



Educational Research and Innovation

AI and the Future of Skills, Volume 2

METHODS FOR EVALUATING AI CAPABILITIES



Educational Research and Innovation

AI and the Future of Skills, Volume 2

METHODS FOR EVALUATING AI CAPABILITIES

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), *AI and the Future of Skills, Volume 2: Methods for Evaluating AI Capabilities*, Educational Research and Innovation, OECD Publishing, Paris, <https://doi.org/10.1787/a9fe53cb-en>.

ISBN 978-92-64-88932-3 (print)
ISBN 978-92-64-82429-4 (pdf)
ISBN 978-92-64-42035-9 (HTML)
ISBN 978-92-64-81732-6 (epub)

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

Photo credits: credits: © Shutterstock/LightField Studios; © Shutterstock/metamorworks; © Shutterstock/Rido; © Shutterstock/KlingSup.

Corrigenda to OECD 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 emerged as a significant area of development. Its integration into various sectors necessitates a comprehensive understanding of its capabilities, especially in relation to human skills. The AI and the Future of Skills (AIFS) project by the OECD's Centre for Education Research and Innovation (CERI) has undertaken this task, aiming to provide a methodological framework for assessing and comparing AI capabilities to human skills. This framework should provide a basis for informed discussions on AI's impact on education, work and society.

The project has undergone two phases of developing a rigorous approach to assessing AI's capabilities. The first phase focused on identifying relevant AI capabilities and the tests best suited to evaluate them. Leveraging insights from various fields including computer science, psychology and education, the project offered a multi-disciplinary perspective on the challenges and prospects of assessing AI.

The second phase, the focus of this report, further refines the methodology of the assessment. It encompasses a range of exploratory AI evaluations to identify most promising practices for systematically and periodically assessing AI. These explorations are threefold. First, by assessing AI capabilities with OECD's education tests using expert judgement, the project explored ways to understanding AI's progress in competencies that are traditionally human – competencies in reading, mathematics and science. Second, the project asked experts to rate AI on real-world occupational tasks, such as those encountered in nursing or product design, to provide critical insights into AI's application potential. By situating AI within these occupational contexts, we gain a clearer picture of its impending impact on the economy. Third, the project considered the vast and evolving benchmarks available in AI research that result from direct assessments of AI systems.

These methods, while promising, are not without their challenges. This report underscores the difficulties in solely relying on expert judgements to evaluate AI. While expert input is valuable, achieving consensus, particularly in novel domains, can be challenging. Moreover, the variability in AI applications and the intricacies of real-world tasks suggest the need for diverse evaluation metrics. Therefore, the project decided to integrate both expert judgements and direct AI measures in its subsequent phase to provide a thorough and balanced evaluation. This integrative approach aims to provide decision-makers with a nuanced understanding of AI's capabilities.

The next project phase intends to produce an integrated assessment framework for AI. This will contain a set of key AI indicators that can serve as reference points for various stakeholders. These indicators, informed by a combination of expert input and direct assessments, will offer guidance for policy formulation and implementation.

As AI continues to evolve, having a clear framework to understand its capabilities becomes crucial. The AIFS project's efforts contribute to this understanding, laying the groundwork for informed decisions in education and employment sectors. This work reflects OECD's commitment to producing rigorous, evidence-based insights that can inform decision-making in the context of AI's continued growth and integration into various sectors.

Acknowledgements

This publication was planned and developed by the OECD's Artificial Intelligence and Future of Skills project team – Stuart Elliott (Project lead), Mila Staneva, Margarita Kalamova, Abel Baret, Nóra Révai, Sam Mitchell, Marc Fuster-Rabella and Aurelija Masiulytė. The report was prepared for publication by Mila Staneva and Aurelija Masiulytė.

This publication would not have been possible without the invaluable contributions of the renowned computer scientists and psychologists who are supporting the project.

Firstly, we would like to express our gratitude to the experts who participated in the assessments or provided advice (in alphabetical order): Phillip L. Ackerman, Guillaume Avrin, Chandra Bhagavatula, Joseph Blass, Fergus Bolger, Jill Burstein, Salvador Carrión Ponz, Anthony G. Cohn, Vincent Conitzer, Ulises Cortes, Pradeep Dasigi, Ernest Davis, Angel de Paula, Marie desJardins, Kenneth D. Forbus, Carlos Galindo, Janice Gobert, Jordi González, Arthur C. Graesser, Yvette Graham, Fredrik Heintz, Jim Hendler, Daniel Hendrycks, José Hernández-Orallo, Jerry R. Hobbs, Lawrence Hunter, Juan Izquierdo-Domenech, Maria Juarez, Aina Juraco Frias, Ryota Kanai, Aviv Keren, Rik Koncel-Kedziorski, Patrick Kyllonen, David Leake, Bao Sheng (Aiden) Loe, Fernando Martinez-Plumed, Aqueasha Martin-Hammond, Cynthia Matuszek, Elena Messina, Antoni Mestre Gascón, Ángel Aso-Mollar, Jose Andres Moreno, Constantine Nakos, Taylor Olson, Rebecca J. Passonneau, Swen Ribeiro, Carolyn Rose, Gene Rowe, Vasile Rus, Britta Rüschhoff, Vijay Saraswat, Areg Mikael Sarvazyan, Brian Scassellati, Wout Schellaert, Jim Spohrer, Mark Steedman, Claes Strannegård, Neset Tan, Tadahiro Taniguchi, Moshe Vardi, Karina Vold, Michael Witbrock, Michael Wooldridge, Hiroshi Yamakawa.

Secondly, 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. Colleagues from the Programme for International Assessment of Adult Competencies (PIAAC) and the Programme for International Student Assessment (PISA) made important contributions to the analysis. Colleagues within the Directorate for Education and Skills 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/workinnovation-productivity-skills> and <https://denkfabrik-bmas.de/>.

Table of contents

Foreword	3
Acknowledgements	4
Executive summary	9
1 Overview	12
Overview of the AI and the Future of Skills project	14
Lessons learnt from the first project stage	16
The second stage of the project	17
Outline of the structure of the report	20
References	22
Notes	23
2 Eliciting expert knowledge: Methods and challenges	24
Methods for eliciting expert judgement	26
Large-scale experiment: How many experts can be engaged and through what incentives?	30
Task framing used to collect expert judgement on AI capabilities	33
Establishing consensus: Quantitative disagreement versus qualitative agreement	35
Conclusions: Challenges and future directions	37
References	38
Notes	39
3 Assessing AI capabilities with education tests	40
Rationale for assessing AI capabilities with education tests	42
Overview of the education tests used	43
Methodology for collecting expert judgement on AI with education tests	47
Results	51
Lessons learnt	58
The way forward	60
References	61
Annex 3.A. Analyses of the PIAAC and PISA studies using an alternative approach	63
Notes	64
4 Occupational tests	65
Rationale for collecting expert judgement on AI with complex occupational tasks	66
Occupational tasks from certification and licensure examinations	67
Selection of occupations and examination tasks	70

The way forward	73
References	75
Notes	77
5 Assessing AI capabilities on occupational tests	78
Collecting expert judgement on performance tests of occupational tasks	79
Evaluation of AI and robotics capabilities on tasks and subtasks	81
Evaluation of AI and robotics capabilities on capability scales	90
The way forward	94
References	95
Annex 5.A. Categories of AI capabilities	96
Notes	98
6 A framework for characterising evaluation instruments of AI performance	99
Characterising AI evaluation instruments	100
Evaluation instrument selection and rating methodology	103
Analysis of rater consistency	104
Analysis of facet values	105
Conclusion	110
References	112
Annex 6.A. Supplementary tables	116
Notes	118
7 AI direct tests: LNE and NIST evaluations	119
The need for systematising AI and robotics evaluations	120
Framework structure	121
Evaluation campaigns of AI capabilities	124
Limitations and uncovered tasks from AI evaluations	138
Conclusion	139
References	140
Annex 7.A. Low functionality levels AI tasks of evaluation campaigns across the three major fields of NLP: computer vision and robotics	142
Annex 7.B. Detailed facet characteristics attributions of the LNE and NIST evaluations	145
Notes	145
8 Towards a synthesis of language capability in humans and AI	146
Benchmark tasks: Narrow versus strong AI	147
Conceptual framework of language competences	148
Mapping major language benchmarks to the human language competence framework	150
Language understanding: AI vs. human	151
Language generation: AI vs. Human	157
Update of AI language competences post ChatGPT release	160
Conclusion	162
References	163
Annex 8.A. Natural Language Processing research areas	165
Notes	166
9 Project goals, constraints and next steps	167
Potential sources of information about AI capabilities	168
Information needed about AI's implications for education and work	172

