3. Taxonomy of cognitive abilities and measures for assessing artificial intelligence and robotics capabilities

Patrick C. Kyllonen, Educational Testing Service, Princeton, NJ, United States

This chapter reviews taxonomies of human cognitive abilities and measures of those abilities. It recalls the history of key models and frameworks, analysing their strengths and weaknesses. It gives special attention to the second order of frameworks, which comprises approximately nine distinct abilities, including fluid, crystallised, spatial and broad retrieval/creativity abilities. The primary factors associated with these nine abilities are discussed, along with sample tests and test items reflecting the different abilities. It proposes two additional abilities: emotional intelligence and collaboration/communication. Finally, it discusses the prospects and feasibility of using human abilities and their associated tests to evaluate machine intelligence, discussing the advantage of having a justification for the selection of tasks.

Introduction

This chapter reviews taxonomies of human cognitive abilities and measures of those abilities. It recalls the history of key models and frameworks, analysing their strengths and weaknesses as a group and relative to one another. It gives special attention to the second order of frameworks, which comprises approximately nine distinct abilities, including fluid, crystallised, spatial and broad retrieval/creativity abilities. The primary factors associated with these nine abilities are discussed, along with sample tests and test items reflecting the different abilities. It proposes two additional abilities: emotional intelligence and collaboration/communication. Finally, it discusses the prospects and feasibility of using human abilities and their associated tests to evaluate machine intelligence, discussing the advantage of having a justification for the selection of tasks.

Various taxonomies of human skills and abilities

There are numerous taxonomies of human skills and abilities based on various approaches for developing them. This section first explores human cognitive abilities, and psychometric and sampling models. All these models acknowledge the phenomenon of positive manifold. First noted by Spearman (1904[1]; 1927[2]), positive manifold is a label for the universality of positive correlations between performance scores on any pair of cognitive tests. There have been many attempts to determine the cause of positive manifold (e.g. general factor or bonds or network), or, if a general factor, then the nature of the general factor. The section ends with a discussion of executive function, cognitive architectures and general artificial intelligence (AI).

From Spearman to Cattell-Horn-Carroll

Perhaps the oldest and most well-known taxonomies come from the human cognitive abilities literature (or sometimes, factor analytic tradition). This began with Spearman (1904[1]; 1927[2]) who analysed correlations among scores (i.e. tallies of numbers correct from a set of items) from various cognitive tests, primarily samples of school-like tasks. That led to the fluid-crystallised (Gf-Gc) mode (Horn and Cattell, 1966[3]) and the extended Gf-Gc model (Horn and Blankson, 2005[4]; 2012[5]); the three-stratum model (Carroll, 1993[6]); and their synthesis in the Cattell-Horn-Carroll (CHC) model (McGrew and Woodcock, 2001[7]). Annex 3.A compares the three models.

The various models within this framework agree on a strong general factor that accounts for 30-70% of the between-person variance in any test score [called “gf” in the Horn-Cattell model, see Carroll (1993[6])]. Consider, for example, math, verbal, science and problem-solving scores in the OECD Programme for International Student Assessment (PISA). These scores show intercorrelations ranging from r = .8 to r = .9. There is a general fluid versus crystallised distinction; these two factors can be highly correlated, but developmental trajectories differ (fluid ability peaks at an earlier age). There are roughly 8 to 10 group (second order) factors, and 80 or so primary (first order) latent factors that account for covariances among test scores.

Vernon’s hierarchical model

There are alternative approaches within the human abilities/factor analytic tradition, or what is sometimes called the psychometric model of intelligence (Hunt, 2011[8]). The hierarchical model of Vernon (1950[9]) is realised in the g-VPR (general factor, verbal-perceptual-memory, image rotation) model (Johnson and Bouchard, 2005[10]; Johnson, te Nijenhuis and Bourchar, 2007[11]). This model does not differ qualitatively from the Carroll (1993[6]) or CHC models. However, it differs in emphases: there is a general factor at a fourth order, three major factors at the third order (verbal, perceptual and image rotation), about nine factors at the second order and numerous primary factors.

Sampling models

Sampling is another alternative to the abilities model. This tradition began with the bond sampling model (Thomson, 1916[12]), where any test samples a set of mental bonds rather than component abilities per se (Tirre, 1994[13]). More recently, the sampling approach is represented in network (van der Maas et al., 2019[14]) and wiring models (Savi et al., 2019[15]).

Executive function

One popular concept relates the general factor to working memory capacity (Conway and Engle, 1996[16]; Kyllonen and Christal, 1990[17]). Working memory capacity can be characterised as executive attention [i.e. one's capacity to control attention, see Engle (2002[18]) and Kane et al. (2001[19])].

This line of findings arguably underlies the executive functioning literature, which has become popular in education circles (Zelazo, Blair and Willoughby, 2016[20]). Executive functioning is defined as "skills related to working memory, inhibitory control and mental flexibility" (Shuey and Kankaraš, 2018[21]). These skills, in turn, are defined as the temporary activation or storage of information while engaged in cognitive processing (Baddeley, 1986[22]; Cowan, 2017[23]); directing or sustaining attention in the face of distractions (Diamond, 2013[24]); and the ability to shift between different mental sets or tasks (Archambeau and Gevers, 2018[25]), respectively.

Cognitive architectures

Other lines of research that fall outside the factor analytic tradition of conventional abilities contribute to a positing or understanding of a human abilities taxonomy. One of these lines of research is computational modelling, or cognitive architectures. These include Adaptive Control of Thought-Rational (ACT-R) (Anderson and Lebiere, 1998[26]; Anderson et al., 2004[27]); Cortical Capacity-Constrained Concurrent Activation-based Production System (4CAPS) (Just and Varma, 2007[28]); Executive Process Interactive Control (EPIC) (Kieras and Meyer, 1997[29]); SOAR (Newell, 1994[30]); and Hypothesis Generation (HyGene) models (Thomas et al., 2008[31]).

These models are designed to simulate human problem solving. As a side effect, their architectures suggest constructs that may be treated as human abilities. For example, ACT-R distinguishes procedural (production rules) and declarative (chunks) memory. It includes specialised perceptual-motor, goal and declarative memory modules, as well as learning processes. HyGene, which is designed for diagnostic reasoning, includes processes of information sampling, derivation of prototypical representations, generation of candidate hypotheses, probability estimation, hypothesis testing and search termination.

Many of these processes and modules map to the lower-order factors in the hierarchical abilities' models. However, they are implemented more precisely with respect to how they function, which is required for a computer simulation.

Artificial intelligence ability taxonomies

Related to cognitive architectures is the more general AI literature. This is not concerned with simulating human cognition but with building intelligent entities more broadly. A popular AI textbook (Russell and Norvig, 2010[32]) includes chapters entitled problem solving; knowledge, reasoning and planning; knowledge representation; probabilistic reasoning; making simple (and complex) decisions; learning; and perception. These chapters include many methods for addressing these topics, such as induction, case-based reasoning and reasoning by analogy, which also map to the abilities identified in the Carroll (1993[6]) model. Identifying the constructs included in AI can inform discussions of human abilities.

Cognitive tests associated with these taxonomies

Each of the literatures reviewed in the previous section is based on empirical research that involves performance of cognitive tests. The most extensive of these is from the cognitive abilities/factor analytic tradition because the associated taxonomies are derived directly from scores on batteries of cognitive tests. This section looks at cognitive tests associated with these taxonomies.

From school tasks to intellectual tasks

Originally, cognitive tests were essentially samples of school tasks (e.g. spelling) along with perceptual (e.g. pitch perception) and memory (e.g. logical, visual, auditory) tasks from the laboratories of experimental psychology (Wissler, 1901[33]; Spearman, 1904[1]).

Over time, new kinds of intellectual tasks were added, such as Thurstone (1938[34]), and World War II led to a considerable expansion of measures. A unit of the US Army (the Army Air Force) developed tests to measure every conceivable mental function (Humphreys, 1947[35]; Damos, 2019[36]). These included tests of verbal and mathematical skills, reasoning, mechanics, judgement, foresight, planning, integration, memory, attention, mental set, perceptual skills, spatial orientation and visualisation, and general information, as well as a set of motion picture tests (Gibson, 1947[37]; Lamkin, Shafer and Gagne, 1947[38]).

Later, Guilford (1950[39]) expanded even this lengthy list to include new measures to fill out his structure of intellect model (Guilford and Hoepfner, 1971[40]). Perhaps the most significant area of expansion was in measures of divergent thinking, or creativity (Guilford, 1950[39]).

Test kits and tool boxes

Educational Testing Service (ETS) produced a kit of cognitive reference tests. They comprised a sample of tests from the most important 46 factors associated with this work (Ekstrom et al., 1976[41]; 1976[42]). These tests, still used widely in research, are available for free or a nominal charge.

Condon and Revelle (2014[43]) and Dworak et al. (2020[44]) produced the International Cognitive Ability Resource (ICAR). This open-source tool measures 19 domains of cognitive ability, including fluid ability (progressive matrices and matrix reasoning, propositional reasoning, figural analogies, letter and number series, abstract reasoning), emotional ability (emotion recognition), mathematical ability (arithmetic), verbal ability (verbal reasoning), creativity (compound remote associates), face-detection (aka the Mooney Test), perceptual ability (melodic discrimination, a perceptual maze task), and judgement (a situational judgement task). This is not as systematic as the ETS resource. However, given the open-source nature of the ICAR, it may eventually become more comprehensive.

