A Test for a Composite General Intelligence

1 Abstract

The proposed Composite General Intelligence Test (CGI) builds on the premise of Process-Overlap Theory (Kovacs and Conway, 2016) and the Three-Stratum Theory (Carroll, 1993) and uses these concepts as a foundation, detailing domain level tests for broad skills which will be evaluated and aggregated to calculate a score correlating to *g* (Cattell, 1963; Carroll, 1993), representing a metric of intelligence for a test participant to be compared to a human intelligence level benchmark.

2 Introduction

2.1 Defining Human Intelligence

To rigorously define general intelligence, we must start with the only universally acknowledged intelligent system: humans. Among various frameworks, IQ remains prominent – despite criticisms regarding bias (Weiss and Saklofske, 2020) – due to its practical value, including correlation with academic performance (Mayes et al., 2009). IQ has been frequently identified with the *g*-factor (Gottfredson, 1997), a statistical construct that *g*-theory describes as a direct psychological attribute (Kovacs and Conway, 2019).

Recent studies have challenged the concept of a single psychological attribute describing intelligence in humans, emphasising instead diverse subtypes (Gardner, 2011) and treating general intelligence as an emergent property (Kovacs and Conway, 2019). Process-Overlap Theory (Kovacs and Conway, 2016) explains intelligent behaviour as emerging from a variety of semi-autonomous cognitive processes, including executive functions like attention control (Baddeley, 1996; Conway, Kane and Engle, 2003; Kovacs and Conway, 2019). Similarly, Carroll's Three-Stratum Theory (1993) presents a hierarchical model, with *g* at the highest stratum, and increasingly narrower cognitive abilities throughout the subsequent two layers, supporting the view of multiple, interrelated cognitive abilities contributing to overall intelligence.

Drawing on these models we propose the following definition of general and human intelligence:

Human intelligence, and by extension general intelligence, is an emergent property of well-defined domains of narrow ('domain') intelligence, which are themselves reducible to psychological or cognitive processes.

2.2 Testing General Intelligence

The CGI test is designed to evaluate machine intelligence against human cognitive abilities, providing a benchmark for general intelligence (Bubeck et al., 2023). The CGI focuses on seven key domains of intelligence, aligning with contemporary theories of intelligence (Carroll, 1993; Kovacs and Conway, 2016) whilst aiming to assess intelligent agents on the basis of the generality of its application, whilst also providing meaningful scoring for applications of narrow intelligence (as in so-called 'weak AI').

This multi-domain approach serves three primary purposes: (1) It enables comprehensive assessment of general intelligence as an emergent property of diverse cognitive processes; (2) It provides a standardised method to benchmark AI progress towards human-level cognition; and (3) It identifies specific areas of machine strength and limitation, informing future AI development.

3 Methodology

A participant in the test takes part in a series of independent tests corresponding to narrow domains of intelligence. Well-established tests will be selected from the literature for each domain, with each test fulfilling a series of requirements to be sufficient and suitable for selection, outlined in Table 1.

Criteria	Importance
Be a sufficient test for the given intelligence domain, either through theoretical proof or an abundance of experimental evidence.	**Required**
Produce a numerical score that can be normalised between 0 and 1 using a distribution of existing test scores.	**Required**
Where relevant, account for cognitive processing speed and scale scores accordingly.	**Required**
Test the agent in a variety of environments of increasing complexity using some measurable and internally consistent definition of complexity (e.g. Levin's *Kt* complexity (Levin, 1973)).	**Required**
Integrate learning and planning within the given domain.	**Required**
Be suitable for completion by both humans and machine entrants with comparable results.	**Required**
Account for cultural and tester bias either implicitly or through corrective means.	**Desirable**
Be completable using a variety of accessible input modalities.	**Desirable**
Performing a measurement of both worst and best performance on given tasks as a more reliable predicator of *g* (Rammsayer and Troche, 2016).	**Desirable**
Resistance to passing through memorisation of test parameters and environments.	**Desirable**

Table 1: Listing the criteria for each domain intelligence test to be included, as well as desirable traits.

