Computer Models Solving Intelligence Test Problems: Progress and Implications*

José Hernández-Orallo
Universitat Politècnica de València,
jorallo@dsic.upv.es

Fernando Martínez-Plumed
Universitat Politècnica de València,
fmartinez@dsic.upv.es

Ute Schmid
University of Bamberg, Germany
ute.schmid@uni-bamberg.de

Michael Siebers
University of Bamberg, Germany
michael.siebers@uni-bamberg.de

David L. Dowe
Monash University, Australia
david.dowe@monash.edu

Abstract

While some computational models of intelligence test problems were proposed throughout the second half of the XXth century, in the first years of the XXIst century we have seen an increasing number of computer systems being able to score well on particular intelligence test tasks. However, despite this increasing trend there has been no general account of all these works in terms of how they relate to each other and what their real achievements are. In this paper, we provide some insight on these issues by giving a comprehensive account of about thirty computer models, from the 1960s to nowadays, and their relationships, focussing on the range of intelligence test tasks they address, the purpose of the models, how general or specialised these models are, the AI techniques they use in each case, their comparison with human performance, and their evaluation of item difficulty.

1 Introduction

AI research can claim some impressive milestones. For example, already in 1959 Arthur Samuel presented a self-learning program which could play checkers (or draughts) [Samuel, 1959]. In 2002 the 1957 prophecy of Herbert Simon that within 10 years a computer would be world's chess champion came true when Deep Blue won against the human chess champion Garry Kasparov [Campbell *et al.*, 2002]. In 2010 IBM's program Watson [Ferrucci *et al.*, 2010; Ferrucci *et al.*, 2013] was the winner of the *Jeopardy!* TV quiz. In 2016, Alphago [Silver *et al.*, 2016] was the first AI system able to master the ancient game Go. However, one can ask whether the mechanism underlying the behaviour of these programs is the same as or similar to the mechanism underlying human intelligent behaviour.

In AI research, the Turing test [Turing, 1950; Oppy and Dowe, 2011] has strongly influenced the philosophical discussions about what intelligence is and has motivated some actual implementations (e.g., Loebner's prize) and many different test variations to tell whether a computer program is intelligent. For example, CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) [von Ahn *et al.*, 2004] are used on web pages to tell humans and machines apart. Typically, CAPTCHAs contain pattern matching tasks (e.g., distorted numbers and letters) that are difficult to solve with current AI technology. However, the tasks that are featured in CAPTCHAs have to be made more complex regularly, as more sophisticated algorithms are devised to identify the presented patterns. Obviously, when a bot can break complex CAPTCHAs, we nevertheless hesitate to ascribe general intelligence to it.

In psychology research, the classical approach to intelligence evaluation is to apply psychometric tests measuring the intelligence quotient (IQ) [Sternberg (ed.), 2000] and other cognitive abilities. These tests are standardised in such a way that humans can be classified as below, about, or above average intelligence. The IQ test problems address a variety of reasoning abilities, for example, solving number series problems, detecting regularities in spatial configurations, or understanding verbal analogies. Some types of problems are rather independent of the subject's educational and cultural background, others depend on background knowledge.

In the early days of AI, the classical IQ test 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. Thus, the question is whether this measurement of mental developmental capabilities leads to a feasible, practical evaluation for AI.

In this work 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 [1965] and going through to Spaun [Eliasmith *et al.*, 2012], 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. 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 to explore the progress and implications of this area

*This paper is an extended abstract of an article in the Artificial Intelligence Journal [Hernández-Orallo *et al.*, 2016].

of research.

2 Historical Account

The relation between AI and psychometrics started more than fifty years ago. As early as 1963, Evans [1965] and Simon and Kotovsky [1963] devised AI programs able to identify regularities in patterns (respectively, analogy tasks and letter series completion problems). Their intention was to better understand some principles of analogy and its presentation as well as analyse how humans solved these kinds of problems and their difficulty.

