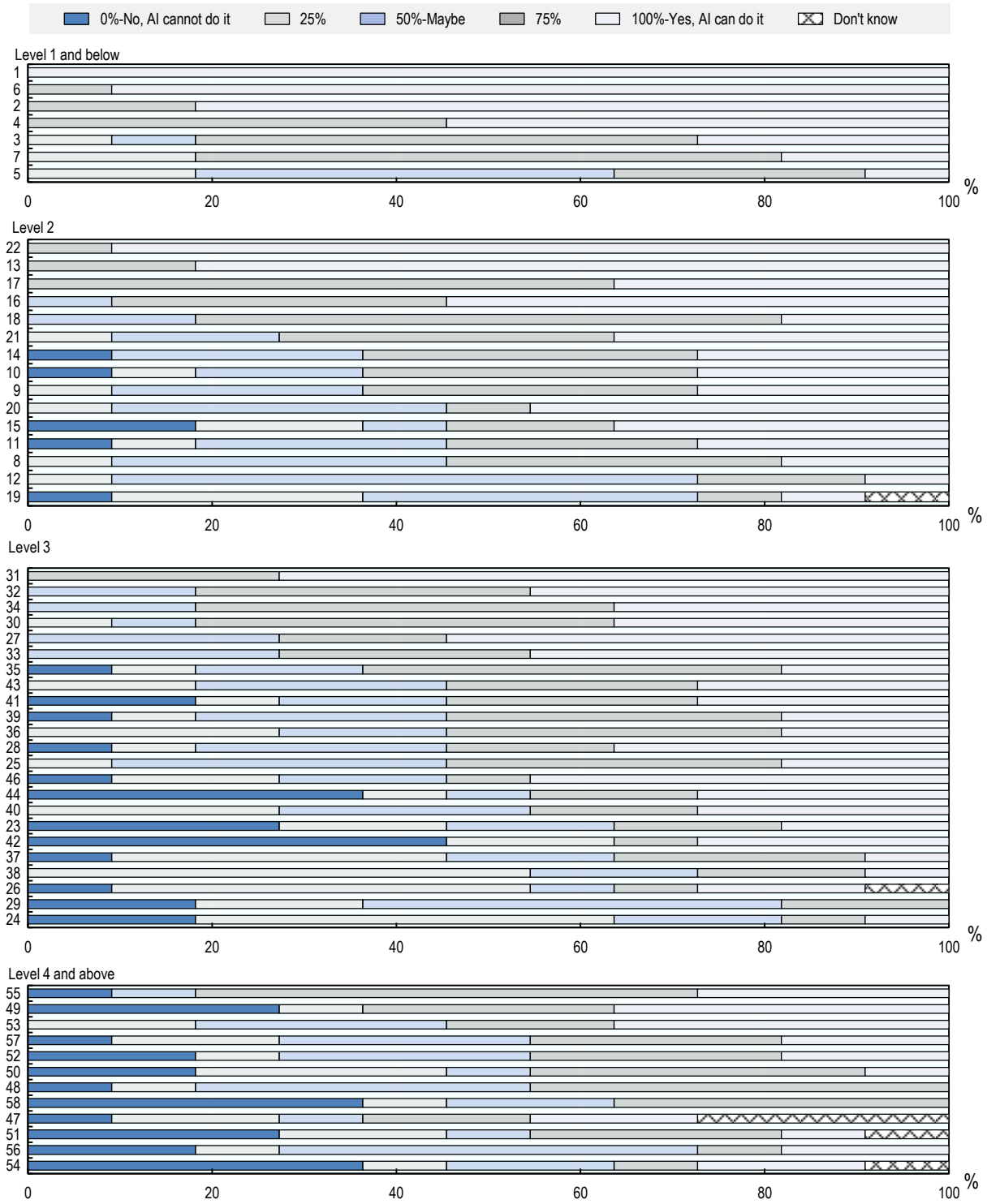


StatLink https://stat.link/lp57m8

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

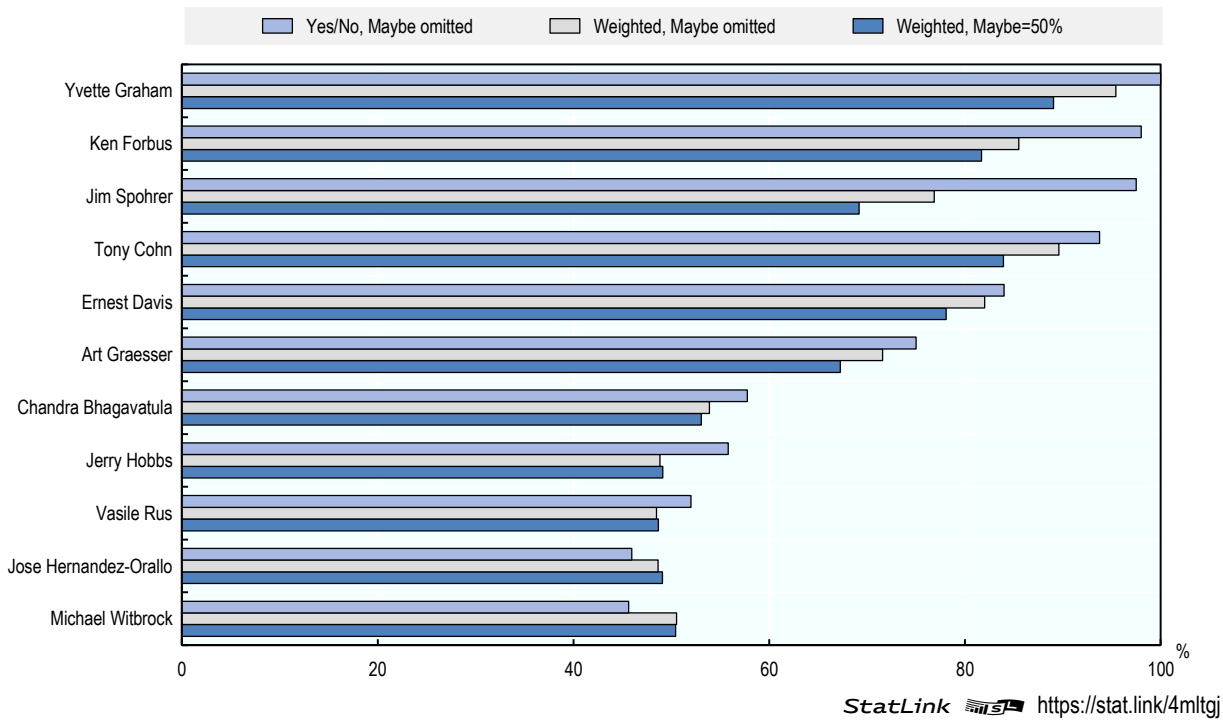


Figure 4.3 shows the average literacy ratings of experts. As in Figure 4.1, average ratings of experts are computed in three ways. First, all ratings are coded as either No (0%) or Yes (100%). In other words, less certain ratings of 25% are counted as 0% and 75%-ratings are counted as 100%. The average 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

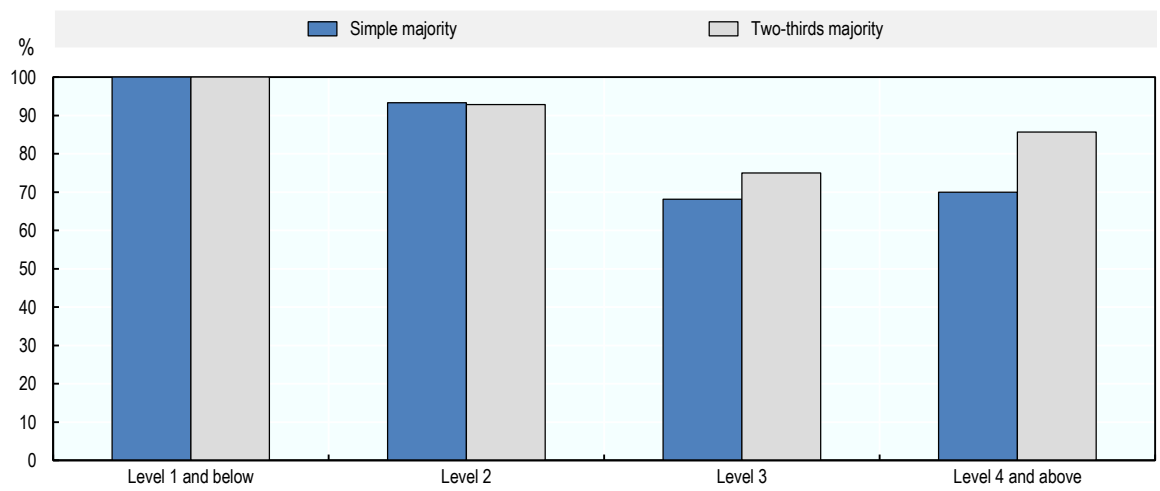
The analysis so far relied on a simple majority rule to determine whether experts rate AI as capable or incapable of answering PIAAC literacy questions correctly. However, as shown in Figure 4.2, some questions received similar shares of opposing ratings. This means that experts' agreement on these questions 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

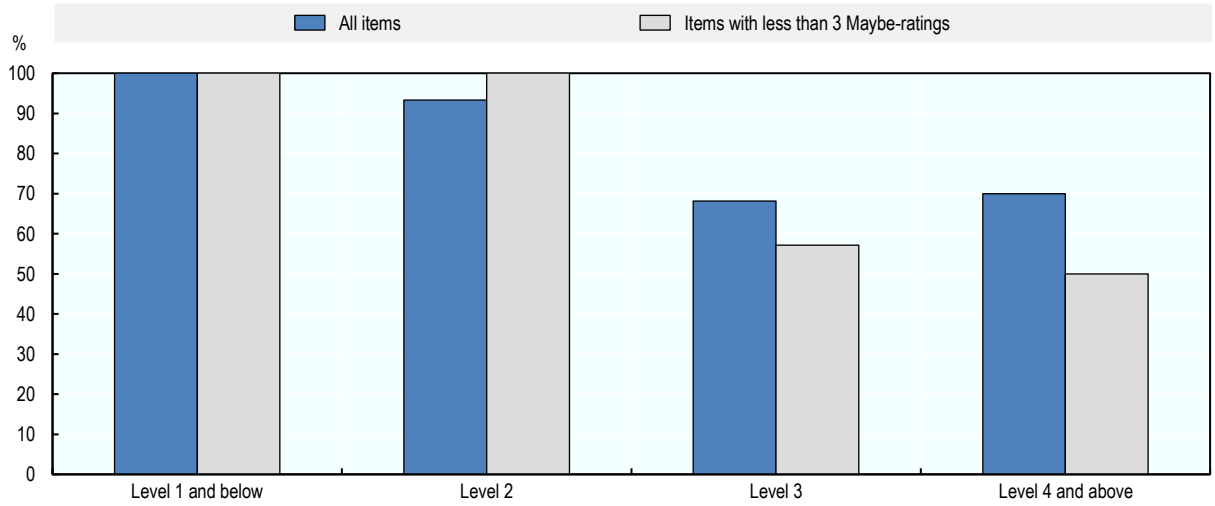
	N	Number of questions with Maybe- or Don't know-ratings:					Share of uncertain ratings
	all items	No Maybe/NA	1 Maybe/NA	2 Maybe/NA	3 Maybe/NA	4+ Maybe/NA	
Level 1 and below	7	5	1	0	0	1	8%
Level 2	15	3	2	3	3	4	22%
Level 3	23	2	2	11	6	2	20%
Level 4 and above	12	1	2	2	4	3	23%
All items	57	11	7	16	13	10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

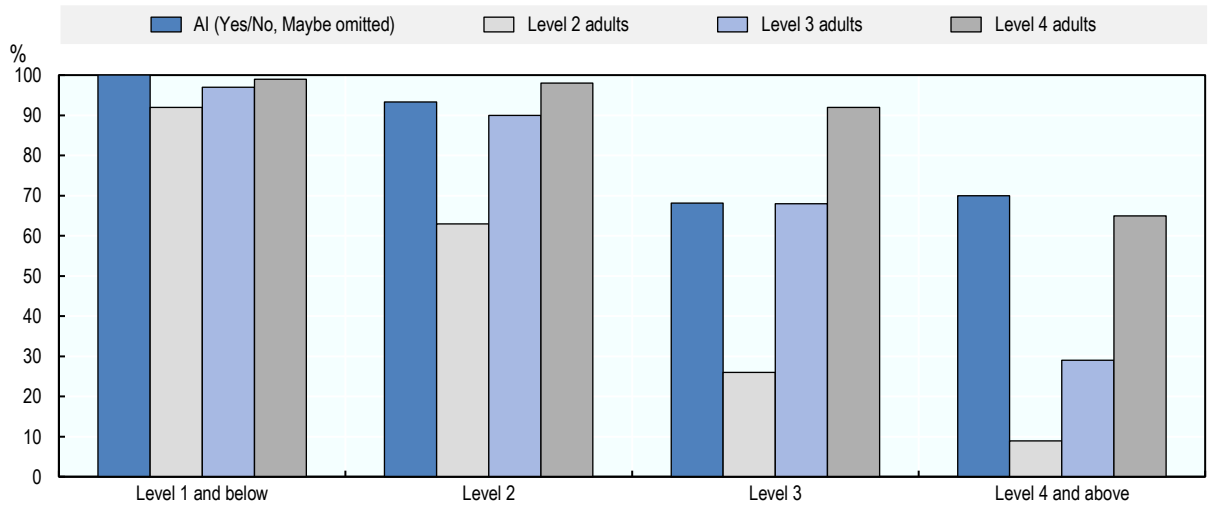
The scoring process for the Survey of Adult Skills uses item response theory to calculate difficulty scores for each question and proficiency scores for each adult. The scores for both questions and people are placed on the same 500-point scale (OECD, 2013[2]). Each adult who takes the test is placed 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

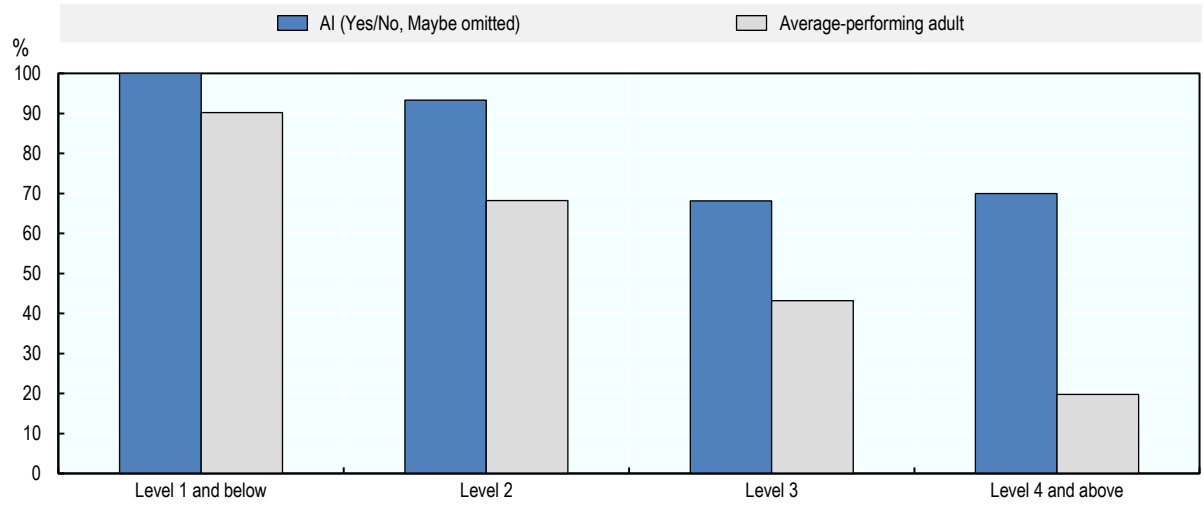


Source: OECD (2012[3]; 2015[4]; 2018[5]), *Survey of Adult Skills (PIAAC) databases*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023).

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.

