# Historical account of computer models solving IQ test problems

**Fernando Martínez-Plumed**[1] and **José Hernández-Orallo**[2] and **Ute Schmid**[3] and **Michael Siebers**[4] and **David L. Dowe**[5]

**Abstract.** In this short paper we summarise our work in [9] where we make a review of what has been done when intelligence test problems have been analysed through cognitive models or particular systems. This work has been motivated by an observed explosion of the number of papers on this topic in recent years. We have made a general account of all these works in terms of how they relate to each other and what their real achievements are. Not only do we aim at analysing the meaning, utility, and impact of these computer models, but also better understanding what these tests measure in machines, whether they are useful to evaluate AI systems, whether they are really challenging problems, and whether they are useful to understand (human) intelligence.

## 1 INTRODUCTION

In the early days of artificial intelligence, the IQ test classical approach to human intelligence evaluation was considered useful not only as a tool for the study of cognitive processes and the development of new techniques, but also for the evaluation of AI systems or even as the goal for AI research. Since then, human psychometric tests have been repeatedly suggested as a much better alternative to most task-oriented evaluation approaches in AI. The question thus is whether this measurement of mental developmental capabilities leads to a feasible, practical evaluation for AI.

In this paper we briefly review our work in [9], where we analysed all the computer models taking intelligence tests (or as many as we could find, about thirty in total), starting with Evans's ANALOGY [6] and going through to Spaun [5], a noteworthy 2.5-million-neuron artificial model brain. This analysis was motivated by an observed explosion in recent years of the number of papers featuring computer models addressing intelligence test problems. featuring computer models addressing intelligence tests problems. We wanted to investigate whether this increase was casual or was motivated by an increasing need of these tests and the computer models solving them. Overall, the main goal of the paper was to understand the meaning, utility, and impact of these computer models taking intelligence tests, and explore the progress and implications of this area of research.

## 2 HISTORICAL ACCOUNT

The relation between artificial intelligence and psychometrics started more than fifty years ago. As early as 1963, Evans [6] and Simon and Kotovsky [18] devised AI programs able to identify regularities in patterns (respectively, analogy tasks and letter series completion problems).

After the initial interest of AI research in IQ test problems, this branch of research sank into oblivion during the twenty of so years. However, since the 1980s, cognitive science research recovered this line of research. Hofstadter developed a series of computational models in the Copycat project [11] with the major goal of understanding analogy. In the 1990s, some cognitive models were proposed to simulate the human cognitive processes that take place when solving inductive inference IQ test problems [2].

In AI, forty years after the work of Evans and Simon & Kotovsky, in 2003, computer programs solving intelligence tests became of interest again. On one hand, Sanghi and Dowe [16] wanted to make a conclusive point about how easy it was to make non-intelligent machines pass intelligence tests. This could have dealt a definitive deathblow to this already ebbing approach. On the other hand, Bringsjord and Schimanski aimed at resuscitating the role of psychometric tests—including not only intelligence tests but also tests about personality, artistic creativity, etc.—in AI [1]. They claimed that psychometric tests should not be dismissed but placed at a definitional, major role for what artificial intelligence is and proposed "psychometric artificial intelligence" (PAI) as a direction of research.

But the fact is that the past ten (and especially five) years (since 2006 and especially 2011) have seen a boom of computational models aimed at solving intelligence test problems. The diversity of goals and approaches has also widened, including the use of intelligence tests for the analysis of what intelligence is, for the understanding of certain aspects of human cognition, for the evaluation of some AI techniques or systems, including robots, and, simply, to have more insights about what intelligence tests really represent. See [9, Section 5] for what we hope has been a complete description of all the computer models that have addressed intelligence tests and related tests, presented in chronological order.

## 3 DISCUSSION

The analysis has not been restricted to performing a survey of all models addressing intelligence tests. Through a comprehensive account of the models we derived a set of criteria aiming at understanding the meaning, utility, and impact of these computer models taking intelligence tests, and explore the progress and implications of this area of research. Furthermore, this analysis helped us to have a better understanding of the relevance and (the limited) connections of these approaches, and draw some conclusions about their usefulness.