After the initial interest of AI research in IQ test problems, this branch of research sank into oblivion for about twenty years. However, since the 1980s, cognitive science research recovered this line of research. Hofstadter developed a series of computational models in the Copycat project [Hofstadter and Mitchell, 1984] with the major goal of understanding analogy. In the 1990s some cognitive models were proposed to simulate the human cognitive processes. Carpenter et al. [1990] addressed Raven's Progressive Matrices [Raven *et al.*, 1992]. Yet again, the goal was to better understand human intelligence and the nature of the tests. Simon et al. [1991] addressed some series completion tasks based on the ideas of Simon and Kotovsky [1963].

In AI, forty years after the work of Evans and of Simon and Kotovsky, in 2003, computer programs solving intelligence tests became of interest again. On one hand, Sanghi and Dowe [2003] wanted to make a conclusive point about how easy it was to make non-intelligent machines pass intelligence tests, which 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 [Bringsjord and Schimanski, 2003]. They claimed that psychometric tests should not be dismissed but placed at a definitional, major role for what AI is and proposed "psychometric artificial intelligence" as a direction of research.

The two previous (opposed) approaches [Bringsjord and Schimanski, 2003; Sanghi and Dowe, 2003] for solving intelligence tests led to an increasing number of works in the area (although some of them were possibly unaware of these opposed views, as they cite neither Bringsjord and Schimanski's paper nor Sanghi and Dowe's paper). For instance, Tomai et al. [2005] revisited Evans's problems but using a more abstract and general approach, based on general-purpose simulation models. Similar to Tomai et al., Lovett et al. addressed three visual problem-solving tasks: geometric analogies [Lovett and Forbus, 2012], Raven's Progressive Matrices [Lovett *et al.*, 2010], and odd-one-out intelligence tests [Lovett and Forbus, 2011], with a major goal of modelling human cognition. A different attempt to address a psychometric test with the aim of understanding humans' cognitive mechanisms was *Phaeaco* [Foundalis, 2006], which focussed on Bongard problems [Bongard, 1970].

A surge of new systems has taken place since 2010. For instance, Sinapov and Stoytchev [2010] started an approach where personalised odd-one-out tasks are taken by a robot in a rich sensorimotor scenario. Schenck et al. [2012] also used an upper-torso humanoid robot to address order completion problems. Another attempt to address Raven's Progressive Matrices and odd-one-out problems using purely iconic visual representations was undertaken by McGreggor, Kunda, and Goel [Kunda *et al.*, 2013]. Klenk et al. [2011] challenged a new kind of test, Bennett's Mechanical Comprehension Tests [Bennett, 1969], with important content about physics and contextual information, by using a cognitive architecture. Turney [2011] introduced a system for analogy perception that recognises lexical proportional analogies. Closely related to the approach of Turney we find the work of Bayoudh et al. [2012]. Ragni and Klein [2011] worked on number series completion applying a general machine learning method. Ragni and Neubert [2014] presented a system for Raven's Progressive Matrices which motivation is to solve the problems in a cognitive way, i.e., explaining why the system fails. Prade, Richard, and Correa [Prade and Richard, 2014] developed a logical representation of the notion of "analogical proportion" for addressing Raven's Progressive Matrices and odd-one-out problems. Both Siebers and Schmid [2012] and Strannegård et al. [2013a] addressed the number series problem with a cognitive model based on the idea of limited working memory. Strannegård et al. [2013b] also address Raven's Progressive Matrices with an anthropomorphic cognitive model using certain problem solving strategies that were reported by high-achieving human solvers. The approach presented by Ohlsson et al. [2013] is one of the few approaches to verbal intelligence tests. Hofmann, Kitzelmann, and Schmid [Hofmann *et al.*, 2014] demonstrate that the inductive programming [Gulwani *et al.*, 2015] system *IGOR2* can be applied to number series problems. Martínez-Plumed et al. [2017] introduced *gErl*, a declarative learning system able to solve several intelligence tests: Raven's Progressive Matrices, odd-one-out problems and letter series completion problems. Finally, with quite a different perspective, Eliasmith et al. [2012] recently produced a 2.5-million-neuron artificial model brain having a similar ability on certain aptitude test questions (some of them similar to the Raven's Progressive Matrices) to what might be found in some humans.

In fact, 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.

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 to draw some conclusions about their usefulness.