The CHC taxonomy is tightly tied to a commercial instrument: the Woodcock Johnson III Tests of Cognitive Abilities (WJ III) (Schrank, 2011[45]) and Woodcock Johnson IV (WJ IV) (Schrank, Mather and McGrew, 2014[46]). Annex 3.B describes the battery factors and tests.

Executive functioning research is associated with tests that measure working memory capacity, inhibitory attentional control, and cognitive and attentional flexibility. Jewsbury, Bowden and Strauss (2016[47]) show how the CHC model can accommodate these measures. Numerous publications describe working memory measures [e.g. Kyllonen and Christal (1990[17]) and Wilhelm, Hilldebrandt and Oberauer (2013[48])].

The US National Institutes of Health (NIH) provides the NIH toolbox, which comprises a set of 100 stand-alone measures to assess cognition, emotion, motor ability and sensation. NIH cognition measures (Zelazo et al., 2013[49]) include attention and executive function (Flanker Inhibitory Control and Attention Test), working memory (list sorting), executive function (Dimensional Change Card Sort), along with episodic memory, language and processing speed measures. These are separated by age range (3-6, 7-17 and 18+).

The cognitive architecture literature has been primarily driven by laboratory tasks from experimental cognitive psychology. Such measures are most often designed to test specific aspects of cognitive theories. They predominantly measure response time (e.g. fact retrieval, lexical decision) and memory recall (e.g. free recall, recognition memory). They tend to be simpler than tests in the psychometric tradition, as they are typically designed for narrower purposes.

The AI literature is voluminous and therefore difficult to characterise. It is equally difficult to characterise the kinds of cognitive measures associated with the literature.

Criteria for establishing taxonomy and suitable tests

An ideal taxonomy for this project would provide a list of human abilities, identified through a methodology or methodologies that enable a strong scientific justification. Such a list would be comprehensive but parsimonious. In this way, there would be minimal conceptual or empirical overlap between abilities. The definition of the ability would also need to be demonstrable or transparent. These, and other, principles are elaborated below.

Comprehensiveness could prove difficult to demonstrate capability with respect to all the abilities proposed. Parsimony would minimise burden in any application exercise, such as rating jobs. Demonstrability or transparency can be measured with processing requirements that are clear and easily understood. Empirically there must be a strong connection between the test and the construct it is intended to measure. Other considerations for determining the suitability of particular tests are their correlation with the factor of interest; amount of time (or number of items) needed to achieve a reliable score; and susceptibility of the test to (contamination with) other factors (test impurity).

Cattell-Horn-Carroll framework

This section reviews one of the lines of literature in more depth: the human cognitive abilities literature from the factor analytic tradition.

The factor analytic tradition in depth

The factor analytic tradition begins with Spearman (1904[1]; 1927[2]). His analyses of cognitive tests found that a general latent factor accounted quite adequately for the correlations among test scores. However, each test score additionally had to contain test-specific variance.

Thurstone (1938[34]) administered a broader sample of tests to a larger group of college students. Through the development of multiple-factor analysis and the concept of simple structure, Thurstone (1934[50]) showed evidence for a set of narrower group factors (verbal comprehension, word fluency, number facility, spatial visualisation, associative memory, perceptual speed, reasoning). He referred to these as primary mental abilities.

Others showed the Spearman and Thurstone findings were compatible through a hierarchical mode (Undheim, 1981[51]; Gustafsson, 1984[52]) or the similar bifactor model (Holzinger and Swineford, 1937[53]; Holzinger and Harman, 1938[54]; Schmid and Leiman, 1957[55]). In other words, performance on a test could be a function of general, group and specific latent factors simultaneously.

Horn and Cattell (1966[3]) proposed two general factors [fluid (Gf) and crystallised (Gc) ability], along with a set of correlated group factors or broad abilities. This line of research (Horn and Blankson, 2005[4]; Horn and Blankson, 2012[5]) culminated in 80 first-order primary mental abilities and 8 second-order abilities. These are Gc; Gf; short-term memory (Gsm) [later, short-term apprehension and retrieval (SAR)]; long-term memory (Gsl) [later, fluency of retrieval from long-term storage (TSR)]; processing speed (Gs); visual processing (Gv); auditory processing (Ga); and quantitative knowledge (Gq).

Carroll (1993[6]) analysed 460 datasets comprising test correlation matrices accumulated over almost a century of research. He reanalysed them using a version of the Schmidt-Leiman procedure and synthesised findings primarily based on informed but subjective judgements of content (and process) overlap. He proposed a three-stratum model (Carroll's "stratum" is synonymous with the more common term of "order") with a general factor at the apex and eight second-stratum factors.

The second-stratum factors were fluid intelligence, crystallised intelligence, general memory and learning, broad visual perception broad auditory perception, broad retrieval ability, broad cognitive speediness and processing speed. Each of the second-stratum factors covered 4 to 14 first-stratum factors. For example, the second-order fluid intelligence covered the first-order (primary) factors: general sequential reasoning, induction, quantitative reasoning and speed of reasoning. These, in turn, were determined by the correlations among the manifest scores from various tests of those factors.

The CHC model was proposed as a synthesis of the Carroll (1993[6]) and Horn and Cattell (1966[3]) models (McGrew and Woodcock, 2001[7]). It has subsequently been revised and expanded regularly with the incorporation of new research findings [e.g. Schneider and McGrew (2018[56])]. However, these remain three distinct frameworks or models. Carroll (2003[57]) updated his model, and Horn did as well (Horn and Blankson, 2012[5]) to accommodate new findings. Still, it is useful to treat them or their synthesis as a common framework. They differ in some details (Carroll, 2003[57]) but are based on mostly common data and methods.

Critiques and modifications of the CHC framework

The CHC model has become a popular framework for the representation of human abilities, partly or perhaps primarily due to its application in school psychology for student cognitive diagnosis. Nevertheless, several important critiques have recently appeared. These include a special issue of *Applied Measurement in Education* (Beaujean and Benson, 2019[58]; Canivez and Youngstrom, 2019[59]; Geisinger, 2019[60]; McGill and Dombrowski, 2019[61]; Wasserman, 2019[62]).

These critiques identify five limitations: over expansiveness; emphasis of group factors over individual factors; mental speed; treatment of quantitative factor; and its combination of two disparate factors. The issues are summarised below.

Over expansiveness

Like many abilities frameworks (Carroll, 1993[6]; Carroll, 2003[57]; Horn and Blankson, 2012[5]), CHC includes too many abilities with scant justification for their inclusion. More replication would be desirable, using a variety of tests (not just those in the WJ III and WJ IV commercial batteries). Users (e.g. teachers, school psychologists) like having many abilities to test to obtain a more complete picture of a student. However, there is a growing awareness that reliability is crucial to distinguish between tests or factors (Haberman, Sinharay and Puhan, 2011[63]). In addition, profile scores (e.g. a set of scores from several tests or factors) are often not justified due to the importance of the general factor. This critique suggests a smaller number of factors than are typically reported are scientifically justified.

General vs. group factors

The general factor can often be shown to be more important in accounting for test score variance than the group (lower order or lower stratum) factor. However, CHC has mostly denied the general factor. It prefers to emphasise the group factors, which are empirically shown to be highly overlapping.

Mental speed

CHC does not treat mental speed in a way consistent with the recent literature on cognitive processing speed [e.g. Kyllonen and Zu (2016[64])]. New psychometric approaches suggest a rethinking of mental speed with respect to abilities models.

Quantitative factor

In the Carroll (1993[6]) framework, and in the CHC, quantitative ability is a lower-order factor of fluid intelligence. However, Wasserman (2019[62]) points to mathematics prodigies as an indicator that quantitative ability might deserve a higher ranking.

Combining disparate functions

Both Carroll (1993[6]) and CHC frameworks combine knowledge retrieval and idea production in a single long-term memory retrieval factor. However, these are disparate functions. Idea production is thought to be the essence of creativity, whereas knowledge retrieval is considered to be a non-creative process.

Despite these criticisms, the CHC model and its constituents [e.g. Carroll (1993[6]) and Horn and Cattell, (1966[3])] may provide enough of a basis for a justifiable taxonomy of human cognitive abilities. In other words, it may satisfy the criteria of being comprehensive, reasonably succinct and transparent in principle.

g-VPR as an alternative to the CHC framework

Other human abilities frameworks are worth considering in addition to the CHC model. The general plus verbal, perceptual and image rotation (g-VPR) model (Johnson and Bouchard, 2005[10]; Johnson, te Nijenhuis and Bourchar, 2007[11]) has been shown to provide a better account of the test score data than the CHC model.

Some prominent researchers such as Hunt (2011[8]) have suggested g-VPR as a viable alternative to the Carroll (1993[6]) or CHC models. However, showing a slightly superior fit for a few datasets is probably not a sufficient reason for claiming the g-VPR framework as a viable alternative. Even Johnson (2018[65]), one of the architects of g-VPR, has argued their model had not "'carved nature at its joints'" in any battery any better than Carroll had. This is because factor analysis spits back at us only what we put into it, and we have no tasks that uniquely measure any one particular ability or skill…" (p. 24).