The normalisation process noted above is a vital part of ensuring that results from each intelligence domain can be properly weighted and incorporated into a general intelligence score (De Amorim, Cavalcanti and Cruz, 2023). Normalisation will use a distribution of test scores drawn from previously obtained test scores drawn from the literature conducted on both humans and machines.

The variable z is computed as a weighted sum of the proposed intelligence domains, with each weighting reflecting the domain's relative importance. These weights are derived through Exploratory Factor Analysis (Cudeck, 2000), applied using test score data from a pre-competition representative sample of humans, drawn from the general population, taking the domain tests independently.

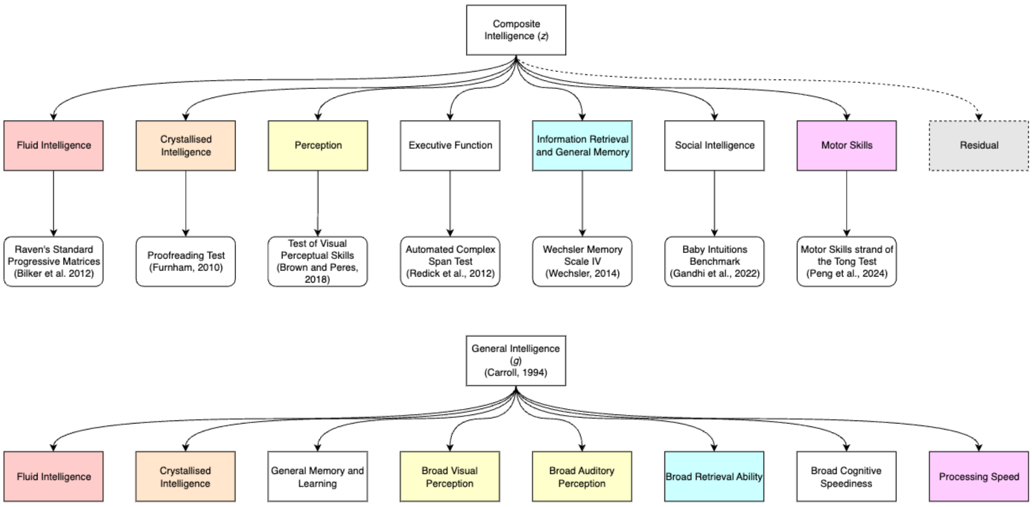

Figure 1: Level II and III of the Three-Stratum Model (Carroll, 1993). Colour matched intelligence domains examine the same areas of intelligence, with additions based on the literature review conducted for this paper; tests for each intelligence domain are marked in dotted and rounded boxes. The weight of the residual intelligence domain will be proportional to the magnitude of the variance not explained by the other intelligence domains and does not form part of the practical test.

