# Executive summary

As artificial intelligence (AI) and robotics technologies continue to expand their scope of applications across the economy, understanding their impact becomes increasingly critical.

The AI and the Future of Skills (AIFS) project at OECD's Centre for Education Research and Innovation (CERI) is developing a comprehensive framework for regularly measuring AI capabilities and comparing them to human skills. The capability measures will encompass a wide range of skills crucial in the workplace and cultivated within education systems. They will establish a common foundation for policy discussions about AI's potential effects on education and work.

The AIFS project has undergone two phases of developing the methodology of the assessment framework. The first phase focused on identifying relevant AI capabilities and existing tests to evaluate them. It drew from a wealth of skill taxonomies and assessments across various disciplines, including computer science, psychology and education.

The second phase, the focus of this report, delves deeper into methodological development. It comprises three distinct exploratory efforts:

## Rating AI on education tests using expert judgement

Education tests offer a valuable means of comparing AI to human capabilities in domains relevant to education and work. The project carried out two studies to explore the use of education tests for collecting expert judgements on AI capabilities. The first study, conducted in 2021/22, followed up an earlier pilot study, asking experts to evaluate AI's performance on the literacy and numeracy tests of the OECD's Survey of Adult Skills (PIAAC). The second study collected expert judgements of whether AI can solve science questions from the OECD's Programme for International Student Assessment (PISA).

### *Purpose*

The studies aimed to refine the assessment framework for eliciting expert knowledge on AI using education tests. They explored different test tasks, response formats and rating instructions, along with two distinct assessment approaches: a "behavioural approach" used in the PIAAC studies, drawing on smaller expert groups engaging in discussions, and a "mathematical approach" adopted in the PISA study, relying more heavily on quantitative data from a larger expert pool.

### *Lessons learnt*

This work showed that there are limits to obtaining robust measures of AI capabilities by surveying experts. Especially in domains that are not the centre of current research, consensus evaluations are hard to reach. In addition, recruiting and engaging experienced experts is costly.

## Rating AI on occupational tests

Two exploratory studies extended the rating of AI capabilities to tests used to certify workers for occupations. These tests present complex practical tasks typical in occupations, such as a nurse moving a paralysed patient, or a product designer creating a design for a new container lid. Such tasks are potentially useful as a way of providing insight into the application of AI techniques in the workplace.

### *Purpose*

The inherent complexity of occupational tasks makes them different from the questions contained in education tests. Occupational tasks require various capabilities, take place in real-world unstructured environments and are often unfamiliar to computer scientists. Consequently, the project had to develop different methods for collecting expert ratings of AI with such tasks. The two studies explored the use of different survey instruments and instructions for collecting reliable and valid expert evaluations on these tasks.

### *Lessons learnt*

Rating AI performance on occupational tasks proved challenging. The rating difficulty related to predicting contextual factors that can potentially affect AI performance and to specifying the underlying capability requirements for each task. On the other hand, the studies suggested the possible use of occupational tasks for better anticipating how occupations might evolve as new AI capabilities emerge. As a result, occupational tasks will be used to understand the implications of AI for work and education rather than for gathering expert judgements on AI capabilities.

## Direct measures of AI capabilities

Recognising the limitations of expert judgement, the project initiated an exploration of measures derived from direct evaluations of AI systems. These benchmark tests offer a diverse range of evaluations but vary in quality, complexity, and target capabilities. To navigate this landscape, the project commissioned experts to explore their uses for the project.

### *Purpose*

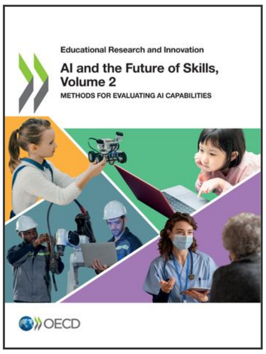
The project needed to find ways to select good-quality benchmarks, categorise them according to AI capabilities and systematise them into single measures. It commissioned experts to work on each of these tasks. Anthony Cohn and José Hernández-Orallo developed a method for describing the characteristics of benchmark tests to guide the selection of existing measures for the project. Guillaume Avrin, Swen Ribeiro and Elena Messina presented evaluation campaigns of AI and robotics and proposed an approach for systematising them according to AI capabilities. Yvette Graham reviewed major benchmark tests in the domain of natural language processing and developed an integrated measure based on the reviewed tests.

### *Lessons learnt*

Using direct measures to develop valid indicators of AI capabilities is a challenging but promising direction because of the large number and variety of direct measures available. The measures evolve rapidly as the field itself, which requires an approach to synthesising them conceptually. In addition, the measures often omit a comparison to human capabilities, which requires additional steps to add this reference. The preliminary work on direct measures suggests ways of addressing these two challenges.

## Future directions

Both expert judgements and direct AI measures are necessary to develop indicators of AI capabilities that are understandable, comprehensive, repeatable and policy relevant. The project’s third phase is working on a concrete approach for developing such indicators in different domains. This approach draws on experts to review, select and synthesise direct AI measures into a set of integrated AI indicators. These will be complemented with measures obtained from expert evaluations in areas where direct AI assessments are lacking. The resulting AI indicators will then be linked to measures of human competences and examples of occupational tasks to derive implications for education and work. They should aid decision-makers in determining necessary policy interventions as AI continues to advance.


Educational Research and Innovation
AI and the Future of Skills, Volume 2
METHODS FOR EVALUATING AI CAPABILITIES
OECD

**From:**

# AI and the Future of Skills, Volume 2

## Methods for Evaluating AI Capabilities

**Access the complete publication at:**

https://doi.org/10.1787/a9fe53cb-en

**Please cite this chapter as:**

OECD (2023), “Executive summary”, in *AI and the Future of Skills, Volume 2: Methods for Evaluating AI Capabilities*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/924ba903-en