What are the most important abilities?

As Carroll (2003[57]) noted in the title of one of his last papers, "Current evidence supports *g* and about ten broad factors." There is considerable agreement across the three major CHC frameworks about those broad factors (see Annex A), although some make distinctions. The major categories are the nine-colour coded distinctions. In addition, one general factor is not listed (because it is at the third stratum). The 80 or so primary (first order) factors are listed in Annex 3.B and Annex 3.C.

General factor and fluid intelligence

The general factor is either identical or close to identical to fluid ability (gf). There is a strong overlap between executive function ability, working memory, attention and gf (Wilhelm, Hilldebrandt and Oberauer, 2013[48]). Most of this research was conducted after Carroll (1993[6]). Still, the primary gf measures are reasoning measures, both quantitative and non-quantitative.

Good examples listed in Annex Figure 3.D.1 are from Carroll's (1993[6]) primary (first order) factors of inductive reasoning, deductive reasoning and quantitative reasoning.

The first primary gf factor of inductive reasoning includes sets (classification tasks, “odd man out” tasks), series (e.g. number, letter, figure series) and matrices tasks. In Raven's progressive matrices (Kyllonen et al., 2019[66]), for example, the goal is to induce a rule or set of rules describing an arrangement of a set of elements then to apply the rule(s) to identify or categorise new elements.

The second primary gf factor of deductive reasoning includes tests for syllogistic reasoning and diagramming relationships using Euler diagrams, as shown in Annex Figure 3.D.1. This example and the other listed illustrate how inductive and deductive reasoning tasks can be implemented in verbal, numerical and spatial content.

The third primary gf factor of quantitative reasoning is illustrated with the Necessary Arithmetic Operations test, which asks respondents to select the operations needed to solve an arithmetic word problem.

All these example tasks (and others listed in Annex 3.D) are singled out because they are good representations of some key primary factors associated with second-order factors. Further discussion of the reasoning factor, the varieties of reasoning and evidence from diverse research traditions can be found in Kyllonen (2020[67]).

Abductive reasoning

Abductive reasoning involves deriving an explanation for a finding or set of facts. Consider the following example taken from a retired form of the GRE: because the process of freezing food consumes energy, many people keep their electric freezers half empty, using them only to store commercially frozen foods. Yet freezers that are half empty often consume more energy than if kept fully stocked.

The example then proposes five possible explanations for the apparent discrepancy. This might be solved with deductive reasoning. However, it follows the form of an abductive reasoning problem in that a phenomenon is presented in search of a cause or explanation. The example presents possible explanations, but abductive reasoning could also involve an open-ended item. In that case, a person would have to retrieve relevant information to come up with an explanation. Consequently, this kind of problem overlaps to some extent with ones in the *broad retrieval ability* category, below.

Crystallised intelligence

Crystallised intelligence, in principle, represents acculturated knowledge but in practice overlaps highly with “verbal ability” (Carroll, 1993[6]). Some of the best example tasks are reading comprehension tests, vocabulary items (open-ended or multiple choice) and cloze tests. A cloze test presents a sentence or paragraph with missing words that need to be provided. This requires knowledge of the topic, vocabulary, grammar rules and the like. Crystallised and fluid intelligence tasks appear to be distinct, but empirically, fluid and crystallised intelligence are highly correlated in individuals. One explanation is that students use reasoning processes in developing verbal knowledge (Marshalek, 1981[68]). Annex Figure 3.D.2 lists examples of tasks.

Broad visual perception

Broad visual perception is commonly called spatial ability. It involves the perception, memory, mental transformation and reasoning about presented or imagined spatial materials. Guilford's blocks test provides an example of imagined spatial materials. Respondents imagine painting a wooden block red, dividing it into 3 x 3 x 3 blocks, then determining the number of blocks with exactly one side painted red. Example items from the most prototypical spatial ability tests covering the most prominent spatial ability primary factors appear in Annex Figure 3.D.3. The factors (and test examples) are spatial visualisation (mental paper folding), closure flexibility (the copying test) and perceptual speed (a picture matching test). Lohman (1979[69]) is a still useful review of this literature.

Broad retrieval ability

Broad retrieval ability is Carroll's label for a set of factors that involve creativity and mental fluency. Prototypical fluency tasks are ones that involve rapidly generating lists of responses that follow a set of rules. This could be generating all the words that begin with "S" and end with "N", or four letter words that do so; an example word fluency item is shown in Annex Figure 3.D.4. An analogous process is figural fluency, such as moving toothpicks around to create a form (see example in Annex Figure 3.D.4). Creativity measures are fluency tests that involve more complex ideas. For example, the consequences test from Christensen, Merrifield and Guilford (1953[70]) (Annex Figure 3.D.4) asks respondents to respond with as many plausible and non-repeating responses as they can in a short interval to prompts such as "What would happen if we didn't have to eat?" or "How can traffic congestion problems be curtailed?"

General memory ability

Carroll (1993[6]) found evidence for a general memory ability factor based on performance on short-term and long-term memory tasks that have been studied in the verbal learning tradition in experimental psychology. These include memory span, associative memory and free recall, as well as a separate visual memory first-order dimension. There was also evidence for a loose learning ability factor.

Memory and learning are obviously important human abilities, but this factor has not been shown to relate uniquely to educational outcomes in the way fluid and crystallised ability have. The factor may represent performance on the peculiar sort of arbitrary memory tasks that psychologists have devised, but not the ability invoked in typical educational learning situations.

Another peculiarity is that simple forward memory span (repeating a string of 7 to 9 digits) seems to invoke an ability different from backward memory span (repeating the string backwards) (Reynolds, 1997[71]). The latter operates more like a working memory test, requiring simultaneous storage and processing (Baddeley, 1986[22]).

Working memory is also highly correlated with fluid intelligence, as reviewed in a previous section. Consequently, it may not be useful to include memory ability factor in a test of AI. Technology may lessen the requirement for memorising arbitrary strings of words and symbols, which is another reason to exclude memory ability from an AI test.

Broad auditory perception

Carroll (1993[6]) found evidence for a distinct broad auditory perception factor, called "listening and hearing" in the Horn-Cattell model, and "auditory processing" in the CHC. These involve speech-sound and general sound discrimination, memory for sound patterns, musical discrimination and the like. These are important human abilities but are more perceptual in nature. They do not seem as pertinent to testing AI as abilities from the other categories.

Psychomotor ability

Psychomotor abilities are important in many jobs and other human activities, such as playing sports and games. Carroll considered this literature outside the scope of his focus on cognitive abilities, but psychomotor ability is represented in the CHC model. Fleishman (1954[72]) provided a taxonomy and set of psychomotor tasks.

Some decades later, Fleishman and Quaintance (1984[73]) and Chaiken, Kyllonen and Tirre (2000[74]) made further comments on psychomotor ability. They suggested a general psychomotor factor that could account for most of the psychomotor tasks. It can be measured with tasks such as multi-limb co-ordination and tracking tasks, such as pursuit motor co-ordination.

Processing speed

Processing speed is an important component of human cognition. Carroll (1993[6]) suggests there was an independent second-order speed factor (i.e. two independent speed factors). The nature of a processing speed factor is a complex topic within cognitive psychology and within the human abilities' literature. This complexity is due to speed-accuracy trade-off, willingness or proclivity to abandon unproductive solution attempts and time management issues generally. In fact, Carroll's two speed factors could be due to interactions among these factors (Kyllonen and Zu, 2016[64]).

It is difficult to imagine how tasks designed to measure a processing speed factor in the human abilities' literature could be used productively to measure AI abilities. A prototypical task is simply an easy version of a fluid or crystallised intelligence test (e.g. an easy vocabulary synonym judgement test). The primary dependent variable is the time it takes to retrieve the answer or solve the simple problem. Thus, little additional information is likely to be obtained by trying to determine machine capabilities on tasks sampled from the set of tests designed to measure human processing speed.

Olfactory, tactile and kinaesthetic abilities

This set of sensory abilities was also considered outside the realm of human cognitive abilities in Carroll (1993[6]). However, these abilities are represented within the CHC framework. This inclusion reflects research attempting to document these abilities within the context of human cognitive abilities. Stankov (2019[75]) summarises the research programme. However, like some of the other factors, this work seems to be outside the central focus of this study, which is primarily based on human cognitive abilities.

Additional abilities

Besides the abilities covered in the previous section, several ability factors could be noted: emotional intelligence, and collaboration and communication.

Emotional intelligence

Emotional intelligence only emerged as a concept with Mayer and Salovey (1993[76]). Therefore, it was not part of thinking about human cognitive abilities at the time of Carroll (1993[6]). Since then, there has been considerable research on the topic.

The literature distinguishes between ratings and performance measures; only the latter would be considered relevant for the *purposes* of testing AI abilities. MacCann et al. (2014[77]) administered an emotional intelligence test battery along with a battery of CHC-type cognitive ability tests (e.g. fluid, crystallised, spatial ability, broad retrieval). They identified first- and second-order emotional intelligence factors based on a set of emotional intelligence measures (two tests of each for emotion perception, emotion understanding and emotion management).

