

2 Eliciting expert knowledge: Methods and challenges

Abel Baret, Nóra Révai, OECD

Gene Rowe, Gene Rowe Evaluations

Fergus Bolger, Anglia Ruskin University

This chapter delves into the methodology of collecting expert judgement in the AI and the Future of Skills project. It provides an overview of the project's journey in refining its methodology and discusses associated challenges and considerations. The chapter begins by exploring the different methods of expert knowledge elicitation based on the research literature and discusses their relevance to the project. It then addresses key questions such as the number of experts required for reliable assessments, the framing of tasks for experts, and the aggregation and interpretation of expert judgements. The chapter concludes by offering points of consideration for the project's long-term trajectory.

This chapter reports on the journey of refining the project's methodology to collect expert judgement on the capabilities of artificial intelligence (AI). As its main approach to developing measures of AI capabilities, the AI and the Future of Skills (AIFS) project initially relied on the judgement of computer scientists to assess these capabilities based on questions in human tests. This idea originates in a need to support the policy community in planning education and employment policies with a sound knowledge of the progress in AI capabilities and how that compares to human skills.

The study chose to focus initially on human tests rather than on direct measures for several reasons. There are many direct measures of AI system performance (benchmarks, competitions, formal evaluation campaigns). However, these often measure performance on specific, narrow tasks. In addition, these are not synthesised into broader capability areas that would be meaningful for policy makers. These direct measures also miss certain skills that are important for humans and do not always allow for a comparison between machine and human performance. Therefore, the OECD decided to develop measures reflecting computer scientists' judgements using human tests as a first approach.

This approach requires establishing a robust methodology for collecting expert judgements that is valid and reliable, and ideally reflects a consensus of the expert community. Such a methodology involves recruiting and engaging the right experts, a well-established process for collecting expert judgement, a well-framed task for experts, an instrument (test questions or tasks) that allows computer scientists to assess AI capabilities correctly and a method that yields the consensual result of experts' judgements.

The precursor of the project was a pilot study in 2016 that asked computer scientists whether AI technology of the time and five years from then could answer the questions in the OECD's Survey of Adult Skills (in the Programme for International Assessment of Adult Competencies [PIAAC]) (Elliott, 2017^[1]). The pilot revealed several strengths but also some weaknesses in its methodology.

The project thus set out to consolidate its method to collect expert judgements in two main ways. First, it explored the literature on expert knowledge elicitation (EKE) and sought advice from experts in this methodological field. A meeting in March 2021 brought together experts to discuss the various methods for EKE and assess their relevance and feasibility for the AIFS project. Second, it conducted a series of exploratory studies in which the project tried out different methods to answer the following questions:

- Which EKE method is the most suitable to collect expert judgement on AI capabilities using human tests?
- How many experts are needed to obtain a reliable assessment of AI capabilities? How can they be identified, recruited and engaged?
- How does the task need to be framed so that experts have a unified understanding of the task and are able to provide a precise judgement of AI capabilities?
- How can we aggregate and interpret the results of expert judgement to obtain a single measure of AI capabilities?

In March 2022, the project held an expert meeting that discussed aspects of the methodology of collecting expert judgement. This included the overall framing of the task and the nature of information needed about the test used, as well as the specific instructions and response formats.

This chapter gives an overview of the different aspects of EKE based on the literature and discusses their application in the context of the project. It describes the evolution of methodologies across the exploratory studies along three main factors: the method of collecting expert judgements, the number of experts involved in the assessment and the framing of tasks for experts. The chapter then discusses the level of consensus in expert judgements on AI capabilities. It concludes with summarising the major developments and challenges in the methodology, offering a few points of consideration for the longer term. The subsequent chapters give details about the methods the project used in the series of exploratory studies.

Methods for eliciting expert judgement

Eliciting expert judgement has its own methodological literature referred to as EKE (O’Hagan et al., 2006^[2]) or Structured Expert Judgement (Cooke, 1991^[3]; Hanea et al., 2021^[4]). This area emerged from the necessity to supplement sparse or missing empirical, scientific evidence with expert judgement that can serve as the basis for decisions and policy making. EKE – defined as structured group techniques for the elicitation of judgements of uncertain quantities – is a relatively new area. However, it is based on earlier techniques for surveying and eliciting expert knowledge and group techniques [e.g. (Rowe, 1992^[5]; von der Gracht, 2012^[6]; Kahneman, Slovic and Tversky, 1982^[7]; Linstone and Turoff, 1976^[8])]. In the case of AI capabilities, the motivation to use expert judgement is due to the scattered and unstructured nature of available direct assessments, and their unsuitability for the policy community (see details in Chapter 9).

Behavioural and mathematical approaches to eliciting expert knowledge

EKE methods attempt to elicit judgements from experts that are as reliable and as valid as possible. This involves aggregating across several different opinions in a carefully managed process that helps reduce individual bias (e.g. resulting from beliefs and cognitive or social dispositions rather than scientific findings) and possible distortions resulting from group interactions. Further, quantitative judgements carry varying degrees of uncertainty, which are important to capture when informing policy decisions. In addition, EKE can elicit qualitative judgements from experts, either as input to decision making in their own right or to support quantitative judgements, for example, as rationales for them.