Next steps for the project	174
References	177
Notes	177

FIGURES

Figure 1.1. Sources of AI assessments	16
Figure 2.1. The effect of incentives on the final response rate	32
Figure 2.2. Self-reported motivation to complete the survey	32
Figure 3.1. PIAAC Literacy and Numeracy – Sample items	45
Figure 3.2. PISA Science – Sample item	46
Figure 3.3. AI literacy performance in 2016 and 2021, by question difficulty	52
Figure 3.4. AI numeracy performance in 2016 and 2021, by question difficulty	53
Figure 3.5. Predicted AI performance on PISA science questions in 2022 by core experts and larger expert group, by question difficulty	54
Figure 3.6. Literacy performance of AI and adults of different proficiency	55
Figure 3.7. Divergence in experts' evaluations in different assessments	56
Figure 3.8. Share of questions that receive more than 20% of uncertain ratings in different assessments	57
Figure 3.9. Experts' ratings of AI and GPT-3.5 performance on PISA science questions	58
Figure 5.1. AI and robotics performance on entire task, by task format	82
Figure 5.2. Distribution of expert ratings of AI and robotics performance on entire task	83
Figure 5.3. Average AI and robotics performance, by expert and expertise	84
Figure 5.4. AI and robotics performance in broad capability domains, by task and expertise	85
Figure 5.5. AI and robotics performance on subtasks, by complexity level and broad capability domain, mid-2022	86
Figure 5.6. Expert descriptors of complexity levels of broad capability domains	87
Figure 5.7. AI capability expert ratings and their comparison to the ratings of the first study	91
Figure 5.8. AI capability expert ratings, by task	92
Figure 6.1. Rater agreement across all facets	105
Figure 6.2. Rater value selection on validity facets	106
Figure 6.3. Raters value selection on consistency facets	108
Figure 6.4. Raters value selection on fairness facets	109
Figure 7.1. Rater values selection on validity facets for eight evaluation campaigns by NIST and LNE	125
Figure 7.2. Rater values selection on consistency facets for eight evaluation campaigns by NIST and LNE	126
Figure 7.3. Rater values selection on fairness facets for eight evaluation campaigns by NIST and LNE	127
Figure 8.1. General relationship between human language and NLP difficulty levels with respect to the input and output format moving from text to speech and vice versa	149
Figure 8.2. Relationship between required minimum human language competence level and state-of-the-art NLP system performance for a sample of NLP tasks	150
Figure 8.3. Example dialogue from the TRAINS corpus	152
Figure 8.4. Sample constituency parse tree of English sentence	155
Figure 9.1. Conceptual scale reflecting AI performance levels	175

TABLES

Table 2.1. Major EKE protocols	27
Table 2.2. Methods used to collect expert judgement in the AIFS project	29
Table 2.3. Response rate	31
Table 2.4. Task framing and response format in the AIFS project	35
Table 4.1. Selected occupations	70
Table 4.2. Selected occupational tasks	72
Table 6.1. Primary testing domain of sampled evaluation instruments	103
Table 6.2. Type of sampled evaluation instruments	104
Table 7.1. Text processing and comprehension high-level task examples and associated evaluation campaigns	129
Table 7.2. Speech processing high-level task examples and associated evaluation campaigns	130

Table 7.3. Recognition high-level task example and associated evaluation campaigns	131
Table 7.4. Motion analysis high-level task examples and associated evaluation campaigns	132
Table 7.5. Locomotion high-level task examples and associated evaluation campaigns	134
Table 7.6. Manipulation task examples and associated evaluation campaigns	136
Table 8.1. NLP research areas with type and level of language competence required for humans	149
Table 8.2. Datasets in the GLUE benchmark	153

Annex Table 3.A.1. List of online figures for Chapter 3	63
Annex Table 5.A.1. Categories of AI capabilities	96
Annex Table 6.A.1. Overview of Evaluation Instruments	116
Annex Table 7.A.1. Low functionality level tasks of evaluation campaigns associated with the NLP field	142
Annex Table 7.A.2. Low functionality level tasks of evaluation campaigns associated with the Computer Vision field	143
Annex Table 7.A.3. Low functionality level tasks of evaluation campaigns associated with the Robotics field	144
Annex Table 8.A.1. Natural Language Processing research areas with at least one benchmark task	165

BOXES

Box 1.1. Types of AI measures discussed in the report	19
Box 2.1. Evolution of assessment instructions in the AIFS project	34
Box 3.1. Use of education tests in AI evaluation	43
Box 3.2. Example items from PIAAC and PISA	45
Box 4.1. Direct assessment of large language models on written professional certification tests	71
Box 7.1. Facet characteristics of the LNE and NIST evaluations vs. those of benchmark tests	124
Box 8.1. Transformer models	148
Box 8.2. Example of anaphora and coreference resolution	152
Box 8.3. GLUE benchmark	153

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.

3 Assessing AI capabilities with education tests

Mila Staneva (OECD), Abel Baret (OECD), Àngel Aso-Mollar (Universitat Politècnica de València), Joseph Blass (Northwestern University), Salvador Carrión Ponz (Universitat Politècnica de València), Vincent Conitzer (Carnegie Mellon University), Ulises Cortes (Universitat Politècnica de Catalunya), Pradeep Dasigi (Allen Institute for AI), Angel de Paula (Universitat Politècnica de València), Carlos Galindo (Universitat Politècnica de València), Janice Gobert (Rutgers University), Jordi González (Universitat Autònoma de Barcelona), Fredrik Heintz (Linköping University), Jim Hendler (Rensselaer Polytechnic Institute), Daniel Hendrycks (Center for AI Safety), Lawrence Hunter (University of Colorado Anschutz Medical Campus), Juan Izquierdo-Domenech (Universitat Politècnica de València), Maria Juarez (Universitat Politècnica de València), Aina Juraco Frias (Universitat Politècnica de València), Aviv Keren (Anyword), Rik Koncel-Kedziorski (Kensho Technologies), David Leake (Indiana University), Bao Sheng (Aiden) Loe (University of Cambridge), Fernando Martinez-Plumed (Universitat Politècnica de València), Aqueasha Martin-Hammond (Indiana University), Cynthia Matuszek (University of Maryland, Baltimore County), Antoni Mestre Gascón (Universitat Politècnica de València), Jose Andres Moreno (Universitat Politècnica de València), Constantine Nakos (Northwestern University), Taylor Olson (Northwestern University), Carolyn Rose (Carnegie Mellon University), Areg Mikael Sarvazyan (Universitat Politècnica de València), Brian Scassellati (Yale University), Wout Schellaert (Universitat Politècnica de València), Claes Strannegård (Chalmers University of Technology), Neset Tan (University of Auckland), Tadahiro Taniguchi (Ritsumeikan University), Karina Vold (University of Toronto), Michael Wooldridge (University of Oxford)

This chapter introduces three exploratory studies that assessed the capabilities of artificial intelligence (AI) through standardised education tests designed for humans. The first two studies, conducted in 2016 and 2021/22, asked experts to evaluate AI's performance on the literacy and numeracy tests of the OECD's Survey of Adult Skills (PIAAC). The third study collected expert judgements of whether AI can solve science questions from the OECD's Programme for International Student Assessment (PISA). The studies aimed to refine the assessment framework for eliciting expert knowledge on AI using established educational assessments. They explored different test formats, response methodologies and rating instructions, along with two distinct assessment approaches. A "behavioural approach" used in the PIAAC studies emphasised smaller expert groups engaging in discussions, and a "mathematical approach" adopted in the PISA study relied more heavily on quantitative data from a larger expert pool. This chapter presents the results of the studies and discusses the advantages and disadvantages of their methodological approaches.

The AI and the Future of Skills (AIFS) project carried out three exploratory studies on the use of education tests for collecting expert evaluations on artificial intelligence (AI). The first two studies used the OECD's Survey of Adult Skills of the Programme for International Assessment of Adult Competencies (PIAAC). PIAAC assesses the proficiency of adults aged 16-65 in three general cognitive skills – literacy, numeracy and problem solving in technology-rich environments (OECD, 2019^[1]).¹ In 2016, a pilot study asked 11 experts to assess whether AI can do the literacy, numeracy and problem-solving tests of PIAAC (Elliott, 2017^[2]). This pilot study served as a stepping stone into the AIFS project. In 2021/22, a follow-up study surveyed 15 computer experts to show how AI capabilities in literacy and numeracy have evolved since the pilot assessment (OECD, 2023^[3]).

A third study in 2022 used test questions from the OECD's Programme for International Student Assessment (PISA) as a measurement tool. PISA assesses the knowledge and skills of 15-year-old students in reading, mathematics and science (OECD, 2019^[4]). The study collected expert judgement on AI capabilities using questions from the science domain. In contrast to the studies using PIAAC, this study attempted to assemble a much larger sample of experts. For this purpose, the study offered different incentives for attracting and engaging experts in the assessment (see Chapter 2).

The three exploratory studies aimed to improve the assessment framework for eliciting expert knowledge on AI using standardised tests designed for humans. To that end, the studies explored the use of different tests, different response formats and different instructions for rating. Moreover, they tested the feasibility of two generally different approaches to assessing expert knowledge for the project purposes. The studies using PIAAC explored the so-called behavioural approach (see Chapter 2). This means they relied on smaller expert groups that could engage in extensive discussions to reach agreement on AI capabilities. By contrast, the study using PISA followed a so-called mathematical approach. This means it relied more heavily on quantitative information from a larger group of experts, under the assumption that aggregation across many judgements reduces random error in those judgements.

The use of education tests for assessing AI capabilities can provide both reliable and policy-relevant AI measures. Education tests provide standardised and objective criteria for assessing AI capabilities. This enables assessing AI with different expert groups and tracking AI progress across time. Moreover, education tests typically target skills that are taught in education institutions and widely used at work. Assessing how AI performs on these skills thus provides insights into AI's potential impacts on education and employment. This information is important for designing education and labour market policies that are responsive to incoming technological changes.

The results of the exploratory studies showed that AI performs well in all three tested domains. In literacy, computer experts expected that AI could solve 80% of the PIAAC questions in 2021. This marks a considerable improvement to the success rate of 55% in literacy obtained in 2016. In numeracy, AI was expected to solve around two-thirds of the PIAAC test questions in 2021/22. In science, experts expected AI to solve PISA science questions with 61% confidence. These results correspond to human performance levels in the middle or at the higher end of the performance spectrum. They show that AI can potentially outperform large shares of the adult and youth population. AI's estimated performance in literacy, for example, is equal to or higher than that of 90% of adults in OECD countries with PIAAC data (OECD, 2023^[3]).