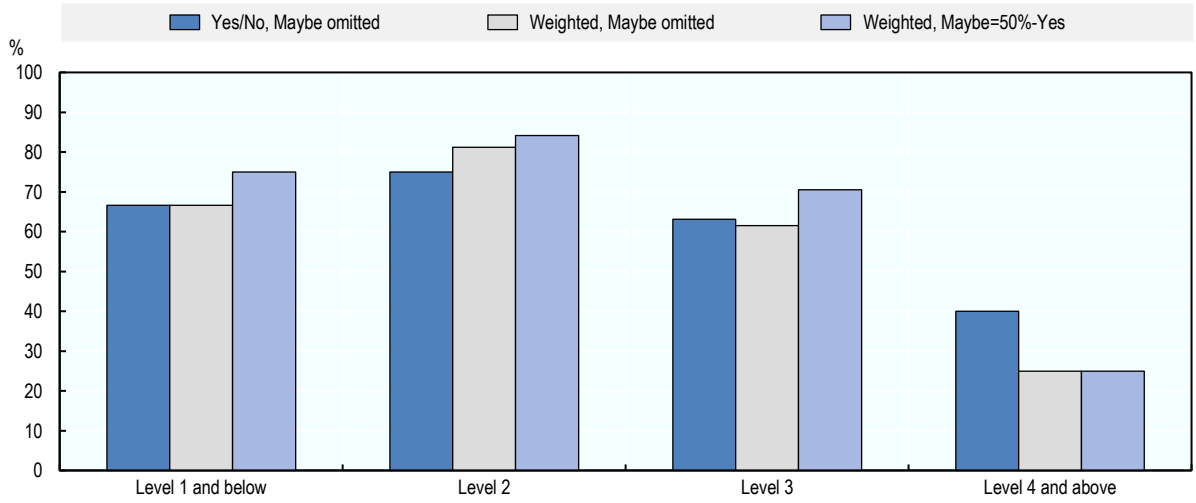
AI numeracy ratings by question difficulty

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

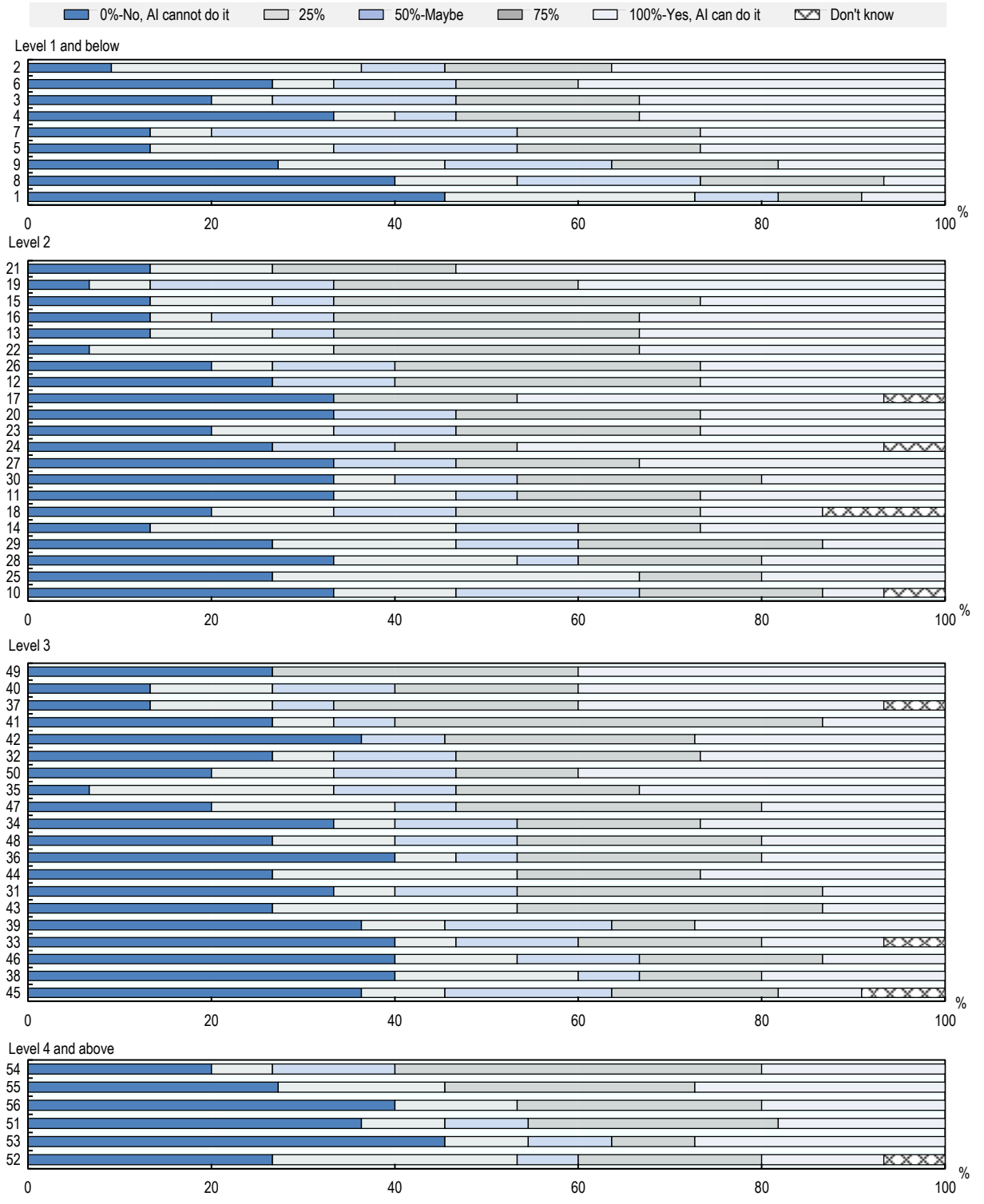
Percentage share of numeracy questions that AI can answer correctly according to the simple majority of experts



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.

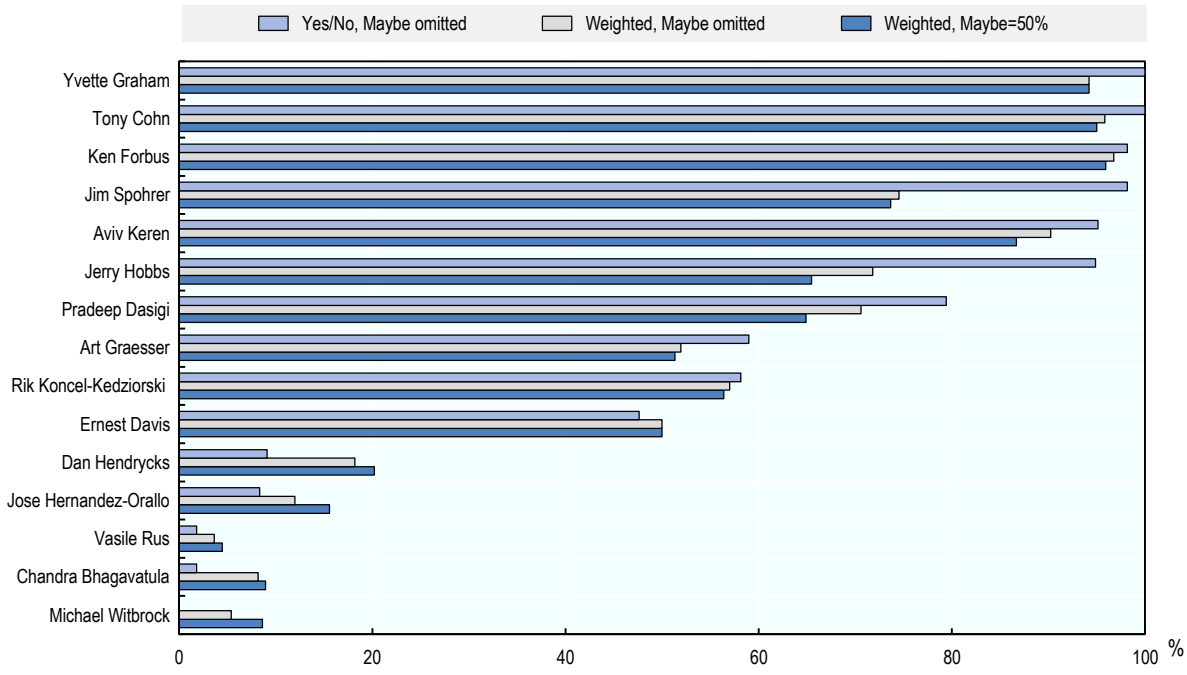
Figure 4.9 shows the distribution of ratings at individual questions by difficulty of questions. It shows that all questions receive both certain negative and certain positive ratings. The shares of these opposing evaluations are often close to each other, indicating that only thin majorities decide 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

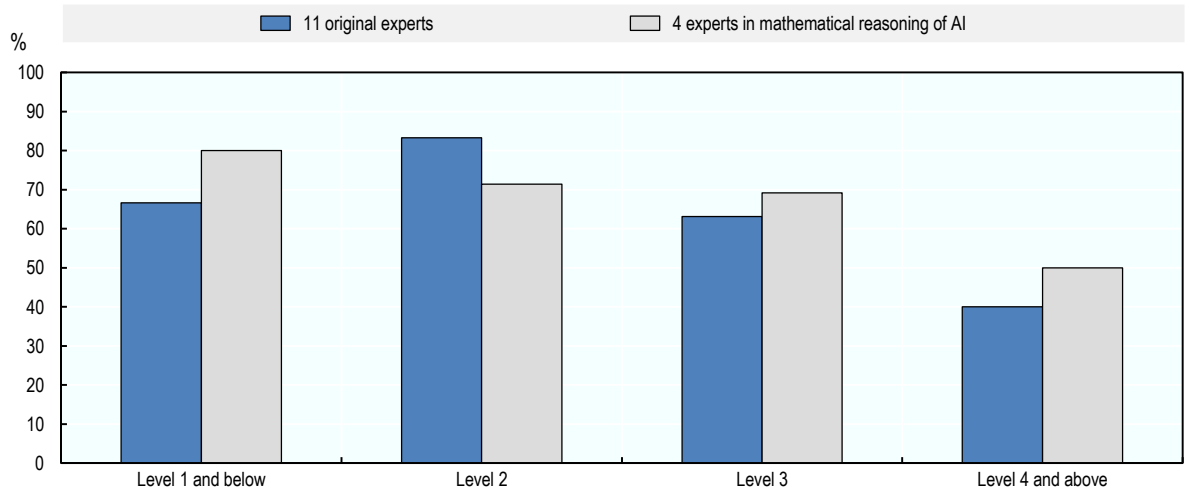


StatLink https://stat.link/j60pqe

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

	N all items	Number of questions on which agreement is reached according to the following rule:					
		Simple majority			Two-thirds majority		
Question difficulty		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.

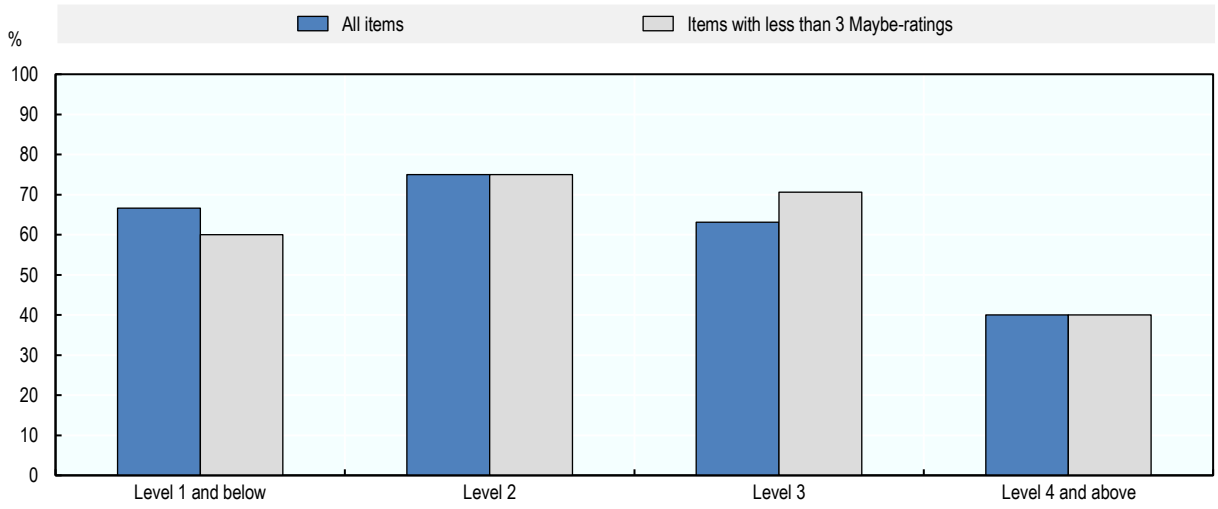
Table 4.4. Experts' uncertainty on numeracy questions

	N	Number of questions with Maybe- or Don't know-ratings:					Share of uncertain ratings
Question difficulty	all items	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.

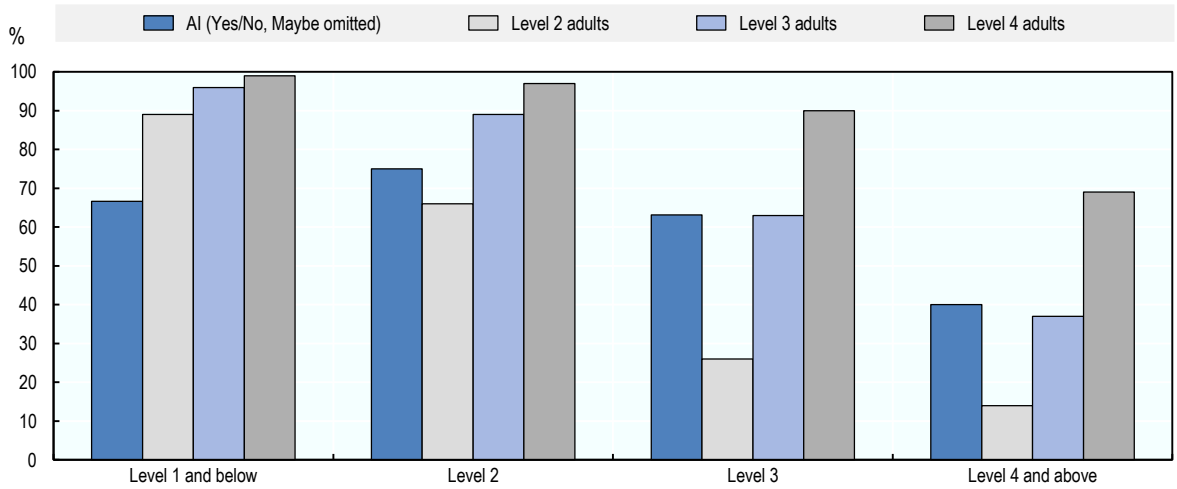
Figure 4.13 compares AI numeracy performance with the average performance of adults at proficiency levels 2, 3 and 4. The AI performance measure shows the share of questions that AI can answer correctly according to the simple majority among the 15 experts. It relies only on positive (75% and 100%) and 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

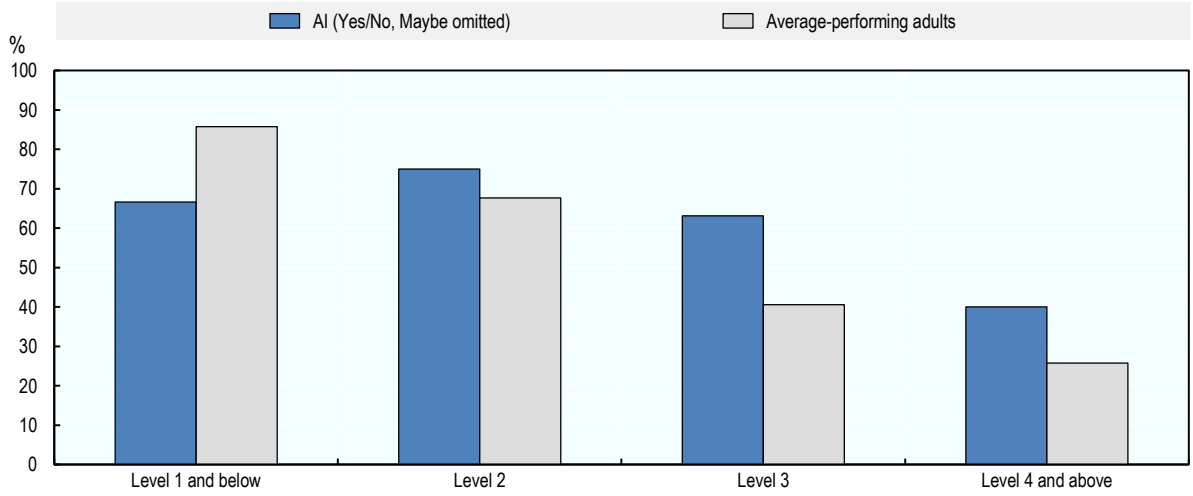


Source: OECD (2012[3]; 2015[4]; 2018[5]), *Survey of Adult Skills (PIAAC) databases*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023).

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*, https://arxiv.org/pdf/2301.09723.pdf (accessed on 28 February 2023). [10]

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*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023). [5]

OECD (2015), *Survey of Adult Skills (PIAAC) database*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023). [4]

OECD (2013), *The Survey of Adult Skills: Reader's Companion*, OECD Publishing, Paris, https://doi.org/10.1787/9789264204027-en. [2]

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. [1]

OECD (2012), *Survey of Adult Skills (PIAAC) database*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023). [3]

Radford, A. et al. (2018), *Improving Language Understanding by Generative Pre-Training*, https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 1 February 2023). [6]

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

5 Changes in AI capabilities in literacy and numeracy between 2016 and 2021

The chapter analyses changes in assessed literacy and numeracy capabilities in artificial intelligence (AI) between 2016 and 2021. To that end, it compares the majority responses of the expert groups that completed the pilot and the follow-up assessments. In addition, it looks at how the AI evaluations of experts who participated in both studies changed over the period. The chapter also studies the level of experts' agreement and the prevalence of uncertain answers in both assessments to compare the quality of group ratings obtained in 2016 and 2021. Subsequently, it analyses experts' projections of how AI capabilities will evolve by 2026 to obtain information on the likely direction of AI progress in the near future.

Tracking advances in artificial intelligence (AI) is important for anticipating the impacts of this technology on work and education. A periodical assessment of AI capabilities can provide information on the direction, pace and content of technological developments in the AI field. This knowledge base can help policy makers develop realistic scenarios about how jobs and skill demand will be redefined and how to reshape education and labour-market policies in response.

This chapter compares results of the 2016 and 2021 assessments to study how AI capabilities in literacy and numeracy developed over the period. In 2016, during a two-day workshop, 11 computer scientists rated potential AI performance on the literacy, numeracy and problem-solving tests of the Programme for the International Assessment of Adult Competencies (PIAAC). They rated whether AI could solve each test question with a Yes, Maybe or No. Three computer scientists also assessed whether AI could solve the questions in ten years' time, that is, in 2026 (Elliott, 2017[1]).

