8. Abilities and skills: Assessing humans and artificial intelligence/robotics systems

Phillip L. Ackerman, Georgia Institute of Technology

This chapter makes recommendations towards an approach for comparing human and artificial intelligence (AI) capabilities. It stresses the need to compare domain knowledge and skills rather than broad, higher-order abilities such as intelligence. The chapter provides the theoretical and empirical foundations of domain knowledge and skills assessments. It discusses methodological challenges arising from humans' use of tools, differences in learning between humans and AI, and the inaccuracy of skills assessments at high performance levels. Finally, the chapter proposes an assessment strategy that draws on tests developed for jobs subject to licensing examination.

Introduction

This chapter argues the pedagogical rather than the psychological method of assessment holds more promise for assessing the respective capabilities of human and artificial intelligence (AI)/robotics systems. A psychological approach assesses underlying abilities identified from research on human intelligence. Conversely, a pedagogical approach seeks to understand and assess specific knowledge and skills of an individual that allow for the successful (or unsuccessful) accomplishment of real-world tasks. The latter approach will usefully allow a contrast between the relevant respective capabilities of both human and AI/robotics systems. These knowledge and skills repertoires typically have both declarative knowledge and procedural knowledge components. In many cases, they also have tacit knowledge involvement [see (Polanyi, 1966/1983[1])].

The chapter provides the theoretical and empirical foundations of such assessments, especially in the domain of certification tests. Different types of knowledge, along with issues of tool use, learning and differentiating between competence and expertise are discussed. Finally, it proposes a strategy for developing a sampling of tests and tasks for comparative assessments.

Intelligence assessment from psychological and pedagogical methods

The mainstream theory and application of human intelligence/abilities assessment are fundamentally mismatched with assessing the adequacy of AI/robotics systems to replace human systems. This mismatch derives from how psychologists have explored the taxonomy of human abilities and used ability assessments to predict individual differences in learning and task/job performance.

Starting with Binet and Simon (1961[2]) in 1905 and 1908, and even with Spearman (1904[3]), abilities researchers and practitioners have primarily sought to determine the factors underlying human capabilities and limitations to help develop respective tests and measurements. The results could then be used to predict individual differences in success or failure in *future* academic and occupational situations. Binet was interested in predicting failures of children in overall academic performance. Spearman, in contrast, wanted to determine the fundamental characteristics of individual differences in a general intellectual ability – sometimes called a “mental engine”.

Tests and measurements derived from both approaches’ attempt to assess the underlying abilities that give rise to individual differences in learning and performance. However, they say little about whether an individual can perform any particular task beyond the tests themselves. For example, Spearman’s advocates commonly use the Raven’s Progressive Matrices test. This measure requires inductive reasoning with non-verbal test content [e.g. Burke (1958[4])]. The test has obvious limitations for predicting academic or job performance [e.g. Vernon and Parry (1949[5])]. Moreover, the test scores provide no useful information about the individual’s knowledge and skills for any other task.

Broad intelligence tests, such as the Stanford-Binet or one of the Wechsler tests, have some advantages over the Raven’s test. Stanford-Binet and Wechsler provide more detailed information about a range of different abilities than the Raven’s test. For example, they assess spatial, verbal and math abilities, as well as some general/cultural knowledge.

Yet, like the Raven’s test, the broad intelligence tests fall short in key areas. For example, they fail to indicate whether the examinee is good at physics, carpentry, medicine, plumbing or just about any other occupation. Furthermore, the design of these tests actually precludes assessment of basic literacy skills.

Information processing tests, such as assessments of simple and choice reaction time, perception, attention and working memory, may attempt to determine individual differences in the “building blocks” for human intelligence. However, they too provide little insight into the knowledge and skills of the individual examinees.

At best, all these tests may simply predict an individual's likelihood to succeed in an academic learning or occupational training programme. The tests resemble Aristotle's depiction of a block of bronze in terms of "potentiality" and "actuality" (Ackerman, 2008[6]). In other words, intelligence tests only assess current performance (and not actual "potential").