To inform decision making, a summary of judgements by groups of experts into a single estimate (or perhaps two or three if there are distinct schools of thought) is more useful than numerous individual judgements. Aggregation across multiple judgements also serves to reduce random error in those judgements. EKE techniques can use behavioural and mathematical aggregation of judgements or a mixture of the two to arrive at a single group judgement (O’Hagan et al., 2006^[2]).

- Behavioural aggregation involves interacting experts – facilitated or otherwise – coming to a consensus.
- Mathematical aggregation means averaging over different individual judgements. This can be done with equal weights given to each expert, or different weights (e.g. “performance weights” based on an assessment of individual expert ability).

The main difference between the two approaches is the degree of interaction between experts. In behavioural approaches, there is usually a high level of interaction among experts (either in a facilitated discussion or freely in a meeting, by e-mail or otherwise). In purely mathematical approaches, experts do not interact with each other (Rowe, 1992^[5]).

Behavioural aggregation can be applied to both qualitative and quantitative judgements and tolerates different schools of thought. If well-managed, it can allow experts to weight themselves in terms of their respective knowledge of an issue (e.g. by moving towards the positions of those with more expertise). However, if not well-managed, the process of behavioural aggregation can lead to biased outcomes resulting from social and cognitive biases such as group polarisation, overconfidence and groupthink (Lichtenstein, Fischhoff and Phillips, 1982^[9]; Myers and Lamm, 1976^[10]; Turner and Pratkanis, 1998^[11]). Mathematical aggregation with equal weights is simple but does not consider individual differences in expertise. Performance weighting has the advantage to account for such differences (Hanea et al., 2021^[4]). However, it has practical difficulties such as obtaining valid performance weights and may risk alienating the experts (Bolger and Rowe, 2015^[12]; Bolger and Rowe, 2015^[13]).

Behavioural and mathematical aggregation represent the two extremes in EKE. In between, other approaches combine both behavioural and mathematical elements. The main steps of different EKE

approaches are commonly recorded as protocols. Table 2.1 describes the major protocols that have been developed to collect expert judgement.

The protocols differ in the degree they use behavioural and mathematical aggregation and thus require varying degrees of interaction among experts. They also differ in the extent and nature of facilitation needed. Facilitators' skills can be key to the successful organisation and running of interactive group processes. They involve the ability to carefully guide discussions to include every issue intended for debate, avoid inserting their own viewpoints and ensure that discussions are not prematurely closed off. Facilitators also need to be able to involve all experts equitably and use continual summarising processes to confirm that all points are accurately understood and collated.

Table 2.1. Major EKE protocols

Group EKE protocol (and Reference)	Description	Aggregation type (MA: mathematical BA: behavioural aggregation)
One-shot surveys	A questionnaire for experts to complete individually, with responses usually averaged (with equal weighting) to indicate group judgement (and distributions used to indicate response variability).	MA
Classical method (CM) (Cooke, 1991 ^[3])	Experts are usually "tested" individually and then their judgements are combined mathematically and unequally according to performance weights based on testing results	MA
Delphi method (Linstone and Turoff, 1976 ^[8] ; Rowe, Wright and Bolger, 1991 ^[14])	Experts complete a survey anonymously and individually, receive the (summarised) responses from a facilitator and revise their responses. This can be repeated in further rounds. Delphi methods vary according to how they are precisely operationalised (e.g. Classical, Policy, Real-Time). Well-suited to online delivery	MA with equal weighting and varying degrees of BA.
Investigate-Discuss-Estimate-Aggregate (IDEA) (Hemming et al., 2018 ^[15])	As in CM, experts first individually make judgements of "seed questions" – for performance weighting – and the target questions. Next there is a (usually online) meeting of all experts with a facilitator to discuss the initial estimates and ensure a common understanding of the judgement task. Finally, the experts make judgements of target questions individually again, which are aggregated using the performance weights.	MA and some BA (although discussion primarily meant for problem clarification).
Nominal Group Technique (NGT) (Delbecq and Van de Ven, 1971 ^[16] ; Delbecq, Van de Ven and Gustafson, 1975 ^[17])	A facilitated group approach that allows face-to-face discussion with individual and anonymised estimations of the solution before and after discussion, and equal-weighted judgement of the final (post-discussion) estimates	BA and MA with equal weighting.
Facilitated group processes (e.g. Sheffield method) (Gosling, 2018 ^[18])	Interactive group processes that are generally held face-to-face (although real-time online processes are also possible). They rely on careful facilitation to ensure focused discussion and equal participant contribution, with the aim being group consensus.	BA

When determining which protocol to choose for a particular application, a number of factors need to be considered.

- Number of experts: behavioural methods are suitable for a small number of experts; mathematical approaches allow for collecting judgement from a large number of experts.
- Range of experts: heterogeneous expert groups (e.g. in terms of disciplinary background) favour a behavioural method where a facilitator can help overcome differences for example in knowledge base and language.
- Number of questions: mathematical aggregation is easier where there are a large number of questions.
- Nature (complexity) of questions: behavioural method is more suitable for complex tasks/questions that require substantive input to ensure a common understanding and more in-depth discussions.
- Nature of response: mathematical methods require quantitative response options sometimes complemented with qualitative responses (e.g. rationales); behavioural methods can be suitable for both quantitative and qualitative responses.

Additional considerations for the method include its cost, feasibility of recruiting and engaging experts, and feasibility of achieving consensus or establishing a single aggregate measure.