We have seen that most approaches are very recent [5, 12, 14, 17, 20, 21, 15, 10]. Is it an indication of relevance? According to the

[1] Universitat Politècnica de València, Spain, email: fmartinez@dsic.upv.es
[2] Universitat Politècnica de València, Spain, email: jorallo@dsic.upv.es
[3] University of Bamberg, Germany, email: ute.schmid@uni-bamberg.de
[4] University of Bamberg, Germany, email: michael.siebers@uni-bamberg.de
[5] Monash University, Australia, email: david.dowe@monash.edu

publication venues, we have seen that they go from mainstream AI to cognitive science, or even psychology, and some of them are in leading conferences and journals in these areas or even in interdisciplinary general outlets. However, it seems that most approaches aim at unveiling general (artificial) intelligence principles in ways that are not necessarily connected to the way humans solve these tests. In fact, there is a wide variety in the techniques used, from more adhoc to more general AI techniques (mostly from machine learning, pattern recognition, automated reasoning, and natural language processing). This suggests that this is attracting more interest in artificial intelligence and cognitive science than in psychology. Overall, some of these models (anthropomorphic or not) have been useful to provide insights and valuable information about how human cognition works.

What about the use of these tests for AI evaluation? Are they becoming more common? It has been recently argued—from human intelligence researchers—that intelligence tests are the right tool to evaluate AI systems [3]. Nonetheless, we have not seen that artificial intelligence has changed its evaluation protocols following this increase of models taking intelligence tests (witch a few exceptions such as [7, 19, 17, 5]). Furthermore, we have seen that even for supposedly general tasks that are designed for evaluation, many approaches have the (understandable) tendency to specialise to the task and hard-wire parts (or most) of the solution. The key issue is thus to consider a greater diversity of problems. Very few approaches address more than one kind of test. Actually, the more specific a test is the easier it is to develop specific solutions. Furthermore, different presentations and difficulty levels should be explored. The categories and overlaps between problems could be assessed via theoretical models, instead of using factor analysis as in psychometrics. In other words, a theoretical alternative to the classification of mental abilities should be considered (see [8, 4]).

There is also a huge diversity in whether performance and difficulty are assessed. We need to be clear that focussing on the overall results of a computer model and comparing them with the results of humans is not very informative about how challenging the problem is. Humans are general-purpose systems and it is not fair to compare them with some systems that are only able to solve one problem—even if the problem comes from an intelligence test. Furthermore, many of these intelligence test problems have been developed for humans, and hence it can be unfair to evaluate AI systems' limitations with anthropocentric measures. Nonetheless, some of the works perform an interesting analysis in terms of difficulty. The purpose is to determine what instances are more difficult, but this is not very related to how challenging the problem is. In fact, focussing on the most difficult problems may even make the system more specialised to the intelligence test task at hand. Some of the previous works have studied whether difficulty is related to the size of the working memory, the size of the pattern, the number of elements that need to be combined or retrieved from background knowledge or the operational constructs needed to solve this problems [18, 2, 20, 21, 13]. These notions of difficulty are much more general and can work independently of the problem and the representation.

## 4 CONCLUSION

Of the approximately 30 papers we have analysed in [9], half of them have appeared in the past five years. We wanted to investigate whether this increase was casual or was motivated by an increasing need of these tests and the computer models solving them. In order to study this we soon realised that computer models addressing intelligence tests may have different purposes and applications: to advance AI by the use of challenging problems (this is the Psychometric AI approach), to use them for the evaluation of AI systems, to better understand intelligence tests and what they measure (including item difficulty) and, finally, to better understand what (human) intelligence is. Furthermore, the use of intelligence tests for AI evaluation has provided very insightful information about what intelligence tests measure and what they do not and, ultimately, about what characterises intelligence in humans.

## REFERENCES


[1] S. Bringsjord and B. Schimanski, 'What is artificial intelligence? Psychometric AI as an answer', in *International Joint Conference on Artificial Intelligence*, pp. 887–893, (2003).