However, the use of intellectual ability tests is fundamentally a matter of prediction of some criterion, such as future academic or occupational performance. These tests do not directly measure the knowledge, skills and abilities needed for any specific task, except in highly limited and artificial circumstances (e.g. mental arithmetic).

The idea of "potentiality" and "actuality" can be transposed to the AI/robotics domain. A central processing unit and a six-axis robot arm have enormous "potential" to perform basic or advanced human tasks/jobs. However, their "actual" capability is limited by lack of software to guide them in such tasks. Thus, the discussion about potential is practically useless in determining what tasks the systems can accomplish today, and likely limited in predicting tasks they will accomplish tomorrow.

The pedagogical method of intelligence assessment

Binet referred to his method of intelligence as the "psychological method" of assessing intelligence. Such an approach assesses higher-order mental abilities (reasoning, memory, comprehension). He contrasted this with the "pedagogical method" [see Ackerman (1996[7])], which assesses "intelligence" by *examining what the individual knows*.

The shortcomings of psychological tests have been known for more than a century. As early as the 1910s, when psychologists developed the first "trade" tests, they recognised the limitations of the psychological method for assessing *current* capabilities [see Chapman (1921[8])]. In the First World War, trade tests aimed to determine which individuals had expertise in trades such as electronics, automotive mechanics, butcher, cook, barber, etc. The pedagogical approach allows individuals to be assigned to a specialty and perform the job without additional training. These tests took a variety of approaches, including self-report background, open-ended tests, multiple-choice questions and hands-on performance. The key theme is that the tests were designed to sample the knowledge and skills of individual examinees. The most successful versions were those that required actual hands-on demonstration of expertise and thus a direct assessment of skills.

Such techniques are still used in competence assessments (see Chapter 9). Similar assessments might be used across a variety of jobs/tasks[1] to compare human and AI/robotics systems for at least two of the three major types of knowledge: declarative, procedural and tacit. These are discussed below.

Declarative knowledge (knowing that)