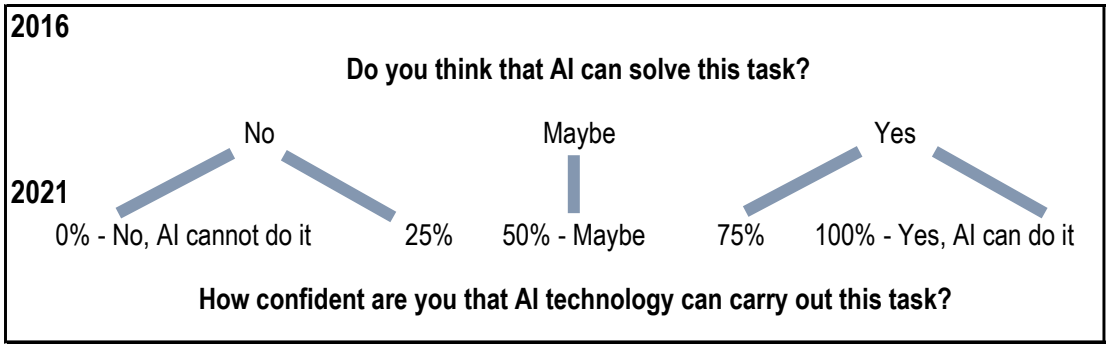
In the follow-up assessment of 2021, 11 experts, six of whom participated in the pilot, assessed AI in literacy and numeracy in an online survey and discussed the results during a four-hour workshop. They followed similar instructions for rating as in 2016. However, they used a different rating scale, ranging from 0% (confident that AI cannot solve the question) to 100% (confident that AI can solve the question). Four other experts in mathematical reasoning of AI completed the numeracy assessment with revised rating instructions. In addition, all experts predicted the evolution of AI capabilities over the next five years.

The chapter first analyses changes in reported AI literacy capabilities since 2016, as well as experts' predictions of how these capabilities will evolve by 2026. The chapter then compares AI numeracy performance ratings provided by the 11 experts in 2016 with those of the 15 experts in 2021 and presents forecasts for AI numeracy performance in the future.

Change in AI literacy capabilities over time

The follow-up assessment aimed at both providing comparability to the 2016 results and improving methods for collecting expert judgements on AI capabilities with PIAAC. Some notable improvements to the 2016 assessment include use of a facilitated discussion technique and a five-point scale to assess experts' ratings and their confidence in these ratings. The scale is presented in Figure 5.1, together with the answer categories used in 2016. While the 0%- and 100%-categories represent confident negative and positive ratings, respectively, the 25%- and the 75%-answers express lower confidence. In the following, 0%- and 25%-ratings are grouped into a single category, as are 75%- and 100%-ratings. This makes the answer-categories used in 2021 comparable to the Yes-, Maybe- and No-categories used in 2016 (see Figure 5.1).

Figure 5.1. Answer categories used in the 2016 and 2021 assessments



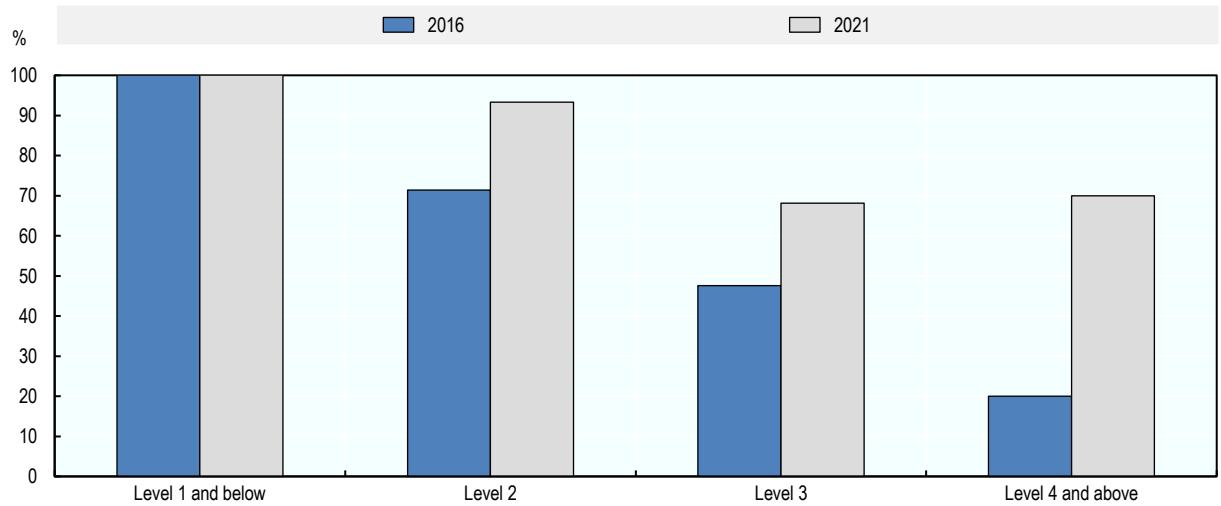
To aggregate experts' ratings into a single AI measure, the study labels each literacy question as solvable or not solvable by AI following the evaluation of most experts. Subsequently, it estimates the share of literacy questions that AI can solve at each level of question difficulty.

Change in AI literacy performance between 2016 and 2021

Figure 5.2 compares the aggregate results in literacy in the 2016 and 2021 assessments.[1] The measures rely on the majority between negative and positive ratings, omitting Maybe-ratings. The results suggest considerable improvement in AI performance in the literacy test. In 2016, AI could correctly answer 71% of Level 2 questions, 48% of Level 3 questions and 20% of questions at Level 4 and above, according to most experts. In 2021, the success rates at these difficulty levels ranged between 93% and 68%. This represents an improvement of 25 percentage points of AI on the entire PIAAC literacy test – from 55% of questions that AI can solve according to the majority of experts in 2016 to 80% in 2021.

Figure 5.2. AI literacy performance in 2016 and 2021, by question difficulty

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



Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

StatLink https://stat.link/z47gt6

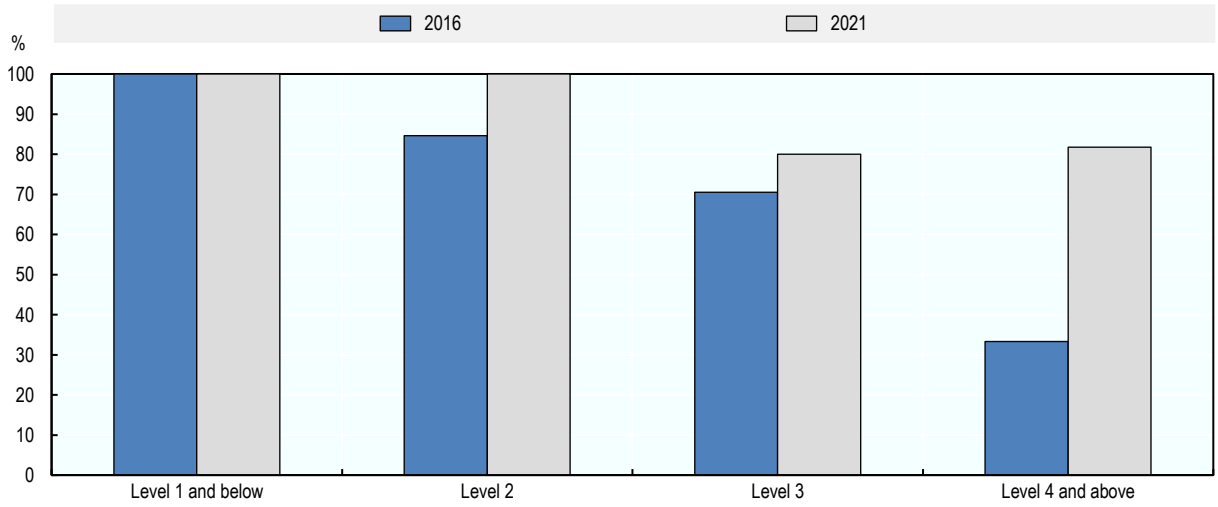
In addition, Figure 5.3 presents AI literacy measures from 2016 and 2021 that include Maybe-ratings as partial Yes-answers. Concretely, Maybe-answers are treated as Yes-votes weighted by 0.5. The vote that exceeds 50% of all ratings, including Maybe-ratings, is then used to determine AI's success on each question. AI literacy scores for both years are higher when counting Maybe as a half Yes. However, the overall picture remains similar. In 2021, AI literacy performance exceeded performance assessed in 2016 at question difficulty Level 2 and higher.

These results reflect progress in the field of natural language processing (NLP) since 2016. As described in Chapter 2, NLP has seen tremendous advances since the introduction of large pre-trained language models in 2018. These include ELMo (Embeddings from Language Models) (Peters et al., 2018[2]), GPT (Generative Pre-Trained Transformer) (Radford et al., 2018[3]), and BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018[4]). These models are trained on unprecedented amounts of data and are featured in systems for specific language tasks. Their introduction has pushed

the state of the art of NLP forward, as measured by various benchmarks and tests for evaluating systems' performance (see Chapter 2).

Figure 5.3. AI literacy performance in 2016 and 2021, counting Maybe as 50%-Yes

Percentage share of literacy questions that AI can answer correctly according to a simple majority of experts, by question difficulty



Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

StatLink https://stat.link/tjqrny

Change in AI literacy performance according to experts participating in both assessments

Comparisons of AI performance ratings over time may be biased by changes in the composition of the expert groups providing the ratings. If expert groups differ in the mix of optimists and pessimists, composition of expertise, or with regard to other characteristics relevant for the assessment, these differences would be reflected in the aggregate AI ratings and wrongly attributed to differences in AI capabilities between both time points. Information on such potential confounding factors is not available. However, an analysis of how the six experts who participated in both assessments changed their ratings over time would account for much of the potential bias related to the use of different expert groups across time points.

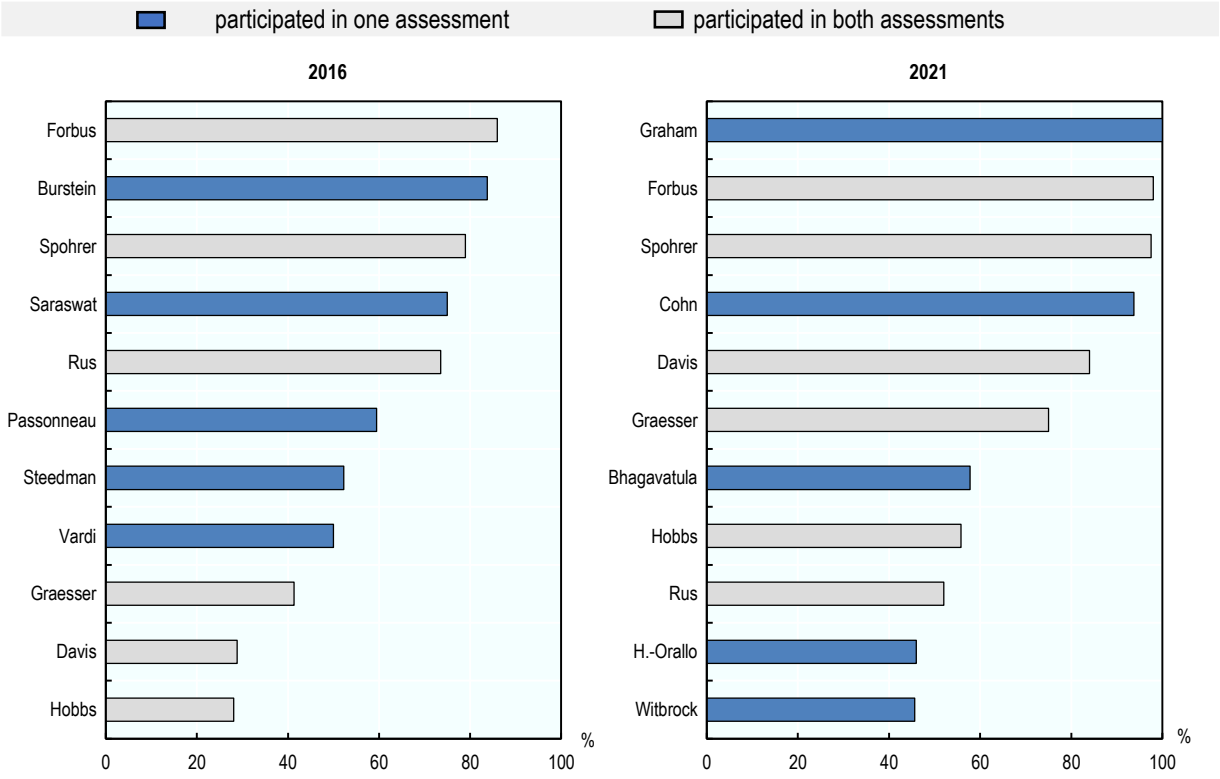
First, Figure 5.4 provides an overview of the average literacy ratings of all experts in the 2016 and 2021 assessments. Overall, the distributions of experts' average ratings are similar across years, with ratings in 2021 being, on average, higher than in 2016. A two-sample independent t-test indicates a significant increase in experts' average ratings ($t(20) = 1.5$; $p = 0.08$).

The blue bars in Figure 5.4 represent the average ratings of the experts who participate in only one assessment. The five experts who participated in the pilot, but did not continue in the follow-up, had middle to high average ratings in 2016 (Moshe Vardi, Mark Steedman, Rebecca Passonneau, Vijay Saraswat and Jill Burstein). Among the new experts in 2021, two had the lowest ratings (Michael Witbrock and José Hernández-Orallo), another two had the highest ratings (Yvette Graham and Antony Cohn) and one expert had a medium overall rating (Chandra Bhagavatula) in their expert group.

Bars in grey show the averages of the six experts who took part in both assessments – Ernest Davis, Ken Forbus, Art Graesser, Jerry Hobbs, Vasile Rus and Jim Spohrer. Their mean ratings represent different opinions on AI in both the 2016 and the 2021 assessments.

Figure 5.4. Average expert ratings in literacy in 2016 and 2021

Averages of Yes and No-answers, Maybe omitted



Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

StatLink https://stat.link/yhig9e

All but one of these experts – Vasile Rus – rated potential AI performance on the literacy test higher in 2021 than in 2016. A paired t-test shows that the "within-person" increase in ratings is significant ($t(5) = 2$; $p = 0.05$).

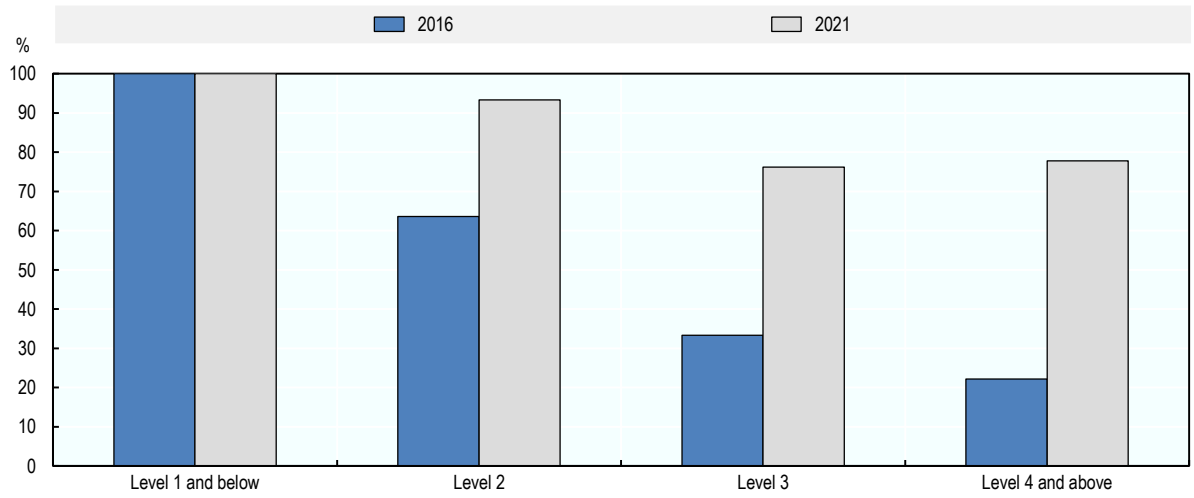
During the discussion, Rus made clear that his decrease in ratings is due to how he interprets the rating exercise rather than to how he evaluates the state of the art in NLP. Concretely, he assumed a more general system for literacy in 2021 that should perform literacy tasks as flexibly as humans. These higher expectations towards the system to be rated explain his lower ratings in 2021.

Figure 5.5 shows the aggregate AI literacy ratings based only on the ratings of the six experts who participated in both assessments. The results are in accordance with those obtained from the full expert groups, showing considerable increase in AI literacy performance in PIAAC over time. Compared to the full groups, the six experts provide somewhat more positive evaluations for 2021 and somewhat more negative ones for 2016 (see Figure 5.2). For 2016, the AI literacy performance estimated from their ratings is at 63% at Level 2, 33% at Level 3 and 22% at Level 4 and above. For 2021, these AI ratings are 93%, 76% and 78%, respectively. Overall, AI performance in the entire literacy test increased by 37 percentage

points following these six experts' judgements – from an estimated success rate at 48% in 2016 to a performance at 85% in 2021.