However, the results also revealed some methodological challenges in collecting expert judgements on AI capabilities with education tests. The main challenge was disagreement among experts, especially in rating AI's potential performance on the PIAAC numeracy test. Disagreement mainly related to ambiguity about how general the computer capabilities being assessed are supposed to be. Some experts assumed general capabilities that should enable successful performance over a wide range of test questions. Others considered narrow systems geared towards solving specific problems. To reach agreement, the experts thus needed clarification on the generality of the underlying capabilities being evaluated. The studies explored different methods to address this issue.

This chapter describes the three exploratory studies and compares their methodologies. The first section discusses the advantages of using education tests to collect expert judgements on AI. The second section introduces the survey instruments. The third section presents the methodology – the methods used for expert knowledge elicitation, including the questionnaires used for the expert surveys. The fourth section presents the main findings and discusses the quality of measures. The fifth section elaborates on the strengths and weaknesses of the approaches used and the sixth section outlines the next steps in this project strand.

Rationale for assessing AI capabilities with education tests

The elicitation of expert ratings on AI capabilities with education tests has a number of methodological advantages:

- Education tests provide a standardised and objective way for eliciting expert knowledge. This enables reproducing an AI assessment with different groups of experts and across time.
- Education tests provide concrete and detailed descriptions of the test tasks. This enables computer experts to make more objective and precise judgements since they do not have to rely on implicit assumptions about the task requirements. This improves the reliability of the assessment.
- Education tests enable comparisons of computer and human capabilities. This can show which skills may become obsolete and which may gain in significance in the years ahead. Moreover, education tests typically assess various socio-demographic characteristics of respondents in addition to their skill proficiency. This enables fine-grained AI-human comparisons within different country contexts, age groups or occupations. Such analysis can show which social groups are particularly vulnerable to automation.
- Education tests offer a graduated progression from simple to complex tasks. This allows for obtaining more nuanced measures of AI performance across different levels of task difficulty.
- Assessing AI capabilities on education tests provides information that is useful for policy making. Both PIAAC and PISA assess key cognitive skills that are used in most social contexts and work situations. These skills strongly affect individuals' ability to participate effectively in the labour market, education and training, and social and civic life (OECD, 2019^[1]). Understanding how AI performs with respect to these skills can thus provide valuable insights into AI's potential impacts on work and life.
- Assessing AI against education tests provides understandable measures. Compared to benchmark tests and evaluations used in AI research, education tests can describe AI capabilities in a way that is meaningful to the general public. In addition, educators and education researchers are typically familiar with the skills assessed in education tests and the ways those skills are developed in education and used at work and in daily life.

Against this background, expert judgement on whether AI can carry out education tests constitutes an important source of information for the AIFS project. This information can complement the overall assessment framework in areas in which results from direct assessments of AI systems are lacking.

Box 3.1. Use of education tests in AI evaluation

Computer scientists employ various education tests to directly assess AI systems' performance. For instance, Hendrycks et al. (2020^[5]) evaluated state-of-the-art AI models, including different configurations of GPT-3 (Generative Pre-trained Transformer), on 57 education tests. The tests cover various disciplines, including mathematics, physics, history, micro econometrics, geography, law, anatomy and philosophy. They span elementary to university-level courses. While most models performed at nearly random levels, the largest GPT-3 achieved an average accuracy of 44% across tests. This performance was lowest in subjects that require quantitative reasoning, such as mathematics and physics, and in subjects related to human values, such as law and ethics.

AI performance on education tests has increased with the introduction of more powerful language models. GPT-3.5, introduced in early 2022, demonstrated strong performance in college-level art history, environmental science, psychology, political studies and writing, among others (OpenAI, 2023^[6]). GPT-4, introduced in March 2023, outperformed GPT-3.5 on most of these exams (Bubeck et al., 2023^[7]; OpenAI, 2023^[6]). However, performance in quantitative subjects remains moderate. A study that evaluated the mathematical capabilities of GPT-4 concluded that while the model performs well in undergraduate-level mathematics, it often fails on graduate-level difficulty (Frieder et al., 2023^[8]).

Some computer scientists have argued that standardised education tests are a useful evaluation tool for AI systems. According to Clark and Etzioni (2016, p. 4^[9]), such tests are “accessible, easily comprehensible, clearly measurable, and offer a graduated progression from simple tasks to those requiring deep understanding of the world”. Additionally, they encompass a broad spectrum of AI capabilities. Hendrycks et al. (2020^[5]) note that education tests require extensive world knowledge and problem solving ability. They thus provide important insights into AI models' overall language understanding abilities.

Overview of the education tests used

The following subsections introduce PIAAC and PISA. They provide information on the approaches these surveys use to assess skills and describe the formats of test questions, as well as the contexts and cognitive strategies they address.

Assessing literacy and numeracy in the Survey of Adult Skills (PIAAC)

The Survey of Adult Skills (PIAAC) is conducted every ten years. The First Cycle took place between 2011 and 2018, collecting data from 39 countries and economies. It surveyed approximately 250 000 respondents, with national samples ranging from about 4 000 to nearly 27 300 (OECD, 2019^[1]). First results from the Second Cycle are expected in 2024.

The survey assesses the proficiency of adults aged 16-65 in literacy, numeracy and problem solving with computers. The pilot study from 2016 focused on all three domains, while the follow-up study from 2021/22 used only the literacy and numeracy assessments of PIAAC. This report covers only results on AI capabilities in literacy and numeracy since they were assessed in both time points. These skills are considered key cognitive skills since they build the foundation for developing higher-order skills (e.g. analytic reasoning and learning-to-learn skills) and for acquiring new knowledge. In technology-rich societies, literacy and numeracy are essential for gaining access to information relevant to everyday life (OECD, 2012^[10]).

PIAAC assesses both skill proficiency and question difficulty on a 500-point scale. Questions are first assigned a difficulty score, which is dependent on the proportion of respondents who complete them successfully. Respondents are then placed on the same scale, based on the number and difficulty of questions they answer correctly. For simplicity, the 500-point scale is broken down into six proficiency/difficulty levels (below Level 1, Levels 1-5). A person with a proficiency score in the middle of the range defining the level can successfully complete tasks at that level approximately 67% of the time. For example, 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^[11]).

The PIAAC literacy test measures adults' ability to understand, evaluate, use and engage with written texts in real-life situations. It contains 57 reading tasks 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^[10]; OECD, 2013^[11]).

Easy literacy tasks (below Level 1 and 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.

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^[10]; OECD, 2013^[11]). It contains 56 tasks that 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”).

Easy numeracy 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.

Box 3.2. Example items from PIAAC and PISA

Figure 3.1 presents example items from the PIAAC literacy and numeracy tests. These sample items are at difficulty Level 3.

Figure 3.1. PIAAC Literacy and Numeracy – Sample items

Panel A literacy

Unit 1 - Question 1/3

Look at the list of preschool rules. Highlight information in the list to answer the question below.

What is the latest time that children should arrive at preschool?

Preschool Rules

Welcome to our Preschool! We are looking forward to a great year of fun, learning and getting to know each other. Please take a moment to review our preschool rules.

- Please have your child here by 9:00 am.
- Bring a small blanket or pillow and/or a small soft toy for naptime.
- Dress your child comfortably and bring a change of clothing.
- Please no jewelry or candy. If your child has a birthday please talk to your child's teacher about a special snack for the children.
- Please bring your child fully dressed, no pajamas.
- Please sign in with your full signature. This is a licensing regulation. Thank you.
- Breakfast will be served until 7:30 am.
- Medications have to be in original, labeled containers and must be signed into the medication sheet located in each classroom.
- If you have any questions, please talk to your classroom teacher or to Ms. Marlene or Ms. Tree.

Panel B numeracy

Look at the graph about the number of births. Click to answer the question below.

During which period(s) was there a decline in the number of births? Click all that apply.

☐ 1957 - 1967

☐ 1967 - 1977

☐ 1977 - 1987

☐ 1987 - 1997

☐ 1997 - 2007

The following graph shows the number of births in the United States from 1957 to 2007. Data are presented every 10 years.

Year	Number of Births
1957	4,300,000
1967	3,520,959
1977	3,326,632
1987	3,809,394
1997	3,880,894
2007	4,315,000

Source: OECD (2012^[10]), *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>.

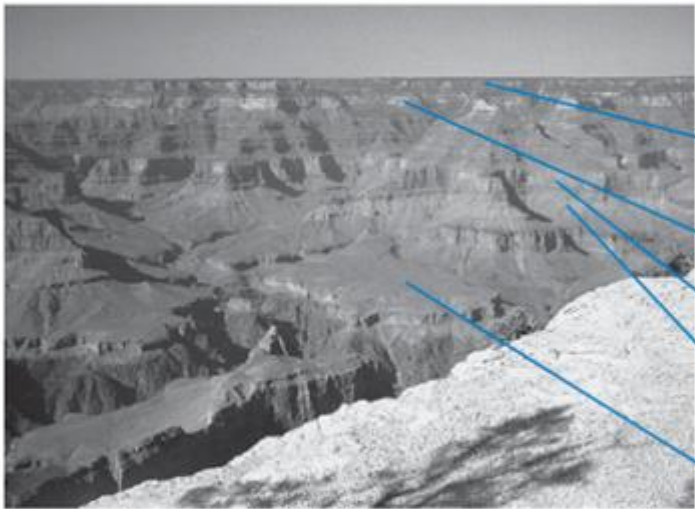
Figure 3.2 shows an example for an item of the PISA science test at difficulty Level 2.

Figure 3.2. PISA Science – Sample item

SCIENCE UNIT 7: THE GRAND CANYON

The Grand Canyon is located in a desert in the USA. It is a very large and deep canyon containing many layers of rock. Sometime in the past, movements in the Earth's crust lifted these layers up. The Grand Canyon is now 1.6 km deep in parts. The Colorado River runs through the bottom of the canyon.