Intelligence Domain	Justification	Test	Justification for Test
Fluid Intelligence	Included in the Carroll's Three-Stratum Model and many other measures of intelligence and well-correlated with human IQ (Jensen, 1998).	Raven's Standard Progressive Matrices (RSPM)	The RSPM test is used to assess mental ability associated with abstract reasoning, i.e. fluid intelligence. (Bilker et al., 2012).
Crystallised Intelligence	Included in the Carroll's Three-Stratum Model and many other measures of intelligence, well-correlated with human IQ (Pietschnig, Voracek and Formann, 2010).	Proofreading (PR) test (Furnham, 2010)	Proofreading tests have been found to be strongly correlated with measures of crystallised intelligence including general knowledge and vocabulary (Furnham, 2010).
Perception	Generalised for non-human intelligence but both visual and auditory perception form a core part of the Three-Stratum Model – practical limitations restrict the testing process to visual perception only.	Test of Visual Perceptual Skills (TVPS-4) (Brown and Peres, 2018)	"The TVPS-4 is a standardized assessment of motor-free visual perception skills" (Brown and Peres, 2018).
Executive Function	Process-overlap theory (Kovacs and Conway, 2016) makes clear the importance of executive function processes to moderating and directing narrower forms of intelligence into achieving goals and expressing narrow intelligence (Baddeley, 1996; Conway, Kane and Engle, 2003; Kovacs and Conway, 2019).	Automated Complex Span test (Redick et al., 2012).	Working memory tests correlate well to executive function processes such as attention control (Baddeley, 1996; Conway, Kane and Engle, 2003; Kovacs and Conway, 2019), and automated complex span tests are widely accepted as some of the most effective tests of working memory available (Redick et al., 2012).
Information Retrieval & General Memory	Included in the Carroll's Three-Stratum Model and many other measures of intelligence, some overlap with crystallised intelligence but reflects importance of information aggregation and processing in general intelligence.	Wechsler Memory Scale – IV (Wechsler, 2009)	"The Wechsler memory scales...have provided detailed assessments of clinically relevant aspects of memory functioning for over half a century" (Holdnack and Drozdick, 2010).
Social Intelligence	Social cognition, 'the ability to think about the minds of other agents' (Gweon, Fan and Kim, 2023), is increasingly understood as a key part of human-computer and human-human interaction, and hence necessary for intelligent cooperation – we hence describe it as a necessary component for intelligent behaviour within multi-agent systems, including the real world.	Baby Intuitions Benchmark (Gandhi et al., 2022).	Designed from the ground up to test machine-based social intelligence, the Baby Intuitions Benchmark is designed to allow for direct comparison between human and machine intelligence in social intelligence and perception of other agent's behaviours.
Motor Skills	Necessary for real-world interactions and environmental manipulation, with proven correlated with *g*-factor general intelligence (Mihaela et al., 2013).	Motor skills strand of the Tong test (Peng et al., 2024)	The Tong Test tests motor skills in dynamic virtual environments of increasing complexity, and explicitly tests motor skills in one strand of the test.

Table 2: Documenting the domains of intelligence included in the CGI test with reasoning for their inclusion, as well as the same for the tests chosen for them. *N.B.* the Perception domain is unique in that sensing and perception of an environment can use a number of different sensing modalities, including many not directly usable by humans, therefore whilst the current version of the CGI test only deals with visual perception updated versions may use approaches like SCAN-3 (Lovett and Johnson, 2010) for auditory perception and sensory fuzzing (Woodlief, Elbaum and Sullivan, 2022) for other modalities.

The competition format of the test will be open to both human and machine participants. Participant scores will primarily consist of their final evaluated z-score from 0-1, but with domain test scores also provided as an assessment of narrow intelligence. While intelligence, as described in this paper, is not regarded as a simple pass-fail endeavour, a score exceeding 90% of the minimum human-attributable z-score (as measured through z-scores computed using the above-defined pre-competition dataset) may be viewed as indicative of comparable human-level intelligence and of proving the test hypothesis that the latent statistical variable z, a function of proposed intelligence domains that correlates with g, exceeds the stated threshold value. This 90% threshold is chosen to allow for natural variation in human performance while still representing a similar level of capability to human-level intelligence.

4 Discussion

The methodology presented is feasible in principle, being based on a theoretical conceptualisation of intelligence yet one well-grounded in existing literature and evidence. The primary practical barriers are factors such as cost due to the extensive testing required, ensuring that all domain tests are reliable, valid, and free of bias (Legg and Hutter, 2007a), and maintaining and updating the test over time as intelligence distributions change in both humans (Pietschnig, Voracek and Formann, 2010) and machines due to progress in AI research. There are also validity challenges in definitively proving a correlation of z with the g-factor, a fact that should be explored in any future work, though accounting for this is to a degree built into the modularity of the test and its integration of intelligence domains independently in a manner that allows for future iterations of the test to be updated with new tests, domains, and particularly non-linear weightings.

While the CGI test allows agents with only narrow intelligence abilities to score, achieving a high overall score can only be achieved by demonstrating high skills levels across a range of domains in which humans are competent. Therefore, contestants must combine the reasoning, memory and language skills offered by machines today (Promma et al., 2025) with perception, social intelligence and motor skills. A machine with this level of holistic intelligence, and indeed to a degree any measurable intelligence in domains like social intelligence, does not yet exist, and so the contest poses a significant challenge. Machine intelligences making use of sensing modalities other than visual perception may also face lower results in this version of the test than others, though improvements in general perception testing and integration into the CGI test would resolve this.

