Intelligence 40 (2012) 77–81

Contents lists available at SciVerse ScienceDirect

Intelligence

# IQ tests are not for machines, yet

David L. Dowe [a], José Hernández-Orallo [b,*]

[a] Computer Science and Software Engineering, Clayton School of Information Technology, Monash University, Clayton, Vic. 3800, Australia
[b] DSIC, Universitat Politècnica de València, València, Spain

ARTICLE INFO

*Article history:*
Received 29 September 2011
Received in revised form 3 November 2011
Accepted 22 December 2011
Available online 21 January 2012

*Keywords:*
Machine intelligence evaluation
IQ tests
Artificial intelligence
Universal tests
Psychometrics
Task difficulty
CAPTCHA

ABSTRACT

Complex, but specific, tasks—such as chess or *Jeopardy!*—are popularly seen as milestones for artificial intelligence (AI). However, they are not appropriate for evaluating the intelligence of machines or measuring the progress in AI. Aware of this delusion, Detterman has recently raised a challenge prompting AI researchers to evaluate their artefacts against IQ tests. We agree that the philosophy behind (human) IQ tests is a much better approach to machine intelligence evaluation than these specific tasks, and also more practical and informative than the Turing test. However, we have first to recall some work on machine intelligence measurement which has shown that some IQ tests can be passed by relatively simple programs. This suggests that the challenge may not be so demanding and may just work as a sophisticated CAPTCHA, since some types of tests might be easier than others for the current state of AI. Second, we show that an alternative, formal derivation of intelligence tests for machines is possible, grounded in (algorithmic) information theory. In these tests, we have a proper mathematical definition of what is being measured. Third, we re-visit some research done in the past fifteen years for effectively measuring machine intelligence—since some assumptions about the subjects and their distribution no longer hold.



## 1. Introduction: The challenge

In February 2011, Douglas K. Detterman announced a challenge (originally to IBM's program Watson (Ferrucci et al., 2010), the recent winner of the *Jeopardy!* TV quiz show at the time) for the whole field of artificial intelligence (AI). AI artefacts should be better measured by classical IQ tests. The challenge goes as follows (Detterman, 2011): "I, the editorial board of *Intelligence*, and members of the International Society for Intelligence Research will develop a unique battery of intelligence tests that would be administered to that computer and would result in an actual IQ score".

Computers are (still) so stupid today, that it seems clear that an average result at IQ tests is far beyond current computer technology. "It is doubtful that anyone will take up this challenge in the near future", Detterman said (Detterman, 2011). But the challenge had already been taken up, in the past.

In 2003, a computer program performed quite well on standard human IQ tests (Sanghi & Dowe, 2003). This was an elementary program, far smaller than Watson or the successful chess-playing Deep Blue (Campbell, Hoane, & Hsu, 2002). The program had only about 960 lines of code in the programming language Perl (accompanied by a list of 25,143 words), but it even surpassed the average score (of 100) on some tests (Sanghi & Dowe, 2003, Table 1).

The computer program underlying this work was based on the realisation that most IQ test questions that the authors had seen until then tended to be of one of a small number of types or formats. Formats such as "insert missing letter/number in middle or at end" and "insert suffix/prefix to complete two or more words" were included in the program. Other formats such as "complete matrix of numbers/characters", "use directions, comparisons and/or pictures", "find the odd

* Corresponding author at: DSIC, Universitat Politècnica de València, Camí de Vera s/n, 46022 València, Spain. Tel.: +34 96 3877007x73585; fax: +34 96 3877359.
*E-mail addresses:* david.dowe@infotech.monash.edu.au (D.L. Dowe), jorallo@dsic.upv.es (J. Hernández-Orallo).

0160-2896/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2011.12.001

man out", "coding", etc. were not included in the program—although they are discussed in Sanghi and Dowe (2003) along with their potential implementation. The IQ score given to the program for such questions not included in the computer program was the expected average from a random guess, although clearly the program would obtain a better "IQ" if efforts were made to implement any, some or all of these other formats.