See the picture below of the Grand Canyon taken from its south rim. Several different layers of rock can be seen in the walls of the canyon.



Limestone A

Shale A

Limestone B

Shale B

Schists and granite

QUESTION 7.1

The temperature in the Grand Canyon ranges from below 0 °C to over 40 °C. Although it is a desert area, cracks in the rocks sometimes contain water. How do these temperature changes and the water in rock cracks help to speed up the breakdown of rocks?

- A. Freezing water dissolves warm rocks.
- B. Water cements rocks together.
- C. Ice smooths the surface of rocks.
- D. Freezing water expands in the rock cracks.

Source: OECD (2009^[12]): *Take the Test: Sample Questions from OECD's PISA Assessments*, <https://doi.org/10.1787/9789264050815-en>.

Assessing science literacy in the Programme for International Student Assessment (PISA)

PISA assesses the knowledge and skills of 15-year-old students in reading, mathematics and science. The assessment has taken place every three years since 2000, with each round testing one of the three subjects in detail and providing basic results for the other two. In addition, some assessment rounds offer optional assessments, for example, on students' familiarity with information and communication technologies (ICT) in 2015 (OECD, 2016^[13]) or on students' well-being in 2018 (OECD, 2019^[4]). The last assessment round to date, in 2022, collected information from more than 80 countries and economies.

The AIFS project focused on the PISA science assessment. It used 20 publicly released test questions from the science domain, which were either used in actual assessments or tested in PISA field trials (see Figure 3.2 in Box 3.2 for an example).² PISA's science questions measure students' scientific knowledge, as well as the ability to use that knowledge to identify scientific issues, explain phenomena scientifically or use scientific evidence. They use multiple-choice or open-ended response formats and are typically presented in text, although some questions contain images, graphics or tables. The questions are designed to resemble a wide variety of real-life situations that involve science and technology. Topics are related to personal (e.g. nutrition), social (e.g. disease control) or global (e.g. climate change) issues and cover the domains of health, natural resources, environmental quality, hazards and the frontiers of science and technology.

Scores in PISA are scaled to fit a normal distribution with a mean of 500 score points and a standard deviation of 100 score points. They do not have a substantial meaning and, theoretically, there is no minimum or maximum score (OECD, 2019^[4]). In the first detailed assessment of science in 2006, around two-thirds of students in OECD countries scored between 400 and 600 score points (OECD, 2007^[14]). Similarly, as in PIAAC, the difficulty of individual questions is given a score on the scale that depends on the proportion of test takers getting the question correct. Student performance is then described by assigning each student a score according to the number and difficulty of questions he or she has answered correctly.

To ease the interpretation of results, PISA summarises science scores into six proficiency levels (Levels 1 to 6). Questions at Levels 1 and 2 are easier, requiring students to recall simple scientific facts or to use common scientific knowledge in drawing or evaluating conclusions. Questions at medium difficulty at Levels 3 and 4 require students to use scientific knowledge to make predictions, provide explanations, recognise relevant questions, and select relevant information from competing data or claims. Hard tasks at Levels 5 and 6 require students to create or use conceptual models to predict or explain scientific phenomena; to understand the design of scientific studies and the hypotheses they test; to use data to evaluate alternative viewpoints or differing perspectives; and to communicate scientific results.

Methodology for collecting expert judgement on AI with education tests

The pilot study using PIAAC was carried out with 11 experts in May 2016 (Elliott, 2017^[21]). Five years later, a second study followed up, using a comparable revised methodology (OECD, 2023^[3]). This follow-up study consisted of two rounds. In December 2021, 11 computer experts completed the literacy and numeracy assessments. Due to diverging ratings in the numeracy domain, a second round of interviews in September 2022 engaged four computer scientists with expertise in mathematical reasoning of AI. This second round applied a modified framework for rating to address expert disagreement.

In 2022, a third study using PISA questions was carried out. The study used the latest modified framework for rating. It first collected judgements from 12 core experts who participated in the previous PIAAC assessment. The study then conducted a large-scale online assessment with new experts to compare the advantages of this approach to the use of smaller groups of experts who engage with the project on a long-term basis. The following sections describe the methodology used in the three studies and how it was improved in the course of the work.

Collecting expert judgement

The methodology for collecting expert judgement in the three exploratory studies progressed from a behavioural to a mathematical approach (Rowe, 1992^[15]). As described in Chapter 2, the behavioural approach relies on a few experts who engage in in-depth discussions to arrive at a consensus judgement on a question. This aims to address questions in their complexity by considering different arguments and

perspectives and to draw on the best of these arguments to build a group judgement. By contrast, the mathematical approach relies on many experts who provide individual judgements without interacting with each other. The goal is to avoid social biases such as social conformity or dominance of influential individuals.

The pilot study with PIAAC came closest to a behavioural approach. Here, the 11 experts made their ratings during a two-day meeting. Materials, containing instructions and the PIAAC questions in literacy, numeracy and problem solving, were provided in advance. Experts were encouraged to study the questions and provide initial comments and reactions prior to the meeting. During the meeting, the experts provided their judgements and discussed salient questions, problematic issues or any ideas and arguments that group members brought up (Elliott, 2017^[2]).

The follow-up study followed a more structured approach. Experts received the PIAAC questions one week in advance. They had then two weeks to provide their ratings in an online survey. During this time, they were able to access the survey at any time via an individualised survey link. Finally, the experts discussed the results in a subsequent four-hour online meeting (OECD, 2023^[3]).

Interaction between experts is a key element of these studies. In the pilot study, experts could freely discuss their evaluations and any other matters related to the assessment. In the follow-up study, experts could communicate with the group via e-mail at any point of the assessment process. Most importantly, they received feedback on the group results from the online survey. During the four-hour workshop, they discussed these results, focusing on questions that received diverging ratings. Afterwards, experts had the opportunity to revise their ratings in response to the feedback received and the group discussion. This interaction was intended to encourage information sharing. Since some experts may have more information on specific AI applications or recent research results, they should be able to share their knowledge with the group.

The study using PISA tested an approach where experts completed the online survey without the possibility to interact and without receiving the test materials in advance. The study started with replicating the approach used in the previous assessment with PIAAC. That is, 12 of the core experts were invited to participate in the study. They received the PISA science questions in advance, rated potential AI performance on them in an online survey and discussed the results in an online meeting. Subsequently, the study invited more than 180 new experts to participate in the AI assessment with PISA. Of these, 63 expressed interest in participating and 33 actually participated in the online survey. These new experts had one month to complete the online survey. During this time, they did not have contact with other participants.

This latter approach served three purposes. First, restricting interaction among group members should account for social biases. Such biases can occur, for example, when only ideas that are broadly acceptable to all group members are discussed, or when a charismatic person imposes his or her opinions on the group (Tversky and Kahneman, 1974^[16]). Second, surveying many experts should better represent opinions and expertise in the scientific community regarding AI capabilities. Third, the approach should offer a faster and less costly way of collecting expert judgement since experts are only completing the online survey.

Response categories

In the pilot study with PIAAC, experts rated whether AI could solve each of the test questions with a Yes, Maybe or No. The subsequent discussion revealed that experts differed in their interpretation of the Maybe category. Some experts used it to express genuine uncertainty about AI's performance, while others used it as a not very certain Yes (Elliott, 2017^[2]).

The follow-up study attempted to gather more nuanced information on the certainty of experts' answers. It used a different question to elicit expert knowledge: "How confident are you that AI technology can carry

out this 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 reflects both experts’ confidence and their rating of the capability of AI. For example, “0% No, AI cannot do it” means that experts are certain that AI cannot carry out the task, while 25% means that experts think that AI probably cannot do it. The “50% – Maybe” category means full uncertainty (OECD, 2023^[3]). The study using PISA assessed experts’ confidence in AI solving the task with the same question. However, the large-scale sample used a continuous scale ranging from 0% to 100% confidence.

Assessing uncertainty in experts’ answers is important for establishing more valid AI measures. Some experts may lack specific knowledge regarding AI’s capabilities on particular tasks. Others may have trouble understanding the test question or the instructions for rating. Accounting for this, for example, by giving uncertain ratings a lower weight in the analysis, can improve measures. Moreover, a high proportion of uncertain ratings on specific questions can draw attention to a lack of clarity of some tasks or to general ambiguity in the field regarding AI’s performance on the tasks. Indicating and excluding such problematic questions can improve the analysis.

Instructions for rating

In making their evaluations, experts needed to consider a hypothetical process of adapting current AI techniques to the specific context of the test questions as no AI systems are tailored for solving PISA or PIAAC. Therefore, existing systems should be adapted for these tests, for example, by training them on relevant examples or by coding information about specific vocabularies, relationships or types of knowledge representation, such as charts and tables. Experts should use identical parameters for this hypothetical development effort in order to provide consistent ratings.

The pilot study using PIAAC defined two such parameters for experts to consider. First, experts were instructed to think of “current” computer techniques, meaning any available techniques addressed sufficiently in the literature. That is, experts were asked to imagine applying available systems instead of creating entirely new ones. Second, the instructions asked experts to consider a development effort that costs up to USD 1 million and takes no longer than one year to implement (Elliott, 2017^[2]).

The follow-up study used the same criteria to define the boundaries of the hypothetical advance preparation of AI systems for the tests. However, after the first assessment round, experts suggested that these parameters should be revised. They generally saw the hypothetical investment of USD 1 million as insufficient and proposed fitting this effort to the size of a major commercial AI development project to better reflect reality in the field. In addition, experts pointed out that PIAAC questions have many and different response types, some of which may be difficult for computers (e.g. clicking an answer). They advised changing the instructions for rating to allow for some hypothetical transformation of the task format. Such transformation should remove trivial hurdles to solving the task with AI, without changing the nature of the capabilities the test attempts to measure (OECD, 2023^[3]). These suggestions were implemented both in the second round of the follow-up study with PIAAC and the study with PISA.