Figure 5.5. AI literacy performance in 2016 and 2021 according to experts who participated in both assessments

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



Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

StatLink https://stat.link/435v6j

Comparison of experts' agreement and confidence across literacy assessments

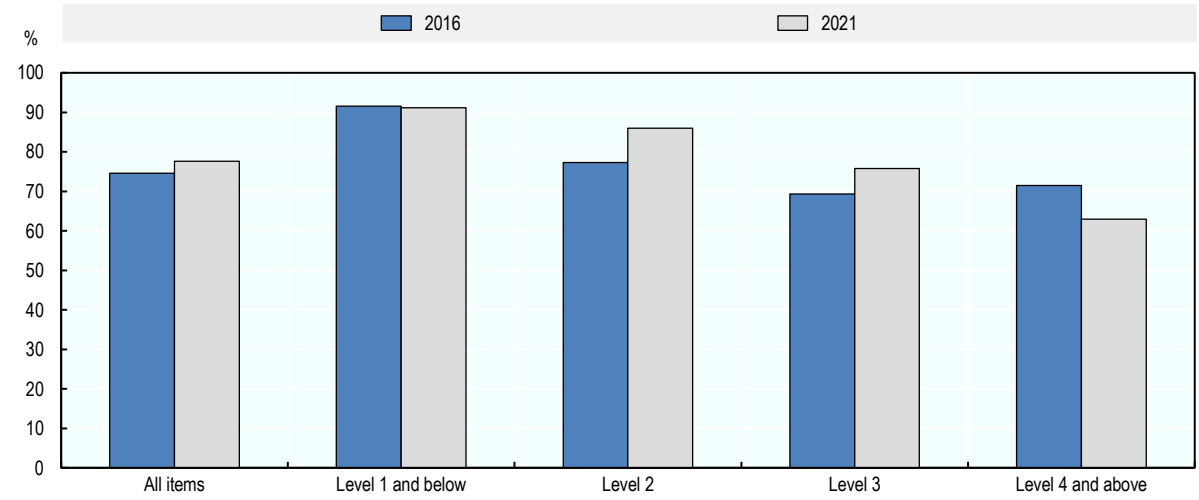
A valid comparison of AI capabilities across time requires sufficient agreement and certainty among experts regarding the state of the art in AI in each time point. This section looks at the level of agreement and the prevalence of uncertain answers in both literacy assessments to compare the quality of group ratings obtained in 2016 and 2021.

Figure 5.6 shows the average size of the majorities reached on the literacy questions of PIAAC in the pilot and follow-up assessments. For example, if an assessment includes ten experts rating two questions, A and B, and if A received six Yes and four No and B got two Yes and eight No, the average majority size would be 70%. This is the mean of the 60% majority reached on A and the 80% majority reached on B. This average is indicative for the overall level of agreement among the experts.

In 2021, the average size of the majorities reached across all literacy questions was 78%, close to the average majority size of 75% achieved in 2016. In both assessments, average agreement was highest at the easiest questions and decreased gradually with question difficulty. On average, across questions, majorities in 2021 are bigger than majorities in 2016 at Level 2 and Level 3 and smaller at Level 4 and above.

Figure 5.6. Average majority size in rating literacy questions in 2016 and 2021

Average size of majorities on literacy questions by question difficulty, Maybe omitted

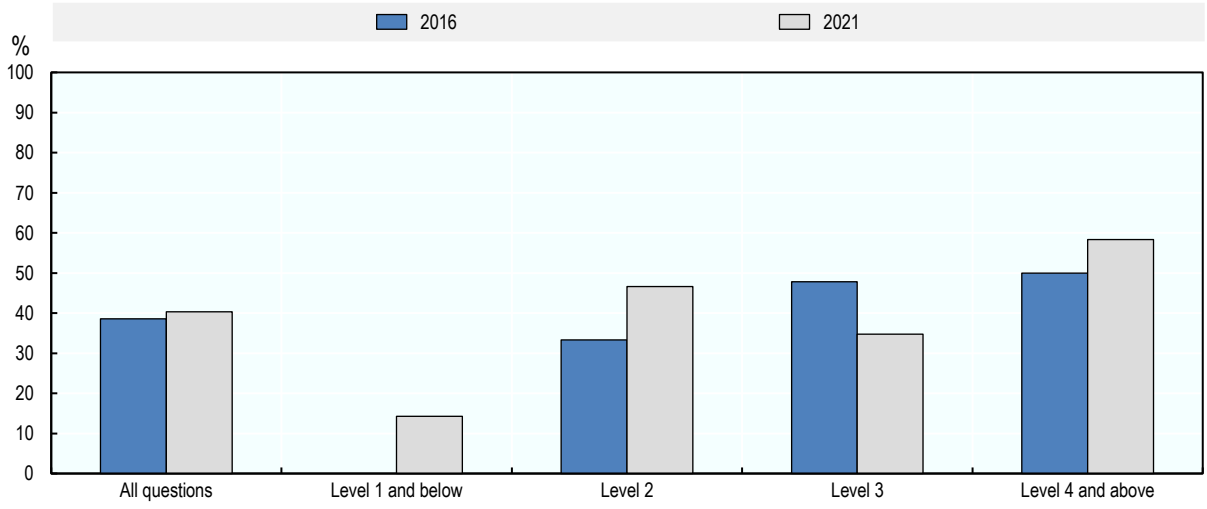


Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

StatLink https://stat.link/ptagik

Figure 5.7. Share of literacy questions that receive three or more uncertain ratings in 2016 and 2021

Share of questions with three or more Maybe- or Don't know-ratings, by question difficulty



Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

StatLink https://stat.link/ky0enx

Although there is high agreement among those with certain ratings, some experts provide uncertain answers to many literacy questions. Figure 5.7 shows the shares of literacy questions at different difficulty levels that receive three or more Maybe- or Don't know-ratings. In both 2016 and 2021, the shares of

questions with uncertainty increase with question difficulty. Uncertainty is similar across assessments, with approximately 40% of all literacy questions in 2016 and 2021 receiving three or more uncertain ratings.

Projections of AI literacy performance for 2026

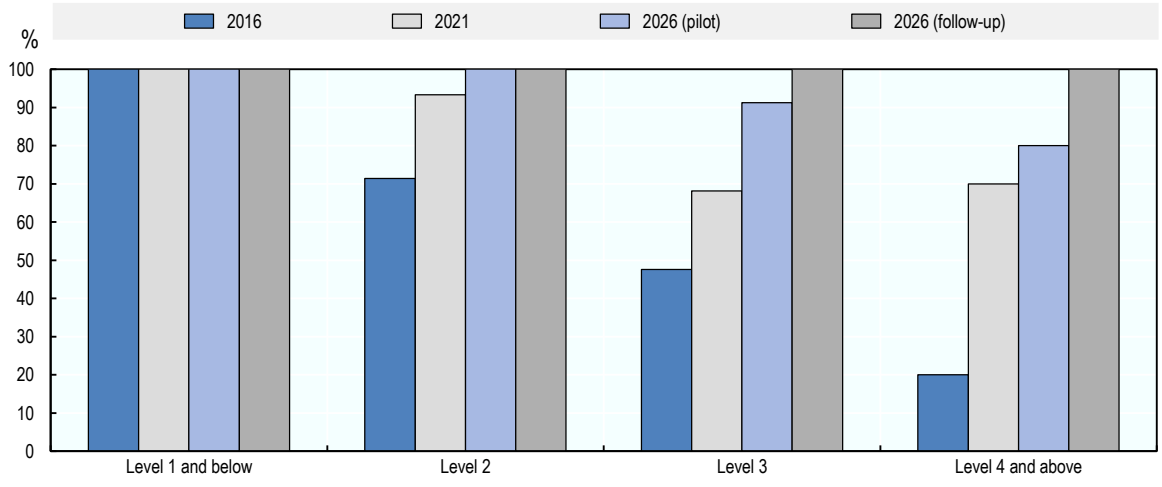
A repeated assessment of AI capabilities provides a sense for the direction and pace of the development of this technology. Another way to obtain information on the progress of AI is to ask experts to predict its capabilities in future. The pilot study asked three experts – Ernest Davis, Ken Forbus and Art Graesser – to rate the potential performance of AI on the literacy questions for 2026. In the follow-up assessment, all 11 experts provided predictions for the same year. The results are shown in Figure 5.8.

According to the majority of experts, computers will perform much better in literacy by 2026. The more recent projections are more optimistic than those made in 2016. Experts in the follow-up assessment expected AI to be able to perform all literacy questions in 2026. The predictions of the three experts in the pilot study suggest an AI performance of 91% at Level 3 and 80% at Level 4 and above for the same year.

Projections over a shorter time horizon are more likely to be precise given the rapid rate of change in AI technology. Moreover, experts pointed out they often provide predictions over three to five years when applying for research grants. Thus, they are more used to thinking of AI progress in terms of shorter time frames.

Figure 5.8. Projected AI literacy performance for 2026, by question difficulty

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



Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

StatLink https://stat.link/crlngt

Change in AI numeracy capabilities over time

The follow-up study assessed AI in numeracy somewhat differently than the pilot study. The pilot study asked 11 experts to rate AI on each of the numeracy questions in PIAAC as part of a two-day assessment workshop (Elliott, 2017[1]). Conversely, the follow-up study collected the judgements of 15 experts in two

assessment rounds. In the first round, 11 experts, 6 of whom took part in the pilot study, rated current AI performance, as well as expected performance for 2026, with regard to each numeracy question. They received instructions similar to those used in the pilot study. In the second round, four experts in mathematical reasoning of AI received more information on the PIAAC test and were asked to conceptualise a single system for addressing all numeracy questions. They subsequently provided ratings of current techniques on each numeracy question. In addition, they provided a single rating on whether AI could carry out the entire numeracy test in five years' time.

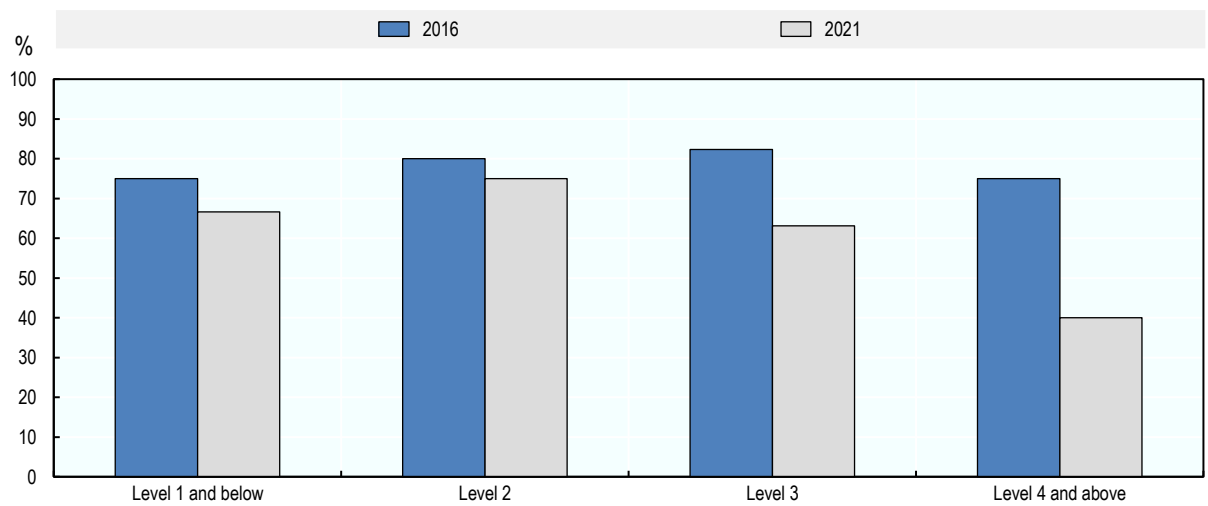
This section compares ratings of current AI capabilities obtained from the 11 experts in 2016 with the aggregate ratings of the 15 experts who participated in the follow-up study. It then presents projection ratings for 2026 by viewing projections in the pilot study, the first and the second round of the follow-up study separately. As in the literacy analyses, experts' answers from the follow-up study are grouped into the three categories of No (0% and 25%), Maybe (50%) and Yes (75% and 100%) to provide comparability to the pilot assessment (see Figure 5.1).

Change in AI numeracy performance between 2016 and 2021

Figure 5.9 compares AI numeracy ratings from the pilot and follow-up assessments, using measures based on Yes- and No-ratings only.[2] The figure shows a decline in assessed AI performance in numeracy at all levels of question difficulty. The decline is smaller at Level 1 and below and Level 2 of question difficulty. It amounts to 19 and 35 percentage points at Level 3 and Level 4 and above, respectively. AI performance on the entire numeracy test has decreased by 14 percentage points between the assessments, according to experts' majority opinion.

Figure 5.9. AI numeracy performance in 2016 and 2021, by question difficulty

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



Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

StatLink https://stat.link/v1umpg

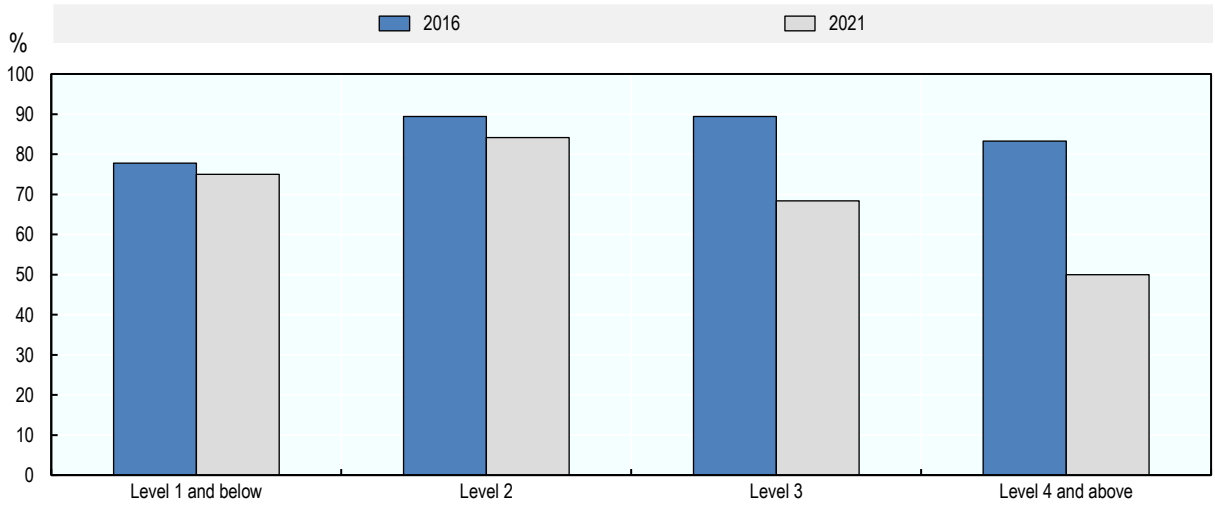
Figure 5.10 provides analogous results based on measures that include Maybe-answers as partial Yes-answers. When adding the Maybe-answers to the Yes-votes, the share of numeracy questions that receive

a majority of positive answers increases at all difficulty levels. However, the gap between AI numeracy performance assessed in 2016 and 2021 remains: performance in 2021 is rated lower than in 2016, particularly at the higher levels of questions difficulty.

Differences in how experts interpret the ratings exercise may drive these counterintuitive findings. As described in Chapter 4, the follow-up assessment instructed experts to rate the capacity of one hypothetical system to solve the PIAAC numeracy test. The discussion showed this led some experts to assume a general AI system for numeracy that should solve diverse mathematical problems similarly to humans. These experts provided more negative ratings, given that current technology is not yet at this stage of generality.

Figure 5.10. AI numeracy performance in 2016 and 2021, counting Maybe as 50%-Yes

Percentage share of numeracy questions that AI can answer correctly according to a simple majority of experts, by question difficulty



Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

StatLink https://stat.link/16hvo4

Change in AI numeracy performance according to experts participating in both assessments

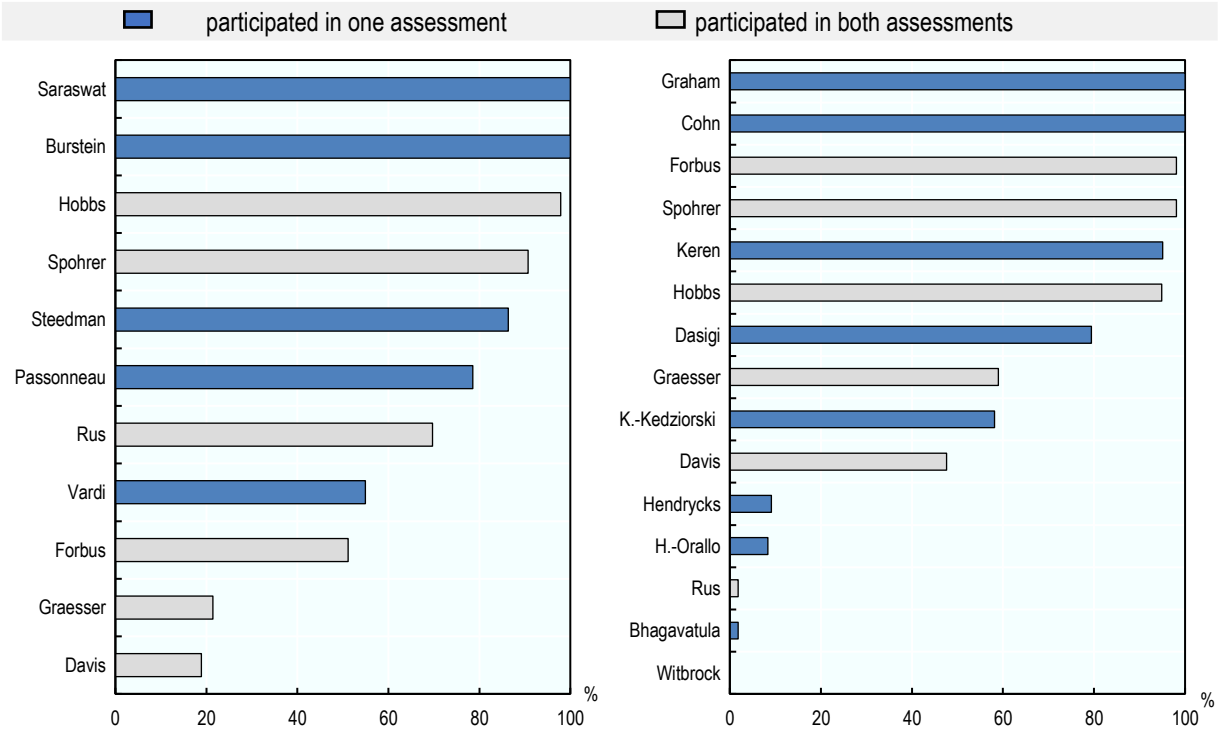
Another explanation for the implausible decline in numeracy may relate to differences in the composition of experts in both assessments. For example, more pessimistic experts or experts with somehow different expertise may have joined the follow-up study. To account for such potential bias, the following analyses draw on ratings only from those six experts who participated in both assessments. The opinions of these experts may not fully represent the relevant expertise in the field. However, a within-person comparison should provide a better sense of the direction of change in AI since it eliminates potential confounding factors related to the use of different expert groups in both assessments.