There already exist today AI systems vastly surpassing humans in narrow domains (Silver et al., 2016; Greg Kochs, 2020), and more recently large language models score highly across a range of exams designed for humans, with multi-modal models showing impressive results when evaluated on vision-related benchmarks (OpenAI et al., 2024), demonstrating proficiency in multiple domains. And this is where the CGI test is strongest: in being able to test in which domains of intelligence AI systems excel.

References

Baddeley, A., 1996. Exploring the central executive. *The quarterly journal of experimental psychology section a* [Online], 49(1), February. Publisher: SAGE Publications, pp.5–28. Available from: `https://doi.org/10.1080/713755608` [Accessed 17 May 2025].

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T. and Zhang, Y., 2023. *Sparks of artificial general intelligence: early experiments with GPT-4* [Online]. arXiv: `2303.12712[cs.CL]`. Available from: `https://arxiv.org/abs/2303.12712`.

Carroll, J., 1993. *Human cognitive abilities: a survey of factor-analytic studies* [Online], Human Cognitive Abilities: A Survey of Factor-analytic Studies no. 1. Cambridge University Press. Available from: `https://books.google.co.uk/books?id=jp9dt4_0_cIC`.

Cattell, R.B., 1963. Theory of fluid and crystallized intelligence: a critical experiment. *Journal of educational psychology* [Online], 54(1), February, pp.1–22. Available from: `https://doi.org/10.1037/h0046743` [Accessed 1 June 2025].

Conway, A.R., Kane, M.J. and Engle, R.W., 2003. Working memory capacity and its relation to general intelligence. *Trends in cognitive sciences* [Online], 7(12), December. Publisher: Elsevier BV, pp.547–552. Available from: `https://doi.org/10.1016/j.tics.2003.10.005` [Accessed 23 May 2025].

Cudeck, R., 2000. 10 - exploratory factor analysis. In: H.E. Tinsley and S.D. Brown, eds. *Handbook of applied multivariate statistics and mathematical modeling* [Online]. San Diego: Academic Press, pp.265–296. Available from: `https://doi.org/https://doi.org/10.1016/B978-012691360-6/50011-2`.

De Amorim, L.B., Cavalcanti, G.D. and Cruz, R.M., 2023. The choice of scaling technique matters for classification performance. *Applied soft computing* [Online], 133, January, p.109924. Available from: `https://doi.org/10.1016/j.asoc.2022.109924` [Accessed 2 June 2025].

Gardner, H., 2011. *Frames of mind: the theory of multiple intelligences*. 2011 pbk. ed. OCLC: 710972961. New York: Basic Books.

Gottfredson, L.S., 1997. Why g matters: the complexity of everyday life. *Intelligence* [Online]. Special Issue Intelligence and Social Policy, 24(1), 1 January, pp.79–132. Available from: `https://doi.org/10.1016/S0160-2896(97)90014-3` [Accessed 10 May 2025].

Greg Kochs, 2020. *AlphaGo - the movie* [Online]. Available from: `https://www.youtube.com/watch?v=WXuK6gekU1Y` [Accessed 1 June 2025].

Kovacs, K. and Conway, A.R., 2016. Process overlap theory: a unified account of the general factor of intelligence. *Psychological inquiry* [Online], 27(3), 2 July, pp.151–177. Available from: `https://doi.org/10.1080/1047840X.2016.1153946` [Accessed 1 June 2025].

Kovacs, K. and Conway, A.R., 2019. What is IQ? life beyond "general intelligence". *Current directions in psychological science* [Online], 28(2), April, pp.189–194. Available from: `https://doi.org/10.1177/0963721419827275` [Accessed 14 May 2025].