EKE methods used in the AIFS project

EKE in the AIFS project involves asking experts about whether AI can answer specific questions or carry out specific tasks. The pilot study, which used the OECD's PIAAC survey, opted for a facilitated face-to-face group discussion over two days. Such an extensive, in-depth discussion was necessary to elicit a feasible and meaningful framing of the rating task for experts, to identify difficulties and agree on the overall approach. However, this method has its trade-offs: it is expensive (travel and accommodation costs for all experts); it limits the number of experts able to participate in the exercise; it is time-consuming without much flexibility (experts cannot choose the best time for themselves to go through the 113 questions of the PIAAC survey); and it leaves little room and time for individual reflection.

For the more recent exploratory studies – an update using the OECD's PIAAC survey, a study using the OECD's Programme for International Student Assessment (PISA) test (see Chapter 3 for details), and studies using selected occupational tests and tasks (see Chapters 4 and 5) – the project tested other methods. In choosing the methods, the project considered the following:

- The selected human tests often involve different areas of computer science, such as natural language processing (NLP), computer vision and robotics. In addition, expertise in other disciplines, such as organisational and industrial psychology, is required to clarify what a test or task involves on the human side. Therefore, the expert group is relatively heterogeneous.
- Unlike many of the EKE tasks reported in the literature, this is not primarily a forecasting task. The capabilities of current technology are only available in a highly technical language, they are scattered, not systematised and evolve rapidly. A high level of expertise is necessary to be aware of and understand current AI capabilities. Projections for the future are generally based on ongoing research grants, which again requires expert knowledge and involvement in research and development.
- Some studies include many questions (PIAAC and PISA tests) and some are complex in nature (occupational tests). Most questions require expertise in several subdomains of computer science (e.g. computer vision and NLP).
- It is important to test whether using a small number of experts as opposed to a large number yields substantively different results.
- It is important to test the feasibility of different approaches in terms of costs, human resources, expert recruitment, etc.
- Reaching consensus is highly desirable given that the task is to gauge current computer capabilities, which should be knowable. While consensus among experts would also facilitate informing the policy community and drawing policy implications, it is vital to draw their attention to existing debates (dissensus) within the computer science community if these exist.

The COVID-19 pandemic (2020-21) prevented the project from organising face-to-face meetings, but made online meetings easier with improved platforms and people getting used to them.

Based on the above considerations, the project opted for testing a combination of mathematical and behavioural methods to elicit experts' judgement on AI capabilities. Table 2.2 summarises the methods used, and the number and background of experts involved.

Table 2.2. Methods used to collect expert judgement in the AIFS project

	EKE method	Experts
PIAAC 2016	Facilitated group discussion: <ul style="list-style-type: none"> • 2 days • In-person 	<ul style="list-style-type: none"> • N=11 • Computer scientists
PIAAC 2021 follow-up 1	Modified Delphi method: <ul style="list-style-type: none"> • Online survey (round 1) • Online group meeting • Online survey (round 2)* 	<ul style="list-style-type: none"> • N=11 • Computer scientists, Cognitive and I/O psychologists
PIAAC 2022 follow-up 2	Modified Delphi method: <ul style="list-style-type: none"> • Online survey (round 1) • Online group meeting • Online survey (round 2)* 	<ul style="list-style-type: none"> • N=4 • Computer scientists
PISA 2022 core experts	Modified Delphi method: <ul style="list-style-type: none"> • Online survey • Online group meeting 	<ul style="list-style-type: none"> • N=12 • Computer scientists, Cognitive and I/O psychologists
PISA 2022 new experts	Online survey	<ul style="list-style-type: none"> • N=170 invited • R=33 respondents • Computer scientists
Occupational 2022	Modified Delphi method: <ul style="list-style-type: none"> • Online survey • Online group meeting 	<ul style="list-style-type: none"> • N=12 • Computer scientists, I/O psychologists

Note: *Completing the same survey again to modify initial judgements was offered to experts, but none of them actually did it. This option was thus dropped from subsequent studies.

With respect to working with a heterogeneous expert group on complex test questions, the project found a **modified Delphi method** as the most appropriate approach for most of the assessments.

Delphi is a structured group technique that consists of at least two rounds of surveys collecting experts' ratings, with feedback on the ratings provided between rounds. The iteration of survey rounds continues until consensus among experts is reached. During each round, experts provide their ratings anonymously and independently from each other. This should reduce potential bias from social conformity or from dominant individuals who impose their opinions on the group. By contrast, the feedback provided after each round should enable social learning and the modification of prior judgements due to new information. This feedback should ultimately increase consensus between experts.

Designing the appropriate method needs to take into account the features of the task of assessing current AI systems' capabilities (described above). Importantly, a range of specialised knowledge is required in sub-fields of AI. For some tests (e.g. PIAAC literacy and PISA reading), all experts are generally aware of the current state of the art in relevant AI domains. Other tests, such as rating occupational tasks, require more specialised knowledge (e.g. in robotics). In either case, individual experts cannot possibly know all existing AI applications, recent research results or other details that may be relevant for the evaluation. For example, only one or a few experts may have knowledge on particular AI systems that can perform a task. To facilitate consensus, experts should be able to communicate such information to the group at any point of the rating process.