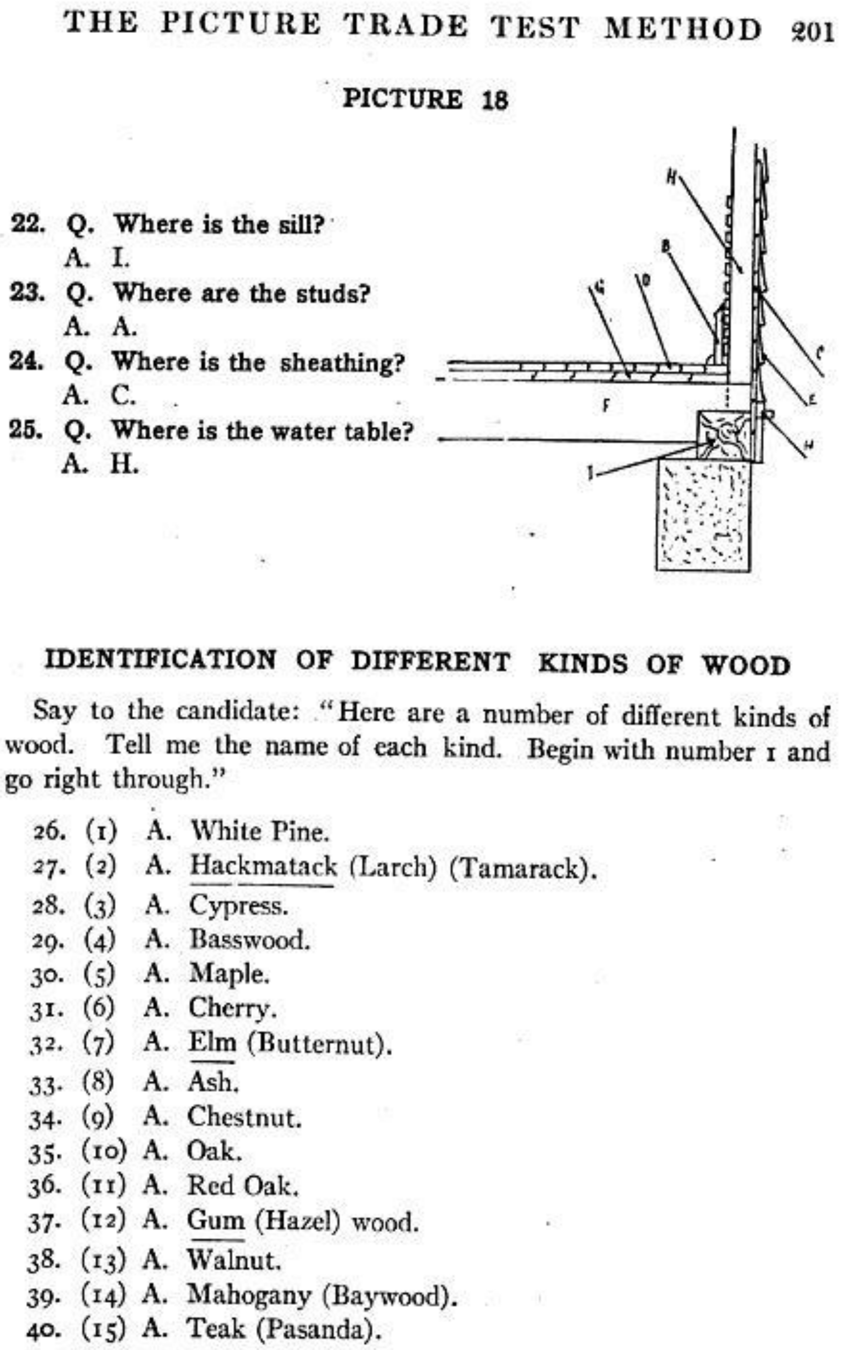
Declarative knowledge is essentially factual knowledge. In earlier eras, such information could be found in an encyclopaedia. Today, the search for facts on a country, historic figures, literature and so on is often easily accessible through an Internet search.

Declarative knowledge may be isolated facts but can also consist of principled and organised information. Examples of organised information include the Periodic Table of chemical elements and classification systems in botany. Adler (1974[9]), for example, outlined five branches of knowledge that mainly represent declarative knowledge (Logic, Mathematics, Science, History and the Humanities, and Philosophy).

A wide range of paper-and-pencil occupational certification tests is designed to assess job domain-specific declarative knowledge. Although multiple-choice tests are common, Carroll (1982[10]) and others have discussed their limitations. Apart from their dependence on literacy/reading comprehension, these tests depend mainly on “recognition” of correct answers.

Conversely, a reliance on “recall” requires deeper knowledge of human examinees and may place similar demands on AI/robotics systems. Other techniques have also been used, such as oral examinations and extensive use of photographs or physical objects (Figure 8.6).

Figure 8.6. Selected item from Carpenter trade test



THE PICTURE TRADE TEST METHOD 201

PICTURE 18

22. Q. Where is the sill?
A. I.

23. Q. Where are the studs?
A. A.

24. Q. Where is the sheathing?
A. C.

25. Q. Where is the water table?
A. H.

IDENTIFICATION OF DIFFERENT KINDS OF WOOD

Say to the candidate: “Here are a number of different kinds of wood. Tell me the name of each kind. Begin with number 1 and go right through.”

26. (1) A. White Pine.
27. (2) A. Hackmatack (Larch) (Tamarack).
28. (3) A. Cypress.
29. (4) A. Basswood.
30. (5) A. Maple.
31. (6) A. Cherry.
32. (7) A. Elm (Butternut).
33. (8) A. Ash.
34. (9) A. Chestnut.
35. (10) A. Oak.
36. (11) A. Red Oak.
37. (12) A. Gum (Hazel) wood.
38. (13) A. Walnut.
39. (14) A. Mahogany (Baywood).
40. (15) A. Teak (Pasanda).