So, apart from random guesses, the program obtains its score from being quite reliable at questions of the "insert missing letter/number in middle or at end" and "insert suffix/prefix to complete two or more words" natures. For the latter "insert suffix/prefix" sort of question, it must be confessed that the program was assisted by a look-up list of 25,143 words. Substantial parts of the program are spent on the former sort of question "insert missing letter/number in middle or at end", with software to examine for arithmetic progressions (e.g., 7 10 13 16 ?), geometric progressions (e.g., 3 6 12 24 ?), arithmetic geometric progressions (e.g., 3 5 9 17 33 ?), squares, cubes, Fibonacci sequences (e.g., 0 1 1 2 3 5 8 13 ?) and even arithmetic-Fibonacci hybrids such as (0 1 3 6 11 19 ?). Much of the program is spent on parsing input and formatting output strings—and some of the program is internal redundant documentation and blank lines for ease of programmer readability.

We can, of course, argue whether this experiment complies with the challenge. Detterman established two levels for the challenge, while only computers passing the second level "could be said to be truly intelligent" (Detterman, 2011). Perhaps this program is only acceptable for the first level of the challenge (where the type of IQ tests is seen in advance by the programmer). However, it is important to note that this toy program was capable of passing tests it had not seen beforehand (which is closer to the second and ultimate level of the challenge). If not directly qualified, the Perl program could at least form an initial benchmark in this challenge.

Of course, the system can be improved in many ways. It was just a 3rd year undergraduate student project, a quarter of a semester's work. With the budget Deep Blue or Watson had, the program would likely excel in a very wide range of IQ tests. But this is not the point. The purpose of the experiment was not to show that the program was intelligent. Rather, the intention was showing that conventional IQ tests are not for machines—a point that the relative success of this simple program would seem to make emphatically. This is natural, however, since IQ tests have been specialised and refined for well over a century to work well for humans.

We are certain that the editorial board of *Intelligence* and some members of the International Society for Intelligence Research can figure out a diverse battery of tests that the average human can pass and the Perl program would fail. (We could probably do this, too.) But an improved version of the program would surely make things more difficult for the board. This spiral is precisely what CAPTCHAs do (von Ahn, Blum, & Langford, 2004). The definition of a CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is basically any problem an average human can pass easily but current computer technology cannot. We are using them everyday on the Internet—e.g., to create a new account or post a comment—, frequently shown as a sequence of deformed letters. They are useful and revealing, but they *cannot* be considered an intelligence test. They are testing humanity—or perhaps rather little more than the ability to recognise deformed Roman-language letters (and then type the matching letters on the console keyboard). In addition, when bots get equipped with tools able to crack them, the CAPTCHAs have to be replaced by more sophisticated ones. As a result, some (brilliant) people struggle with them.

Summing up, there are IQ tests no machine can pass nowadays, but a selection of the 'machine-unfriendly' IQ tests would have no particular relation with their ability to measure intelligence well in humans, but rather just their ability to discriminate between humans and state-of-the-art machines, as CAPTCHAs do. This selection (or battery) of IQ tests would need to be changed and made more elaborate year after year as AI technology advances.

## 2. Measuring machine intelligence

So, what are the right intelligence tests for machines? The Turing test (originally just conceived as an imitation game) (Oppy & Dowe, 2011; Turing, 1950) has been the answer to many philosophical questions about thinking machines, but it is not an intelligence test in that it is not able to evaluate machines and humans on the same scale (or even on different scales). In the end, the Turing test is also a test of humanity, rather than intelligence. Also, it can be argued that the Turing test has set the goal of AI on a philosophical, misguided dimension. However, all said, in our opinion, the Turing test has had a limited impact on how the discipline of AI has evolved over the years.