Earlier research MacCann and Roberts (2008[78]) examined the Situational Test of Emotional Understanding (STEU) and the Situational Test of Emotion Management (STEM). STEU presented items such as the following: *Hasad tries to use his new mobile phone. He has always been able to work out how to use different appliances, but he cannot get the phone to function. Hasad is most likely to feel? a) distressed, b) confused, c) surprised, d) relieved or e) frustrated.* The related STEM measure comprises similar types of items. From the standpoint of this study, this factor does represent something distinct from say, crystallised intelligence: it includes a component on reasoning about emotions.

There are other kinds of emotional intelligence measures. These include determining the emotional state of someone photographed (Baron-Cohen et al., 2001[79]; Olderbak et al., 2021[80]) or of someone expressing emotion through language (Scherer and Scherer, 2011[81]; Hellwig, Roberts and Schulze, 2020[82]). The empathic agent paradigm asks test takers to study how another person depicted in vignettes tends to act in situations, and then to apply that knowledge to predict how that person will react in a new situation (Hellwig, Roberts and Schulze, 2020[82]). All these measures make clear that a second-order emotional intelligence factor is an important human cognitive abilities factor distinct from the others discussed here.

Collaboration and communication

Woolley et al. (2010[83]) found evidence for “collective intelligence”, meaning that some teams of individuals performed better than other teams across a diverse set of team tasks. Team tasks included brainstorming, planning a shopping trip, group typing, group matrix reasoning and group moral reasoning. This “team effect” was independent of individual abilities on the team (e.g. as indicated by how they performed the task alone). Instead, collective intelligence seemed related to members’ emotional intelligence – their ability to read their teammates’ emotions, goals and intentions.

The future economy will likely put a premium on teamwork, collaboration and communication (Deming, 2017[84]). Thus, it would seem important to determine the possibility of assessing teamwork skills. PISA 2015 also included a collaborative problem-solving measure (OECD, 2017[85]). In reviewing small groups research (a branch of social psychology), Larson (2010[86]) concluded that some tasks exhibited *synergy.* This is defined as the situation in which a team outperforms the best member of the team, or at least does as well as the best member.

Tasks exhibiting synergy include a letters-to-numbers problem-solving task and a hidden-profile decision-making task. In the latter, different team members are provided overlapping but distinct knowledge about choices. Successful team performance depends on members sharing and considering their common and unique knowledge to arrive at a group decision. It is not clear if a scenario could be set up to evaluate, say, a machine’s ability to collaborate, but tasks drawn from this category suggest at least possibilities to consider.

Feasibility of a human abilities framework for assessing artificial intelligence and robotics

This section explores the viability of a human abilities framework to assess AI and robotics.

*The psychometric tradition: CHC and O*NET*

An abilities framework in the psychometric tradition (such as the CHC) has already proven viable. The US Department of Labor rates job requirements with respect to abilities similar to the kinds of abilities listed in the CHC framework through the Occupational Network, or O*NET (National Center for O*NET Development, n.d.[87]); (National Research Council, 2010[88]); (Peterson et al., 1999[89]).

O*NET, an occupational analysis system in the United States, collects ratings on job demands annually. Ratings fall into a variety of categories. These include tasks, generalised work activities, knowledge, education and training, work styles and work contexts).

Significantly, for assessing AI and robotics, ratings are collected on the ability involvement (importance and level) for over 950 occupations (Fleisher and Tsacoumis, 2012[90]). It surveys 52 abilities, while eight occupational analysts provide ratings for every occupation. The abilities are grouped into the categories of cognitive, psychomotor, physical and sensory-perceptual.

The framework is based on Fleishman, Costanza and Marshall-Mies (1999[91]) and Fleishman and Quaintance (1984[73]), but the cognitive part is largely consistent with the CHC framework. Cognitive abilities include oral and written comprehension and expression, fluency of ideas, originality, problem sensitivity, deductive and inductive reasoning, information ordering, category flexibility, mathematical reasoning, number facility, memorisation, speed of closure, flexibility of closure, perceptual speed, spatial orientation, visualisation, selective attention and time sharing. In addition, O*NET surveys perceptual and motor factors such as reaction time, auditory attention and speech recognition. It regularly publishes job descriptions with respect to their standings on these factors.

The current O*NET system does not collect judgements related to emotional intelligence or to collaboration ability, but such ratings could potentially be included. Ability ratings using an abilities framework can be collected for occupational requirements and importance. Thus, abilities could be potentially useful constructs on which to collect ratings regarding machine capabilities.

Three useful concepts to consider for machine intelligence

There are significant differences between human abilities and machine abilities. However, the language and set of concepts have evolved over a century of abilities testing. These may still be useful in considering issues in machine intelligence. Many of these concepts are captured in the Standards for Educational and Psychological Testing (AERA, APA and NCME, 2014[92]). Three – “construct irrelevant variance”, “teaching to the test” and “construct underrepresentation” – are discussed below.

Construct irrelevant variance

Construct irrelevant variance is a test fairness issue. It refers to a test intended to measure a construct, such as mathematics. However, performance can be affected by other constructs, such as the ability to understand a diagram or language abilities. If performance is affected by factors that the test is not intended to measure, then a test cannot be considered a fair measure of the construct. This is a major fairness concern motivating accommodations for individuals who might have difficulties with aspects of the test that are not the target of assessment. Consider, for example, sight-impaired individuals.

Teaching to the test

Teaching to the test refers to the notion of instruction related to incidental test features that are not features shared generally with respect to the broader construct the test is intended to measure. Teaching to the test is likely to enhance performance on a test without enhancing the level of the construct the test is designed to assess. Psychometrically, this situation is revealed to the extent that one’s performance on a particular test is not consistent with performance on other related tests, a situation sometimes referred to as model misfit.

Construct underrepresentation

Construct underrepresentation refers to a situation in which the set of tests to measure a construct does not reflect the full range of attributes or skills in the construct definition. Here, test performance might only indicate the level of the underlying trait or ability possessed by the individual. However, the test only captures a part of the larger construct. For example, a vocabulary test is a useful indicator of general verbal ability. However, verbal ability should reflect a broader set of skills than vocabular, such as paragraph comprehension or responding to general knowledge questions.

Conclusions

An abilities framework such as the hierarchical model (Carroll, 1993[6]) or the CHC (Schneider and McGrew, 2018[56]) are useful frameworks for evaluating human abilities and are likely to be useful for evaluating machine abilities as well. There is general agreement among various models at some of the major human performance distinctions. The second-stratum factors (Carroll, 1993[6]) and their equivalents in the CHC and Horn-Cattell (1966[3]) models are a useful level for evaluation of the type envisaged for AI and robotics. These might be supplemented by two additional factors – emotional intelligence and collaboration/communication ability. On the other hand, some included second-order factors, such as processing speed, psychomotor ability and sensory abilities, might be less central for understanding machine intelligence in the context of a project designed to evaluate likely future work requirements.

A solid body of evidence can be used both to identify measures of the various second-stratum factors and to evaluate the appropriateness of those measures as good indicators of those factors. Good measures from the standpoint of human abilities have several qualities. First, they produce a reliable assessment of ability in individuals. Second, their scores are highly correlated with the factor of interest (i.e. high factor loadings in a factor analysis). Third, they have high average correlations with scores from other measures of the factor (i.e. they have high centrality with respect to the construct of interest).

High factor loadings and high centrality are also related to the concept of transferability of skills. Two tasks that are highly correlated should share common skills. Conversely, lower correlations and lower factor loadings indicate less commonality with respect to skill requirements. They suggest lower transfer relations between the tasks from a training perspective.

There are many differences between machine and human intelligence. However, the evolved language used to describe tests and their relationships to the abilities intended to be measured is useful for describing issues in machine intelligence. Concepts such as reliability, validity, fairness, measurement invariance, construct representativeness and others may help clarify issues in evaluating machine intelligence in the same way they have for measuring human intelligence.

References

AERA, APA and NCME (2014), *Standards for Educational and Psychological Testing*, American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education, Washington, DC. [92]

Anderson, J. et al. (2004), “An integrated theory of the mind”, *Psychological Review*, Vol. 111/4, pp. 1036-1060, https://doi.org/10.1037/0033-295X.111.4.1036. [27]

Anderson, J. and C. Lebiere (1998), *The Atomic Components of Thought*, Lawrence Erlbaum Associates Publishers, Mahwah, NJ. [26]

Archambeau, K. and W. Gevers (2018), “(How) Are executive functions actually related to arithmetic abilities?”, in Henik, A. and W. Fias (eds.), *Heterogeneity of Function in Numerical Cognition*, Academic Press, Cambridge, MA, https://doi.org/10.1016/C2016-0-00729-5. [25]

Baddeley, A. (1986), *Working Memory*, Oxford University Press, Oxford. [22]

Baron-Cohen, S. et al. (2001), “The “Reading the Mind in the Eyes” test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism”, *Journal of Child Psychology and Psychiatry*, Vol. 42, pp. 241-251, http://dx.doi.org/10.1111/1469-7610.00715. [79]

Beaujean, A. and N. Benson (2019), “The one and the many: Enduring legacies of Spearman and Thurstone on intelligence test score interpretation”, *Applied Measurement in Education*, Vol. 32/3, pp. 198-215, https://doi.org/10.1080/08957347.2019.1619560. [58]