Legg, S. and Hutter, M., 2007a. Tests of machine intelligence. In: M. Lungarella, F. Iida, J. Bongard and R. Pfeifer, eds. *50 years of artificial intelligence: essays dedicated to the 50th anniversary of artificial intelligence* [Online]. Berlin, Heidelberg: Springer, pp.232–242. Available from: `https://doi.org/10.1007/978-3-540-77296-5_22` [Accessed 8 May 2025].

Lovett, B.J. and Johnson, T.L., 2010. Test review: r. w. keith SCAN-3 for adolescents and adults: tests for auditory processing disorders. san antonio, TX: pearson, 2009. *Journal of psychoeducational assessment* [Online], 28(6). tex.eprint: https://doi.org/10.1177/0734282909353341, pp.603–607. Available from: `https://doi.org/10.1177/0734282909353341`.

Mayes, S.D., Calhoun, S.L., Bixler, E.O. and Zimmerman, D.N., 2009. IQ and neuropsychological predictors of academic achievement. *Learning and individual differences* [Online], 19(2), June. Publisher: Elsevier BV, pp.238–241. Available from: `https://doi.org/10.1016/j.lindif.2008.09.001` [Accessed 23 May 2025].

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, Ł., Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, J.H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, Ł., Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Peres, F.d.A.B., Petrov, M., Pinto, H.P.d.O., Michael, Pokorny, Pokrass, M., Pong, V.H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M.B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C.J., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W. and Zoph, B., 2024. *GPT-4 technical report* [Online]. arXiv, 4 March. arXiv: `2303.08774[cs]`. Available from: `https://doi.org/10.48550/arXiv.2303.08774` [Accessed 1 June 2025].

Pietschnig, J., Voracek, M. and Formann, A.K., 2010. Pervasiveness of the IQ rise: a cross-temporal meta-analysis. *PLoS ONE* [Online], 5(12), 22 December. Ed. by A.V. García, e14406. Available from: `https://doi.org/10.1371/journal.pone.0014406` [Accessed 14 May 2025].

Promma, W., Imjai, N., Usman, B. and Aujirapongpan, S., 2025. The influence of AI literacy on complex problem-solving skills through systematic thinking skills and intuition thinking skills: an empirical study in thai gen z accounting students. *Computers and education: artificial intelligence* [Online], 8, June, p.100382. Available from: `https://doi.org/10.1016/j.caeai.2025.100382` [Accessed 3 June 2025].

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D., 2016. Mastering the game of go with deep neural networks and tree search. *Nature* [Online], 529(7587), 28 January, pp.484–489. Available from: `https://doi.org/10.1038/nature16961` [Accessed 1 June 2025].

Weiss, L.G. and Saklofske, D.H., 2020. Mediators of IQ test score differences across racial and ethnic groups: the case for environmental and social justice. *Personality and individual differences* [Online], 161, July, p.109962. Available from: `https://doi.org/10.1016/j.paid.2020.109962` [Accessed 1 June 2025].

Woodlief, T., Elbaum, S. and Sullivan, K., 2022. Semantic image fuzzing of AI perception systems. *Proceedings of the 44th international conference on software engineering* [Online]. ICSE '22: 44th International Conference on Software Engineering. Pittsburgh Pennsylvania: ACM, 21 May, pp.1958–1969. Available from: `https://doi.org/10.1145/3510003.3510212` [Accessed 3 June 2025].

Bibliography

Baghaei, P. and Tabatabaee, M., 2015. The c-test: an integrative measure of crystallized intelligence. *Journal of intelligence* [Online], 3(2), 25 May, pp.46–58. Available from: `https://doi.org/10.3390/jintelligence3020046` [Accessed 26 May 2025].

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. and Plumb, I., 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* [Online], 42(2). tex.eprint: https://acamh.onlinelibrary.wiley.com/doi/pdf/10.1111/1469-7610.00715, pp.241–251. Available from: `https://doi.org/https://doi.org/10.1111/1469-7610.00715`.

Bilker, W.B., Hansen, J.A., Brensinger, C.M., Richard, J., Gur, R.E. and Gur, R.C., 2012. Development of abbreviated nine-item forms of the raven's standard progressive matrices test. *Assessment* [Online], 19(3), September, pp.354–369. Available from: `https://doi.org/10.1177/1073191112446655` [Accessed 3 June 2025].