The problem with AI cannot be found in a wrong interpretation of the notion of intelligence. In fact, the very discipline of AI adheres to the mainstream concept of intelligence as, e.g., "a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience" (Gottfredson, 1997). All these mentioned abilities are mirrored in subdisciplines of AI, such as automated reasoning, planning, problem solving, knowledge representation, natural language processing, machine learning, etc. However, the advances in developing evaluation devices for all these subdisciplines, and across subdisciplines, have not been particularly encouraging. In more than fifty years of history of AI, only a few ad-hoc, scattered and specific testing mechanisms have been developed in each of these subdisciplines. This is one of the problems, if not the most important one, in AI—the lack of proper measuring devices to evaluate its progress.

Some early works in the 1990s looked for better alternatives. Dowe and Hajek (1997, 1998) presented an enriched Turing test with compression exercises, which, in the end, aim at measuring the ability of inductive inference à la MML (Wallace & Boulton, 1968). The Minimum Message Length (MML) principle advocates for (two-part) compression as a way to perform inductive inference (or induction—i.e., learning) and, ultimately, intelligence (Dowe, 2011, sec. 7.3).

Similarly, Hernández-Orallo (2000a); Hernández-Orallo and Minaya-Collado (1998) devised a new test—known as the *C*-test—with sequence-completion exercises very similar

to those found in some IQ tests. Unlike IQ tests, the exercises were created from computational principles, and its complexity ($k$, basically the number of instructions of the shortest program generating the sequence) was mathematically quantified using constructs derived from Kolmogorov complexity—a theory closely related to MML (Wallace & Dowe, 1999). The *C*-test was built by including exercises whose complexity range was $k = 7..14$, an interval that was deemed appropriate for humans since items with $k<7$ looked trivial, while items with $k>14$ were really challenging (to us). With the limitations of an amateur experimentation procedure, the test was administered to humans, and the results strongly correlated (certainly as many other difficult tasks do) with the results of some IQ tests which were also administered to the same subjects (Hernández-Orallo & Minaya-Collado, 1998, Appendix A). For the authors, intelligence was no longer (only) what the IQ tests measure, but (also) a *precise mathematical concept*. The new pathway looked so promising that even mathematically-derived 'factorisations' of intelligence were suggested (Hernández-Orallo, 2000b).

But once that 'stupid' Perl program (Sanghi & Dowe, 2003) was able to pass some IQ tests, the idyllic picture blurred. It must be said that the idea of connecting the exercises found in many IQ tests with the mathematical theory of inductive inference based on (algorithmic) information theory was not affected by the experiment with the Perl program. However, the research plan was hit by a fatal torpedo, since this new battery of mathematically-devised *C*-tests looked fairly easy to crack. In fact, the toy Perl program managed to solve some problems of high $k$ complexity and, overall, can score slightly better than humans on the *C*-tests.

The previous quest for alternative approaches to intelligence measurement and our dismissal of conventional IQ tests for machine evaluation should not be misunderstood. There must be few people in the area of AI who have advocated more than us for the hybridisation of psychometrics, comparative cognition and machine intelligence measurement. But, in our opinion, the experiment with this small program in Perl has already shown that IQ tests have become specialised for humans, and many difficulties may arise if these tests are used to measure intelligence outside this 'normative' population. This is easier to accept from the point of view of comparative (animal) cognition since human IQ tests are not commonly used for evaluating animals. Since 2003, this conviction is also shaping up in AI: human IQ tests are not for machines. The reason is that, in our opinion, many things are taken for granted in IQ tests which cannot be assumed for machines. Current IQ tests are *anthropocentric*. Some important elements of intelligence which are shared by almost all humans are not evaluated by regular IQ tests. We can see some of these elements in 'non-regular' tests which are designed for people with intellectual disabilities or mental retardation, for children or for non-human animals. But even these kinds of tests (including the *C*-tests) assume many things about the subject.