Canivez, G. and E. Youngstrom (2019), “Challenges to the Cattell-Horn-Carroll theory: Empirical, clinical and policy implications”, *Applied Measurement in Education*, Vol. 32/3, pp. 232-248, https://doi.org/10.1080/08957347.2019.1619562. [59]

Carroll, J. (2003), “The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors”, in Nyborg, H. (ed.), *The Scientific Study of General Intelligence: Tribute to Arthur R. Jensen*, Elsevier Science/Pergamon Press, Oxford. [57]

Carroll, J. (1993), *Human Cognitive Abilities: A Survey of Factor-analytic Studies*, Cambridge University Press, New York. [6]

Chaiken, S., P. Kyllonen and W. Tirre (2000), “Organization and components of psychomotor ability”, *Cognitive Psychology*, Vol. 40/3, pp. 198-226, https://doi.org/10.1006/cogp.1999.0729. [74]

Christensen, P., P. Merrifield and J. Guilford (1953), *Consequences form A-1*, Sheridan Supply, Beverly Hills, CA. [70]

Condon, D. and W. Revelle (2014), “The International Cognitive Ability Resource: Development and initial validation of a public-domain measure”, *Intelligence*, Vol. 43/March-April, pp. 52-64, http://dx.doi.org/10.1016/j.intell.2014.01.004. [43]

Conway, A. and R. Engle (1996), “Individual differences in WM capacity: More evidence for a general capacity theory”, *Memory*, Vol. 4, pp. 577-590. [16]

Cowan, N. (2017), “The many faces of working memory and short-term storage”, *Psychonomic Bulletin & Review*, Vol. 24, pp. 1158-1170, https://doi.org/10.3758/s13423-016-1191-6. [23]

Damos, D. (2019), *Technical Review and Analysis of the Army Air Force Aviation Psychology Program Research Reports*, Air Force Personnel Center, Randolph, TX, http://dx.doi.org/10.13140/RG.2.2.35387.36641. [36]

Deming, D. (2017), “The growing importance of social skills in the labor market”, *Quarterly Journal of Economics*, Vol. 132/4, pp. 1593-1640, http://dx.doi.org/10.3386/w21473. [84]

Diamond, A. (2013), “Executive functions”, *Annual Review of Psychology*, Vol. 64, pp. 135-168, https://doi.org/10.1146/annurev-psych-113011-143750. [24]

Dworak, E. et al. (2020), “Using the international cognitive ability resource as an open-source tool to explore individual differences in cognitive ability”, *Personality and Individual Differences*, Vol. 169/109906, https://doi.org/10.1016/j.paid.2020.109906. [44]

Ekstrom, R. et al. (1976), *Manual for Kit of Factor-referenced Cognitive Tests*, Educational Testing Service, Princeton, NJ. [42]

Ekstrom, R. et al. (1976), *Kit of Factor-referenced Cognitive Tests*, Educational Testing Service, Princeton, NJ. [41]

Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264284395-en. [96]

Engle, R. (2002), “Working memory capacity as executive attention”, *Current Directions in Psychological Science*, Vol. 11/1, pp. 19-23, https://doi.org/10.1111/1467-8721.00160. [18]

Fleisher, M. and S. Tsacoumis (2012), “O*NET analyst occupational abilities ratings: Analysis cycle 12 results”, *HumRRO Research Report*, National Center for O*NET Development, Raleigh, NC. [90]

Fleishman, E. (1954), “Dimensional analysis of psychomotor abilities”, *Journal of Experimental Psychology*, Vol. 48/6, pp. 437-454, https://doi.org/10.1037/h0058244. [72]

Fleishman, E., D. Costanza and J. Marshall-Mies (1999), “Abilities”, in Peterson, N. et al. (eds.), *An Occupational Information System for the 21st Century: The Development of O*NET*, American Psychological Association, Washington, DC. [91]

Fleishman, E. and M. Quaintance (1984), *Taxonomies of Human Performance: The Description of Human Tasks*, Academic Press, New York. [73]

Frey, C. and M. Osborne (2017), “The future of employment: How susceptible are jobs to computerization?”, *Technological Forecasting and Social Change*, Vol. 114/January, pp. 254-280, https://doi.org/10.1016/j.techfore.2016.08.019. [94]

Geisinger, K. (2019), “Empirical considerations on intelligence testing and models of intelligence: Updates for educational measurement professionals”, *Applied Measurement in Education*, Vol. 32/3, pp. 193-197, https://doi.org/10.1080/08957347.2019.1619564. [60]

Gibson, J. (ed.) (1947), “Aptitude tests”, *Motion Picture Testing and Research Report No. 7*, U.S. Government Printing Office, Washington, DC. [38]

Gibson, J. (ed.) (1947), “Motion picture testing and research”, *Research Report*, No. 7, US Government Printing Office, Washington, DC. [37]

Guilford, J. (1950), “Creativity”, *American Psychologist*, Vol. 5/9, pp. 444-454. [39]

Guilford, J. and R. Hoepfner (1971), *The Analysis of Intelligence*, McGraw-Hill Book Co., New York. [40]

Guilford, J. and J. Lacey (1947), *Printed Classification Tests Report 5*, U.S. Government Printing Office, Washington, DC. [95]

Gustafsson, J. (1984), “A unifying model for the structure of intellectual abilities”, *Intelligence*, Vol. 8/3, pp. 179-203, https://doi.org/10.1016/0160-2896(84)90008-4. [52]

Haberman, S., S. Sinharay and G. Puhan (2011), “Reporting subscores for institutions”, *British Journal of Mathematical and Statistical Psychology*, Vol. 62/1, pp. 70-95, https://doi.org/10.1348/000711007X248875. [63]

Hellwig, S., R. Roberts and R. Schulze (2020), “A new approach to assessing emotional understanding”, *Psychological Assessment*, Vol. 32/7, pp. 649-662, http://dx.doi.org/10.1037/pas0000822. [82]

Holzinger, K. and H. Harman (1938), “Comparison of two factorial analyses”, *Psychometrika*, Vol. 3, pp. 45-60. [54]

Holzinger, K. and F. Swineford (1937), “The bi-factor method”, *Psychometrika*, Vol. 2, pp. 41-54. [53]

Horn, J. and A. Blankson (2012), “Foundations for better understanding of cognitive abilities”, in Flanagan, D. and P. Harrison (eds.), *Contemporary Intellectual Assessment: Theories, Tests and Issues, 3rd ed.*, The Guilford Press, New York. [5]

Horn, J. and N. Blankson (2005), “Foundations for better understanding of cognitive abilities”, in Flanagan, D. and P. Harrison (eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues, 2nd ed.*, The Guilford Press, New York. [4]

Horn, J. and R. Cattell (1966), “Refinement and test of the theory of fluid and crystallized general intelligences”, *Journal of Educational Psychology*, Vol. 57/5, pp. 253-270, https://doi.org/10.1037/h0023816. [3]

Humphreys, L. (1947), “Tests of intellect and information”, in Guilford, J. and J. Lacey (eds.), *Printed Classification Tests Report*, Government Printing Office, Washington, DC. [35]

Hunt, E. (2011), *Human Intelligence*, Cambridge University Press, New York. [8]

Jewsbury, P., S. Bowden and M. Strauss (2016), “Integrating the switching, inhibition, and updating model of executive function with the Cattell-Horn-Carroll model”, *Journal of Experimental Psychology: General*, Vol. 145/2, pp. 220-245, http://dx.doi.org/10.1037/xge0000119. [47]

Johnson, W. (2018), “A tempest in a ladle: The debate about the roles of general and specific abilities in predicting important outcomes”, *Journal of Intelligence*, Vol. 6/2, p. 24, http://dx.doi.org/10.3390/jintelligence6020024. [65]

Johnson, W. and T. Bouchard (2005), “The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized”, *Intelligence*, Vol. 33/4, pp. 393-416, https://doi.org/10.1016/j.intell.2004.12.002. [10]

Johnson, W., J. te Nijenhuis and T. Bourchar (2007), “Replication of the hierarchical visual-perceptual-image rotation model in de Wolff and Buiten’s (1963) battery of 46 tests of mental ability”, *Intelligence*, Vol. 35, pp. 69-81, https://doi.org/10.1016/j.in. [11]

Just, M. and S. Varma (2007), “The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition”, *Cognitive, Affective, & Behavioral Neuroscience*, Vol. 7/3, pp. 153-191, https://doi.org/10.3758/CABN.7.3.153. [28]

Kane, M. et al. (2001), “A controlled-attention view of WM capacity”, *Journal of Experimental Psychology: General*, Vol. 130, pp. 169-183, http://dx.doi.org/10.1037//0096-3445.130.2.169. [19]

Kieras, D. and D. Meyer (1997), “An overview of the EPIC architecture for cognition and performance with application to human-computer interaction”, *Human-Computer Interaction*, Vol. 12, pp. 391-438, https://doi.org/10.1207/s15327051hci1204_4. [29]

Kyllonen, P. (2020), *Reasoning Abilities*, Oxford Research Encyclopedia of Education, Oxford. [67]

Kyllonen, P. and R. Christal (1990), “Reasoning ability is (little more than) working-memory capacity?!”, *Intelligence*, Vol. 14/4, pp. 389-433, https://doi.org/10.1016/S0160-2896(05)80012-1. [17]