Framing the rating exercise

The studies using PIAAC instructed experts to imagine a single hypothetical AI system for solving each test domain. However, experts did not always follow this rule. Some viewed different question types within a test (e.g. numeracy questions containing tables) as separate, narrow problems and evaluated AI’s capacity to solve them independently from each other. That is, they considered different systems for different problems. By contrast, other experts viewed a test as a general challenge for AI to process multimodal inputs in various settings. They considered one system solving all test questions, including similar tasks that are not part of the test (OECD, 2023^[3]).

How experts saw the scope of the test affected their judgements. The ones who focused on narrow problems generally gave more positive ratings than those who focused on general challenges. This led to diverging evaluations. The divergence in experts' rating was most pronounced in the follow-up PIAAC numeracy assessment. The numeracy test contains more diverse question types, including graphs, images, tables and maps, compared to the literacy and science tests that consist mostly of text inputs. This increased the ambiguity about the types of question formats that a hypothetical system is supposed to master.

To address this issue, the studies needed to define the full range of problems that AI is supposed to solve in a test domain. However, providing such information is not trivial. It requires defining all types of tasks that humans who master the test are expected to solve, and that machines should also be able to solve to be assigned the same underlying capability. Therefore, several other steps were taken to improve expert agreement in the follow-up study with PIAAC and the study with PISA.

First, the studies provided experts with information from the assessment frameworks of PIAAC and PISA. These documents describe the conceptual frame of the assessments. They define the underlying skills targeted by the assessments and describe the types and formats of the test questions. This information was synthesised and supplemented by nine example survey questions of low, medium and high difficulty to exemplify the scope of the test and how general the capabilities required for solving it should be. It was provided to experts prior to the online survey in the assessment with the four experts in mathematical reasoning, and at the onset of the online survey in the large-scale assessment.

Second, the studies asked experts to describe a high-level approach for solving each test using the information from the assessment framework and the example tasks. Subsequently, they were asked to rate the potential success of their imagined approach on each question of the test. Encouraging experts to think of a single system that can tackle all problems in a test was intended to provide a common ground for the evaluation. It should also facilitate understanding and communication among experts since it enables them to review the arguments and considerations of their peers.

Additional questions

In addition to how experts assess AI on the test, the survey asked a number of other questions. All online surveys contained open-ended questions asking experts to explain their ratings of AI performance on each question. The goal was to collect additional qualitative information on the rationales behind the ratings. At the end of each survey, experts could report any difficulties in understanding or answering the questions in a domain or leave any comments or suggestions.

The studies also asked experts to predict the performance of AI on the tests in future. These projections were collected in order to explore ways of tracking AI progress over time. The pilot study using PIAAC asked experts to predict AI performance ten years in the future. The follow-up study using PIAAC and the study using PISA used a shorter time frame of five years. The discussion showed that experts are more confident in making predictions over a shorter time horizon given the rapid rate of change in AI technology. They are also used to projecting AI research trends over three to five years when applying for research grants.

One challenge in using tests developed for humans is that they take for granted capabilities that most humans share, such as short-term memory or object recognition. This may result in misleading AI measures if computers fail the tests because they lack such capabilities rather than because they lack the primary capabilities being assessed. To tackle this problem, the follow-up study using PIAAC included an additional 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)?" (OECD, 2023^[3]). This question was intended to specify the elements of tasks that are easy for machines to perform in order to collect more

precise information on computer performance. However, only a few experts made use of this question, which led the OECD team to abandon it in subsequent assessments.

Constructing aggregate measures

The studies used two aggregation approaches to construct single measures for AI literacy, numeracy and science performance from the individual expert ratings (OECD, 2023^[3]). The first approach relies on the majority opinion of experts. It labels each test question as solvable or not solvable by AI based on what most experts judged. Questions on which experts cannot reach majority agreement are excluded from the analysis. The aggregate measures of AI performance show the percentage share of test questions in a domain that AI could solve according to the majority of computer experts. These measures are comparable to human scores that show the expected probability of respondents of successfully completing test items. As another advantage, they are robust, relying on experts' consensus understanding of AI capabilities.

A second approach constructs final measures by averaging across all experts' ratings. That is, the aggregate AI measures are computed by taking the mean of experts' ratings on each question and then averaging these mean ratings across all questions in a domain. The advantage of these measures is that they reflect all experts' opinions about AI capabilities. However, they are harder to interpret and not comparable to human scores as they show the average confidence of experts that AI can solve the test.

The follow-up study with PIAAC used the "majority" rule to aggregate experts' ratings. This is in line with the behavioural approach for eliciting expert judgements that focuses on discourse and consensus building among experts. By contrast, the PISA science assessment, which follows the mathematical approach for expert knowledge elicitation, averages all experts' ratings to arrive at final AI measures. This reflects the goal of the mathematical approach to build measures representing a broad spectrum of expertise and opinions in the expert community.

In the following, results from the studies with PIAAC are presented by following the "majority" rule, while results from the PISA assessment are computed with the "average" rule. Annex 3.A presents analyses from each study using the alternative approach. All measures are presented for different levels of question difficulty to provide a more detailed picture of AI performance on the tests.

Results

This section outlines the assessed performance of AI on the PIAAC literacy, PIAAC numeracy and PISA science questions. It also evaluates the quality of these AI performance metrics through several indicators of validity and reliability.

To facilitate a more direct comparison, answer categories from the 2021/22 PIAAC literacy and numeracy assessments were aligned with the Yes, Maybe and No categories from the 2016 study. That is, ratings of 0% and 25% were combined into a No-category, and ratings of 75% and 100% were treated as Yes. The aggregate measures then show the share of test questions, for which the majority of experts give a Yes. In contrast, the AI measures obtained with PISA indicate the average level of experts' confidence in AI's ability to successfully complete the test tasks.

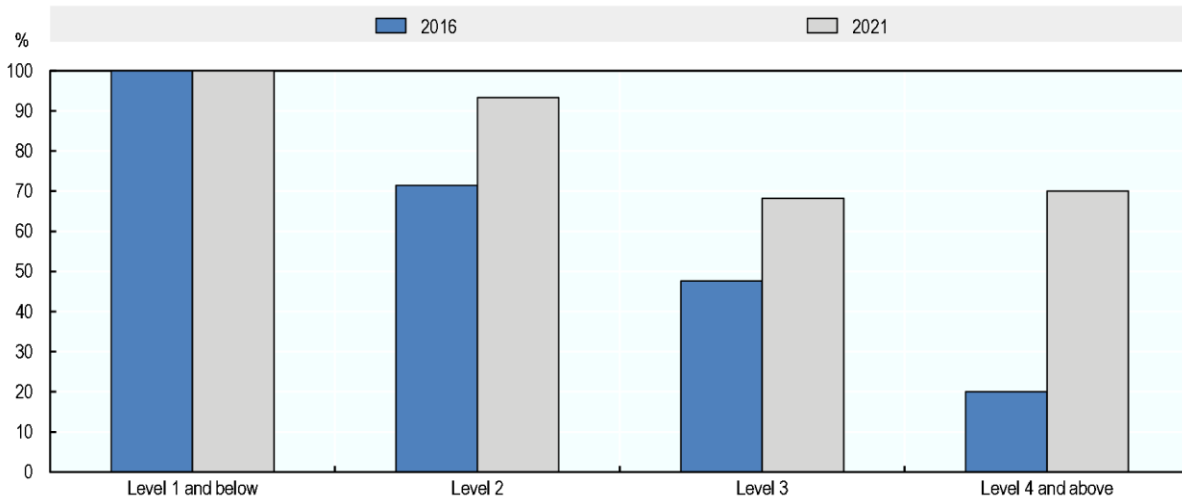
AI capabilities in literacy, numeracy and science

Figure 3.3 shows results from the pilot and follow-up studies with PIAAC for literacy. It indicates a clear improvement of AI literacy capabilities from 2016 to 2021. According to the majority of experts, AI's potential performance on the test increased at all difficulty levels. The increase amounts to 25 percentage points across all questions, moving from 55% to 80% between 2016 and 2021. These findings align well with the significant advances in natural language processing (NLP) that have occurred since 2016. These

include the advent of large pre-trained language models like GPT-2 and GPT-3, predecessors to ChatGPT (Radford et al., 2018^[17]). The coherence between experts' judgements and known progress in AI capabilities suggests that experts have a solid grasp of the task at hand.

Figure 3.3. 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 (OECD, 2023^[3]), *Is Education Losing the Race with Technology?*, Figure 5.2, <https://doi.org/10.1787/73105f99-en>.