## 3. Intelligence tests for machines must be universal

If we really want to measure intelligence instead of humanity and fully understand what we are really measuring, the Turing test is clearly not an option—and neither is a diverse battery of specific tasks such as (e.g.) chess, *Jeopardy* or multiplying large numbers together, even though these tasks were once (considered) the exclusive domain of "intelligent" humans. We do believe that some human IQ tests are much better than any other test that has been developed so far in AI, so we agree with Detterman insofar as "there is a better alternative: test computers on human intelligence tests" (Detterman, 2011). But the above-mentioned works (Dowe & Hajek, 1997, 1998; Hernández-Orallo, 2000a; Sanghi & Dowe, 2003) seem to bring some evidence that this alternative cannot be the definitive answer. Regular IQ tests cannot be the guide for AI. This was, in fact, one of the first tracks that the then young discipline of AI explored in the 1960s. IQ tests were used as inspiration for constructing intelligent systems (Evans, 1964), but this line of research stalled and has had no significant progress since that date, despite several subsequent attempts by other authors. More recently, for example, the idea of using IQ tests as a benchmark to drive the discipline of AI has been made explicit in Bringsjord and Schimanski (2003). However, their impact has been very limited since only AI systems which are specialised to the particular test interface and choice of symbolic representation can be evaluated.

We think there is one main reason for this apparent dead end. The measurement of intelligence for machines must be more holistic, since it cannot take anything for granted. We can assume even fewer things than in animal intelligence evaluation. The solution must be found then in the universality of the concept of intelligence. This means that if a single test is not able to measure the intelligence (or other cognitive abilities) of non-human animals and humans precisely, it is very difficult to expect that this test will measure machine intelligence accurately. This leads to the notion of *universal* test, a test which must be valid for humans, non-human animals, machines, hybrids and collectives, of any degree of intelligence. This does not mean, of course, that these universal tests should replace IQ tests for evaluating humans. There are many possible instruments, or vehicles (Jensen, 1998, chap. 10), to measure the same construct or ability, and we should use the best instrument according to the construct and the kind of subject. For intelligence and humans, IQ tests seem to be a very good instrument, as we could also use brain monitoring and neuro-imaging techniques, or any other kind of physical or genetic traits. For some animals, very specialised and sophisticated instruments have been developed, as well. For some machines we could also use customised tests for a particular 'series' or 'architecture', or we could even inspect their programs. We clearly do not need to use the same test for all of them, *with the proviso that* we are given information about the kind of subject we are evaluating. However, for any unknown machine for which we do not have any information whatsoever, tests must be *universal*.

The crucial point for understanding why universality is so significant is the realisation that machines are a much more heterogeneous set of subjects. The only constraints for a machine are computational resources and the aptitudes of its programmers. Any imaginable (computable) behaviour is possible. It could behave as a human, it could behave as a rat or it could behave as something we have never seen

before. In addition, machines cannot be grouped into populations, or species. The mere notion of an 'average' machine is ridiculous, because there is no normative population of machines. Only if we are given some information about the examinee, can we use more efficient, customised tests, as we do with human IQ tests. But if we are not given any information at all about the subject, our measuring ruler must be as universal as possible, at the risk of losing precision or efficiency. This follows a natural principle of measurement: measuring is less efficient and more difficult the less we know about the subject.

The lack of normative populations also makes some well-established psychometric techniques infeasible. We cannot derive the difficulty of an item by how well *machines* perform on it unless we find a reasonable choice for the machine population. In fact, we can program machines which fail at the easy items while acing at the difficult ones. This means that test items have to be very carefully designed in such a way that we need to know what they represent and what their *intrinsic* difficulty is.

This limits the applicability of some powerful tools such as Spearman's theorem on the indifference of the indicator (Spearman, 1927), which states that any kind of test, any measurement instrument, is perfectly useful for measuring intelligence provided that it correlates with the '*g* factor'. With machines, it is not clear how the '*g* loading' (correlation between a test and *g*) can be derived, since correlations require a population. In fact, it is not clear at all whether a *g* factor exists for machines.