Kyllonen, P. et al. (2019), “General fluid/inductive reasoning battery for a high-ability population”, *Behavior Research Methods*, Vol. 51/2, pp. 502-522. [66]

Kyllonen, P. and J. Zu (2016), “Use of response time for measuring cognitive ability”, *Journal of Intelligence*, Vol. 4/4, p. 14, https://doi.org/10.3390/jintelligence4040014. [64]

Landauer, T. (1986), “How much do people remember? Some estimates of the quantity of learned information in long-term memory”, *Cognitive Science*, Vol. 10/4, pp. 477-493, https://doi.org/10.1207/s15516709cog1004_4. [93]

Larson, J. (2010), *In Search of Synergy in Small Group Performance*, Psychology Press, London. [86]

Lohman, D. (1979), “Spatial ability: A review and reanalysis of the correlational literature”, *Technical Report*, No. 8, DTIC AD A075972, Stanford Aptitude Research Project, School of Education, Stanford University, CA, https://apps.dtic.mil/sti/p. [69]

MacCann, C. et al. (2014), “Emotional intelligence is a second-stratum factor of intelligence: Evidence from hierarchical and bifactor models”, *Emotion*, Vol. 14/2, pp. 358-374, https://doi.org/10.1037/a0034755. [77]

MacCann, C. and R. Roberts (2008), “New paradigms for assessing emotional intelligence: Theory and data”, *Emotion*, Vol. 8, pp. 540-551, http://dx.doi.org/10.1037/a0012746. [78]

Marshalek, B. (1981), “Trait and process aspects of vocabulary knowledge and verbal ability”, *Technical Report*, No. 15, DTIC AD A102757, Stanford Aptitude Research Project, School of Education, Stanford, University, CA, https://apps.dtic.mil/dtic/tr. [68]

Mayer, J. and P. Salovey (1993), “The intelligence of emotional intelligence”, *Intelligence*, Vol. 17/4, pp. 433-442, http://dx.doi.org/10.1016/0160-2896(93)90010-3. [76]

McGill, R. and S. Dombrowski (2019), “Critically reflecting on the origins, evolution, and impact of the Cattell-Horn-Carroll (CHC) model”, *Applied Measurement in Education*, Vol. 32/3, pp. 216-231, http://dx.doi.org/10.1080/08957347.2019.1619561. [61]

McGrew, K. and R. Woodcock (2001), *Technical Manual: Woodcock-Johnson III*, Riverside Publishing, Itasca, IL. [7]

National Center for O*NET Development (n.d.), *O*NET Online*, website, https://www.onetonline.org (accessed on 1 December 2020). [87]

National Research Council (2010), *A Database for a Changing Economy: Review of the Occupational Information Network (O*NET)*, National Academies Press, Washington, DC. [88]

Newell, A. (1994), *Unified Theories of Cognition*, Harvard University Press, Cambridge, MA. [30]

OECD (2017), *PISA 2015 Results (Volume V): Collaborative Problem Solving*, PISA, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264285521-en. [85]

Olderbak, S. et al. (2021), “Reliability generalization of tasks and recommendations for assessing the ability to perceive facial expressions of emotion”, *Psychological Assessment, Advance online publication*, https://doi.org/10.1037/pas0001030. [80]

Peterson, N. et al. (1999), *An Occupational Information System for the 21st Century: The Development of O*NET*, American Psychological Association, Washington, DC. [89]

Reynolds, C. (1997), “Forward and backward memory span should not be combined for clinical analysis”, *Archives of Clinical Neuropsychology*, Vol. 12/1, pp. 29-40, https://doi.org/10.1016/S0887-6177(96)00015-7. [71]

Russell, S. and P. Norvig (2010), *Artificial Intelligence: A Modern Approach, 3rd edition*, Pearson, London. [32]

Savi, A. et al. (2019), “The wiring of intelligence”, *Perspectives on Psychological Science*, Vol. 14/6, pp. 1034-1061, https://doi.org/10.1177%2F1745691619866447. [15]

Scherer, K. and U. Scherer (2011), “Assessing the ability to recognize facial and vocal expressions of emotion: Construction and validation of the emotion recognition index”, *Journal of Nonverbal Behavior*, Vol. 35/4, pp. 305-326, https://doi.org/10.1007/s10919-011-0115-4. [81]

Schmid, J. and J. Leiman (1957), “The development of hierarchical factor solutions”, *Psychometrika*, Vol. 23, pp. 53-61, https://doi.org/10.1007/BF02289209. [55]

Schneider, W. and K. McGrew (2018), “The Cattell–Horn–Carroll theory of cognitive abilities”, in Flanagan, D. and E. McDonough (eds.), *Contemporary Intellectual Assessment: Theories, Tests and Issues*, The Guilford Press, New York. [56]

Schrank, F. (2011), “Woodcock-Johnson III tests of cognitive abilities”, in Davis, A. (ed.), *Handbook of Pediatric Neuropsychology*, Springer Publishing Company, New York. [45]

Schrank, F., N. Mather and K. McGrew (2014), *Woodcock-Johnson IV Tests of Achievement*, Riverside, Rolling Meadows, IL. [46]

Shuey, E. and M. Kankaraš (2018), “The Power and Promise of Early Learning”*, OECD Education Working Papers*, No. 186, OECD Publishing, Paris, https://dx.doi.org/10.1787/f9b2e53f-en. [21]

Spearman, C. (1927), *The Abilities of Man*, MacMillan, Basingstoke, UK. [2]

Spearman, C. (1904), “’General intelligence’ objectively determined and measured”, *American Journal of Psychology*, Vol. 15, pp. 201-293, https://doi.org/10.2307/1412107. [1]

Stankov, L. (2019), “Diminished ‘g’: Fluid and crystallized intelligence and cognitive abilities linked to sensory modalities”, in McFarland, D. (ed.), *General and Specific Mental Abilities*, Cambridge Scholars Publishing, Cambridge, UK. [75]

Thomas, R. et al. (2008), “Diagnostic hypothesis generation and human judgment”, *Psychological Review*, Vol. 115/1, pp. 155-185, https://doi.org/10.1037/0033-295X.115.1.155. [31]

Thomson, G. (1916), “A hierarchy without a general factor”, *British Journal of Psychology*, Vol. 8, pp. 271-281, http://dx.doi.org/10.1111/j.2044-8295.1916.tb00133.x. [12]

Thurstone, L. (1938), *Primary Mental Abilities*, University of Chicago Press. [34]

Thurstone, L. (1934), “The vectors of the mind”, *Psychological Review*, Vol. 41, pp. 1-32, https://doi.org/10.1037/h0075959. [50]

Tirre, W. (1994), “Bond sampling theory of human abilities”, in Sternberg, R. (ed.), *Encyclopedia of Human Intelligence*, Macmillan, New York. [13]

Undheim, J. (1981), “On Intelligence III: Examining developmental implications of Catell’s broad ability theory and of an alternative neo-Spearman model”, *Scandinavian Journal of Psychology*, Vol. 22/4, pp. 243-249, https://doi.org/10.1111/j.1467-9450.1981.tb00400.x. [51]

van der Maas, H. et al. (2019), “The network approach to general intelligence”, in *General and Specific Mental Abilities*, Cambridge Scholars Publishing, Cambridge, UK. [14]

Vernon, P. (1950), *The Structure of Human Abilities*, Methuen, London. [9]

Wasserman, J. (2019), “Deconstructing CHC”, *Applied Measurement in Education*, Vol. 32/3, pp. 249-268, https://doi.org/10.1080/08957347.2019.1619563. [62]

Wilhelm, O., A. Hilldebrandt and K. Oberauer (2013), “What is working memory capacity, and how can we measure it?”, *Frontiers in Psychology*, Vol. 4, p. 433, http://dx.doi.org/10.3389/fpsyg.2013.00433. [48]

Wissler, C. (1901), “The correlation of mental and physical tests”, *The Psychological Review: Monograph Supplements*, Vol. 3/6, pp. i-62, https://doi.org/10.1037/h0092995. [33]

Woolley, A. et al. (2010), “Evidence for a collective intelligence factor in the performance of human groups”, *Science*, Vol. 330, p. 686, http://dx.doi.org/10.1126/science. 1193147. [83]

Zelazo, P., C. Blair and M. Willoughby (2016), *Executive Function: Implications for Education (NCER 2017-2000)*, National Center for Education Research, Institute of Education Sciences, U.S. Department of Education, Washington, DC, http://ies.ed.gov/. [20]

Zelazo, P. et al. (2013), “II. NIH toolbox cognition battery (CB): Measuring executive function and attention”, *Monographs of the Society for Research in Child Development*, Vol. 78/4, pp. 16-33, http://dx.doi.org/10.1111/mono.12032. [49]

Annex 3.A. Comparison of second-order factors in three hierarchical abilities models

Annex Table 3.A.1. Second-order factors in three hierarchical abilities models

Carroll's (1993) 3-stratum model	Horn-Cattell's Gf-Gc model (1966, 2021)	CHC Model
2F Fluid intelligence	Gf Reasoning under novel conditions[2]	Gf Fluid reasoning[4]
	Gq Quantitative mathematical	Gq Quantitative knowledge
2C Crystallized intelligence[5]	Gc Acculturational knowledge[3]	Gc Comprehension knowledge[4]
		Gkn Domain-specific knowledge
		Grw Reading and writing
2Y General Memory/Learning	SAR/Gsm Short-term apprehension, retrieval[2]	Gsm Short-term-memory[4]
2V Broad Visual Perception	Gv Visualization and spatial orientation	Gv Visual processing[4]
2U Broad Auditory Perception	Ga Listening and hearing	Ga Auditory processing[4]
2R Broad Retrieval Ability	TSR/Glm Long-term storage and retrieval[3]	Glr Long-term storage and retrieval[4]
2S Broad Cognitive Speediness	Gs Speed of thinking	Gs Processing speed[4]
2T Processing Speed		Gt Reaction and decision speed
		Gps Psychomotor speed
		Gp Psychomotor abilities
		Go Olfactory abilities
		Gh tactile abilities
		Gk Kinesthetic abilities

Note: [1] (Horn and Blankson, 2012[5]); [2] Decline with age; [3] Do not decline with age; [4] Appears in WJ III/WJ IV commercial tests; [5] 2H combines 2F & 2C.
Source: Adapted from (Carroll, 1993[6]) (Horn and Cattell, 1966[3]) (Schneider and McGrew, 2018[56]).