For this reason, in contrast to a classical Delphi approach, a high degree of interaction among experts is more suitable for assessing AI capabilities on tests/tasks. Thus, after the first round, the project organised a three-hour online meeting in all exploratory studies. This meeting allowed experts to discuss the feedback they received on the survey results, exchange ideas and share references to recent research results. After the meeting, experts were invited to revise their judgements provided in the survey based on the group discussion.

Overall, the experts and the project team were satisfied with the modified Delphi method in at least two regards. Experts appreciated exchanging references and discussing ideas. Meanwhile, the project team could elicit the group's overall assessment of AI capabilities. This was true even if there was disagreement in the ratings experts provided through the survey.

However, one key feature of the Delphi method did not prove feasible. Although the project team asked experts to revise their responses if their views have changed, they did not go back to the survey to modify their ratings after the meeting. This may be because the interactions during the meeting provided an opportunity for experts to explain their judgements and reconsider them in light of a better understanding of the scope of the task. On numerous occasions, they expressed how they would modify their judgement with the new understanding at the meeting. The experts had numerous test questions to review, and it took time to provide judgements on each of them. Consequently, it was generally not practical to push them to do more than a one-time survey and one meeting given that they often provided a modified judgement at the meeting already. As a result, the quantitative (mathematical) aggregation of expert judgements needed to be complemented by the qualitative aggregation resulting from the meeting (see Chapters 3 and 5 for more details).

Large-scale experiment: How many experts can be engaged and through what incentives?

The pilot study with PIAAC, as well as its follow-up, relied on a core group of 10-15 experts who have worked closely with the project team from the outset. To test the feasibility of involving substantially more experts and if this would yield different and/or more robust results, the project conducted a large-scale version with a different assessment – the PISA science assessment (see Chapter 3). Having a large sample of computer scientists willing to invest substantially in providing judgements was expected to be challenging. The team thus tested different strategies in approaching and engaging experts. The goal of the experiment was to answer the following questions:

- How many experts can be identified and contacted within a limited timeframe?
- What response rate can be expected?
- Is an incentive necessary to engage experts? And if so, which one is the most effective?

Recruiting and engaging experts: Outreach and incentives

The first challenge was to identify a large number of experts with the appropriate background. The list of experts had to cover all relevant domains of computer science (e.g. NLP, computer vision, reasoning) and demonstrate diversity, in particular with respect to gender and geographical coverage. To compile the list, the project used snowballing, starting with recommendations from its already engaged small group of experts (henceforth “core experts”). In addition, the team identified and scanned the webpages of relevant research laboratories, conference attendee lists, and public and private organisations. Some 170 experts were selected (of whom 119 were recommended by our core experts) and contacted. In addition, the project reached out to 19 graduate students. Overall, the final list included 111 males and 59 females and covered 19 countries.

The second challenge was to convince the experts to participate in the study, i.e. to try to achieve a high response rate. The project tested different incentives to determine the most effective way to engage experts. All graduate students were offered a EUR 250 honorarium. Meanwhile, the experienced computer scientists were randomly distributed in four groups of 11 participants that had different incentives to complete the survey:

1. *Honorarium group*: receiving an EUR 800 honorarium;

2. *Co-authorship group*: offered to be co-authors of a future report;
3. *Honorarium + Co-authorship group*: receiving both incentives;
4. *No incentive group*¹.

To reach out to experts, the project used a foot-in-the-door technique. As such, it drew on evidence that people are more likely to agree to a request when they have already made a commitment to a similar action (Freedman and Fraser, 1966^[19]). The first e-mail briefly presented the project, and issued invitations to participate in a survey (without giving many details) and to join the project community. The e-mail also mentioned the name of the core expert who recommended them, when applicable. It asked about experts' interest to learn more about the project and the survey. To experts recommended by the project's core experts, the first e-mail also offered an online call to discuss the project. Experts who answered the first e-mail and expressed interest received a second e-mail with detailed information about the survey and their respective incentives.

Results: Response rate and effects of incentives

Table 2.3 shows the response rates. A quarter of the targeted experts showed initial interest (i.e. responded to the first e-mail); 77% of these respondents were experts recommended by the project's core experts. This shows the importance of snowballing and referrals. Slightly less than the half of experts who showed initial interest actually completed the survey. Most experts who did not complete the survey informed us of their withdrawal, and typically referred to lack of time or interest in the survey. Among them, 15 nonetheless expressed interest in meeting with the team to learn about the project.

Table 2.3. Response rate

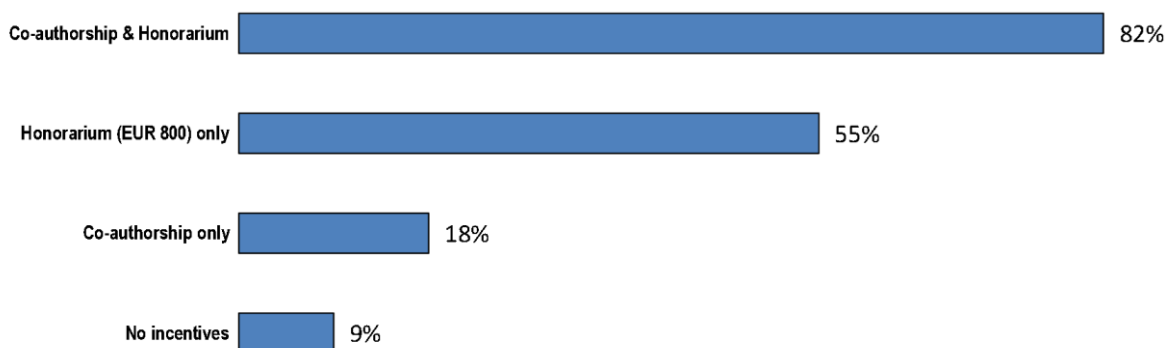
	Experienced experts		Graduate students	Total
	Recommended by core experts	Identified		
Total number targeted	119	51	NA*	189+
Answered the first e-mail (as percent of those targeted)	34 (28.6%)	10 (19.6%)	19	63
Filled in the survey	13	5	15	33
Final response rate (with respect to those who answered the first e-mail)	38%	50%	79%	52%


Note 1: *Graduate students were initially reached via a form sent to university forums and the project's core experts.