We have seen that most approaches are very recent. Is it an indication of relevance? Regarding 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 ad-hoc to more general AI techniques (mostly from machine learning, pattern recognition, automated reasoning, and natural language processing). This suggests that there is 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. This is especially the case when results of a model and humans coincidence even though the model was not conceived to follow exactly what humans do. Nonetheless, a systematic disagreement in results or ability correlations may also be very informative. In fact, these studies can be very useful to better understand what intelligence tests really measure and to better interpret the correlations between abilities found in humans.

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 [Detterman, 2011]. Nonetheless, we have not seen that artificial intelligence has changed its evaluation protocols following this increase of models taking intelligence tests (with a few exceptions such as [Sinapov and Stoytchev, 2010; Schenck, 2013; Eliasmith *et al.*, 2012]). 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. Therefore, for intelligence tests to be useful evaluation tools for AI, several things must be considered. Instead of a collection of problems, a collection of instance generators developing new problems instances is required. The collection must consist of many different problems, in order to avoid the big switch approach, that is having a small program dispatching problems to extremely specialised approaches. Moreover, brand-new problems could be generated by the combination of existing ones or by the development of more abstract problem generators (instead of instance generators). 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 endeavoured (see [Dowe and Hernández-Orallo, 2014; Hernández-Orallo, 2016; Hernández-Orallo, 2017].

There is also a huge diversity in how 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 [Simon, 1963; Carpenter *et al.*, 1990; Strannegård *et al.*, 2013a; Martínez-Plumed *et al.*, 2017]. These notions of difficulty are much more general and can work independently of the problem and its representation.

One way or the other, there seems to be an agreement that there will be an increasing number of machines in the near future which show a range of cognitive abilities, and that we will require evaluation mechanisms for them [Hernández-Orallo, 2017]. These mechanisms will have to give scores for several cognitive abilities so that we can compare them with humans and other animals. As a realisation of how much work needs to be done here, the development of well-grounded tests for humans, animals, robots, agents, animats, hybrids, swarms, etc., has been proposed in an emerging new (but integrating) discipline, dubbed 'universal psychometrics' [Dowe and Hernández-Orallo, 2014], which may inherit and integrate many important concepts from psychometrics, cognition, and AI.

4 Conclusion

This paper was motivated by an observed explosion of the number of papers featuring computer models addressing intelligence test problems. Among the around 30 papers we have analysed, half of them have appeared since 2011. 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. When we began our investigation we soon realised that computer models addressing intelligence tests 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.

Note that if any or all of the above reasons were spurious, its negation would still be most interesting. Namely, if intelligence tests were not challenging, were not useful for the evaluation of AI systems, did not measure the purported abilities beyond humans, or were useless to understand what human intelligence is, then this would represent important insights for psychometrics, cognitive science, and artificial intelligence. In fact, the authors of this paper may have different stances on some of these issues, which partially explains why all of us agree that this area of research is worth being pursued and empowered in the near future.

Another motivation for this paper was that there seems to be a limited connection between these works, as many of them seem to ignore results and ideas already present in previous approaches. This includes the (mis)understanding of what a computer model passing an intelligence test really means. We hope that this work could facilitate and encourage future computer models taking intelligence test problems to link with and build upon previous research.

Finally, a very ambitious goal would be to create a repository or generator of all these problems. We know that many intelligence tests are not publicly available, and many of the approaches we have surveyed here have used alternative formulations because of this. It would be very useful for AI to arrange these problems, record the results of computer models and humans over them and organise competitions. This benchmark should be broad (including a wide range of tests), standard (using some kind of general protocol for inputs and outputs), characterised (accompanied with a catalogue of information about their difficulty, the abilities they cover, etc.), available (on a web or problem library), and renewed (new items are generated or disclosed so that systems cannot rote-learn them).

References

[Bayoudh *et al.*, 2012] Meriam Bayoudh, Henri Prade, and Gilles Richard. Evaluation of analogical proportions through kolmogorov complexity. *Knowledge-Based Systems*, 29(0):20–30, 2012. Artificial Intelligence 2010.

[Bennett, 1969] George K. Bennett. *Bennett mechanical comprehension test*. Psychological Corporation, 1969.

[Bongard, 1970] Mikhail Moiseevich Bongard. *Pattern Recognition*. Spartan Books, 1970.