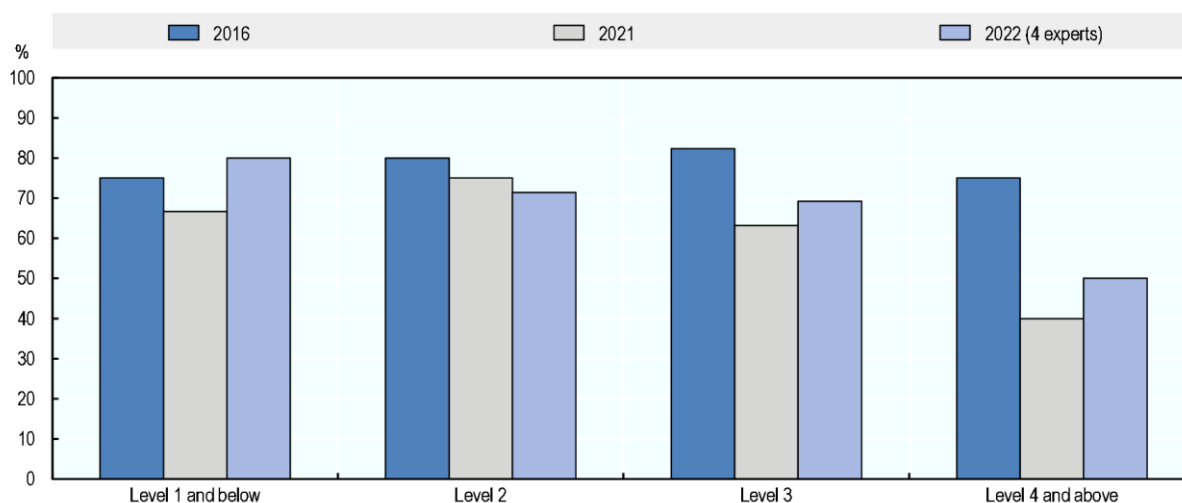
The results of the numeracy assessment are less straightforward. Following the same aggregation approach, Figure 3.4 shows a decline in AI's performance on the numeracy questions between 2016 and 2021/22. The decline is most pronounced at question difficulty Level 3 and Level 4 and above – 16 and 35 percentage points, respectively.

In the follow-up assessment, the 11 experts who completed the first assessment round in 2021 and the 4 mathematical reasoning experts who re-assessed numeracy in 2022 provided similar aggregate ratings. This suggests that neither the assessment modifications nor the shift in expertise significantly impacted group ratings on numerical skills.

These counter-intuitive results have to do with strong disagreement among experts in the follow-up study. Two opposing groups emerged. In the first round, five experts evaluated AI negatively on almost all questions, while four other experts provided mostly positive ratings. In the second round, one expert had overly negative ratings, another had mostly negative ratings and the other two were in the middle. This led to thin majorities, often determined by a single vote, and resulting in arbitrary conclusions on AI's capabilities.

Figure 3.4. 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, 2017^[2]), *Computers and the Future of Skill Demand*, <https://doi.org/10.1787/9789264284395-en>, and (OECD, 2023^[3]), *Is Education Losing the Race with Technology?*, <https://doi.org/10.1787/73105f99-en>.

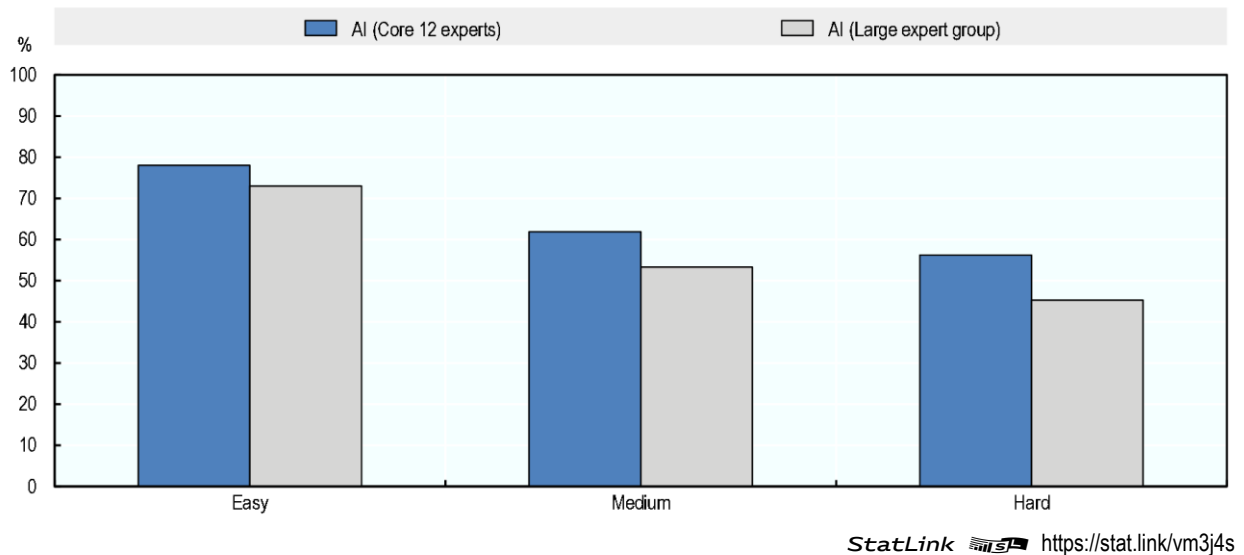
StatLink  <https://stat.link/2ryum7>

Figure 3.5 presents results from the assessment using PISA science questions, distinguishing between the 12 core experts and the larger expert group. It shows that both groups of experts have similar high confidence in AI solving easier questions, and lower confidence for more difficult ones. Overall, AI is expected to solve easy questions (Levels 1 and 2) at 78% confidence in the core expert group and at 73% confidence in the large expert group, which decreases to 62% and 53%, respectively, for questions of medium difficulty (Levels 3 and 4), reaching 56% and 45%, respectively, for hard questions (Levels 5 and 6).

The science questions consist mostly of text inputs. Therefore, the similarity of the results to those obtained with the PIAAC literacy test is not surprising. It reflects the strong performance of NLP systems in question-answering and text generation. That both the small- and the large-scale assessments produce similarly high ratings in this domain suggests that both the behavioural and the mathematical approaches to collecting expert judgement are effective in obtaining plausible evaluations from experts.

Figure 3.5. Predicted AI performance on PISA science questions in 2022 by core experts and larger expert group, by question difficulty

Average of experts' confidence in AI solving PISA science questions



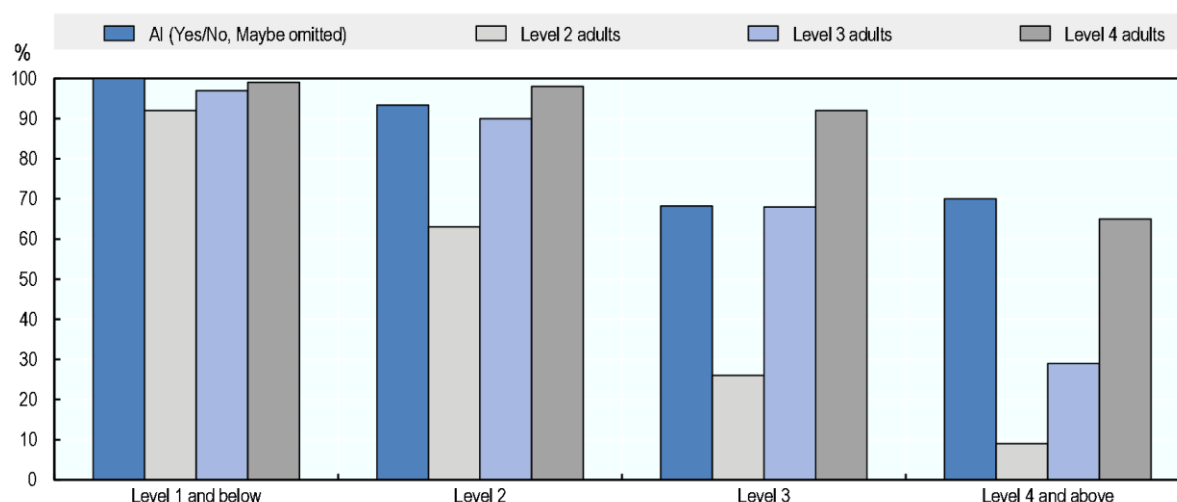
Comparison with human performance

As discussed above, the use of education tests for measuring AI capabilities offers the benefit of detailed comparisons to human performance. This provides insights into the potential impact of AI on essential skills used in educational and work settings. Tests like PIAAC go a step further in linking performance scores of respondents to various socio-economic and demographic characteristics. This allows, for example, for nuanced analyses of skill performance across countries, occupations, education levels or age groups.

Figure 3.6 illustrates how AI and human capabilities compare in literacy. The assessed AI performance by experts is compared to three proficiency levels of adult respondents. Adults at each proficiency level are expected to complete successfully 67% of the questions at that level. They have higher probability of success at easier questions, and lower chances to answer harder questions. The figure shows that expected AI performance resembles that of Level 3 adults. That is, AI is expected to solve about two-thirds of the Level 3 questions and almost all Level 1 and 2 questions. At Level 4, expected performance is actually closer to that of Level 4 adults, at 70%. However, this latter result should be interpreted with caution due to the small number of questions at that level.

Figure 3.6. 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: Adapted from (OECD, 2023^[18]), *Is Education Losing the Race with Technology?*, Figure 4.6, <https://doi.org/10.1787/73105f99-en>.

PIAAC data show that most adults have literacy skills below Level 3. 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^[11]). This suggests that AI can potentially outperform a large proportion of the population on the PIAAC literacy test.