Figure 5.11 presents the individual average ratings of all experts who completed the pilot and the follow - up assessments. It shows that experts' average ratings in 2021 are more variable than average ratings in 2016. Ratings in 2021 are also, on average, lower than the ratings in 2016. However, a two- sample independent t-test shows this difference is not significant ($t(24) = 0.89$, $p = 0.19$).

A look at the six experts who participated in both assessments (bars in grey) shows that four of them rated AI performance in numeracy higher in 2021 than in 2016. These experts were Ken Forbus (51% in 2016 and 98% in 2021), Jim Spohrer (91% and 98%), Art Graesser (21% and 59%) and Ernest Davis (19% and 48%). One expert – Jerry Hobbs – had a high average rating of AI numeracy capabilities in both years (98% in 2016 and 95% in 2021). Another expert – Vasile Rus – rated AI low on almost all numeracy questions in the 2021 assessment (2%), which marked a sharp decline to his evaluation in 2016 (70%). He explained the decline by the higher degree of generality assumed for the hypothetical system to be rated in 2021 (see also above). A paired t-test shows that this "within-person" difference in ratings is not significant ($t(5) = 0.49$; $p = 0.65$).

Figure 5.11. Average expert ratings in numeracy in 2016 and 2021

Averages of Yes and No-answers, Maybe omitted



Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

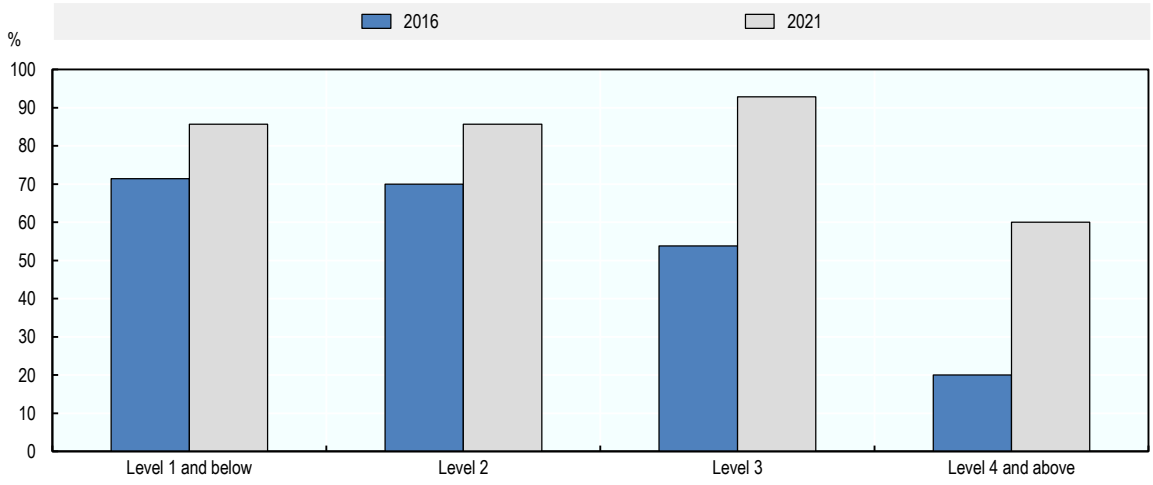
StatLink https://stat.link/cgas53

Figure 5.12 presents aggregate AI numeracy measures for 2016 and 2021 based on the ratings of the six experts who completed both assessments. In contrast to the ratings of the full expert groups, those of the six experts suggest an increase in potential AI performance on the numeracy test between 2016 and 2021. According to the majority opinion of the six experts, AI was expected to complete around 70% of questions at Level 2 and below, 54% of Level 3 questions and 20% of questions at Level 4 and above. For 2021, the corresponding performance ratings were assessed at 86%, 93% and 60%, respectively.

These results show that changes in AI numeracy capabilities over time are hard to define since findings are not robust to different specifications of the expert groups judging these capabilities.

Figure 5.12. AI numeracy performance in 2016 and 2021 according to experts who participated in both assessments

Percentage share of numeracy questions that AI can answer correctly according to a simple majority of experts, by question difficulty; measures use Yes/No-ratings, Maybe omitted.



Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

StatLink https://stat.link/ze5w7v

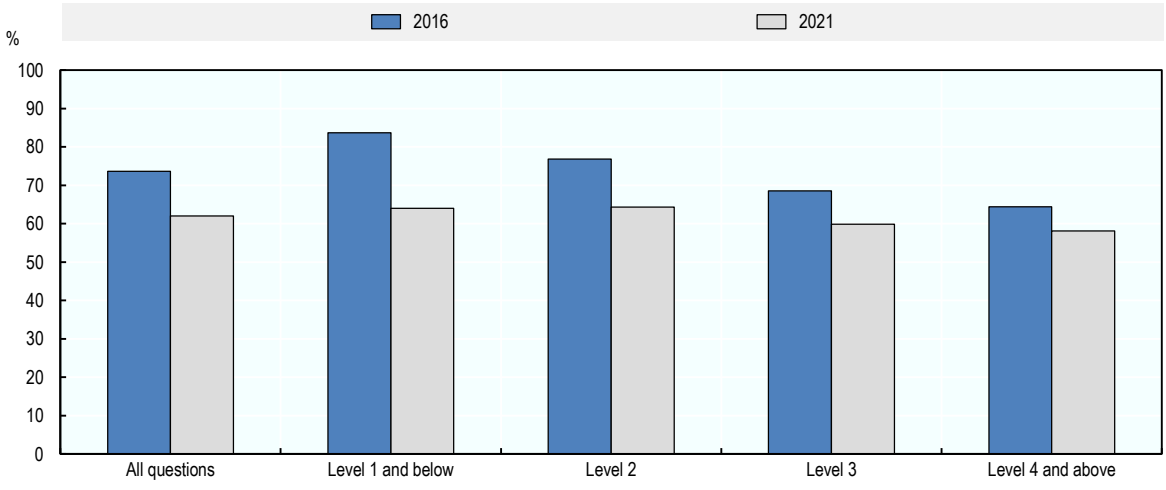
Comparison of experts' agreement and confidence across numeracy assessments

As shown in Chapter 4, experts in the 2021 assessment disagreed about AI performance in numeracy. Two opposing groups emerged of five experts who evaluated AI negatively on almost all numeracy questions and four who provided mainly positive ratings (see Figure 5.11). This hindered consensus building in the quantitative evaluation of AI. In the following, consensus and experts' confidence in numeracy ratings in the two assessments are compared. This can show whether disagreement is specific to the follow-up assessment or characteristic of the assessment of numeracy capabilities altogether.

Figure 5.13 shows the average size of the majorities reached on the numeracy questions of PIAAC in both assessments. In the follow-up, the majority opinion regarding AI numeracy capabilities included, on average across all questions, 62% of the experts who provided a positive or negative evaluation. This majority share is similar at different levels of question difficulty. By contrast, experts' agreement in numeracy was higher in the pilot study. Across all questions on average, 74% of the experts with ratings different than "Maybe" or "Don't know" formed the majority. This share is highest at Level 1 and below, at 84%, and decreases gradually to 64% at Level 4 and above.

Figure 5.13. Average majority size in rating numeracy questions in 2016 and 2021

Average size of majorities on numeracy questions by question difficulty; Maybe omitted

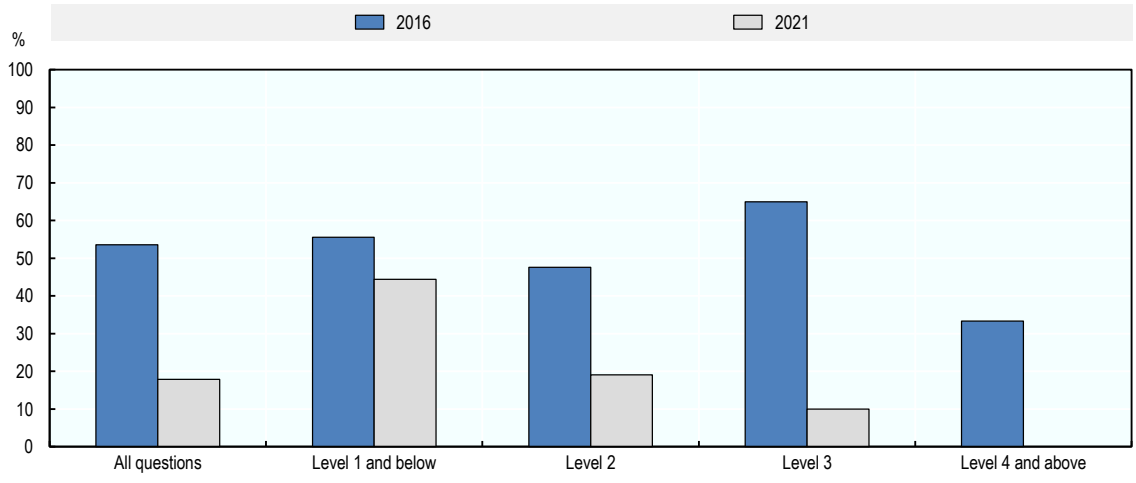


Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

StatLink https://stat.link/b3cak0

Figure 5.14. Share of numeracy questions that receive three or more uncertain ratings in 2016 and 2021

Share of questions with three or more Maybe- or Don't know-ratings, by question difficulty



Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

StatLink https://stat.link/uh2kqs

In addition, Figure 5.14 provides information on the prevalence of uncertain ratings in the pilot and follow-up assessments. It shows, that, in 2016, around half of the numeracy questions received three or more uncertain answers. In the follow-up assessment, uncertainty was lower, with only 18% of all questions

having at least three Maybe- or Don't know-answers. Uncertainty varied with question difficulty. In 2016, the share of questions with uncertainty was bigger at the first three levels of question difficulty and smaller at Level 4 and above. In the follow-up assessment, uncertainty was highest at the easiest questions and lowest at the hardest questions.

Overall, the numeracy assessment in 2016 is characterised by higher agreement – similar to the one achieved in the literacy domain in the same year – and low certainty in ratings. By contrast, experts in the follow-up assessment had more opposing views on AI numeracy capabilities but expressed more certainty in their evaluations. As described in Chapter 4, disagreement in the follow-up study had different reasons. Among others, these included varying interpretations of the rating instructions and differing assumptions about the systems to be rated. This may have driven the decline in experts' agreement compared to 2016.

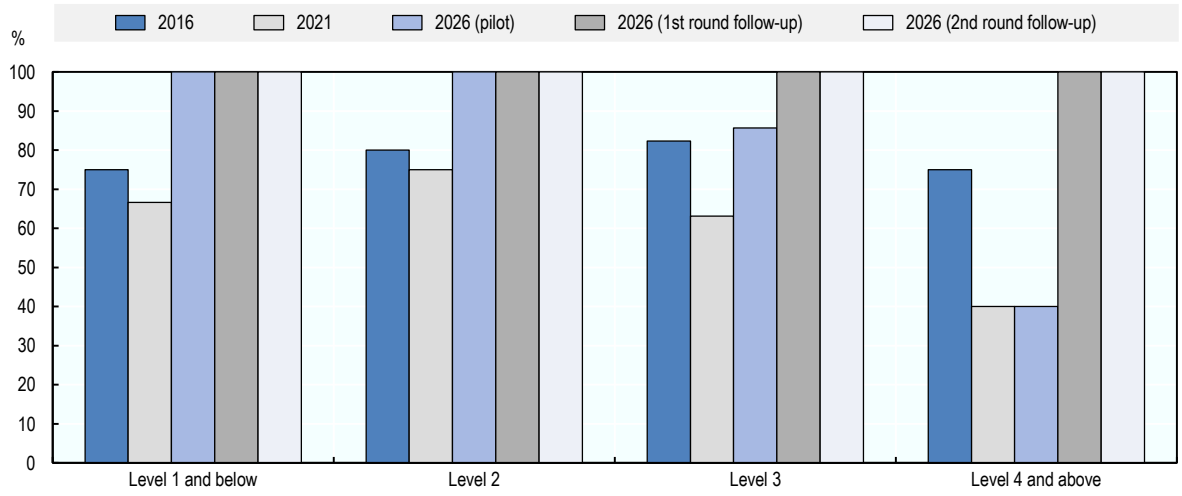
An alternative explanation may have to do with technological developments related to numeracy capabilities. In 2016, large language models have not yet been widely applied for mathematical reasoning. The limited information on AI's capabilities related to quantitative reasoning may have caused uncertainty in the pilot assessment, but also consensus, as it is easier to agree on issues for which information is concise. Since 2016, the research field has expanded. The availability of more information relevant for the numeracy assessment may have decreased uncertainty in 2021. However, it may have increased disagreement as it is harder to agree on issues for which there is much and novel information.

Projections of AI numeracy performance for 2026

Figure 5.15 compares the numeracy ratings for 2016 and 2021 with experts' projections for 2026. This can provide a sense of the likely development of AI numeracy capabilities.

Figure 5.15. Projected AI numeracy performance for 2026, by question difficulty

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



Source: Adapted from Elliott, S. (2017[1]), *Computers and the Future of Skill Demand*, https://doi.org/10.1787/9789264284395-en.

StatLink https://stat.link/9ibetj

Three experts in 2016 rated potential AI performance on the numeracy test in 2026 – Ernest Davis, Ken Forbus and Art Graesser. They were most sceptical with regard to AI's numeracy capabilities in 2016, as

shown in Figure 5.11. Similarly, they were doubtful about AI's performance in ten years' time, providing majority ratings of 86% at Level 3 and 40% at Level 4 and above. The latter is below the full group's rating of current AI numeracy performance for 2016.

All 11 experts who first rated AI in numeracy in the follow-up study provided projections for each of the test questions. Their majority rating suggests 100% successful AI performance in numeracy in 2026. The four experts in mathematical reasoning who completed the second assessment round provided only one rating for future AI performance. These ratings also indicate a 100% success rate of AI in numeracy in 2026.

References

Devlin, J. et al. (2018), "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". [4]

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]

Peters, M. et al. (2018), "Deep contextualized word representations". [2]

Radford, A. et al. (2018), *Improving Language Understanding by Generative Pre-Training*, https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 1 February 2023). [3]

Annex 5.A. Supplementary figures

Annex Table 5.A.1. List of online figures for Chapter 5

Table Number	Table Title
Figure A5.1	Comparison of average and majority expert opinions on AI capabilities in literacy in 2016 and 2021
Figure A5.2	Comparison of average and majority expert opinions on AI capabilities in literacy in 2016 and 2021, counting Maybe as 50%
Figure A5.3	Comparison of average and majority expert opinions on AI capabilities in numeracy in 2016 and 2021
Figure A5.4	Comparison of average and majority expert opinions on AI capabilities in numeracy in 2016 and 2021, counting Maybe as 50%

StatLink https://stat.link/uanq7b

Notes

[1] The results for 2016 reported in this study differ from those in Elliott (2017[1]) because the studies use different approaches to aggregate experts' ratings. The pilot study by Elliott (2017[1]) computes the aggregate literacy and numeracy measures by taking the mean of experts' ratings on each question and then averaging these mean ratings across questions. This measure has the advantage of reflecting all experts' opinions about AI capabilities. By contrast, the follow-up study classifies each question as solvable or not solvable by AI according to the majority of experts' ratings and estimates the share of questions marked as solvable. This aggregation approach disregards minority opinions. However, its measures are more easily interpretable, and rely only on questions with majority agreement.

Figures A5.1 and A5.2 in Annex 5.A offer additional analyses that compare the results from both assessments following the aggregation rule used in 2016. They show that using the average of experts' ratings as a measure of AI performance on PIAAC produces results similar to those following the majority rule. At each level of question difficulty, AI performance in literacy assessed with the average expert rating is higher in 2021 than in 2016. At Level 4 and above, the increase in literacy performance indicated by the "average" approach is smaller than the increase produced with the "majority" approach. This may have to do with the small number of questions and the bigger disagreement at this level, producing arbitrariness in results.

[2] The results for numeracy for 2016 differ from those reported by Elliott (2017[1]) because they rely on the majority rating and not the average rating of experts (see note 1 above). Figures A5.3 and A5.4 in Annex 5.A present results for 2016 and 2021 obtained by averaging ratings across experts and questions, as done in Elliott (2017[1]). Similar to the results relying on the majority of experts' ratings, the results relying on averages show lower AI numeracy performance in 2021 than in 2016. However, the decline in AI performance indicated by the "average" approach is smaller than the one indicated by the "majority" approach at Level 3 and higher. Questions at these difficulty levels received similar shares of Yes- and No-votes in both assessments. This leads to arbitrary results when using the majority vote as AI's success on these questions is usually decided by a difference of only one vote. By contrast, the "average" approach produces measures close to 50% at these levels since it averages similar shares of 0% and 100% ratings, ignoring that these evaluations reflect disagreement rather than medium AI performance.

6 Implications of evolving AI capabilities for employment and education

This chapter summarises the results of the assessment of artificial intelligence (AI) capabilities in literacy and numeracy and discusses their implications for policy. The chapter first considers likely impacts of developing computer capabilities on employment. For this purpose, it analyses the use of literacy and numeracy at work and the proficiency of workers who use these skills on a daily basis. It then discusses the education implications of AI advancements. In particular, it highlights the need for developing skills in the population that are beyond those of AI technology. The chapter also touches upon the importance of supplying workers with diverse skills, including digital skills, to help them cope with occupational changes resulting from the use of technology.