[Bringsjord and Schimanski, 2003] Selmer Bringsjord and Bettina Schimanski. What is artificial intelligence? Psychometric AI as an answer. In *International Joint Conference on Artificial Intelligence*, pages 887–893, 2003.

[Campbell *et al.*, 2002] Murray Campbell, A. Joseph Hoane Jr, and Feng-hsiung Hsu. Deep Blue. *Artificial Intelligence*, 134(1-2):57–83, 2002.

[Carpenter *et al.*, 1990] Patricia A. Carpenter, Marcel A. Just, and Peter Shell. What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological review*, 97:404–431, 1990.

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

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

[Eliasmith *et al.*, 2012] Chris Eliasmith, Terrence C. Stewart, XXuan Choo, Trevor Bekolay, Travis DeWolf, Yichuan Tang, and Daniel Rasmussen. A large-scale model of the functioning brain. *Science*, 338(6111):1202–1205, 2012.

[Evans, 1965] Thomas G. Evans. A heuristic program to solve geometric-analogy problems. In *Proc. SJCC*, volume 25, pages 327–339, 1965. vol. 25.

[Ferrucci *et al.*, 2010] David Ferrucci, David Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, et al. Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59–79, 2010.

[Ferrucci *et al.*, 2013] David Ferrucci, Anthony Levas, Sugato Bagchi, David Gondek, and Erik T. Mueller. Watson: Beyond jeopardy! *Artificial Intelligence*, 199:93–105, 2013.

[Foundalis, 2006] Harry E. Foundalis. Phaeaco: A cognitive architecture inspired by Bongard's problems. *Doctoral thesis, Indiana University, Bloomington*, 2006.

[Gulwani *et al.*, 2015] Sumit Gulwani, Jose Hernandez-Orallo, Emanuel Kitzelmann, Stephen H. Muggleton, Ute Schmid, and Benjamin Zorn. Inductive programming meets the real world. *Communications of the ACM*, 58(11):90–99, 2015.

[Hernández-Orallo *et al.*, 2016] José Hernández-Orallo, Fernando 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.

[Hernández-Orallo, 2016] José Hernández-Orallo. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, pages 1–51, 2016.

[Hernández-Orallo, 2017] José Hernández-Orallo. The measure of all minds: Evaluating natural and artificial intelligence. 2017.

[Hofmann *et al.*, 2014] Jacqueline Hofmann, Emanuel Kitzelmann, and Ute Schmid. Applying inductive program synthesis to induction of number series a case study with IGOR2. In *KI 2014: Advances in AI*, pages 25–36. Springer, 2014.

[Hofstadter and Mitchell, 1984] Douglas R. Hofstadter and Melanie Mitchell. The copycat project, 1984.

[Klenk *et al.*, 2011] Matthew Klenk, Kenneth Forbus, Emmett Tomai, and Hyeonkyeong Kim. Using analogical model formulation with sketches to solve Bennett mechanical comprehension test problems. *Journal of Experimental & Theoretical Artificial Intelligence*, 23(3):299–327, 2011.

[Kunda *et al.*, 2013] Maithilee Kunda, Keith McGreggor, and Ashok K. Goel. A computational model for solving problems from the Ravens Progressive Matrices intelligence test using iconic visual representations. *Cognitive Systems Research*, pages 22–23, 47–66, 2013.

[Lovett and Forbus, 2011] Andrew Lovett and Kenneth Forbus. Cultural commonalities and differences in spatial problem-solving: A computational analysis. *Cognition*, 121(2):281–287, 2011.

[Lovett and Forbus, 2012] Andrew Lovett and Kenneth Forbus. Modeling multiple strategies for solving geometric analogy problems. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 701–706, 2012.

[Lovett *et al.*, 2010] Andrew Lovett, Kenneth Forbus, and Jeffrey Usher. A structure-mapping model of Ravens Progressive Matrices. In *Proceedings of CogSci*, volume 10, pages 2761–2766, 2010.

[Martínez-Plumed *et al.*, 2017] Fernando Martínez-Plumed, César Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. A computational analysis of general intelligence tests for evaluating cognitive development. *Cognitive Systems Research*, 43:100–118, 2017.