Brown, T. and Peres, L., 2018. An overview and critique of the test of visual perception skills – fourth edition (TVPS-4). *Hong kong journal of occupational therapy* [Online], 31(2). tex.eprint: https://doi.org/10.1177/1569186118793847, pp.59–68. Available from: `https://doi.org/10.1177/1569186118793847`.

Bryson, J.J., Diamantis, M.E. and Grant, T.D., 2017. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial intelligence and law* [Online], 25(3), September, pp.273–291. Available from: `https://doi.org/10.1007/s10506-017-9214-9` [Accessed 12 May 2025].

Button, G., Coulter, J., Lee, J.R. and Sharrock, W., 1995. *Computers, minds and conduct*. [Accessed 10 May 2025].

Dombrowski, S., McGill, R. and Canivez, G., 2017. Hierarchical exploratory factor analyses of the woodcock-johnson IV full test battery: implications for CHC application in school psychology. *School psychology quarterly* [Online], 33, July, pp.235–250. Available from: `https://doi.org/10.1037/spq0000221`.

Furnham, A., 2010. Proofreading as an index of crystallised intelligence. *Educational psychology* [Online], 30(6), October, pp.735–754. Available from: `https://doi.org/10.1080/01443410.2010.506005` [Accessed 2 June 2025].

Gandhi, K., Stojnic, G., Lake, B.M. and Dillon, M.R., 2022. *Baby intuitions benchmark (BIB): discerning the goals, preferences, and actions of others* [Online]. arXiv, 11 February. arXiv: `2102.11938[cs]`. Available from: `https://doi.org/10.48550/arXiv.2102.11938` [Accessed 22 May 2025].

Google DeepMind, 2016. *Match 1 - google DeepMind challenge match: lee sedol vs AlphaGo* [Online]. 9 March. Available from: `https://www.youtube.com/watch?v=vFr3K2DORc8` [Accessed 29 May 2025].

Gweon, H., Fan, J. and Kim, B., 2023. Socially intelligent machines that learn from humans and help humans learn. *Philosophical transactions of the royal society a: mathematical, physical and engineering sciences* [Online], 381(2251), 24 July. Publisher: The Royal Society. Available from: `https://doi.org/10.1098/rsta.2022.0048` [Accessed 22 May 2025].

Hernandez-Orallo, J., 1998. A formal definition of intelligence based on an intensional variant of algorithmic complexity, 8 July.

Hernández-Orallo, J., 2001. On the computational measurement of intelligence factors. In: Meystel, A. and Messina, E.R. *Measuring the performance and intelligence of systems: proceedings of the 2000 PerMIS workshop, august 14-16, 2000*. Google-Books-ID: yAhRAAAAMAAJ. The Institute.

Hernández-Orallo, J. and Dowe, D.L., 2010. Measuring universal intelligence: towards an anytime intelligence test. *Artificial intelligence* [Online], 174(18), December, pp.1508–1539. Available from: `https://doi.org/10.1016/j.artint.2010.09.006` [Accessed 2 June 2025].

Holdnack, J.A. and Drozdick, L.W., 2010. Using WAIS-IV with WMS-IV. *WAIS-IV clinical use and interpretation* [Online]. Elsevier, pp.237–283. Available from: `https://doi.org/10.1016/B978-0-12-375035-8.10009-6` [Accessed 3 June 2025].

Huang, J. and Li, O., 2024. *Measuring the IQ of mainstream large language models in chinese using the wechsler adult intelligence scale* [Online]. Available from: https://www.authorea.com/doi/full/10.36227/techrxiv.171778886.62839657?commit=f530ea69de68e61523480091048428fd413595ad [Accessed 14 May 2025].

Jensen, A.R., 1998. *The g factor: the science of mental ability*. 1. publ, Human evolution, behavior, and intelligence. Westport, Conn.: Praeger. 648pp.