Note 2: Statistical tests (chi-square) did not yield a significantly different response rate between the group recommended by experts and those identified by the team.

Unsurprisingly, the combination of incentives had the strongest effect on completion, while the group receiving no incentives had a low response rate (Figure 2.1). Clearly, money matters most. More than half of participants from the Honorarium group completed the survey, as opposed to less than one in five among those offered co-authorship.

Figure 2.1. The effect of incentives on the final response rate



StatLink  <https://stat.link/g5k6i0>

At the end of the survey, experts were asked about their motivation to answer via a multiple choice of four options (Figure 2.2). The “*interest in the nature of the assessment and the test items*” appeared to be the strongest self-reported motivation factor, whereas *co-authorship opportunity* was the least important self-reported factor.

Figure 2.2. Self-reported motivation to complete the survey



StatLink  <https://stat.link/chife0>

In sum, the experiment highlighted several challenges of engaging a large number of experts in such a time-consuming activity. First, it is difficult to identify many experts with such specific expertise. Crowdsourcing experts would not ensure the level of expertise needed for the task. Second, it is difficult to engage them in completing a long and complex survey. Offering money (if possible, together with co-authorship) may ensure acceptable response rates. However, this is obviously a costly measure, especially with large samples.

Computer scientists with the high level of expertise required for this task are generally very busy. In addition, those working in industry often do not feel comfortable participating in such an exercise because they are bound by business secrecy. Moreover, the financial incentives might need to be larger to be effective for this group.

Survey length could in principle be reduced using incomplete block design, i.e. each respondent only answers a smaller subset of the questions. However, this method requires a large sample of experts to ensure that missing values can be reliably estimated. Many questions require several domains of expertise at the same time (e.g. they include a visual element and require language understanding necessitating

expertise in both computer vision and NLP). Therefore, sorting questions based on subdomain expertise is not possible.

Conducting a large-scale survey repeatedly to collect expert judgement in the domain of AI capabilities is therefore highly challenging, if feasible at all.

Task framing used to collect expert judgement on AI capabilities

This section discusses the challenge of framing the rating task for experts in a way that ensures common understanding and reliable assessment of AI capabilities.

Task framing and instructions

The pilot study that used the PIAAC test asked experts to give a rating (Yes, No or Maybe) of AI systems' ability to solve each test question after one year development and a cost limit of USD 1 million. The latter parameters were defined in the meeting of the pilot study (Elliott, 2017^[1]) to specify what it means to rate current technology even if no off-the-shelf system is available.

The same instruction was kept for the first exploratory study that updated the PIAAC pilot in 2021. However, some limitations of this instruction emerged. These included experts interpreting the scope of abilities covered in the assessments differently. For instance, some experts focused on AI systems' narrow ability to answer the given set of questions, while others imagined that the questions were representative of a broad underlying capability (see Chapter 3 and OECD (2023^[20])). Some also judged the USD 1 million parameter as unrealistic with regard to commercial AI development projects in the field. These limitations suggested the need for developing a finer framing for the assessments.

To address the above concerns, the project team created a **framework document** for the subsequent exploratory studies that gives more details on the assessments and describes the characteristics of the test questions:

- what the test measures in terms of human skills (e.g. literacy skills)
- how the test measures this skill (e.g. multiple choice questions about simple comprehension of a text)
- factors affecting question difficulty (e.g. interpretation required)
- scoring rubrics used to evaluate test takers' performance.

The framework document also included examples of test items to give experts a sense of what to expect from the assessment. The examples can be considered a representative set of training data that define the scope of abilities. This helps experts imagine a machine learning system that could be developed.

Instructions were also changed to account for the problems mentioned above (see Box 2.1 and Table 2.4 for the evolution of instructions and task framing). In particular, the description of “current computer techniques” changed and the scope of abilities was specified through examples. The prompt to imagine an AI system that answers the questions clarified that we need one integrated system as opposed to fine-tuned systems for each question.

Box 2.1. Evolution of assessment instructions in the AIFS project

Extract from PIAAC 2016 and 2021 assessment instruction

[...] You will be asked to evaluate the capacity of AI technologies to correctly answer the PIAAC questions. In making your judgement, please consider the following:

- Please consider “current” computer techniques, meaning any available techniques that have been addressed in the literature that we can describe their capabilities and limitations.
- Please consider techniques that might need “reasonable advance preparation”, [...] thinking about a development team receiving detailed information about the types of questions included in the test and being given one year and USD 1 million funding to build and refine a system to work with such questions using current techniques.