[Ohlsson *et al.*, 2013] Stellan Ohlsson, Robert H. Sloan, Gyrgy Turán, and Aaron Urasky. Verbal IQ of a four-year old achieved by an AI system. In *COMMONSENSE 2013, 11th International Symposium on Logical Formalizations of Commonsense Reasoning*, page 6, 2013.

[Oppy and Dowe, 2011] Graham Oppy and David L. Dowe. The Turing Test. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*, pages Stanford University, http://plato.stanford.edu/entries/turing–test/, 2011.

[Prade and Richard, 2014] Henri Prade and Gilles Richard. From analogical proportion to logical proportions: A survey. In *Computational Approaches to Analogical Reasoning: Current Trends*, volume 548 of *Studies in Computational Intelligence*, pages 217–244. Springer Berlin Heidelberg, 2014.

[Ragni and Klein, 2011] Marco Ragni and Andreas Klein. Predicting numbers: an AI approach to solving number series. In *KI 2011: Advances in Artificial Intelligence*, pages 255–259. Springer, 2011.

[Ragni and Neubert, 2014] Marco Ragni and Stefanie 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*, pages 351–370. Springer, 2014.

[Raven *et al.*, 1992] J. C. Raven, John H. Court, and John Raven. *Manual for Raven's Progressive Matrices and Vocabulary Scale*. San Antonio, TX: Psychological Corporation, 1992.

[Samuel, 1959] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959.

[Sanghi and Dowe, 2003] Pritika Sanghi and David L. Dowe. A computer program capable of passing IQ tests. In *4th Intl. Conf. on Cognitive Science (ICCS'03), Sydney*, pages 570–575, 2003.

[Schenck *et al.*, 2012] Connor Schenck, Jivko Sinapov, and Alexander Stoytchev. Which object comes next? Grounded order completion by a humanoid robot. *Cybernetics and Information Technologies*, 12(3):5–16, 2012.

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

[Siebers and Schmid, 2012] Michael Siebers and Ute Schmid. Semi-analytic natural number series induction. In *KI 2012: Advances in Artificial Intelligence*, pages 249–252. Springer, 2012.

[Silver *et al.*, 2016] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[Simon *et al.*, 1991] Tony Simon, David Klahr, and Allen Newell. SCSoar: Pattern induction in series completion problem solving. In *European Soar Workshop*, page 17, Cambridge, 1991.

[Simon, 1963] Kenneth Simon, Herbert A.; Kotovsky. Human acquisition of concepts for sequential patterns. *Psychological Review*, 70(6):534, 1963.

[Sinapov and Stoytchev, 2010] Jivko Sinapov and Alexander Stoytchev. The odd one out task: Toward an intelligence test for robots. In *Development and Learning (ICDL), 2010 IEEE 9th International Conference on*, pages 126–131. IEEE, 2010.

[Sternberg (ed.), 2000] R. J. Sternberg (ed.). *Handbook of intelligence*. Cambridge University Press, 2000.

[Strannegård *et al.*, 2013a] Claes Strannegård, Mehrdad Amirghasemi, and Simon Ulfsbücker. An anthropomorphic method for number sequence problems. *Cognitive Systems Research*, 2223:27–34, 2013.

[Strannegård *et al.*, 2013b] Claes Strannegård, Simone Cirillo, and Victor Ström. An anthropomorphic method for progressive matrix problems. *Cognitive Systems Research*, 2223(0):35–46, 2013.

[Tomai *et al.*, 2005] Emmet Tomai, Andrew Lovett, Kenneth Forbus, and Jeffrey Usher. A structure mapping model for solving geometric analogy problems. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 2190–2195, Stressa, Italy, 2005. Cognitive Science Society, Inc.

[Turing, 1950] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[Turney, 2011] Peter D. Turney. Analogy perception applied to seven tests of word comprehension. *Journal of Experimental & Theoretical Artificial Intelligence*, 23(3):343–362, 2011.

[von Ahn *et al.*, 2004] Luis von Ahn, Manuel Blum, and John Langford. Telling humans and computers apart automatically. *Communications of the ACM*, 47(2):56–60, 2004.