Of course we can calculate the *g* loading and the item difficulty for humans, i.e. normed on a human subject sample, and extrapolate to other kinds of individuals. Those tests with lower *g* loading can be considered more specific. Therefore, specialised, non-intelligent machines are expected to eventually be able to solve these tasks. Apart from being anthropocentric, this rationale does not seem to work. It is true that many tasks machines perform well at nowadays have relatively low *g* loading in humans (e.g. arithmetic). However, there are some other cognitive tests at which machines can perform well but which have very high *g* loading in humans, such as inductive inference tasks (e.g. series completion, such as the *C*-tests). Conversely, CAPTCHAs typically use tasks with presumably very low *g* loading to tell humans and computers apart, such as reading distorted text. In fact, these tasks can be performed by humans *without thinking*. All this does not necessarily entail that the '*g* construct' may not work for machines, but just the well-known fact that factor correlations and loadings cannot be extrapolated between dissimilar distributions—here, humans and machines.

One could (even) go so far as to imagine an elaborate model of hierarchical clusters of models with single (and even multiple) latent factors, as per the statistical theory in (e.g.) (Edwards & Dowe, 1998; Wallace, 1995; Wallace & Freeman, 1992) (Wallace, 2005, sec. 6.9) (Dowe, 2008, sec. 0.2.4, p537, col. 2) and possibly also as per Jensen (1998), chap. 10, p337, Fig. 10.3, but even this would still leave our models and their resultant latent factors dependent upon the underlying population being modelled—and it is not only unclear how best to sample from the machine population, but computing history shows that this population (whatever it is) seems to be changing quite rapidly. This may (all) suggest that the *g* loading of a cognitive test should be determined theoretically and not empirically.

A convenient theoretical, a priori approach to constructing intelligence tests can be based on formal computational definition of the cognitive abilities involved, and likewise a mathematical, intrinsic, derivation of task complexity. For example, the theory of inductive inference based on (algorithmic) information theory (Wallace & Boulton, 1968; Wallace & Dowe, 1999) *models* inductive inference ability, closely related to Solomonoff's earlier notions of algorithmic probabilistic prediction (Solomonoff, 1964). This allows for the a priori generation of exercises of different complexity, where this complexity is derived in a mathematical way (Hernández-Orallo, 2000a,b) and not obtained in an empirical way over a population. Determining the complexity of tasks in this way would then be much more informative for AI (and also for the research on human and non-human animal intelligence). Ultimately, AI would know exactly what it is aiming at.

## 4. The future calls for working together

As happens in any scientific discipline, any extension over the objects and phenomena of interest leads to further research opportunities and discoveries, but also raises the risk of some parts of the established paradigm being refuted or revised for the more general situation. Universal tests are not only strictly necessary for machine evaluation, but they can also be a very important research tool for detecting inconsistencies and for the generalisation of scales and procedures.

The common use of CAPTCHAs in a wide range of situations is an indicator that things will become more complex in the future. An astonishing plethora of bots, avatars, animats, swarms and other kinds of virtual agents will require a thoughtfully-designed battery of tests. Many of these tests will have to be universal, especially when no information about the subject is given beforehand (not even information of whether it is a human or a machine). Universal tests will become even more imperative when machines get close to and ultimately beyond human intelligence (Hernández-Orallo & Dowe, 2010; Solomonoff, 1985).