Extract from PISA 2022 and PIAAC follow-up assessment instruction – Imagined AI system

[...] The questions are presented in different formats (including pictures, texts and numbers) and are designed to resemble real-life tasks in work and personal life. You will be asked to:

- briefly describe a high-level approach for an AI system built to answer the questions on the PISA science assessment
- evaluate the likely performance of that AI system on different questions from PISA.

To help you understand the domain, a document describing the framework for the PISA science test and providing a set of ten example questions was provided to you beforehand. [...]

In designing the high-level approach for your AI system, you had to consider any “current” computer techniques [...]. The point is that the design for your imagined AI system should involve the application of existing AI techniques, not research to develop new approaches.

Extract from Occupational tasks assessment instruction

[...] You will be asked to evaluate the capacity of AI technologies to carry out several occupational tasks. In making your judgement for each task:

- Please consider “current” computer techniques [...].
- Please consider techniques that might need “reasonable advance preparation”. You can consider possible AI systems involving any level of development effort as long as the work involves established AI techniques.

Analyses of the results and comments obtained from experts in the exploratory studies highlighted a better understanding of the task and the type of AI systems they should consider than in previous studies. However, group discussions and feedback on ratings were still necessary to remove remaining misunderstandings and share additional precisions on the AI systems they envisaged.

Question phrasing and response format

The project also explored different possibilities of asking the questions and response formats and their implications for the reliability of experts’ judgements, and the analysis and interpretation of data. The expert meeting organised in March 2021 discussed the advantages and disadvantages of:

- simple categorical questions (Yes/No/Maybe)
- Likert scale questions with probabilities of whether AI systems can solve the question (with or without detailed rubrics for each level on the scale)

- open-ended questions to elicit the rationale of expert judgements.

In addition, the project considered ways to elicit experts' confidence in their judgement, which can be important information to communicate to the policy community. Experts at the March 2021 meeting (including computer scientists and psychologists with survey expertise) endorsed the use of a scale that simultaneously captured experts' judgement of AI capabilities and confidence in their judgements. The question "*How confident are you that your AI system could carry out this task?*" with a **Likert or continuous scale of probabilities** and a "Don't know" option received overall positive feedback from experts. The analysis of the quantitative results and how the subsequent meeting helped finetune experts' judgement and increase levels of certainty, are discussed in Chapter 3.

Experts also agreed to provide **rationales and comments** following their answers. This allowed them to express uncertainty and complement their answers with clarifications and/or references. Such qualitative information was also valuable for the team to better understand quantitative judgements and to prepare the group discussions following the ratings. Table 2.4 summarises the instructions and response formats used across the exploratory studies.

Table 2.4. Task framing and response format in the AIFS project

	Task framing and instructions	Response format (scale)
PIAAC 2016	<ul style="list-style-type: none"> • No framework document • USD 1M + 1 year development 	<ul style="list-style-type: none"> • Yes / No / Maybe • No rationale
PIAAC 2021 follow-up 1	<ul style="list-style-type: none"> • Framework document • USD 1M + 1 year development 	<ul style="list-style-type: none"> • Discrete probabilities (0%; 25%; 50%; 75%, 100%) • Rationale
PIAAC 2022 follow-up 2	<ul style="list-style-type: none"> • Framework document • Imagined AI system based on existing techniques 	<ul style="list-style-type: none"> • Discrete probabilities (0%; 25%; 50%; 75%, 100%) • Rationale
PISA 2022 core experts	<ul style="list-style-type: none"> • Framework document • Imagined AI system based on existing techniques 	<ul style="list-style-type: none"> • Discrete probabilities (0%; 25%; 50%; 75%, 100%) • Rationale
PISA 2022 new experts	<ul style="list-style-type: none"> • Framework document • Imagined AI system based on existing techniques 	<ul style="list-style-type: none"> • Continuous probabilities (0-100%) • Rationale
Occupational 2022	<ul style="list-style-type: none"> • Framework document • Current techniques with reasonable advanced preparation 	<ul style="list-style-type: none"> • Discrete probabilities (0%; 25%; 50%; 75%, 100%) • Rationale

Finally, to track technological advances and evolution over time, the 2016 pilot study asked experts about the projected capability of AI systems solving similar tasks over the short term (5 years) and long term (10 to 20 years). Experts felt more confident about short-term projections as they could link them to ongoing research projects. In addition, grant applications typically require five-year projections (Elliott, 2017^[11]). Projections provide comparative data (an assessment in five years can be compared to projections), particularly for longer periods. Questions about future AI capabilities in the exploratory studies were limited to a five-year projection for the PIAAC 2021 follow-up 2022 assessments.

Establishing consensus: Quantitative disagreement versus qualitative agreement

As one of its most important objectives, the AIFS exploratory studies tested whether and with what method it is possible to establish consensus among experts. Consensus can be an indicator of data quality and the usefulness of expert judgement to inform policy decisions.

Consensus or agreement among experts can be measured via quantitative and qualitative methods (for a full review, see von der Gracht (2012^[6])). The AIFS project primarily used simple mathematical aggregations and comparisons of ratings as quantitative methods:

- Simple and two-thirds majority: more than half (or two-thirds) of experts gave the same rating (e.g. said “Yes, AI can solve this task”). Can be adapted to discrete or continuous scales by setting a threshold-point for decision.
- Interquartile range (IQR), standard deviations, coefficients of variation: measures of dispersion. The higher the value is, the more ratings are spread around the mean or median. A low value can be an indicator of consensus across experts’ ratings.