Source: Chapman (1921[8]).

Procedural knowledge (knowing how)

Procedural knowledge consists of sequences of actions (e.g. baking a cake, operating a table saw). In some cases, procedural knowledge of a sequence can be represented as declarative knowledge (e.g. a musical score or a recipe for a meal). However, for procedural knowledge, the action sequence must be performed to assess the individual's competence.

Thus, someone who can write down from memory the score from a Beethoven sonata could only be said to have declarative knowledge of the sonata. Someone who can play it competently in real time on a piano could be said to have procedural knowledge of the sonata (regardless of whether that person could write down the score).

The task list to acquire a barber's licence in the state of New York (United States) offers a suitable example of procedural knowledge assessment. Among 14 areas of job tasks, for example, the examination tests "haircutting techniques". This, in turn, has the following subcomponents: "comb the head, taper the nape area, sideburns, top, back & sides, arching, re-drape, and prepare for finish [the hairstyle]" (Government of New York, 2020[11]).

Tacit knowledge (knowing with)

Human adults are said to have acquired various degrees of knowledge that cannot be easily subsumed under either declarative or procedural knowledge categories. Polanyi (1966/1983[1]) described this as "tacit knowing". This kind of knowledge is not typically articulated. Frequently, it is not even accessible through personal introspection.

Broudy (1977[12]) called this type of knowledge as "knowing with". He proposed that, for educated people, this knowledge would be what the individual "thinks, perceives and judges with everything that he has studied in school, even though he cannot recall these learnings on demand" (Broudy, 1977, p. 12[12]).

These concepts of tacit knowledge share similarities with Gestalt principles of organisation [e.g. Köhler (1947[13])]. However, Bransford and Schwartz (2000[14]) offer more explicit examples of tacit knowledge in the context of transfer of knowledge/training from one domain to another. For a discussion, see Ackerman (2008[6]).

Job/task-relevant tacit knowledge may contribute to success in some occupations. However, it is not yet possible to develop assessments that reveal individual differences in tacit knowledge. This is mainly because it is difficult to articulate this type of knowledge to begin with.

Task-independent components of tacit knowledge may exist, such as determining how to manage or motivate particular employees or interact with customers. However, there may also be task/job-specific elements, too. Ultimately, attention to this type of knowledge may be necessary to compare human and AI/robotic systems for effectiveness.

Tool use

One of the most salient elements in the human history of work has been the development and application of tools. The first tools augmented human muscles (e.g. the lever, pulley). The next tools augmented sensory and perceptual limitations (e.g. telescopes, radar). Finally, tools were developed to augment cognitive limitations (calculators and computers).

Most recently, of course, are the tools made available by the Internet. These tools provide both declarative knowledge of the sort available in news websites or Wikipedia. However, they also entail information relevant to the acquisition of procedural knowledge (such as explanatory YouTube videos).

Historically, ability assessments have limited the availability of most tools for completion of problems. Notable exceptions include paper and pencil for group tests, and hand-held calculators for tests like the Scholastic Assessment Test (SAT) [e.g. Ackerman (2018[15])].

Tools in occupations serve different uses. They may be required for day-to-day task completion (such as for a carpenter, electrician or surgeon). Tools may also serve mainly to augment individual capabilities (such as computerised spell checkers and grammar advisers for writing tasks).

For many jobs, removing access to such tools might make required tasks difficult or impossible to complete. In these cases, individuals might need an entirely new set of skills or to relearn an old set of skills. This could happen, for example, if a dentist desired to diagnose the presence or absence of a cavity without being able to employ X-ray technology.

On one hand, AI/robotics systems could be considered to be tools rather than agents. On the other hand, if AI/robotics systems are considered to be independent agents compared to humans, what would be a "fair" comparison? Would AI/robotics systems be assessed solely by what can be accomplished with a common set of "tools"? Would they be assessed without reference to other computerised sources (e.g. natural language processors or external databases)? It will be difficult to establish the boundary conditions for allowing either system to make use of tools during the assessment.