Annex 3.B. The WJ III and WJ IV set of factors and tests

Annex Table 3.B.1. WJ III and WJ IV set of factors and tests

Test name	Factor name	Sub-Factor	Description of task requirements
Numerical Reasoning	Gf	Quantitative Reasoning	Determine numerical sequences and a two-dimensional numerical pattern.
Concept Formation	Gf	Induction	Identify rules that make up geometric figures after being exposed to concepts.
Analysis Synthesis	Gf	General Sequential Reasoning	Analyse the structure of an incomplete logic puzzle and solve the missing parts.
Block Rotation	Gv	Mental Rotation, Visualisation	Choose geometric designs that match another design which have been physically rotated to a different position.
Spatial Relations	Gv	Spatial Relations	Select the component parts of whole shape.
Picture Recognition	Gv	Visual Memory	Study five images, remember them and recognise them in a larger set of other arranged images.
Visual Matching	Gs	Perceptual Speed	Quickly find and circle two identical numbers in a row of six numbers in 3 minutes.
Decision Speed	Gs	Mental Comparison Speed	Quickly analyse a row of images and mark two images that are the most closely related in 3 minutes.
Cross out	Gs	Perceptual Speed & Rate of Test Taking	Mark drawings that are identical to the first drawing in the row in 3 minutes.
Rapid Picture Naming	Gs	Naming Facility	Quickly name a series of pictures as fast as possible.
Retrieval Fluency	Glr	Ideational Fluency	State as many words from specified categories as possible in 1 minute.
Visual Auditory Learning: Delayed	Glr	Associative Memory	Recall and relearn (after a 30-minute to 8-day delay) symbols presented in.
Visual Auditory Learning	Glr	Associative Memory	Translate visual symbols after being given orally presented words that are associated with them.
Memory For Names	Glr	Associative Memory	Remember an increasingly large number of names of novel cartoon characters.
Memory For Names: Delayed	Glr	Associative Memory	Recall and relearn (after a 30-minute to 8-day delay) names of novel cartoon.
Sound Blending	Ga	Phonetic Coding Synthesis	Listen to a series of individual syllables, individual phonemes, or both that form words and name the complete words.
Incomplete Words	Ga	Phonetic Coding Analysis	Listen to words with one or more phonemes missing and name the complete words.
Sound Patterns	Ga	Speech-Sound Discrimination	Indicate whether pairs of complex sound patterns are the same or different. The patterns may differ in pitch, rhythm, or sound content.
Auditory Working Memory	Gsm	Working Memory	Listen to a mixed series of words and digits and then to rearrange them by first saying the words in order and then the numbers.
Numbers Reversed	Gsm	Working Memory	Repeat a series of random numbers backward
Memory For Words	Gsm	Memory Span	Repeat lists of unrelated words in the correct sequence
Memory For Sentences	Gsm	Memory Span	Repeat complete sentences.
Picture Vocabulary	Gc	Lexical Knowledge	Name familiar and unfamiliar pictured objects.
Verbal Comprehension	Gc	Language Development & Lexical Knowledge	Name familiar and unfamiliar pictured objects and then say words similar in meaning to word presented, say words that are opposites in meaning to the word presented, and complete phrases with words that complete analogies.

General Information	Gc	General Information	Provide characteristics of objects by responding to questions, such as “Where would you find…?” and “ What you would do with…?”.
Academic Knowledge	Gc	General Information	Provide information about biological and physical sciences, history, geography, government, economics, art, music and literature.
Oral Comprehension	Gc	Listening Ability	Listen to a short passage and orally supply the word missing at the end of the passage.
Story Recall	Gc	Listening Ability	Listen to a short passage and describe the details.
Verbal Attention (WJIV only)	Gsm	Working memory capacity	Listen to a series of numbers and animal words mixed together and answer questions regarding the sequence.
Number Series (WJIV only)	Gf	Quantitative reasoning	Participants have to identify the correct number in a series of numbers that correctly completed the series. Ex. (2, 4, ?, 8, 10,…)
Letter-Pattern Matching (WJIV only)	Gs	Perceptual speed	Quickly find and circle identical letters and patterns.
Visualisation (WJIV only)	Gv	Mental rotation, Visualisation	Identify two sets of 2D pieces that form a specific shape; also identify two sets of 3D rotated blocks that match another shape.
Phonological Processing (WJIV only)	Ga	Phonetic coding, Word fluency	Name words that begin with a certain sound; also use parts of words to create new ones.
Nonword Repetition (WJIV only)	Ga	Phonetic coding	Listen to a nonsense word and repeat the word exactly.
Segmentation (WJIV only)	Ga	Phonetic coding	Listen to words and break them into syllables and phonemes.

Note: *Appears in WJ IV, not WJ III.
Source: Adapted from (Schrank, 2011[45]); (Schrank, Mather and McGrew, 2014[46]).

Annex 3.C. Factor hierarchy in Carroll's three-stratum model of human cognitive abilities

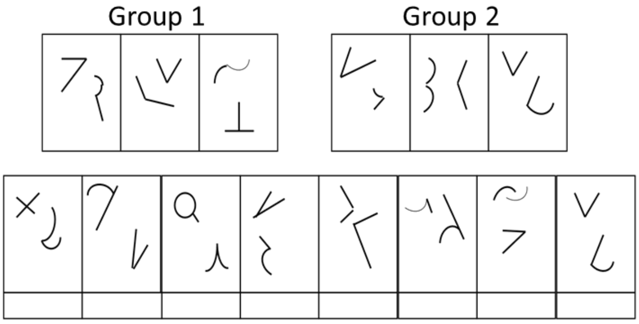
Annex Table 3.C.1. Factor hierarchy in Carroll's three-stratum model

Stratum III	Stratum II	Stratum I
general intelligence	fluid intelligence	general sequential reasoning induction quantitative reasoning speed of reasoning
	crystallised intelligence	language development lexical knowledge learning ability phonetic coding communication ability oral production and fluency (seven more)
	general memory & learning	memory span associative memory meaningful memory free recall memory visual memory learning abilities broad visual perception
	broad visual perception	visualisation spatial relations coding speed flexibility of closure perceptual speed spatial scanning (six more)
	broad auditory perception	speech-sound discrimination general sound discrimination resistance to auditory stimulus distortion temporal tracking memory for sound patterns musical discrimination and judgement (five more)
	broad retrieval ability	ideational fluency associational fluency expressional fluency naming fluency word fluency originality/creativity (three more)
	broad cognitive speediness	rate of test taking numerical facility perceptual speed
	processing speed	simple reaction time choice reaction time semantic processing time mental comparison speed

Annex 3.D. Sample items

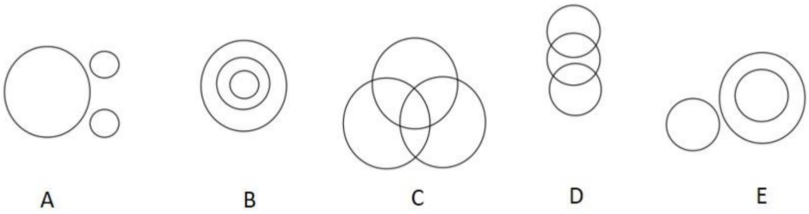
Annex Figure 3.D.1. Example fluid intelligence test items

Figure sets: A test of the Induction factor within the Fluid intelligence domain



Note: Other Induction test examples include figural, verbal, or numerical sets, series, and matrices tests. The task is to classify the 8 items below into Groups 1 or 2 by inducing the rule from the exemplars above.

Diagramming Relationships: A test of the Sequential (Deductive) Reasoning factor within the Fluid intelligence domain

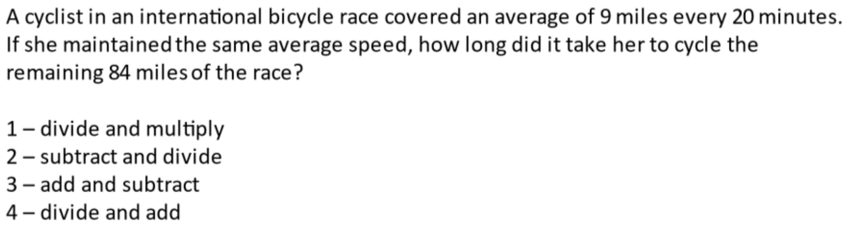


Pines, trees, stones

A B C D E

Note: Other examples include logical deductions. The task is to choose the Euler diagram that reflects the relationships among the listed entities.