These simple measures provided an effective way to compare experts’ ratings across the different assessment scales and allowed for a straightforward analysis and interpretation of results. Other, more complex measures can be used, such as Kappa and Kendall’s coefficient of concordance. Kappa in the exploratory studies generally indicated low levels of agreement.

The project revisited methods for collecting and aggregating expert judgement to increase consensus. This was only partially achieved from the 2016 pilot study to the 2021/22 assessments: agreement on the literacy questions increased but not on the numeracy questions. Importantly, there was still no overall consensus among experts: the ratings showed considerable variations (see Chapter 3 and Chapter 5) in any of the exploratory studies. This could be partly due to the difference between raters’ domain of expertise, which affects their judgements. For example, NLP experts might not be aware of all the technological advances in the vision domain. This, in turn, could negatively bias their judgements on questions involving computer vision. As another explanation for lack of consensus, information included in the task framing for experts was still not enough for a common understanding of how AI capabilities on the questions should be rated. Although additional information in task framing could help increase consensus, there are practical limits in the amount of preparatory information that respondents are willing to review when the rating task itself is already quite long.

Group discussions and an *analysis of experts’ rationales* have provided substantive qualitative data to understand consensus/dissensus among experts and identify their reasons. In the 2016 PIAAC group discussion, experts agreed on common challenges of current AI systems, such as the difficulty to deal with multimodal questions or the likely overfitting of systems (i.e. systems are fine-tuned to solve a specific set of questions) (Elliott, 2017^[1]). The 2021/22 follow-up assessments showed a stronger consensus on several aspects of AI state of the art that became apparent in the group discussions. In the PIAAC 2021 repeat, the quantitative analysis of expert ratings showed disagreement across experts on AI capabilities to solve the numeracy questions. However, the rationales provided in the survey and the group discussion showed overall agreement about AI systems’ capabilities to solve the PIAAC numeracy questions (OECD, 2023^[20]).

The discrepancy between the level of consensus in the quantitative and qualitative analysis of experts’ judgements can be largely explained by two factors. First, the limitations of task framing described above (see Chapter 3). Second, the differences in computer scientists’ domain expertise and knowledge of the latest AI performance measures. Experts tended to base their judgements on the direct measures of AI performance that they know and that are relevant to the given set of questions. Naturally, one expert cannot know all the thousands of such measures and cannot follow their rapid evolution. However, interactions during the follow-up meeting allowed them to exchange references and reconsider their judgement in view of the evidence shared by others.

Overall, the exploratory studies have shown that despite several revisions and improvements, reaching consensus was not possible based on a purely quantitative analysis of expert judgements. When quantitative analysis was complemented with qualitative information, however, a global consensus was possible in most cases.

Conclusions: Challenges and future directions

This chapter described the processes and methods for engaging experts and collecting their judgement on AI capabilities using human tests/tasks. The refinements to the methodology since the pilot study explored only a small set of configurations discussed in the literature. Nevertheless, the explorations highlighted some key challenges of this approach.

First, collecting expert judgements on such complex assessments proved to be more resource-intensive than anticipated. This was true both in terms of financial costs and the time commitment required from experts and the project team. Part of the problem is the natural limit on the number of people with both the appropriate expertise and the interest to engage with this work. Identifying enough experts and raising their interest to engage with this work require substantial time from the project team. The project tested various incentives to engage experts and found that some (particularly money and a combination of several incentives) work but are costly.

The project's goal is to regularly update the measures of AI capabilities once they are developed to inform the policy community. Using human tests/tasks, such as the OECD's educational tests (PIAAC, PISA) and occupational tasks, means the project would need to collect expert judgements regularly (e.g. every two-five years). The methodological explorations described above indicate this will be very difficult, if feasible at all with a large number of experts given the limited interval between assessments and the resources available. On the positive side, the EKE literature and the exploratory assessments suggest a smaller group of experts' judgements gives similar aggregate results to that of a larger group (see Chapter 3). However, the team has recognised that engaging even a smaller group of experts on a regular basis would be substantially more time-consuming and expensive than originally believed.

Second, it is challenging to formulate tasks that provide valid and reliable expert judgement and yield an acceptable level of quantitative consensus in cases where experts agreed qualitatively. The project worked with experts to reflect on the task framing and instructions and improved its methodology through multiple exploratory assessments. Despite trying several different techniques and achieving expert agreement on qualitative descriptions of current AI capabilities, the project could not get adequate agreement in experts' quantitative judgements of those capabilities.

Overall, the explorations concluded that using expert judgement to establish measures of AI capabilities has limits. The project therefore began to explore using direct measures of AI systems originating from benchmark tests, competitions and formal evaluations. This seemed to be a natural choice for two reasons. First, a huge amount of such measures exists in the field of computer science and they are constantly growing in number. Second, experts participating in the exploratory studies continuously referred to such direct measures when making their judgements. Thus, it was straightforward to rely directly on these measures instead of their judgements. Despite the shift of focus from expert judgement to synthesising direct measures, the former remains relevant in certain domains where direct measures are not available.

Alternative pathways to develop AI measures led the project to rely on experts in different ways. Experts' role in identifying and interpreting the results of available direct measures became stronger than providing their judgements about likely performance on specific tasks. New roles involve experts from different domains to work together, build a shared understanding of the project goal and collectively develop tools. Examples of such co-construction are the development of the facets to characterise benchmarks (see Chapter 6), the classification of formal evaluation campaigns (see Chapter 7), the design of the occupational assessment (Chapter 5) and the development of AI capability scales (Chapter 9).