[2] P. A. Carpenter, M. A. Just, and P. Shell, 'What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test', *Psychological review*, **97**, 404–431, (1990).

[3] D. K. Detterman, 'A challenge to Watson', *Intelligence*, **39**(2-3), 77–78, (2011).

[4] D.L. Dowe and J. Hernández-Orallo, 'How universal can an intelligence test be?', *Adaptive Behavior*, **22**(1), 51–69, (2014).

[5] C. Eliasmith, T. C. Stewart, X. Choo, T. Bekolay, T. DeWolf, C. Tang, and D. Rasmussen, 'A large-scale model of the functioning brain', *Science*, **338**(6111), 1202–1205, (2012).

[6] T. Evans, 'A heuristic program to solve geometric-analogy problems', in *Proc. SJCC*, volume 25, pp. 327–339, (1965). vol. 25.

[7] S. Griffith, J. Sinapov, V. Sukhoy, and A. Stoytchev, 'A behavior-grounded approach to forming object categories: Separating containers from noncontainers', *Autonomous Mental Development, IEEE Transactions on*, **4**(1), 54–69, (2012).

[8] J. Hernández-Orallo and D. L. Dowe, 'Measuring universal intelligence: Towards an anytime intelligence test', *Artificial Intelligence*, **174**(18), 1508–1539, (2010).

[9] J. Hernández-Orallo, F. Martínez-Plumed, Ute Schmid, Michael Siebers, and David L. Dowe, 'Computer models solving intelligence test problems: Progress and implications', *Artificial Intelligence*, **230**, 74 – 107, (2016).

[10] J. Hofmann, E. Kitzelmann, and U. Schmid, 'Applying inductive program synthesis to induction of number series a case study with IGOR2', in *KI 2014: Advances in AI*, pp. 25–36. Springer, (2014).

[11] D. R. Hofstadter and M Mitchell. The copycat project, 1984.

[12] A. Lovett and K. Forbus, 'Modeling multiple strategies for solving geometric analogy problems', in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pp. 701–706, (2012).

[13] F. Martínez-Plumed, C. Ferri, J. Hernández-Orallo, and M. J. Ramírez-Quintana, 'A computational analysis of general intelligence tests for evaluating cognitive development', *submitted (second revision)*, (2016).

[14] H. Prade and G. Richard, 'From analogical proportion to logical proportions: A survey', in *Computational Approaches to Analogical Reasoning: Current Trends*, volume 548 of *Studies in Computational Intelligence*, 217–244, Springer Berlin Heidelberg, (2014).

[15] M. Ragni and S. Neubert, 'Analyzing ravens intelligence test: Cognitive model, demand, and complexity', in *Computational Approaches to Analogical Reasoning: Current Trends*, volume 548 of *Studies in Computational Intelligence*, 351–370, Springer, (2014).

[16] P. Sanghi and D. L. Dowe, 'A computer program capable of passing IQ tests', in *4th Intl. Conf. on Cognitive Science (ICCS'03), Sydney*, pp. 570–575, (2003).

[17] C. Schenck, *Intelligence tests for robots: Solving perceptual reasoning tasks with a humanoid robot*, Master's thesis, Iowa State Univ., 2013.

[18] H. A. Simon and K. Kotovsky, 'Human acquisition of concepts for sequential patterns.', *Psychological Review*, **70**(6), 534, (1963).

[19] J. Sinapov and A. Stoytchev, 'The odd one out task: Toward an intelligence test for robots', in *Development and Learning (ICDL), 2010 IEEE 9th International Conference on*, pp. 126–131. IEEE, (2010).

[20] C. Strannegård, M. Amirghasemi, and S. Ulfsbücker, 'An anthropomorphic method for number sequence problems', *Cognitive Systems Research*, **2223**, 27–34, (2013).

[21] C. Strannegård, A. Nizamani, A. Sjöberg, and F Engström, 'Bounded Kolmogorov complexity based on cognitive models', in *Artificial General Intelligence*, eds., K. U. Kühnberger, S. Rudolph, and P. Wang, volume 7999 of *LNCS*, 130–139, Springer Berlin Heidelberg, (2013).