Necessary Arithmetic Operations: A test of the Quantitative Reasoning factor within the Fluid intelligence domain



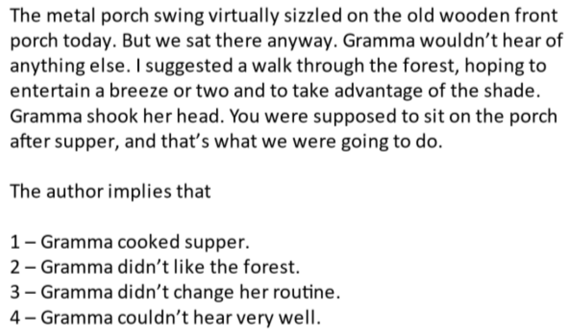
A cyclist in an international bicycle race covered an average of 9 miles every 20 minutes. If she maintained the same average speed, how long did it take her to cycle the remaining 84 miles of the race?

1 – divide and multiply
2 – subtract and divide
3 – add and subtract
4 – divide and add

Note: Other examples include mathematics word problems. The task is to indicate which operations would be required to solve the problem.
Source: Ekstrom et al. (1976[41]).

Annex Figure 3.D.2. Example crystallised intelligence test items

Reading Comprehension: A test of Reading Comprehension within the Crystallized intelligence domain



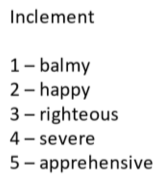
The metal porch swing virtually sizzled on the old wooden front porch today. But we sat there anyway. Gramma wouldn't hear of anything else. I suggested a walk through the forest, hoping to entertain a breeze or two and to take advantage of the shade. Gramma shook her head. You were supposed to sit on the porch after supper, and that's what we were going to do.

The author implies that

1 – Gramma cooked supper.
2 – Gramma didn't like the forest.
3 – Gramma didn't change her routine.
4 – Gramma couldn't hear very well.

Note: The task is to select the best characterization of the passage from the choices given.

Vocabulary: A test of Lexical Knowledge within the Crystallized intelligence domain

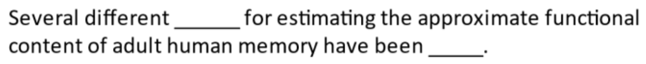


Inclement

1 – balmy
2 – happy
3 – righteous
4 – severe
5 – apprehensive

Note: The task is to identify the closest synonym to the target word.

Cloze: A test of Cloze Ability within the Crystallized intelligence domain

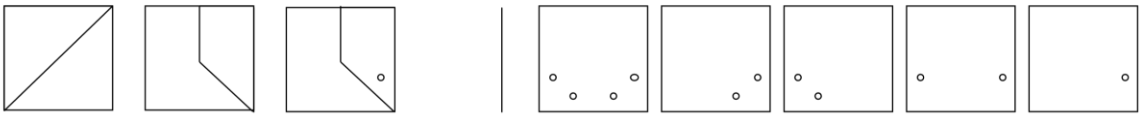


Several different ______ for estimating the approximate functional content of adult human memory have been _____.

Note: The task is to fill in the blanks through inferencing.
Source: First two panels, (Ekstrom et al., 1976[41]); third panel, (Landauer, 1986[93]).

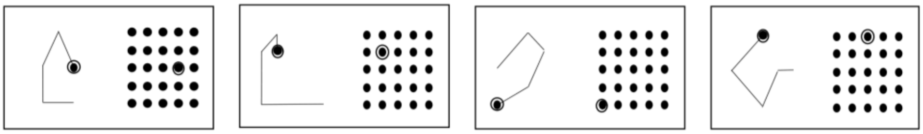
Annex Figure 3.D.3. Example broad visual perception test items

Paper Folding: A test of Spatial Visualization within the Broad Visual Perception domain



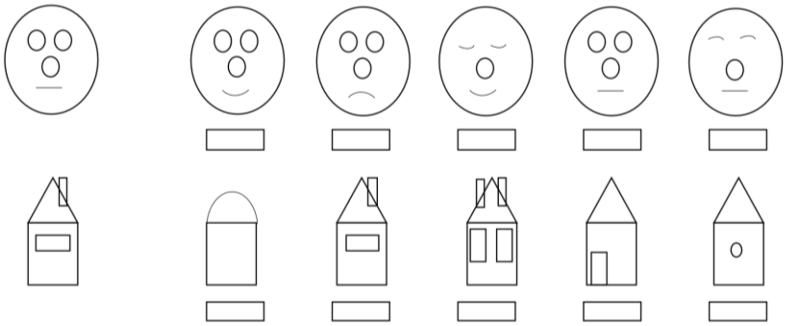
Note: The task is to select the unfolded diagram from the options on the right based on the pattern of folding and punched holes in the depiction on the left.

Copying: A test of Closure Flexibility within the Broad Visual Perception domain



Note: The task is to copy the image on the left by connecting the appropriate dots on the right.

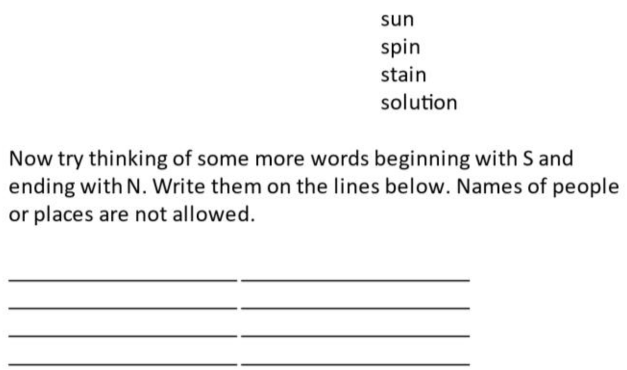
Identical Pictures: A test of Perceptual Speed within the Broad Visual Perception domain



Note: The task is to select the picture on the right that matches the target picture on the left.
Source: Ekstrom et al. (1976[41]).

Annex Figure 3.D.4. Example broad retrieval ability test items

Word Beginnings and Endings: A test of Word Fluency within the Broad Retrieval Ability domain

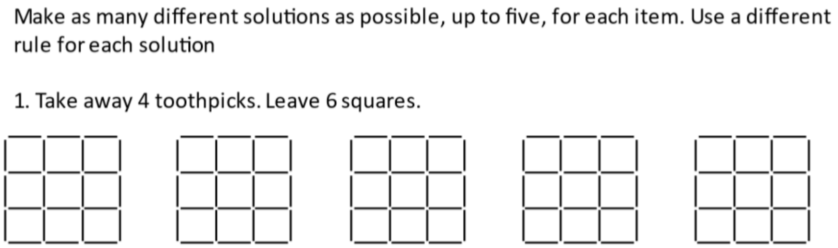


sun
spin
stain
solution

Now try thinking of some more words beginning with S and ending with N. Write them on the lines below. Names of people or places are not allowed.

Note: The task is to generate as many words as possible within a time limit that meet the constraints given.

Matchsticks: A test of Figural Flexibility within the Broad Retrieval Ability domain

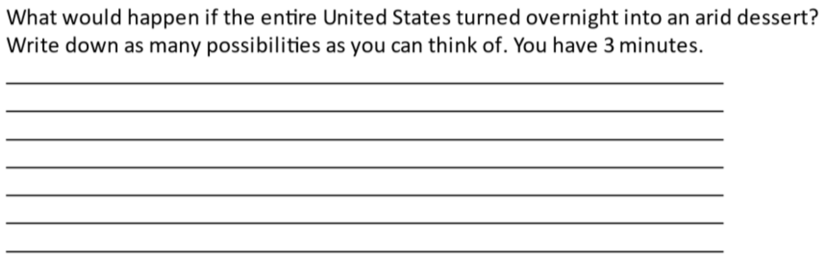


Make as many different solutions as possible, up to five, for each item. Use a different rule for each solution

1. Take away 4 toothpicks. Leave 6 squares.

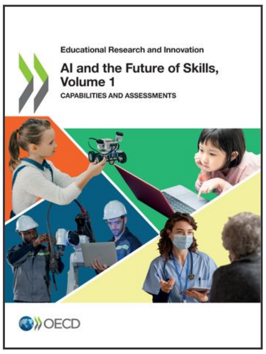
Note: The task is to generate as many solutions as possible within a time limit that meet the constraints given.

Consequences: A test of Creativity within the Broad Retrieval Ability domain



What would happen if the entire United States turned overnight into an arid dessert? Write down as many possibilities as you can think of. You have 3 minutes.

Note: The task is to generate as many implications as possible within a time limit. Scores are based on the number of unique, on topic responses given in 3 minutes.
Source: First two panels, (Ekstrom et al., 1976[41]); third panel, (Christensen, Merrifield and Guilford, 1953[70]).



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:

Kyllonen, Patrick C. (2021), “Taxonomy of cognitive abilities and measures for assessing artificial intelligence and robotics capabilities”, in OECD, *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/feecd512-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.