Over the past three years, the AIFS project has developed a core group of committed experts and a set of methods that allow for obtaining valid and reliable expert judgements across several domains. The rest of the report will describe in more detail the exploratory assessments and other approaches to summarise the state-of-the-art AI capabilities.

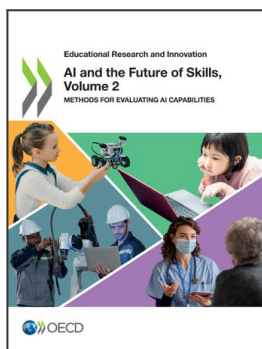
References

- Anderson, B. (ed.) (2018), “A practical guide to structured expert elicitation using the IDEA protocol”, *Methods in Ecology and Evolution*, Vol. 9/1, pp. 169-180, <https://doi.org/10.1111/2041-210x.12857>. [15]
- Bolger, F. and G. Rowe (2015), “The aggregation of expert judgment: Do good things come to those who weight?”, *Risk Analysis*, Vol. 35/1, pp. 5-11, <https://doi.org/10.1111/risa.12272>. [12]
- Bolger, F. and G. Rowe (2015), “There is data, and then there is data: Only experimental evidence will determine the utility of differential weighting of expert judgment”, *Risk Analysis*, Vol. 35/1, pp. 21-26, <https://doi.org/10.1111/risa.12345>. [13]
- Cooke, R. (1991), *Experts in Uncertainty: Opinion and Subjective Probability in Science*, Oxford University Press. [3]
- Delbecq, A. and A. Van de Ven (1971), “A group process model for problem identification and program planning”, *The Journal of Applied Behavioral Science*, Vol. 7/4, pp. 466-492, <https://doi.org/10.1177/002188637100700404>. [16]
- Delbecq, A., A. Van de Ven and D. Gustafson (1975), *Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes*, Scott Foresman and Company. [17]
- Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, <https://doi.org/10.1787/9789264284395-en>. [1]
- Freedman, J. and S. Fraser (1966), “Compliance without pressure: The foot-in-the-door technique”, *Journal of Personality and Social Psychology*, <https://doi.org/10.1037/h0023552>. [19]
- Gosling, J. (2018), “SHELF: The Sheffield Elicitation Framework”, in *International Series in Operations Research & Management Science, Elicitation*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-65052-4_4. [18]
- Hanea, A. et al. (eds.) (2021), *Expert Judgement in Risk and Decision Analysis*, Springer Cham, <https://doi.org/10.1007/978-3-030-46474-5>. [4]
- Kahneman, D., P. Slovic and A. Tversky (eds.) (1982), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press. [7]
- Lichtenstein, S., B. Fischhoff and L. Phillips (1982), “Calibration of probabilities: The state of the art to 1980”, in *Judgment under Uncertainty*, Cambridge University Press, <https://doi.org/10.1017/cbo9780511809477.023>. [9]
- Linstone, H. and M. Turoff (eds.) (1976), *The Delphi Method: Techniques and Applications*, Addison-Wesley, <https://doi.org/10.2307/3150755>. [8]
- Myers, D. and H. Lamm (1976), “The group polarization phenomenon.”, *Psychological Bulletin*, Vol. 83/4, pp. 602-627, <https://doi.org/10.1037/0033-2909.83.4.602>. [10]
- 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>. [20]

- O'Hagan, A. et al. (2006), *Uncertain Judgements: Eliciting Expert Probabilities*, John Wiley. [2]
- Rowe, G. (1992), "Perspectives on Expertise in the Aggregation of Judgments", in G. Wright and F. Bolger (eds.), *Expertise and Decision Support*, Plenum, https://doi.org/10.1007/978-0-585-34290-0_8. [5]
- Rowe, G., G. Wright and F. Bolger (1991), "Delphi: A reevaluation of research and theory", *Technological Forecasting and Social Change*, Vol. 39/3, pp. 235-251, [https://doi.org/10.1016/0040-1625\(91\)90039-j](https://doi.org/10.1016/0040-1625(91)90039-j). [14]
- Turner, M. and A. Pratkanis (1998), "Twenty-five years of groupthink theory and research: Lessons from the evaluation of a theory", *Organizational Behavior and Human Decision Processes*, Vol. 73/2-3, pp. 105-115, <https://doi.org/10.1006/obhd.1998.2756>. [11]
- von der Gracht, H. (2012), "Consensus measurement in Delphi studies", *Technological Forecasting and Social Change*, Vol. 79/8, pp. 1525-1536, <https://doi.org/10.1016/j.techfore.2012.04.013>. [6]

Notes

¹ To ensure ethical treatment, after completing the survey, all respondents received both the EUR 800 honorarium and co-authorship.



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:

Baret, Abel, *et al.* (2023), "Eliciting expert knowledge: Methods and challenges", in OECD, *AI and the Future of Skills, Volume 2: Methods for Evaluating AI Capabilities*, OECD Publishing, Paris.

DOI: <https://doi.org/10.1787/4b900c5c-en>

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Extracts from publications may be subject to additional disclaimers, which are set out in the complete version of the publication, available at the link provided.

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