Quality of AI measures

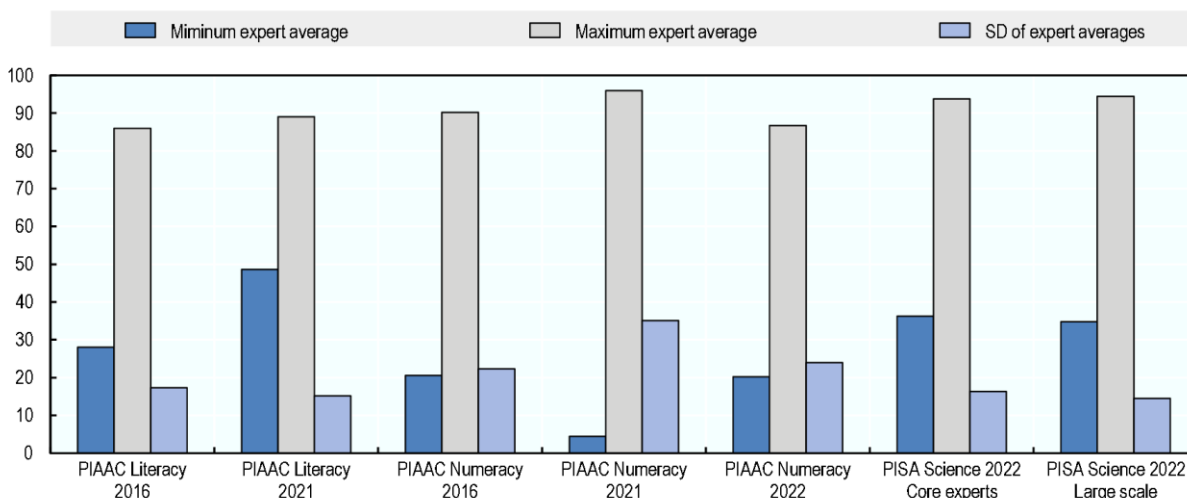
Disagreement among experts

If experts strongly diverge in their ratings, the assessment instrument likely lacks objective and clear criteria for rating, necessary for ensuring consistent results. In other words, the assessment instrument would not be reliable. This section looks at the diversity in experts' ratings, as it provides insights into inter-rater reliability.


Figure 3.7 shows the minimum and maximum average expert rating, as well as the standard deviation (SD) in these averages across the assessments. The average expert ratings are the means of each expert's ratings within an assessment. The figure shows that the highest variability in experts' overall judgements is in the numeracy domain (SD of 22.3 in 2016, 35.1 in 2021 and 24.0 in 2022), while SDs in literacy and science vary between 14.5 and 17.3. This reflects the strong disagreement in opinions in numeracy. In the first assessment round of the follow-up study with PIAAC, experts were uncertain how to interpret the scope of the numeracy tasks that an AI is supposed to master. Some assumed narrow tasks, while others focused on the entire range of tasks contained in the numeracy test. The result was two groups of experts with opposing opinions. Specifying the scope of tasks and clarifying the instructions for rating in the second round of the numeracy assessment resulted in agreement in the group discussion. However, there was still considerable variability in numerical ratings (OECD, 2023^[3]).

Figure 3.7. Divergence in experts' evaluations in different assessments

Minimum and maximum average expert rating and standard deviation of average expert ratings



Source: Adapted from (Elliott, 2017^[2]), *Computers and the Future of Skill Demand*, <https://doi.org/10.1787/9789264284395-en>, and (OECD, 2023^[18]), *Is Education Losing the Race with Technology?*, <https://doi.org/10.1787/73105f99-en>.

StatLink  <https://stat.link/zou86t>

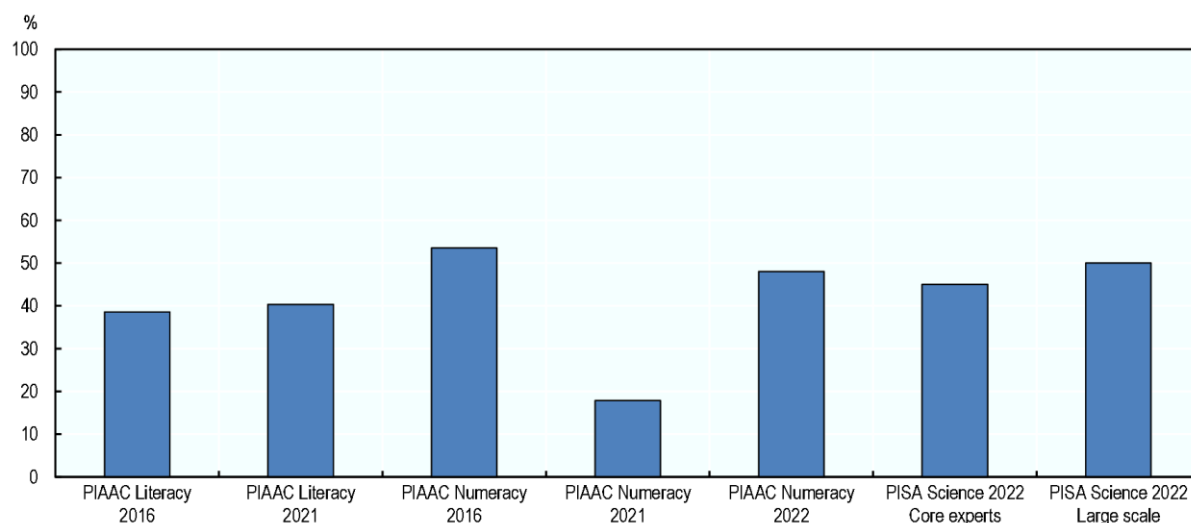
Uncertainty among experts

The degree to which experts are certain in their evaluations is instructive for the validity of measures. A high level of uncertainty among experts would suggest that the resulting indicators may not measure what they intend to measure. This would call for refining of assessment methodologies and the evaluation process.

Figure 3.8 shows the share of questions in each assessment that receive at least 20% of uncertain ratings. Uncertain ratings include the Maybe- and Don't know-categories. In the study with PISA, which uses a continuous scale for assessing experts' confidence, uncertainty is defined as confidence ratings in the 40-60% range. The share of questions receiving more than 20% of uncertain ratings is high in all assessments (between 39% and 54%). One exception is the first round of the follow-up numeracy assessment (18%). The 11 experts here expressed more certainty in their evaluations but had more opposing views on AI numeracy capabilities. Overall, the observations on experts' uncertainty show that obtaining valid ratings from experts is hard. Experts are not always knowledgeable about AI's potential to solve concrete tasks.

Figure 3.8. Share of questions that receive more than 20% of uncertain ratings in different assessments

Share of questions receiving at least 20% of Maybe-, Don't know-ratings or ratings within the 40-60% certainty range on PISA science questions



Source: Adapted from (Elliott, 2017^[2]), *Computers and the Future of Skill Demand*, <https://doi.org/10.1787/9789264284395-en>, and (OECD, 2023^[3]), *Is Education Losing the Race with Technology?*, <https://doi.org/10.1787/73105f99-en>.

StatLink  <https://stat.link/gno8v3>

Testing AI directly

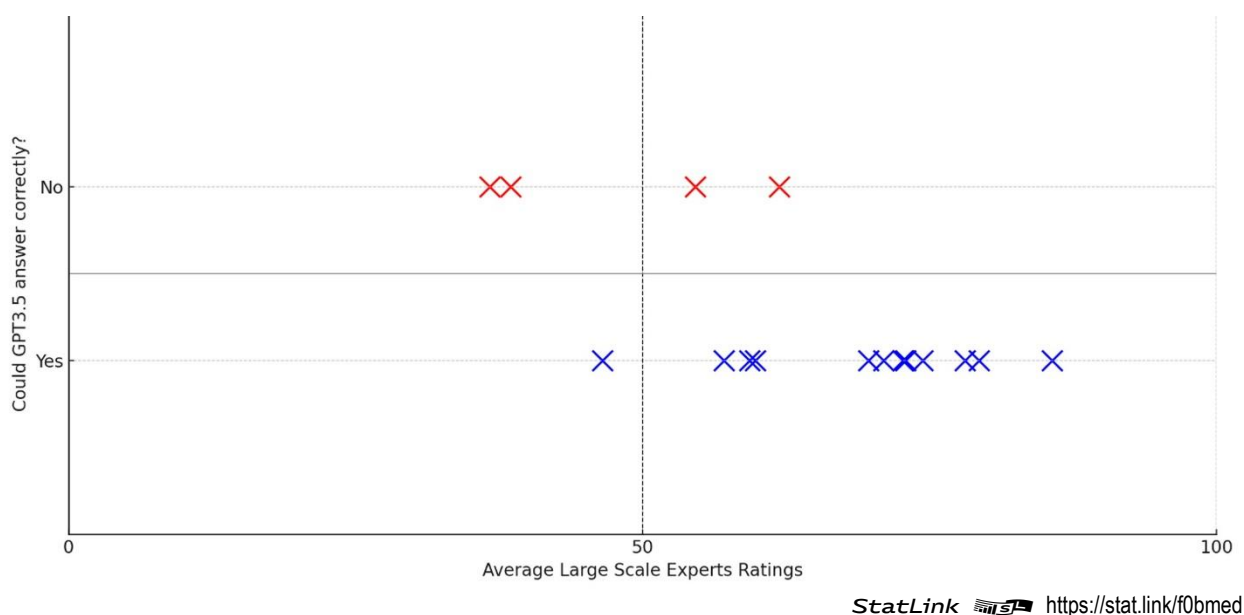
An effective way to evaluate the validity of the AI measures obtained from experts is to compare them to actual performance of state-of-the-art AI systems. In 2023, the project commissioned AI researchers to assess the performance of GPT-3.5 and GPT-4 on PISA reading, mathematics and science questions (OECD, 2023^[18]). Sixteen of the science questions were also evaluated by the experts. In the following, the expert ratings on these questions are compared to GPT-3.5 performance (Ye et al., 2023^[19]). This model was released in March 2022, before the assessment took place, and was an integral part of the first version of ChatGPT, released in November 2022.

GPT-3.5 performed well on easier questions, and less so on harder ones. It solved all of the test questions at Levels 1 and 2 and fewer questions at higher levels of difficulty. Overall, it could solve 12 of the 16 items. Figure 3.9 compares these results to the confidence ratings of experts. All questions that GPT-3.5 could solve have received high confidence ratings by experts, except for one. This latter question received an average rating of 47%, indicating uncertainty among experts regarding AI performance. Out of the four questions that GPT-3.5 could not solve, two were incorrectly rated as likely solvable. However, these ratings are again closer to uncertainty (55% and 62%).