The preceding chapters described artificial intelligence (AI) capabilities in literacy and numeracy assessed in 2016 and 2021 using expert evaluations with the OECD Survey for Adult Skills (PIAAC). This chapter summarises the results and discusses their implications for policy. It considers likely impacts of developing computer capabilities on employment by looking at the use of literacy and numeracy skills at work. The chapter also discusses the education implications of AI advancements. This focuses on the need for developing skills in the population that are beyond those of technology and the importance of diversifying the skill set of people to enable them to compete, but also to work together, with AI.

Summary of results

The study asked 11 computer experts to rate the capacity of AI to solve the questions of PIAAC's literacy and numeracy tests. Four experts in mathematical reasoning of AI provided additional ratings in numeracy. AI's likely performance on the tests was determined by looking at the majority expert opinion on each question.

Assessment of current AI capabilities in literacy and numeracy

According to most experts, AI can perform high on the PIAAC literacy test. It can solve most of the easiest questions, which typically involve locating information in short texts and identifying basic vocabulary. It can also master many of the harder questions, which require understanding rhetorical structures and navigating across larger chunks of text to formulate responses (see Chapter 4). Overall, AI is expected to solve between 80-83% of all literacy questions, depending on how the majority response of experts is calculated (see Table 6.1).

Table 6.1. Summary of AI and adults' performance in PIAAC

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

	Literacy	Numeracy
AI measures		
Yes/No	80%	66%
Weighted	81%	67%
Weighted + Maybe	83%	73%
Quality of AI measures		
Agreement	78% average majority	62% average majority
Uncertainty	20% Maybe/Don't know responses	12% Maybe/Don't know responses
Adults' performance		
Average adults	50%	57%
Level 2 adults	41%	52%
Level 3 adults	67%	74%
Level 4 adults	86%	90%

Source: OECD (2012[1]; 2015[2]; 2018[3]), *Survey of Adult Skills (PIAAC) databases*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023).

This evaluation rests on high consensus among experts. The group judgements of whether AI can solve each PIAAC literacy question were supported by 78% of experts, across questions, on average (see Table 6.1). However, many votes were excluded from the group response as they were less informative of AI's potential outcome on the PIAAC test. These were the Maybe- and Don't know-answers to the question

of whether AI can correctly solve an item. These responses amounted to 20% of all responses in the literacy assessment (see Table 6.1).

The analyses set out in Chapter 4 compared AI literacy performance to that of adults at varying proficiency levels. PIAAC assesses respondents' proficiency and questions' difficulty on the same levels, going from low (Level 1 and below) to high proficiency/difficulty (Levels 4-5). Respondents with proficiency at a given level have a 67% chance of correctly completing the items at this level, higher chance of success at lower levels of difficulty and lower chances of success at more difficult levels (see Chapter 3).

According to the evaluation of experts, AI can perform similar to or better than Level 3 adults at all levels of question difficulty in literacy (see Figure 4.6, Chapter 4). This is also indicated by the overall success rate of AI on the literacy test, estimated at 80%, which is between that of Level 3 and Level 4 adults (see Table 6.1). This suggests that AI can potentially outperform a large proportion of the population on the PIAAC literacy test. Across the OECD countries that participated in PIAAC, on average, 35% of adults are proficient at Level 3 and 54% score below this level; only 10% of adults perform better than Level 3 in literacy (OECD, 2019, p. 44[4]).

AI performs less well in numeracy – according to the 15 experts who completed the numeracy assessment. Following their majority vote, AI can answer around two-thirds of the easier and intermediate numeracy questions of PIAAC, and less than half of the hardest questions (see Chapter 4). This amounts to an overall success rate of 66-73% on the entire numeracy test, depending on the type of aggregate measure (see Table 6.1).

AI's estimated success rate in numeracy is beyond that of Level 2 adults (see Table 6.1). However, AI could not outperform these adults at each level of question difficulty, as shown in Chapter 4. According to most experts, AI would score similar to low-performing adults at Level 1 and below. The estimated performance at Level 2 of question difficulty is close to that of Level 2 adults. At Level 3 and above, AI's outcome expected by experts corresponds to adults' proficiency at Level 3.

These results should be interpreted with caution, given the high disagreement among experts in the numeracy domain. Two groups of experts with opposing opinions emerged: five experts rated AI's potential performance on most of the numeracy questions low, and five believed that AI could answer most questions. As a result, thin majorities determined AI's capacity to solve the numeracy questions. On average, across questions, the majority opinion on AI performance was supported by 62% of exerts (see Table 6.1). This may have produced arbitrary results since AI's success on single questions was often decided by a difference of only one vote.

The disagreement among experts relates to ambiguity about the generality of the evaluated AI. Some experts imagined narrow AI solutions for separate PIAAC questions. Others considered general AI systems that can reason mathematically and process all types of numeracy questions simultaneously, including similar questions that are not part of the test. These considerations affected experts' evaluations: the latter experts gave lower ratings than the former. However, the group discussion revealed much agreement behind this divergence in ratings. Both groups seemed to agree the test can be solved by developing a number of narrow AI solutions, while an AI with general mathematical abilities is still out of reach of current technology.

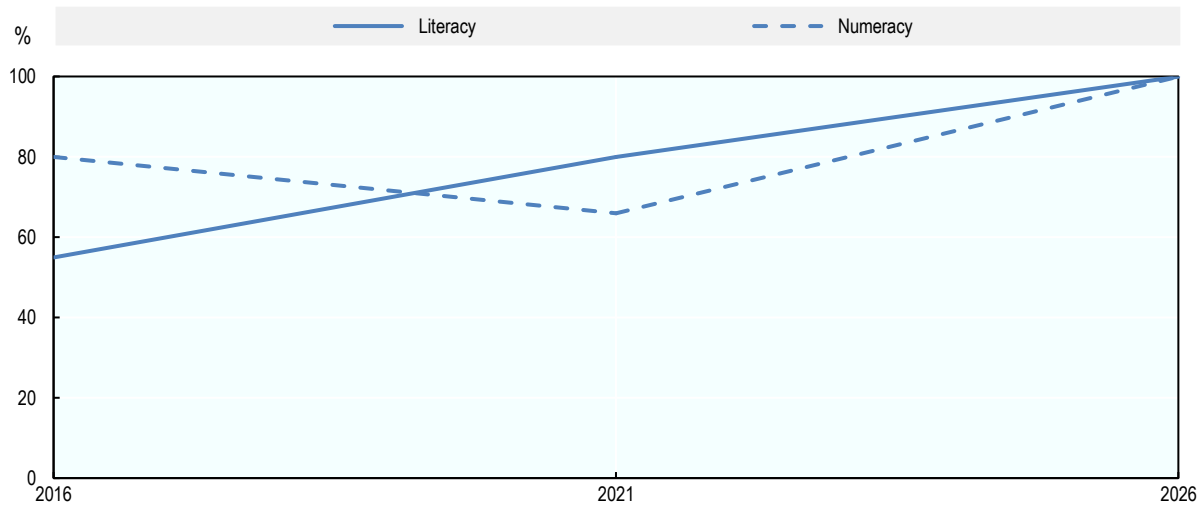
Development of AI literacy and numeracy capabilities over time

This study follows up a pilot assessment of AI capabilities with PIAAC conducted in 2016 (Elliott, 2017[5]). The pilot study asked 11 computer scientists to rate AI capabilities with respect to PIAAC's literacy, numeracy and problem-solving tests. The assessment approach used in the follow-up study is comparable to that of the pilot.

The comparison of both assessments reveals a considerable improvement in AI's literacy capabilities since 2016 (see Figure 6.1). In 2016, experts assessed the potential success rate of AI on the literacy test at 55%. By contrast, experts in 2021 expected AI to solve 80% of the test correctly. Chapter 5 showed that expected AI literacy performance increased at all levels of question of difficulty: from 71% to 93% at Level 2, from 48% to 68% at Level 3, and from 20% to 70% at Level 4-5, while performance at Level 1 and below remained at 100%.

Figure 6.1. Success rate of AI in PIAAC according to experts' assessments

Percentage of PIAAC questions that current AI and AI in 2026 can answer correctly according to most experts



StatLink https://stat.link/brmyfj

These results reflect the technological developments in AI in the period since the pilot assessment. The introduction of large pre-trained language models in 2018 has pushed the state of the art in natural language processing (NLP) forward considerably. These models are trained once, on a large corpus of data, and then used as base models for developing NLP systems for specific tasks and domains. Their success lies in the huge amounts of training data used, as well as in the application of advanced architectures, such as Transformers (Russell and Norvig, 2021[6]). The latter allow the models to capture relations over longer paragraphs and to "learn" the meaning of words in context.

Given these technological advancements and the heavy investment and research in NLP (see also Box 6.1), experts judged that AI's literacy capabilities will continue to develop. They expect that AI will be able to solve all literacy questions in PIAAC by 2026.

The numeracy assessment produced implausible results of declining AI numeracy capabilities over time (see Figure 6.1). This may have to do with methodological issues: disagreement among experts in the domain or varying interpretations of the rating exercise across studies. Another reason may relate to the more limited scope of research on AI's mathematical capabilities on problems such as PIAAC at the time of the pilot assessment. This lack of information may have led experts to overestimate AI's likely performance on the numeracy test in 2016.

Box 6.1. ChatGPT as an example of AI literacy capabilities

In November 2022 – roughly one year after the 11 core experts completed the online assessment, and three months after the four experts in mathematical reasoning re-assessed AI in numeracy – ChatGPT was released. ChatGPT is an AI chatbot developed by OpenAI, a prominent AI research laboratory. Its ability to answer diverse questions and interact in a human-like manner attracted huge public attention. Next to mimicking conversation, ChatGPT performs a variety of tasks, such as composing poetry, music and essays, or writing and debugging code. It demonstrated for the first time to the broader public what state-of-the-art language models are capable of.

ChatGPT relies on a somewhat upgraded version of GPT-3, the model that the experts often considered in their evaluation of AI capabilities. GPT-3 "learns" language in a self-supervised manner, by predicting tokens in a sentence based on the surrounding text. By contrast, ChatGPT is built upon the InstructGPT models of the GPT-3.5 series, which are trained to follow instructions using reinforcement learning from human feedback. Specifically, GPT-3 is fine-tuned with data containing human-written demonstrations of the desired output behaviour. Subsequently, the model is provided with alternative responses ranked by human trainers. The ranks act as reward signals to train the model to predict which outputs humans would prefer. This makes the model better at following a user's intent (Ouyang et al., 2022[7]).

However, the model still has important limitations. It can produce plausible-sounding but incorrect responses (OpenAI, 2023[8]). It can generate toxic or biased content and respond to harmful or inappropriate instructions, although it has been trained to refuse such requests. In addition, the model often fails to respond to incomplete inquiries by asking clarifying questions.

The discussion with experts suggested the numeracy capabilities of AI are unlikely to have changed much between 2016 and 2021. During the period, constructing mathematical models from tasks that require general knowledge and are expressed in language or images has received less research attention. The interest and investment by companies were also limited, focused on specific areas of mathematical reasoning, such as verifying software.

However, this has recently begun to change. In 2021, benchmarks for mathematical reasoning of AI were released (Hendrycks et al., 2021[9]; Cobbe et al., 2021[10]). These allow researchers to train and test models in solving mathematical problems of various kinds. In addition, multimodal models that process information in different formats have received more attention (Lindström and Abraham, 2022[11]). These models are particularly relevant for solving the types of numeracy questions contained in PIAAC since the questions use formats as diverse as images, graphs, tables and text. These trends led experts to expect that AI will advance considerably in numeracy and could solve all PIAAC numeracy questions by 2026 (see Figure 6.1).

In sum, AI's performance in literacy estimated by experts increased by 45% between 2016 and 2021 and is expected to increase by a further 20% by 2026. AI's potential performance on the PIAAC's numeracy test is unlikely to have changed much between 2016 and 2021. However, experts expect it to reach a ceiling by 2026. For comparison, the literacy and numeracy skills of adults change much more slowly (see Chapter 2). Over 13 to 18 years – between the 1990s and 2010s – the distribution of literacy skills in the adult population and the working population has, on average, changed marginally across 19 countries and economies. The same goes for numeracy skills compared over five to nine years across seven countries.

Policy implications of evolving AI capabilities in literacy and numeracy

Fast-evolving AI capabilities in key skill domains raise questions about whether AI will substitute workers in jobs and what implications this would have for education systems. Before turning to these questions, several limitations of this analysis for formulating clear policy conclusions should be discussed.

First, the analysis focuses on the technological capabilities of AI and not on AI deployment in the economy. Whether and how evolving AI technology in the domains of NLP and mathematical reasoning will be adopted in the workplace depends on many factors. Among these are the costs of the technologies, capital investment, regulation and social acceptance (Manyika et al., 2017[12]).

Second, the study assesses AI capabilities in only two skill domains – literacy and numeracy. However, workers use a variety of skills to perform a variety of tasks in occupations. An assessment of AI capabilities across the full range of skills used in the workplace will be needed to determine the exact impacts of AI on employment.

Third, the comparison of AI and human performance in PIAAC should not imply that AI can carry out all kinds of everyday literacy and numeracy tasks as flexibly as adults at a corresponding level of proficiency. In fact, some experts criticised that education tests applied on AI do not necessarily capture the general underlying capabilities that would allow for performing a wide range of similar tasks (as they do when applied on humans). However, this problem – known as overfitting to the test – is common to all tests for evaluating AI. The study attempted to decrease the risk of overfitting by providing experts with more information on the underlying skills that PIAAC is supposed to measure (see Chapter 3).

Despite the limitations of the analysis, it is safe to conclude that the rapid development of AI capabilities laid out here will have important effects on employment, and in particular, on the employment of workers with low literacy and numeracy skills. This may rebound on education systems as they will be increasingly expected to equip people with the skills needed to work in the digitised economy.

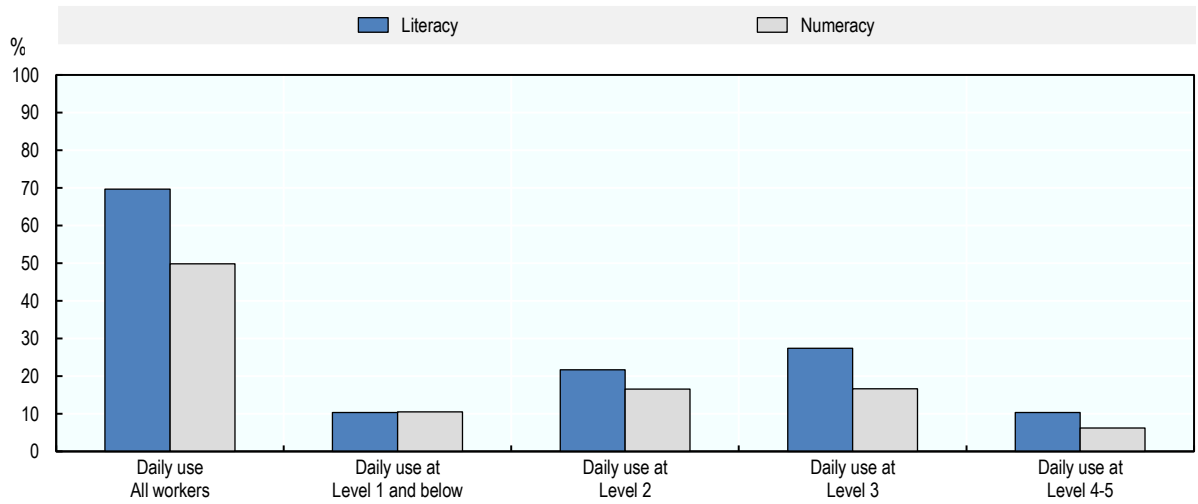
Implications for employment

How evolving AI will affect employment depends not only on how AI capabilities compare to human skills, but also on how skills are used in the economy. AI would substantially affect demand for workers if it can reproduce those skills that are in high demand in the economy. The study chose to assess AI with PIAAC precisely because this test measures key information-processing competencies that are an integral part of work.

Figure 6.2 shows that 70% of workers use literacy skills on a daily basis at work. These workers can be affected by advancing AI capabilities if their literacy proficiency is comparable to or below that of computers. According to computer experts, AI performance in literacy is beyond proficiency Level 3 in PIAAC. Approximately 27% of workers use literacy daily at Level 3 of proficiency. Another 32% perform literacy tasks on a daily basis, having a literacy proficiency below Level 3. Taken together, AI could affect the literacy-related tasks of 59% of the workforce.

Figure 6.2. Percentage of workers at different proficiency levels who use literacy and numeracy on a daily basis at work

Percentage shares from all workers



Note: The use of literacy skills includes reading books; professional journals or publications; manuals or reference materials; diagrams, maps or schematics; financial statements; newspapers or magazines; directions or instructions; letters, memos or mails. The use of numeracy skills includes the use of advanced math or statistics; preparing charts, graphs or tables; use of simple algebra or formulae; calculating costs or budgets; using or calculating fractions or percentages; using a calculator. The bars show the percentage shares of all workers who report performing at least one of these practices daily at work.
Source: OECD (2012[1]; 2015[2]; 2018[3]), *Survey of Adult Skills (PIAAC) databases*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023).