But how can these tests be constructed? We think they must be based upon many of the techniques which have been developed for intelligence measurement in psychometrics, comparative psychology, animal cognition, information theory and AI. The study of this possibility is the very goal of the project "Anytime Universal Intelligence" (http://users.dsic.upv.es/proy/anynt/). One of the outcomes of this project has been a mathematical setting for constructing universal tests—built upon solid, non-anthropomorphic foundations (Hernández-Orallo & Dowe, 2010). Naturally, there are also many difficulties along the way—as some preliminary tests on machines and humans are showing (Hernández-Orallo, Dowe, España-Cubillo, Hernández-Lloreda, & Insa-Cabrera, 2011; Insa-Cabrera, Dowe, España-Cubillo, Hernández-Lloreda & Hernández-Orallo, 2011; Insa-Cabrera, Dowe & Hernandez-Orallo, 2011). For example, some simple "a priori algorithms"—the terminology used by Detterman for the second level of the challenge—can score better than humans

at a diverse, randomly-generated and previously unknown set of tasks.

One natural consequence of aiming at universality is that tests need to be adaptive to an unknown subject of unexpected characteristics, and the complexity of items must be determined in advance in order to use them on demand, as tests do in the area of computerised adaptive testing. Many other notions from psychometrics and (comparative) psychology also apply, such as the notions of a task being discriminative, the use of rewards and, of course, the use of (latent) factor analysis and other statistical tools for experimental results. In the end, most of the knowledge and techniques in the science of human intelligence are directly applicable. Yet, as we point out in this paper, only a few concepts and techniques (would) need to be re-considered or revised. Ultimately, humans are a perfect source for refutation of tests for machines and vice versa.

What seems clear to us is that a new generation of intelligence tests for machines will require an enormous effort. This will most likely involve several decades and a myriad of researchers from many disciplines. But the quest is so fascinating that it is already growing in popularity (Biever, 2011; Kleiner, 2011). We share Detterman's "hope" that "it may someday happen […] that the fields of artificial and human intelligence will grow closer together, each learning from the other" (Detterman, 2011). We have been stirring up this hope for over fifteen years. Everyone is welcome on board.

## Acknowledgements

We thank the anonymous reviewers for their comments, which have helped to significantly broaden the scope of this paper. This work was supported by the MEC projects EXPLORA-INGENIO TIN 2009-06078-E, CONSOLIDER-INGENIO 26706 and TIN 2010-21062-C02-02, and GVA project PROMETEO/2008/051.

## References

Biever, C. (2011). Ultimate IQ: One test to rule them all. *New Scientist*, *211*, 42–45.

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

Campbell, M., Hoane, A. J., & Hsu, F. (2002). Deep Blue. *Artificial Intelligence*, *134*, 57–83.

Detterman, D. K. (2011). A challenge to Watson. *Intelligence*, *39*, 77–78.

Dowe, D. L. (2008). Foreword re C. S. Wallace. *The Computer Journal*, *51*, 523–560 (Christopher Stewart WALLACE (1933–2004) memorial special issue).

Dowe, D. L. (2011). MML, hybrid Bayesian network graphical models, statistical consistency, invariance and uniqueness. In P. S. Bandyopadhyay, & M. R. Forster (Eds.), *Handbook of the philosophy of science. Philosophy of Statistics*, *Volume 7*. (pp. 901–982) : Elsevier.

Dowe, D. L., & Hajek, A. R. (1997). A computational extension to the Turing test. *Proceedings of the 4th conference of the Australasian Cognitive Science Society, University of Newcastle, NSW, Australia*.

Dowe, D. L., & Hajek, A. R. (1998). A non-behavioural, computational extension to the Turing test. *Intl. Conf. on computational intelligence & multimedia applications, Gippsland, Australia* (pp. 101–106).

Edwards, R., & Dowe, D. L. (1998). Single factor analysis in MML mixture modelling. *2nd Pacific–Asia conference on knowledge discovery and data mining, Melbourne, Australia, LNAI Series 1394* (pp. 96–109). : Springer.

Evans, T. G. (1964). A program for the solution of a class of geometric-analogy intelligence-test questions. Technical report DTIC document 1964, also appeared in 1968. In M. Minsky (Ed.), *Semantic information processing* (pp. 271–353). Cambridge, Massachussets: MIT Press.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., et al. (2010). Building Watson: An overview of the DeepQA project. *AI Magazine*, *31*, 59–79.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24*, 13–23.