Competence vs. expertise

There are numerous examples of thorough assessment procedures for assessing "competence" of human jobs/tasks across many occupations. However, there is a significant limitation associated with using such assessments for comparing AI/robotics and human capabilities. This is because most competence assessment instruments and procedures are designed to determine whether the examinee has a "minimum competence" for certification.

Such certifications range from law and health care (medical doctors, nurses, psychologists) to plumbing, electricians, taxi-cab drivers and practice in a variety of other occupations. Depending on the country, they could even include obtaining a high-school diploma. This is a reasonable threshold for determining whether the individual can perform a task at an acceptable level. However, when two individuals have obtained a passing grade, such assessments do not ordinarily determine whether one is more "expert" than the other.

Some assessments set a higher threshold than "acceptable" performance (e.g. Board Certification for medical doctors in the United States). However, rather than assigning differential scores to individuals, these assessments mainly just raise the passing threshold to a higher standard.

Ideally, assessments for comparing AI/robotics systems against human operators should distinguish between the two on continuous-graded scales. If not, they should at least be capable of reliable and valid assignment of categorical ratings, such as "novice, apprentice, journeyman and expert" Chapman (1921[8]).

It may be difficult to use such assessments in a manner that allows for rank-ordering of expertise among human examinees. Various stakeholders (e.g. unions, individual employees) may resist disclosure of information beyond the certification. (For example, consider the doctor who received only a "barely passing" score on a Board-certification exam.) The employees may also fear the information might be used in ways that are deleterious or cannot be anticipated, such as when clients or customers do not know how to evaluate varying levels of performance on the certification exams. For example, in the United States, airline pilots are re-certified periodically to maintain their licences. However, in accordance with union agreements, all other data beyond whether they pass or fail are scrubbed after re-certification.

One could adapt many certification or competence assessments to assess the relative strengths and weaknesses of individual humans or AI/robotics systems. However, the traditional psychometric approach provides the maximal discrimination at the competent/not competent cut-off levels in these certification assessments. (In this sense, it is similar to how the original IQ tests were designed to differentiate more precisely at lower levels of performance. They give little attention to distinguishing among higher-ability individuals). Thus, a redesign of such assessments might be needed prior to use. This would remove ceiling-effect limitations and make more precise evaluations at the higher-end of performance.

The "critical incident technique" (Flanagan, 1954[16]) develops/adapts assessments with greater focus on measuring varying levels of expertise rather than competence. This allows the measurement professionals to generate assessment scenarios that have historically been associated with exceptionally high/low levels of effectiveness for specific tasks/jobs. These scenarios can then be adapted for future assessments.

The technique draws on descriptions of situations or scenarios with an outsized effect on the success or failure of a job task. From the collection of critical incidents, assessments can be designed to go beyond the "basic" requirements of the tasks/jobs, which typically focus on day-to-day requirements. Rather, they can focus on a more holistic assessment. These would look at situations that are likely to have greater impacts on performance than are identified through a more traditional job analysis.

Learning

The speed of acquisition of knowledge and skills (e.g. learning) of the respective systems is another consideration. The respective strengths and weaknesses of human and AI/robotics systems likely diverge in terms of the underlying characteristics for tasks of different learning demands.

For some high-knowledge tasks, such as radiological diagnoses, humans can only attain successful performance through long and varied training and practice. Conversely, AI/robotics systems only need access to a substantial database to find patterns and draw accurate conclusions. As a result, humans may be highly limited in contrast to AI/robotics systems.

Jobs/tasks that require relatively little training/learning for accomplishment pose an interesting contrast between human and AI/robotics systems. Human operators bring a basic repertoire of knowledge and skills to such tasks, such as sorting mail or "bussing" a restaurant table. Yet AI/robotics systems may be especially challenged for these same tasks.