StatLink https://stat.link/zdks8g

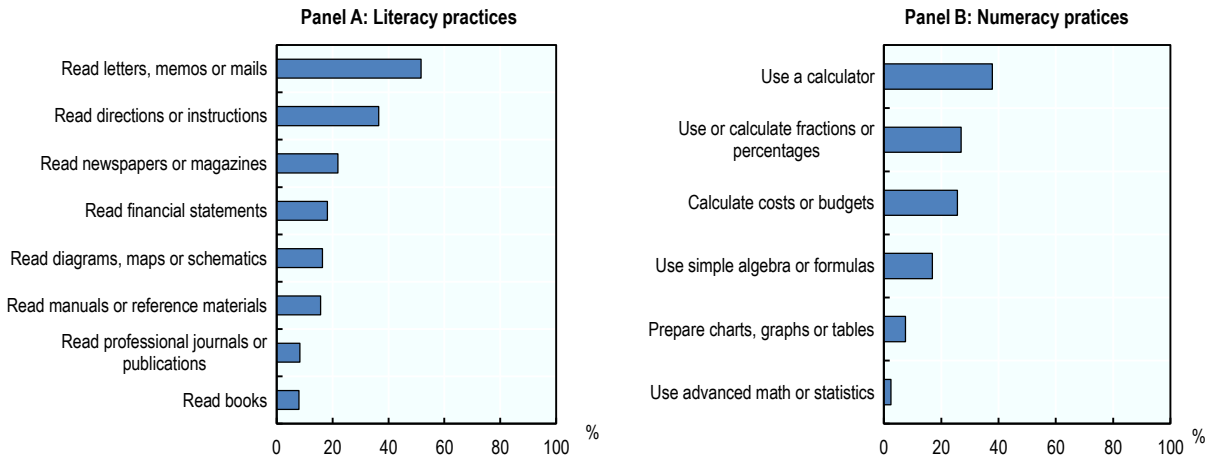
Similarly, Figure 6.2 shows that 50% of workers perform numeracy tasks daily at work. The AI numeracy capabilities assessed by experts are beyond those of adults at proficiency Level 2 on most PIAAC questions and close to those of Level 3 adults on some questions. Across the 39 countries and economies, on average, 27% of the workforce uses numeracy on a daily basis at or below Level 2 proficiency. A share of 44% uses numeracy at Level 3 or below that level. AI can negatively affect the employment of these workers if numeracy tasks constitute a substantial part of their daily work.

AI's impact on employment also depends on the difficulty of tasks performed in occupations. As this study shows, AI performs better on literacy and numeracy tasks that are easier for humans and worse on tasks that are difficult for humans. Thus, AI is more likely to affect workers with easier tasks – independent of their skill proficiency.

Figure 6.3 indicates that more workers perform easy than hard tasks at work. Across the 39 countries participating in PIAAC, on average, 52% of workers read memos or mails, 37% read directions and instructions, and 22% read newspapers and magazines each day. A smaller share of the workforce reads longer text such as professional journals (8%) or books (8%) each day. Similarly, simple numeracy skills are used more broadly than complex numeracy skills. Between 26% and 38% of workers, across all countries, on average, calculate costs or budgets, use a calculator, fractions or percentages daily at work. By contrast, only 3% use advanced math and statistics, 8% prepare charts and graphs, and 17% use simple algebra or formulae each day. This shows that literacy and numeracy tasks that are potentially easy for AI are more prevalent in the economy, even though the workers that perform them may be more proficient than computers.

Figure 6.3. Daily use of literacy and numeracy practices at work

Percentage of workers reporting using a practice on a daily basis



Source: OECD (2012[1]; 2015[2]; 2018[3]), *Survey of Adult Skills (PIAAC) databases*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023).

StatLink https://stat.link/8epnr0

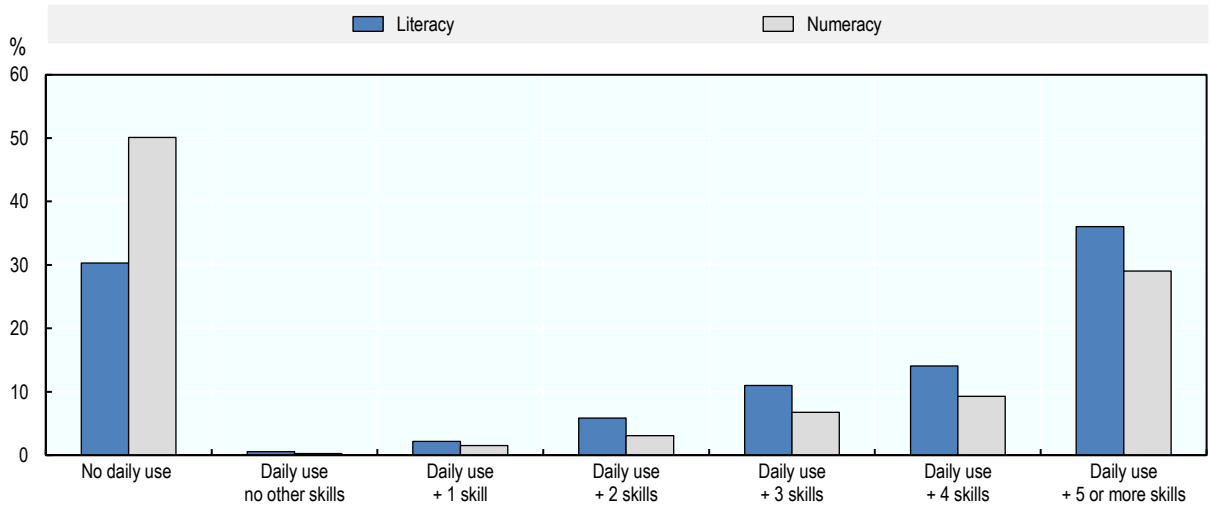
AI's potential to automate jobs further depends on the skill mix jobs require. Jobs that involve a diverse set of skills are more sheltered from automation as it is less likely that AI reproduces many and different skills of workers at once. If AI capabilities come close to reproducing some skills in a rich set of skill requirements, workers will still be needed for other skills. By contrast, workers who use only one or a few skills at work intensively can be completely replaced by machines with the corresponding capability.

Next to literacy and numeracy skills, the Survey of Adult Skills (PIAAC) collects information on the performance of a variety of practices at work (OECD, 2013[13]). These include, for example, the frequency of writing documents, solving complex problems, working with a computer or instructing, teaching or training people. Based on this information, Figure 6.4 explores how workers combine literacy and numeracy with the following generic skills at work: writing, digital skills, problem solving, learning at work, influencing skills, co-operative skills, organising skills and physical skills.[1] Although these skills do not cover all possible skills used in the workplace, they give a glimpse of how workers use skills in concert in their jobs.

Figure 6.4 shows the percentage shares of all workers who do not use literacy or numeracy on a daily basis, and who use these skills daily alone or in combination with other generic skills. The results show that the biggest shares of workers with daily use of literacy and numeracy use a diverse skill mix at work. Across all countries participating in PIAAC, on average, 36% of all workers combine literacy and 29% of workers use numeracy with at least five other generic skills in their daily work routines. By contrast, less than 1% of the workforce reports using literacy or numeracy with none of the other generic skills daily. However, some workers combine literacy and numeracy with only a few other skills – 20% of workers use literacy and 12% use numeracy with up to three other skills at work.

Figure 6.4. Daily use of literacy and numeracy at work together with other skills

Percentage shares from all workers



Note: Daily use of literacy and numeracy together with the following skills: writing, digital skills, problem solving, learning at work, influencing skills, co-operative skills, organising skills and physical skills. See note 1 at the end of this chapter.
Source: OECD (2012[1]; 2015[2]; 2018[3]), *Survey of Adult Skills (PIAAC) databases*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023).

StatLink https://stat.link/2prxu9

The above results suggest that advances in AI with regard to literacy and numeracy can negatively impact employment since literacy and numeracy skills are widely used at work. This holds particularly for the employment of workers who use these skills at a proficiency below that of machines, who perform easy tasks manageable for AI or who use only a few other skills intensively at work. These workers constitute a considerable share of the workforce.

However, AI may change the nature of work in ways that do not affect the aggregate demand for labour. A standard economic view is that long-term effects of AI on employment may be positive due to increases in productivity. AI is expected to improve productivity in companies by performing particular tasks faster and more accurately than human workers. This would leave the latter time to concentrate on more important tasks that may include creativity, management or critical thinking. This, in turn, would enable companies to produce more at a lower cost. Lower costs are expected to increase the demand for the products. This would boost labour demand both in the AI-using companies and in other companies connected to the value chain (OECD, 2019[14]).

In addition, AI is expected to create demand for new tasks – tasks related to adoption and use of machines in the workplace. In future, more workers will be needed for producing data, developing AI applications, operating AI systems and analysing their outputs. An OECD study that analyses job-postings data for 2012-18 in four countries – Canada, Singapore, the United Kingdom and the United States – shows increasing demand for AI-related skills (Squicciarini and Nachtigall, 2021[15]). In the United States, for example, the total number of AI-related job openings increased from around 20 000 in 2012 to almost 150 000 in 2018. In particular, skills related to data mining and classification, NLP and deep learning are more often advertised on line.

Moreover, AI can lead to the emergence of completely new occupations and industries by enabling the creation of new products and services. This technology has been seen as an “invention of a method of

invention" (Cockburn, Henderson and Stern, 2018[16]). This means it is expected to accelerate the process of innovation at an unprecedented rate. What makes AI a trigger of innovation is its wide applicability: learning algorithms have many potential new uses in a variety of sectors and occupations. In addition, AI is increasingly used in science to formulate hypotheses, search and systemise information, or identify hidden patterns in high-dimensional data (Bianchini, Müller and Pelletier, 2022[17]). This can facilitate scientific discovery and the creation of novelty.

How advancing AI will reshape work and the demand for skills remains an open question. What is certain is that workers will need new skills to meet future demands – skills that allow them to compete and work together with AI. This raises questions about the role of education in preparing people for the future.

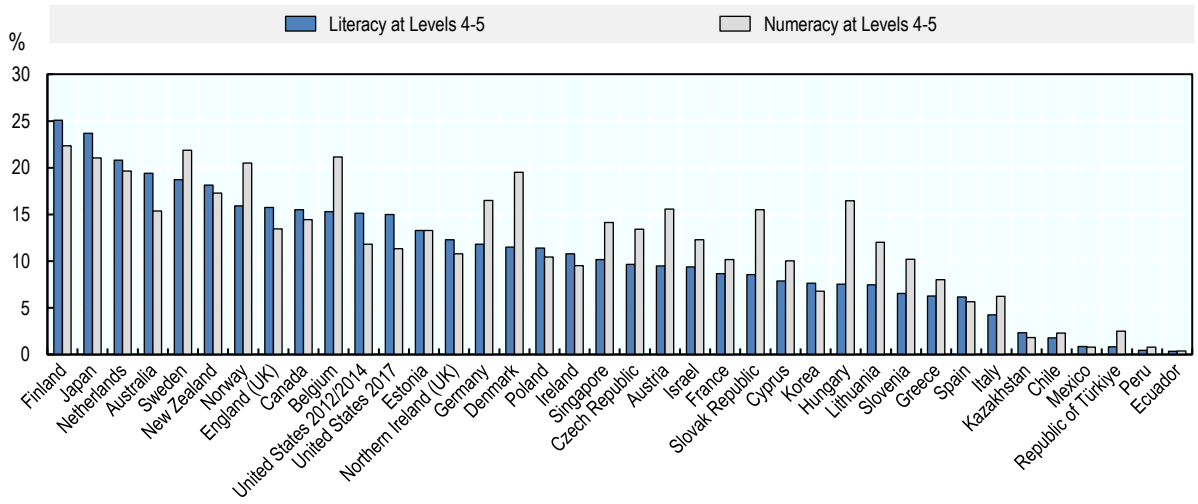
Implications for education

Technological change pressures education systems to supply the economy with a suitably skilled workforce. As one likely response, education would attempt to increase the skill level of the workforce beyond that of computers. In the domains of literacy and numeracy, this would mean to lift up the working population to the highest proficiency levels – Levels 4 and 5. This proficiency would enable workers to understand, interpret and critically evaluate complex texts and multiple types of mathematical information. Developing such skills is not only relevant for outperforming AI in reading and mathematical tasks. Much more importantly, strong literacy and numeracy skills build the foundation for developing other higher-order skills, such as analytic reasoning and learning-to-learn skills. They also ease the access to new knowledge and know-how (OECD, 2013[13]).

Chapter 2 showed the foundation skills of the working population have not changed substantially in the past decades. Of course, future efforts to improve the literacy and numeracy skills of workers could be more successful. In particular, countries with high proportions of highly proficient adults in their workforce can serve as examples of good practice. Other countries can extract formative lessons and borrow promising policies from the comparison with these high performers to increase their skills stock.

Figure 6.5 shows the shares of workers with literacy and numeracy proficiency at Levels 4-5 across the 39 countries and economies that participated in PIAAC. The best ranking country in literacy – Finland – has 25% of adults at literacy proficiency Levels 4-5, followed by Japan with 24% and the Netherlands with 21%. In numeracy, the three top ranking countries – Finland, Sweden and Belgium – have between 21-22% of workers at Levels 4-5. This shows that even the best performers to date cannot supply more than a quarter of their workforce with the literacy and numeracy skills needed to outperform AI. For median performers these shares are much smaller – 10% of workers at Level 4-5 in literacy (Singapore) and 12% in numeracy (Lithuania). These countries would have to double the shares of highly proficient workers in literacy and numeracy to reach their best-performing peers.

Figure 6.5. Proportion of workers with high literacy and numeracy proficiency



Source: OECD (2012[1]; 2015[2]; 2018[3]), *Survey of Adult Skills (PIAAC) databases*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023).

StatLink https://stat.link/6km89z

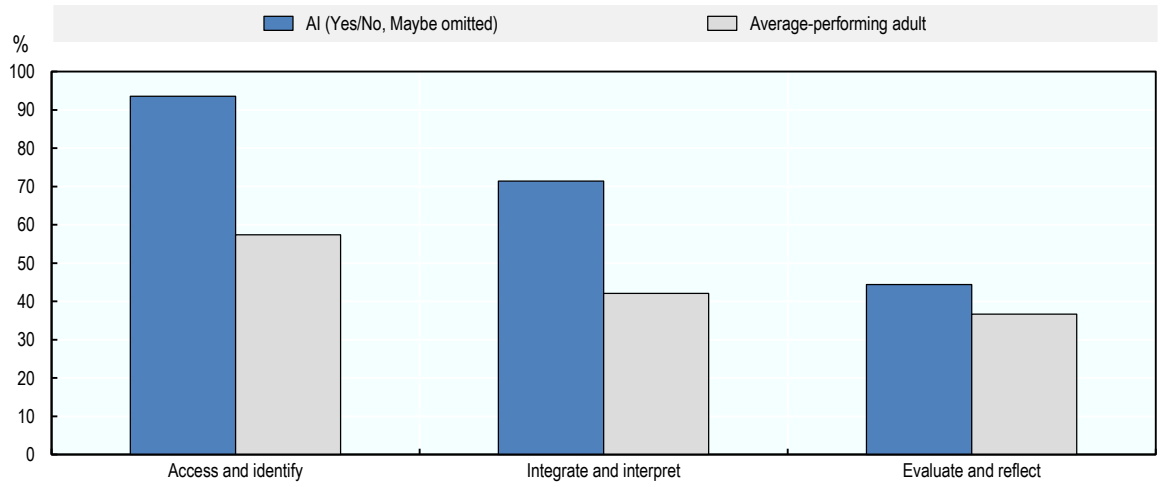
Another goal of education providers could be to trigger those components of foundation skills that prove hard for AI. As shown in Chapter 3, literacy and numeracy are complex constructs. Literacy, for example, involves the use of three cognitive strategies. Individuals should be able *to access and identify* information in texts; to *integrate and interpret* the relations between parts of the text, such as those of cause/effect or problem/solution; and to *evaluate and reflect* on information from texts using own knowledge or ideas (OECD, 2012[18]). Figure 6.6 shows that not all of these literacy sub-skills are easy for AI. According to experts' evaluations, AI is expected to solve 94% of the questions that require accessing and identifying information and 71% of the questions that involve integrating and interpreting relations in the text. Expected performance on questions containing evaluation and reflection is lower, at 44%.

These findings reflect the technological developments in NLP. As Chapter 2 showed, state-of-the-art AI excels on Question-Answering tasks, such as those of the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016[19]; Rajpurkar, Jia and Liang, 2018[20]) and the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018[21]; Wang et al., 2019[22]). These benchmarks test the ability of systems to answer questions related to texts by accessing/identifying the information containing the right answer. Progress was also registered in natural language inference (NLI) (Storks, Gao and Chai, 2019[23]). This is the task of "understanding" the relationship between sentences, which comes close to the "integrate and interpret" tasks in PIAAC. By contrast, AI still struggles with language tasks that require logical reasoning and common knowledge (see, for example Yu et al. (2020[24])). This could explain the low expert ratings on PIAAC questions that require evaluating and reflecting on texts.

However, evaluation and reflection on information in texts is also more challenging for humans. An average respondent in PIAAC has a 37% probability of successfully completing questions in this category, compared to a 57% probability of success on questions requiring the cognitive strategy "access and identify" and 43% on questions involving the "integrate and interpret" strategy. Strengthening people's ability to evaluate and reflect on texts would not only give them an important advantage over machines. This skill would also enable them to cope with the information overload of the digital age and determine the accuracy and credibility of sources against the background of spreading fake news and misinformation.

Figure 6.6. Literacy performance of AI and average adults by cognitive strategy required in PIAAC questions

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[1]; 2015[2]; 2018[3]), *Survey of Adult Skills (PIAAC) databases*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023).

StatLink https://stat.link/2ckahz

In Figure 6.1, projections to 2026 suggest that AI systems will likely soon be able to perform the full range of literacy and numeracy tasks on PIAAC. If this is correct, then the objective for education may need to change substantially. With more capable systems, even high proficiency in literacy and numeracy may no longer be sufficient to allow people to compete with AI. In that context, it seems more plausible that adults will begin to work regularly with AI systems for performing literacy and numeracy tasks. AI systems may help them carry out the tasks more effectively than they could do on their own. As a result, the focus of education may need to shift towards teaching students how to use AI systems effectively.