Hernández-Orallo, J. (2000). Beyond the Turing test. *Journal of Logic, Language and Information*, *9*, 447–466.

Hernández-Orallo, J. (2000). On the computational measurement of intelligence factors. In A. Meystel (Ed.), *Performance metrics for intelligent systems workshop* (pp. 1–8). Gaithersburg, MD, U.S.A: National Institute of Standards and Technology.

Hernández-Orallo, J., & Dowe, D. L. (2010). Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, *174*, 1508–1539.

Hernández-Orallo, J., Dowe, D. L., España-Cubillo, S., Hernández-Lloreda, M. V., & Insa-Cabrera, J. (2011). On more realistic environment distributions for defining, evaluating and developing intelligence. In J. Schmidhuber, K. Thórisson, & M.L. (Eds.), *Artificial general intelligence 2011* (pp. 82–91). (LNAI series, Springer volume 6830).

Hernández-Orallo, J., & Minaya-Collado, N. (1998). A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. *Proc. intl symposium of engineering of intelligent systems (EIS'98), February 1998, La Laguna, Spain* (pp. 146–163). : ICSC Press.

Insa-Cabrera, J., Dowe, D. L., España-Cubillo, S., Hernández-Lloreda, M. V., & Hernández-Orallo, J. (2011). Comparing humans and AI agents. In J. Schmidhuber, K. Thórisson, & M. L. (Eds.), *Artificial general intelligence 2011* (pp. 122–132). (LNAI series, Springer volume 6830).

Insa-Cabrera, J., Dowe, D. L., & Hernandez-Orallo, J. (2011). Evaluating a reinforcement learning algorithm with a general intelligence test. In J. M. J. A. Lozano, & J. A. Gamez (Eds.), *Current topics in artificial intelligence. CAEPIA 2011*. : Springer (LNAI Series 7023).

Jensen, A. (1998). *The g factor: The science of mental ability*. Westport: Praeger.

Kleiner, K. (2011). Who are you calling bird-brained? An attempt is being made to devise a universal intelligence test. *Economist*, *398*, 82.

Oppy, G., & Dowe, D. L. (2011). The Turing test. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. : Stanford University http://plato.stanford.edu/entries/turing-test/.

Sanghi, P., & Dowe, D. L. (2003). A computer program capable of passing I.Q. tests. In P. P. Slezak (Ed.), *Proceedings of the joint international conference on cognitive science, 4th ICCS international conference on cognitive science & 7th ASCS Australasian Society for Cognitive Science (ICCS/ASCS-2003)* (pp. 570–575). (Sydney, NSW, Australia).

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and Control*, *7*, 1–22.

Solomonoff, R. J. (1985). The time scale of artificial intelligence: Reflections on social effects. *Human Systems Management*, *5*, 149–153.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460.

von Ahn, L., Blum, M., & Langford, J. (2004). Telling humans and computers apart automatically. *Communications of the ACM*, *47*, 56–60.

Wallace, C. S. (1995). *Multiple factor analysis by MML estimation.* Technical report TR 95/218 Dept. of Computer Science, Monash University.

Wallace, C. S. (2005). *Statistical and inductive inference by minimum message length.* : Springer-Verlag.

Wallace, C. S., & Boulton, D. M. (1968). An information measure for classification. *The Computer Journal*, *11*, 185–194.

Wallace, C. S., & Dowe, D. L. (1999). Minimum message length and Kolmogorov complexity. *The Computer Journal*, *42*, 270–283 (Special issue on Kolmogorov complexity).

Wallace, C. S., & Freeman, P. R. (1992). Single factor analysis by MML estimation. *Journal of the Royal Statistical Society (Series B)*, *54*, 195–209.