There is no suitable taxonomy of knowledge and skills that has a high prevalence or is nearly universal among human adults. It clearly runs the gamut – from gross body movements (e.g. walking, running, picking up widely different objects) and a variety of manual dexterity tasks (e.g. eating, bathing, cooking), to basic personal tool use (e.g. toothbrush, knife, fork, spoon or chopsticks) and occupational tool use (e.g. screwdriver, hammer).

Similarly, although basic literacy skills are not nearly universal in many areas, humans have a wide prevalence of skills for recognising letters and numbers. This means many individuals can learn mail-sorting skills without substantial investment in training. Conversely, machine learning of mail-sorting was a significant challenge to the early AI/robotics community. This was especially the case with hand-addressed mail rather than mail with bar-coded address labels. In addition, most jobs/tasks also require understanding of natural language and other near-universal human skills. Such differences in competence must be considered in assessing the overall effectiveness of AI/robotics systems in comparison to human systems.

Recommendations

The OECD wants to use taxonomies of human abilities as a starting point, presumably as a tool to provide the foundation for sampling of particular assessments to compare human and AI/robotics systems. However, the disconnection between ability taxonomies [e.g. Carroll (1993[17])] and real-world job knowledge/skills is a significant impediment for such an approach.

The problem does not seem insurmountable. The following approach, followed in sequence, might be suitable:

- **Survey OECD countries for jobs that require explicit testing**

Survey OECD countries to find which jobs require explicit testing. Of particular interest are tests that combine domain-specific declarative knowledge assessments (e.g. written or oral tests) with hands-on procedural knowledge assessments (i.e. that require more than completion of course work or apprenticeship/supervised clinical hours).

- **Determine nature of analysis (jobs vs. tasks)**

Determine whether the analysis will take place at the broader level of "jobs" or the narrower level of "tasks" within jobs.

- **List jobs/tasks and sort by ability/skill**

Obtain a list of these jobs/tasks, and then have subject matter experts sort the jobs/tasks by similarity of underlying ability/skill demands.[2]

- **Conduct a cluster analysis**

Subject the aggregated sorting data to cluster analysis (Hartigan, 1975[18]; Arabie and Carroll, 1980[19]) to determine groups of similarly situated jobs/tasks.

- **Create representative set of jobs/tasks for use/adaptation**

Sample from the obtained clusters and then create a representative set of jobs/tasks from which certification/competence assessments can be used/adapted for comparing human and AI/robotics systems.

It will not be easy to compare humans and AI/robotics head-to-head on a job level. This is because many jobs involve a multitude of tasks, some of which may be idiosyncratic to a particular individual and/or a particular day. Comparing humans and AI/robotics head-to-head will therefore likely need to start at the task level. At a later stage, accounting for an increasing number of tasks may approximate the majority of particular job requirements.

References

Ackerman, P. (2018), "Intelligence as potentiality and actuality", in Sternberg, R. (ed.), *The Nature of Human Intelligence*, Cambridge University Press, Cambridge, UK. [15]

Ackerman, P. (2008), "Knowledge and cognitive aging", in Craik, F. and T. Salthouse (eds.), *The Handbook of Aging and Cognition: Third Edition*, Psychology Press, New York. [6]

Ackerman, P. (1996), "A theory of adult intellectual development: Process, personality, interests, and knowledge", *Intelligence*, Vol. 22/2, pp. 227-257, https://doi.org/10.1016/S0160-2896(96)90016-1. [7]

Adler, M. (1974), “The circle of learning”, in *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., Chicago, IL. [9]

Arabie, P. and J. Carroll (1980), “MAPCLUS: A mathematical programming approach to fitting the ADCLUS model”, *Psychometrika*, Vol. 45/2, pp. 211-235, https://doi.org/10.1007/BF02294077. [19]

Binet, A. and T. Simon (1961), “Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux (Année Psychologique, 11, 191-336); and Le développement de l’intelligence chez les enfants (Année Psychologique, 14, 1-94)”, in Jenkins, J. and D. Paterson (eds.), *Studies in Individual Differences: The Search for Intelligence*, Appleton-Century-Crofts, New York, https://doi.org/10.1037/11491-000. [2]