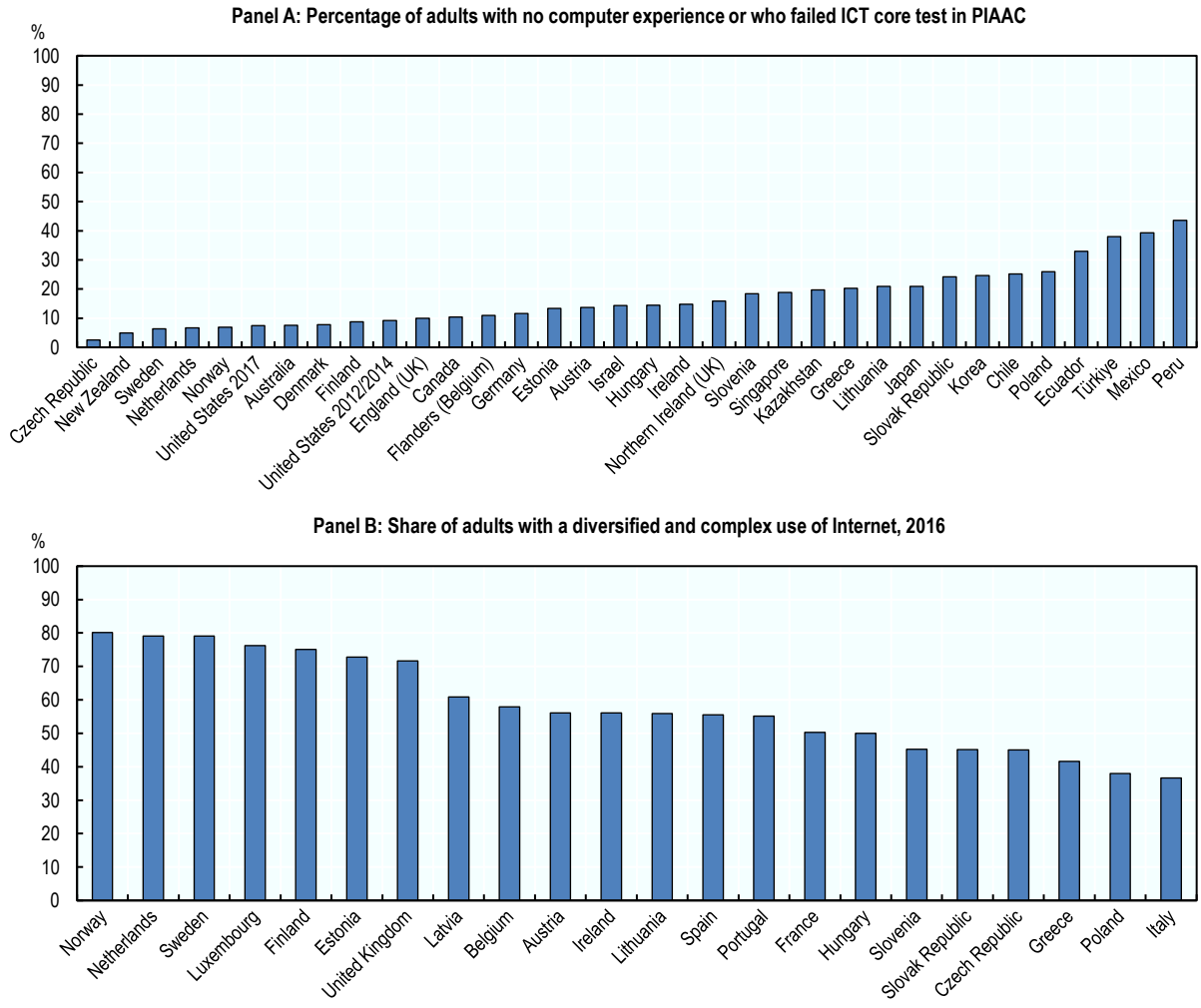
Education systems should also seek to strengthen the digital skills of individuals. These skills would help individuals meet the demands of increasingly digitised workplaces and seize the opportunities brought about by technological advances. Figure 6.7 shows two indicators of the availability of digital competencies in the population (OECD, 2019[25]). The first is the share of adults who are insufficiently familiar with computers. These adults either reported having no prior computer experience in PIAAC or could not perform basic computer tasks (e.g., using a mouse or scrolling through a webpage) to take part in PIAAC's computer-based assessment (OECD, 2019[4]). The second indicator is the share of adults with diversified and complex use of the Internet. It is based on previous analysis of the OECD of data from the European Community Survey on Information and Communication Technologies (ICT) Usage in Households and by Individuals and covers fewer countries (OECD, 2019[25]).

The figure shows that both indicators vary widely across countries. Among countries with available data, Norway, the Netherlands and Sweden have around 80% of the population equipped with skills allowing for diverse and complex Internet use (Panel B). In these countries, as well as in New Zealand and the Czech Republic, less than 7% of the population cannot work with computers (Panel A). By contrast, in Greece and Poland, around 40% of the population can perform many and complex online activities, and one-fifth and one-quarter, respectively, cannot use computers at all. In Peru, the share of adults who cannot use computers exceeds 40%. These latter countries must up-skill large proportions of their adult population to

meet the skill needs emerging from technological change. Another possibility is that the lack of digital skills in these countries slows the spread of new technologies in their economies. This could have negative effects on competitiveness, productivity, innovation and, eventually, on employment.

The current ICT skills of adults shown in Figure 6.7 reflect massive change over the past four decades. Although formal data are not available, in 1980 – before the widespread adoption of computers, the Internet and smartphones – most adults in all countries would likely have failed the ICT core test in PIAAC. They would also have likely said they made no use of the early Internet.

Figure 6.7. Digital skills of adults



Source: Adapted from OECD (2019[4]), *Skills Matter: Additional Results from the Survey of Adult Skills*, Figure 2.15, https://doi.org/10.1787/1f029d8f-en, and OECD (2019[25]), *OECD Skills Outlook 2019: Thriving in a Digital World*, Figure 4.16, https://doi.org/10.1787/df80bc12-en.

StatLink https://stat.link/xryga5

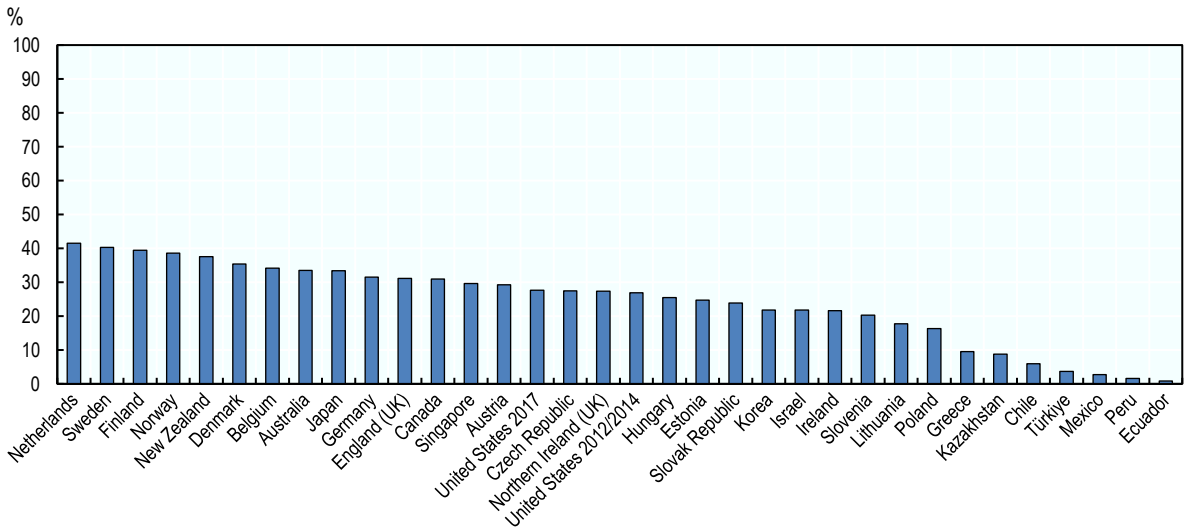
As argued above, the use of diverse skills at work can shelter workers from automation. Therefore, education systems should aim at equipping people with a well-rounded skill set. This would enable people

to adapt to potential changes in their occupations induced by technology. It would also ease their mobility between occupations since diverse skills apply in different work contexts.

Figure 6.8 shows the proportion of working adults with solid skills in three key areas – literacy, numeracy and problem-solving in technology-rich environments.[2] Concretely, the figure shows the proportion of workers with literacy and numeracy skills at Level 3 or above and problem-solving skills at Level 2 or above (see also OECD (2019[25])). At 42%, the Netherlands has the highest proportion of working adults with strong skills in all three domains. However, in nine of the participating countries, the share of workers with a well-balanced skill mix is lower than 20%.

Figure 6.8. Proportion of workers with a well-balanced skill set

Proportion of workers with literacy and numeracy skills at Level 3 and above and problem-solving skills in technology-rich environments at Level 2 and above



Source: OECD (2012[1]; 2015[2]; 2018[3]), *Survey of Adult Skills (PIAAC) databases*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023).

StatLink https://stat.link/lfn861

In sum, AI advances in key cognitive skills will likely pose a challenge for education. Many education systems will initially need to up-skill substantial parts of the population to help them keep up with improving AI capabilities in literacy and numeracy. As AI capabilities in cognitive areas continue to improve, education systems may need to substantially shift their approach to focus far more on working with powerful AI systems that have high levels of literacy and numeracy skills. In addition, education will be increasingly expected to strengthen many other skills, including digital skills, to help people develop strong, diverse skill sets. Such skill sets can help people avoid the risks and benefit instead from the opportunities of the AI revolution.

A new approach to assessing AI

This study provided an example of how AI capabilities improve with respect to two key cognitive skills of humans – literacy and numeracy. These skills were selected because they are of key importance at work and in everyday life, and the foundation for acquiring additional skills and knowledge (OECD, 2013[13]).

Yet, these skills are hard to cultivate, compared to others seen as necessary in the digitised future, such as digital skills. Over the last few decades, literacy and numeracy skills have not changed substantially in most countries, while it did not take long for most people to learn to use a computer or the Internet.

The study showed that AI has developed strong capabilities in literacy and numeracy. According to experts, these capabilities are likely to improve further over the next five years. This raises questions about the possible impacts of advancing AI – about how it will change the ways key skills are used in the workplace and taught in education. Ultimately, to understand how AI will affect future skill use and skill needs, the assessment of AI capabilities should go beyond the general cognitive skills addressed in PIAAC. This would require information on the full range of skills used in occupations and on the proficiency of people with respect to these skills.

This exploratory project is part of a bigger effort by the OECD to assess AI. The AI and the Future of Skills (AIFS) project is developing a comprehensive and authoritative approach to regularly measuring AI capabilities and comparing them to human skills. The capability measures will cover various skill domains that are crucial for work and important in education.

Expert ratings of AI on education tests are an important tool in this approach. Over the past years, the project has repeated and extended the gathering of these expert judgements. For example, the project explored the use of a large-scale expert survey to assess potential AI performance on the PISA science test. It also collected expert judgement on whether AI can perform occupational tests from vocational training and education.

More recently, the AIFS project started to use information from direct tests of AI systems. These include benchmarks, competitions and formal evaluation campaigns that apply AI techniques directly to various kinds of tasks, producing success or failure. The project is developing an approach to inventorying and selecting high-quality direct tests for the assessment. It is working to develop an approach to synthesise the information from such evaluation tests into indicators of AI performance that are understandable and policy relevant.

To help policy makers understand the implications of the AI measures, the project will link them to existing taxonomies of occupational tasks (e.g. ESCO (European Commission, n.d.[26]), O*NET (National Center for O*NET Development, n.d.[27])). These taxonomies provide a way of systematically considering the range of skills needed for performing work tasks and the way these different skills are brought together in occupations.

Moreover, the project will map the AI performance measures to information about the skill proficiency of workers. With the rapid development of AI across a wide range of skill areas, such an approach can systematically identify which skills will likely become obsolete and which may become more significant for work and in education.

The project's first methodology report described its initial work (OECD, 2021[28]). Successive volumes in this series will describe the development of the set of AI measures and the project's explorations in their use. Armed with this information, policy makers can better understand the implications of AI for education and work.

References

Bianchini, S., M. Müller and P. Pelletier (2022), "Artificial intelligence in science: An emerging general method of invention", *Research Policy*, Vol. 51/10, p. 104604, https://doi.org/10.1016/j.respol.2022.104604. [17]

Cobbe, K. et al. (2021), "Training Verifiers to Solve Math Word Problems". [10]

Cockburn, I., R. Henderson and S. Stern (2018), *The Impact of Artificial Intelligence on Innovation*, National Bureau of Economic Research, Cambridge, MA, https://doi.org/10.3386/w24449. [16]

Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/9789264284395-en. [5]

European Commission (n.d.), *The ESCO Classification*, https://esco.ec.europa.eu/en/classification (accessed on 24 February 2023). [26]

Hendrycks, D. et al. (2021), "Measuring Mathematical Problem Solving With the MATH Dataset". [9]

Lindström, A. and S. Abraham (2022), "CLEVR-Math: A Dataset for Compositional Language, Visual and Mathematical Reasoning". [11]

Manyika, J. et al. (2017), *A Future That Works: Automation, Employment, and Productivity*, McKinsey Global Institute (MGI). [12]

National Center for O*NET Development (n.d.), *O*NET 27.2 Database*, https://www.onetcenter.org/database.html (accessed on 24 February 2023). [27]

OECD (2021), *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/5ee71f34-en. [28]

OECD (2019), *Artificial Intelligence in Society*, OECD Publishing, Paris, https://doi.org/10.1787/eedfee77-en. [14]

OECD (2019), *OECD Skills Outlook 2019: Thriving in a Digital World*, OECD Publishing, Paris, https://doi.org/10.1787/df80bc12-en. [25]

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 (2018), *Survey of Adult Skills (PIAAC) database*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023). [3]

OECD (2015), *Survey of Adult Skills (PIAAC) database*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023). [2]

OECD (2013), *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*, OECD Publishing, Paris, https://doi.org/10.1787/9789264204256-en. [13]

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. [18]

OECD (2012), *Survey of Adult Skills (PIAAC) database*, http://www.oecd.org/skills/piaac/publicdataandanalysis/ (accessed on 23 January 2023). [1]

OpenAI (2023), *Introducing ChatGPT*, https://openai.com/blog/chatgpt (accessed on 23 February 2023). [8]

Ouyang, L. et al. (2022), "Training language models to follow instructions with human feedback". [7]

Rajpurkar, P., R. Jia and P. Liang (2018), "Know What You Don't Know: Unanswerable Questions for SQuAD". [20]

Rajpurkar, P. et al. (2016), "SQuAD: 100,000+ Questions for Machine Comprehension of Text". [19]

Russell, S. and P. Norvig (2021), *Artificial Intelligence: A Modern Approach*, Pearson. [6]

Squicciarini, M. and H. Nachtigall (2021), "Demand for AI skills in jobs: Evidence from online job postings", *OECD Science, Technology and Industry Working Papers*, No. 2021/03, OECD Publishing, Paris, https://doi.org/10.1787/3ed32d94-en. [15]

Storks, S., Q. Gao and J. Chai (2019), "Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches". [23]

Wang, A. et al. (2019), "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems". [22]

Wang, A. et al. (2018), "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". [21]

Yu, W. et al. (2020), "ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning". [24]

Notes

[1] The use of each skill is assessed with a number of variables from PIAAC: *writing* – frequency of writing letters, memos or mails; articles; reports; or of filling in forms*; digital skills* – frequency of using the Internet for mail; for finding work-related information; for conducting transactions; frequency of using spreadsheets; Microsoft Word; programming languages; or online real-time discussions; *problem solving* – frequency of solving complex problems at work; *learning at work* – frequency of learning from co-workers/supervisors; of learning-by-doing; keeping up to date; *influencing skills* – frequency of teaching people; giving presentations; selling; advising people; influencing people; negotiating with people; *co-operative skills* – co-operating with co-workers more than half of the time; *organising skills* – planning others' activities; *physical skills* – working physically for longer time. A skill is used daily if the respondent reports daily use of at least one of the activities used to measure the skill.

[2] See note 1 in Chapter 1.

Educational Research and Innovation

Is Education Losing the Race with Technology?

AI'S PROGRESS IN MATHS AND READING

Advances in artificial intelligence (AI) are ushering in a large and rapid technological transformation. Understanding how AI capabilities relate to human skills and how they develop over time is crucial for understanding this process.

In 2016, the OECD assessed AI capabilities with the OECD's Survey of Adult Skills (PIAAC). The present report follows up the earlier study, collecting expert judgements in 2021 on whether computers can solve the PIAAC literacy and numeracy tests. It is part of a comprehensive ongoing project on assessing AI.

This study shows that AI could potentially outperform large shares of the population on PIAAC – 90% of adults in literacy and 57-88% of adults in numeracy. AI's literacy capabilities had improved considerably since the 2016 assessment. According to experts, AI will solve the entire literacy and numeracy tests by 2026.

These findings have important implications for employment and education. Large shares of the workforce use literacy and numeracy skills daily at work with a proficiency comparable or below that of computers. AI could affect the literacy- and numeracy-related tasks of these workers. In this context, education systems should strengthen the foundation skills of students and workers and teach them to work together with AI.

Federal Ministry
of Labour and Social Affairs

PRINT ISBN 978-92-64-45137-7
PDF ISBN 978-92-64-92037-8

9 789264 451377