Kaufman, A.S., Choi, D., Kapoor, H. and Kaufman, J.C., 2022. A brief history of IQ testing: fixed vs. malleable intelligence. In: R.J. Sternberg and D.D. Preiss, eds. *Intelligence in context: the cultural and historical foundations of human intelligence* [Online]. Cham: Springer International Publishing, pp.59–92. Available from: https://doi.org/10.1007/978-3-030-92798-1_4 [Accessed 14 May 2025].

Keith, R., 2009. *SCAN-3: c tests for auditory processing disorders for children*. San Antonio, TX: Pearson.

Korteling, J.E.(, Van De Boer-Visschedijk, G.C., Blankendaal, R.A.M., Boonekamp, R.C. and Eikelboom, A.R., 2021. Human- versus artificial intelligence. *Frontiers in artificial intelligence* [Online], 4, 25 March, p.622364. Available from: https://doi.org/10.3389/frai.2021.622364 [Accessed 15 May 2025].

Legg, S. and Hutter, M., 2007b. Universal intelligence: a definition of machine intelligence. *Minds and machines* [Online], 17(4), 1 December, pp.391–444. Available from: https://doi.org/10.1007/s11023-007-9079-x [Accessed 27 May 2025].

Levin, L.A., 1973. On the notion of a random sequence [Online]. 14(5), pp.1413–1416. Available from: https://cir.nii.ac.jp/crid/1574231875382165120.

Mihaela, P., Gabriela, G., Catalin, P., Gabriel, P. and Nicolae, E., 2013. Relationship between general intelligence and motor skills learning specific to combat sports. *Procedia - social and behavioral sciences* [Online], 84, July, pp.728–732. Available from: https://doi.org/10.1016/j.sbspro.2013.06.635 [Accessed 1 June 2025].

Peng, Y., Han, J., Zhang, Z., Fan, L., Liu, T., Qi, S., Feng, X., Ma, Y., Wang, Y. and Zhu, S.-C., 2024. The tong test: evaluating artificial general intelligence through dynamic embodied physical and social interactions. *Engineering* [Online], 34, March, pp.12–22. Available from: https://doi.org/10.1016/j.eng.2023.07.006 [Accessed 26 May 2025].

Rammsayer, T. and Troche, S., 2016. Validity of the worst performance rule as a function of task complexity and psychometric g: on the crucial role of g saturation. *Journal of intelligence* [Online], 4(1), 16 March, p.5. Available from: https://doi.org/10.3390/jintelligence4010005 [Accessed 1 June 2025].

Redick, T.S., Broadway, J.M., Meier, M.E., Kuriakose, P.S., Unsworth, N., Kane, M.J. and Engle, R.W., 2012. Measuring working memory capacity with automated complex span tasks. *European journal of psychological assessment* [Online], 28(3), 1 January, pp.164–171. Available from: `https://doi.org/10.1027/1015-5759/a000123` [Accessed 26 May 2025].

Schaul, T., Togelius, J. and Schmidhuber, J., 2011. *Measuring intelligence through games* [Online]. arXiv, 6 September. arXiv: `1109.1314[cs]`. Available from: `https://doi.org/10.48550/arXiv.1109.1314` [Accessed 2 June 2025].

Stewart, J., Anthony, L., Batty, A.O., Nakamura, K., Nicklin, C., McLean, S. and Tomaru, K., 2025. Can we reliably score meaning recall vocabulary tests using AI? a comparison of human vs. AI scoring. *Computer assisted language learning* [Online], 9 April, pp.1–23. Available from: `https://doi.org/10.1080/09588221.2025.2481397` [Accessed 26 May 2025].

Stoeckel, T., McLean, S. and Nation, P., 2021. LIMITATIONS OF SIZE AND LEVELS TESTS OF WRITTEN RECEPTIVE VOCABULARY KNOWLEDGE. *Studies in second language acquisition* [Online], 43(1), March, pp.181–203. Available from: `https://doi.org/10.1017/S027226312000025X` [Accessed 26 May 2025].

Wechsler, D., 2009. *WMS-IV: wechsler memory scale*. PsychCorp.