Bransford, J. and D. Schwartz (2000), “Rethinking transfer: A simple proposal with multiple implications”, *Review of Research in Education*, Vol. 24, pp. 61-100, https://doi.org/10.3102/0091732X024001061. [14]

Broudy, H. (1977), “Types of knowledge and purposes of education”, in Anderson, R., R. Spiro and W. Montague (eds.), *Schooling and the Acquisition of Knowledge*, Erlbaum, Hillsdale, NJ. [12]

Burke, H. (1958), “Raven’s progressive matrices: A review and critical evaluation”, *Journal of Genetic Psychology*, Vol. 93/2, pp. 199-228, https://doi.org/10.1080/00221325.1958.10532420. [4]

Carroll, J. (1993), *Human Cognitive Abilities: A Survey of Factor-analytic Studies*, Cambridge University Press, New York. [17]

Carroll, J. (1982), “The measurement of intelligence”, in Sternberg, R. (ed.), *Handbook on Human Intelligence*, Cambridge University Press, New York. [10]

Chapman, J. (1921), *Trade Tests: The Scientific Measurement of Trade Proficiency*, Henry Holt and Company, New York. [8]

Flanagan, J. (1954), “The critical incident technique”, *Psychological Bulletin*, Vol. 51/4, pp. 327-358, https://doi.org/10.1037/h0061470. [16]

Government of New York (2020), *Barber License Examination*. [11]

Hartigan, J. (1975), *Clustering Algorithms*, John Wiley and Sons, New York. [18]

Köhler, W. (1947), *Gestalt Psychology*, Liveright Publishing, New York. [13]

Polanyi, M. (1966/1983), *The Tacit Dimension*, Peter Smith, Gloucester, MA. [1]

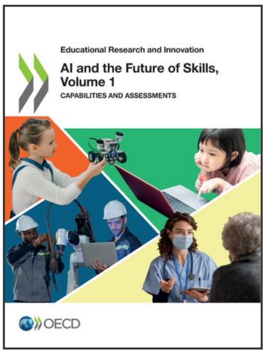
Spearman, C. (1904), “’General intelligence’, objectively determined and measured”, *American Journal of Psychology*, Vol. 15, pp. 201-293, https://doi.org/10.2307/1412107. [3]

Vernon, P. and J. Parry (1949), *Personnel Selection in the British Forces*, University of London Press, London. [5]

Notes

[1] “Jobs/tasks” is meant to capture the range of cognitively demanding activities associated with a particular occupation. One can consider “tasks” as the lowest-meaningful component of a job, and a “job” as typically subsuming a number of different tasks required of the employee. That is, in a particular job, the employee would be expected to perform a range of tasks. Job analysis and task analysis procedures have typically been used to decompose the requirements of a job into identifiable elements. These individual elements can be described at several levels of analysis. For example, psychologists have used many job and task analyses to hypothesise the underlying ability, knowledge and skill requirements for a particular occupation.

[2] Traditionally, with a set of 50 or so items to be sorted, each subject matter expert will create 8-12 sets of items that are conceptualised to be similar (within sets) and different (between sets). The resulting “data” for each subject matter expert will be a co-occurrence matrix (0’s indicate a pair of items is in different sets, and 1’s indicate the pair of items is in the same set). Data are then aggregated across the subject matter experts. In this way, each entry in the co-occurrence matrix ranges from 0 (all experts agree the items are different) to n, where n represents the total number of experts (all experts agree the items are similar).



From:

AI and the Future of Skills, Volume 1

Capabilities and Assessments

Access the complete publication at:

https://doi.org/10.1787/5ee71f34-en

Please cite this chapter as:

Ackerman, Phillip L. (2021), “Abilities and skills: Assessing humans and artificial intelligence/robotics systems”, in OECD, *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/497be8f9-en

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

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