Overall, these findings show that experts, although uncertain in many matters, can correctly assess the capabilities of current state-of-the-art of AI technology.

Figure 3.9. Experts' ratings of AI and GPT-3.5 performance on PISA science questions

Means of experts' ratings on 16 PISA science questions, by correct and incorrect response by GPT-3.5



Lessons learnt

The results of the online assessments and, above all, the discussions with experts revealed a number of challenges in using education tests to collect expert judgement on AI. The project team worked together with the experts to develop methodological solutions. This section outlines the major takeaways from this work.

Quantitative disagreement, qualitative agreement

It proved difficult to obtain coherent expert ratings on AI's capability to solve the education tests. Part of this difficulty related to differences in how experts interpreted the scope of the test questions that a hypothetical system is supposed to master. Discussions with experts showed that some were inclined to rate AI on subsets of similar test questions, similar to how AI systems are typically trained and evaluated in practice. Other experts evaluated current systems' capability to tackle all test questions in a domain at once. Overall, experts agreed that developing tailored solutions for narrow tasks is easier than developing general systems that can tackle all types of questions, including similar questions that are not part of the test. The project team attempted to specify the assumed scope of the test in order to translate this qualitative agreement into coherent numerical ratings.

The proposed new instructions for rating were tested in a second round of the follow-up study with the PIAAC numeracy test and in the study with PISA. The ratings obtained with this method from the assessment with PISA did not diverge strongly, as shown in Figure 3.7. However, the new framing has only partially reduced disagreement in the PIAAC numeracy assessment. That is, compared to the 11 experts who completed the first round of the follow-up numeracy assessment, the 4 experts in mathematical reasoning who re-assessed numeracy with the new framing showed lower – but still substantial – variation in their ratings.

The discussion with the four experts showed that they clearly interpreted the scope of tasks that a hypothetical system is supposed to solve from the instructions. However, they differed in the time frame

which they used for the evaluation. Experts were instructed to consider a hypothetical engineering effort to develop a system for the test using state-of-the-art AI techniques. The experts with the highest ratings argued that, given rapid advancements in the field, such an effort would produce the desired results within less than one year.³ By contrast, the expert with the lowest ratings focused on the current state of AI systems, which were not able to solve the numeracy test at the time. However, he agreed that systems will likely reach this stage within a year.

Overall, the methodological changes introduced in the rating exercise have increased clarity and consensus about AI capabilities. However, there is still need for improvement. The instructions need to reflect the fast pace of AI progress by using shorter time frames for rating. This would enable more precise evaluations of the state of the art of AI capabilities.

Literacy easier to rate than numeracy

Experts seemed at ease considering the application of NLP systems on the PIAAC literacy test. During group discussions, they noted that the literacy questions are similar to real-world tasks addressed by existing applications. In addition, benchmark tests used for evaluating NLP systems, such as the Stanford Question Answering Dataset (SQuAD) (Rajpurkar, Jia and Liang, 2018_[20]), often contain similar problems and tasks. Therefore, experts saw PIAAC as an appropriate tool for evaluating potential AI performance in language processing.

By contrast, the 11 experts who first rated AI in numeracy in the follow-up study 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. Until 2021, AI research had paid less attention to mathematical reasoning of AI because of its relatively lower applicability and commercial use. In addition, the mathematical tasks typically addressed in research – automated theorem proving and math word problems (i.e. quantitative problems stated in text), among others – are different from the ones in PIAAC. This made it challenging for experts to rate potential AI performance on the test. For the four experts in mathematical reasoning, the evaluation was easier due to their better understanding of the domain.

This suggests that expert knowledge elicitation on AI capabilities is more feasible in domains that are an established application domain in AI research. In less prominent or novel domains, experts have more limited information on research results and existing systems, unless they have specialised knowledge in the relevant domain.

More experts do not add value to the results

The study with PISA showed that the behavioural and mathematical approaches for expert knowledge elicitation produce similar results on AI capabilities. Expert ratings obtained with the mathematical approach also show similar variability compared to the ratings of the core experts. In addition, the similarity of these ratings with the performance of the contemporaneous GPT-3.5 system on the PISA science questions provides some evidence of their validity.

However, the advantages of this approach – obtaining robust measures that reflect the opinions of a large number of experts – do not outweigh its disadvantages. As described in Chapter 2, recruiting a large sample of experts proved challenging. Among the 189 computer scientists who were contacted by the project team, 63 expressed interest in participating in the survey, and only 33 actually completed it. Monetary incentives played a strong role in this process, suggesting that a repeated large-scale assessment of AI capabilities will be costly to implement.

As a result of these explorations, the project has chosen to rely on input from small groups of familiar experts for future activities involving expert knowledge elicitation. The study with PISA confirmed the robustness of this approach to the use of many experts.

The way forward

The exploratory studies using PIAAC and PISA have important implications for the methodology of the project. They identified limits to obtaining robust measures of AI capabilities by surveying experts. Consensus evaluations are hard to obtain, especially in domains that are not the centre of current research. This was the case for AI quantitative reasoning at the time of the PIAAC numeracy assessment. In addition, the assessment is time-consuming, both for the experts who need to invest several hours to provide ratings and participate in discussions, and for the project staff who devoted substantial time to recruit and engage experts. This led the project to test out the use of available direct measures of AI, which are discussed in Chapters 6 to 8.

However, expert judgement remains an indispensable part of the methodology. It is needed for reviewing, selecting and interpreting existing measures of AI. Measures obtained from expert evaluations can also complement the overall assessment framework in areas in which results from direct assessments of AI systems are lacking. For example, research interest and investment in automating particular tasks may be limited because their practical applicability and economic benefits may not be immediately clear. Other tasks may receive less research attention because they are still clearly out of the reach of current state-of-the-art technologies. Expert judgement can help fill such gaps by providing information on how far AI is from performing such tasks. In this way, the approach can contribute to a comprehensive assessment of AI capabilities across a wide range of human skills.

References


- Bolger, G. (ed.) (1992), *Perspectives on Expertise in the Aggregation of Judgments*, Plenum. [15]
- Bubeck, S. et al. (2023), “Sparks of Artificial General Intelligence: Early experiments with GPT-4”. [7]
- Clark, P. and O. Etzioni (2016), “My Computer Is an Honor Student — but How Intelligent Is It? Standardized Tests as a Measure of AI”, *AI Magazine*, Vol. 37/1, pp. 5-12, <https://doi.org/10.1609/aimag.v37i1.2636>. [9]
- 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>. [24]
- Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, <https://doi.org/10.1787/9789264284395-en>. [2]
- Frieder, S. et al. (2023), “Mathematical Capabilities of ChatGPT”. [8]
- Hendrycks, D. et al. (2020), “Measuring Massive Multitask Language Understanding”. [5]
- Hendrycks, D. et al. (2021), “Measuring Mathematical Problem Solving With the MATH Dataset”. [22]
- Lewkowycz, A. et al. (2022), “Solving Quantitative Reasoning Problems with Language Models”. [23]
- 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>. [3]
- OECD (2023), *Putting AI to the test: How does the performance of GPT and 15-year-old students in PISA compare?*, OECD Education Spotlights, No. 6, OECD Publishing, Paris, <https://doi.org/10.1787/2c297e0b-en>. [18]
- OECD (2019), *PISA 2018 Results (Volume I): What Students Know and Can Do*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/5f07c754-en>. [4]
- 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>. [1]
- OECD (2016), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264266490-en>. [13]
- OECD (2016), “PISA 2015 test items”, in *PISA 2015 Results (Volume I): Excellence and Equity in Education*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264266490-15-en>. [21]
- OECD (2013), *The Survey of Adult Skills: Reader’s Companion*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264204027-en>. [11]
- 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>. [10]
- OECD (2009), *Take the Test: Sample Questions from OECD’s PISA Assessments*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264050815-en>. [12]

- OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264040014-en>. [14]
- OpenAI (2023), *GPT-4 Technical Report*, <https://cdn.openai.com/papers/gpt-4.pdf>. [6]
- 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). [17]
- Rajpurkar, P., R. Jia and P. Liang (2018), "Know What You Don't Know: Unanswerable Questions for SQuAD". [20]
- Tversky, A. and D. Kahneman (1974), "Judgment under Uncertainty: Heuristics and Biases", *Science*, Vol. 185/4157, pp. 1124-1131, <https://doi.org/10.1126/science.185.4157.1124>. [16]
- Ye, J. et al. (2023), "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models". [19]

Annex 3.A. Analyses of the PIAAC and PISA studies using an alternative approach

Annex Table 3.A.1. List of online figures for Chapter 3

Figure Number	Figure Title
Figure A3.1	AI literacy performance in 2016 and 2021, following the “average” approach
Figure A3.2	AI numeracy performance in 2016 and 2021/22, following the “average” approach
Figure A3.3	AI performance on PISA science questions in 2022, following the “majority” approach

StatLink  <https://stat.link/glv4ed>

Notes

¹ See Note 2 in Chapter 1 of this volume.

² All items used in this AI assessment are sourced from the publicly released examples of the PISA 2006 and 2015 editions (OECD, 2009^[12]; OECD, 2016^[21]). The publicly released items contain limited information about students' performance on the questions. However, they include information on question difficulty and the sub-skills involved.

³ Prior to the assessment, which took place September 2022, the field of mathematical reasoning of AI has taken major steps. In 2021, the MATH dataset, a leading benchmark for mathematical reasoning, was released (Hendrycks et al., 2021^[22]). Between 2021 and 2022, several large language models fine-tuned for quantitative problems were launched (Lewkowycz et al., 2022^[23]; Drori et al., 2021^[24]). In addition, major AI labs were close to developing multimodal systems that can process both images and text. Experts referred to these developments, reflecting on the likelihood of AI solving the numeracy test in the near future.