AI results for the Atari 2600 games: difficulty and discrimination using IRT

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

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

Abstract

We analyse the behaviour of around 40 learning techniques for one of the most popular general-purpose AI benchmarks in the recent years: the Arcade Learning Environment, based on the Atari 2600 games. We use item response theory, and logistic models in particular, to create item characteristic curves to determine which games in the benchmark are more difficult but also more discriminating. We interpret these results and suggest how ALE can be used and the results interpreted according to the IRT parameters and curves.

1 Introduction

The evaluation of AI systems and its overall progress has traditionally been based on milestones for specific tasks, where superhuman performance has been shown for more and more complex tasks. Of particular impact have been the success for chess in the 2000s with Deep Blue against the then human chess champion Garry Kasparov [Campbell *et al.*, 2002], the 2010s IBM's program Watson winning the *Jeopardy!* TV quiz [Ferrucci *et al.*, 2010; Ferrucci *et al.*, 2013], or, very recently, AlphaGo [Silver *et al.*, 2016] beating human champions on the ancient game of Go.

However, the evaluation in AI is now paying attention to systems that solve several tasks at a time [Hernández-Orallo, 2016a; Hernández-Orallo, 2017]. In particular, a very popular setting for general-purpose evaluation today is the collection of games or tasks under an interactive scenario, where agents can perceive and act, and are rewarded when they are right. Many different platforms have recently appeared for that [Hernández-Orallo *et al.*, 2017].

One benchmark that has become particularly popular in the past years is the Arcade Learning Environment [Bellemare *et al.*, 2015], a collection of Atari 2600 games that is usually tacked by reinforcement learning algorithms or search methods for planning. The popularity of this benchmark has increased significantly since [Mnih *et al.*, 2015] introduced a combination of deep learning and Q-learning, known as DQN, which was able to perform better than humans for many of the games, where learning was just performed by observing the screen and receiving rewards (the score), without any other given representation or description of the game, just learning to play from scratch.

The popularity of this benchmark has produced a good number of results that can now be analysed in hindsight and used to better understand the benchmark, and general-purpose AI evaluation overall. In this paper, we analyse the data from many research papers in the past three years using item response theory, a powerful technique from psychometrics, which allows us to determine which games are not only more or less difficult, but, more importantly, which are more discriminating, so that an increase in performance in these games represents an overall increase in the rest. Also, taking into account the long training and evaluation times of recent computing-demanding algorithms, any understanding of what the key games are (in order to reduce the size of the benchmark, specially in the hyperparameter search) can imply an important contribution for AI researchers.

The paper is organised as follows. Section 2 presents IRT as an appropriate tool for AI evaluation. Section 3 describes the experimental methodology used in terms of AI techniques and Atari 2600 games used. Section 4 focuses on the IRT models and parameters estimated and how they can be used to select those more informative games. Section 5 discusses the findings of the previous sections.

2 Item response theory in AI

Item response theory (IRT) [Embretson and Reise, 2000; De Ayala, 2009] has been mainly used in educational testing and psychometric evaluation in which examinees' ability is measured using a test with several questions (i.e., items). In essence, IRT is a set of mathematical models that describe the relationship between a latent trait of interest and respondents answers to individual items, where the probability of a response for an item is a function of the examinee's ability (or proficiency) and some item's parameters. There are models developed in IRT for different kinds of response, but we will focus on the dichotomous models. In dichotomous models the response can be either correct or incorrect. Multiple choice items (more than two options) can be also considered dichotomous since they can still be scored as correct/incorrect.

Let U_{ij} be a binary response of a respondent j to item i, with $U_{ij} = 1$ for a correct response and $U_{ij} = 0$ otherwise. Let θ_j be the ability or proficiency of j. Now, assuming that

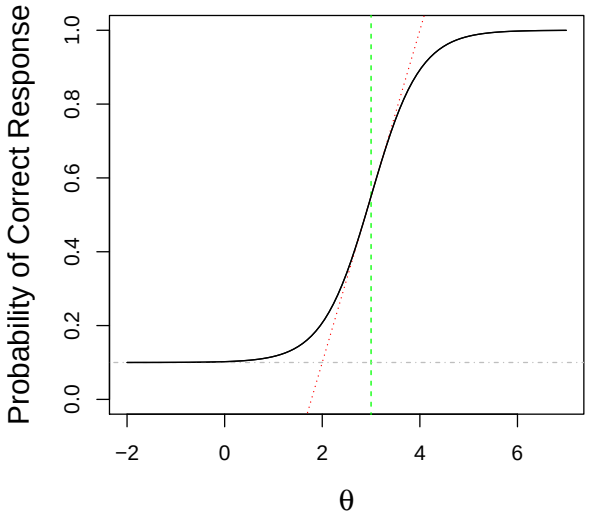

Figure 1: Example of a 3PL IRT model (in black), with slope $a = 2$ (discrimination, in red), location parameter $b = 3$ (difficulty, in green) and guessing parameter $c = 0.1$ (chance, in grey).

the result only depends on the ability and no longer on the particular respondent, we can express the response as a function of i alone, i.e. U_i. For the basic 3-parameter (3PL) IRT model, the probability of a correct response on an item given the examinee's ability is modelled as a logistic function:

$$P(U_i = 1|\theta_j) = c_i + \frac{1 - c_i}{1 + exp(-a_i(\theta_j - b_i))} \quad (1)$$

The above model provides for each item its *Item Characteristic Curve (ICC)* (see Figure 1 as an example) characterised by the parameters:

- *Difficulty* (b_i): it is the location parameter of the logistic function and can be seen as a measure of item difficulty. When $c_i = 0$, then $P(U_i = 1|b_i) = 0.5$.
- *Discrimination* (a_i): it indicates the steepness of the function at the location point. A high value suggests that an small change in ability can result in a big change in the item response. It can be computed as the maximum slope $a_i(1 - c_i)/4$;
- *Guessing* (c_i): it represents the probability of a correct response by a respondent with very low ability ($P(U_i = 1| - \infty) = c_i$). This is usually associated to a result given by chance.

The basic IRT model can be simplified to two parameters (e.g., assuming that $c_i = 0$), or just one parameter (assuming $c_i = 0$ and a fixed value of a_i, e.g. $a_i = 1$) models.

The ability of an individual (latent trait) is estimated based on his/her responses to discriminating items with different levels of difficulty. Respondents who tend to correctly answer the most difficulty items will be assigned to high values of ability. Difficulty items in turn are those correctly answered only by the most proficient respondents. Notice that subject traits and item characteristics are referenced to the same scale.

Straightforward methods based on maximum-likelihood estimation (MLE) can be used to estimate either the item parameters (when examinee abilities are known) or the abilities (when item parameters are known). A more difficult, but common, situation is the estimation when both the item parameters and respondent abilities are unknown. In this situation, an iterative two-step method (Birnbaum's method [Birnbaum, 1968]) can be adopted:

- Step (1) Start with initial values for abilities θ_j (e.g., random values or the number of correct responses) and estimate the model parameters;
- Step (2) Adopt the estimated parameters in the previous step as known values and estimate the abilities θ_j.

In this method, item parameters and respondent abilities are simultaneously estimated only based on a set of observed responses to items, with no previous knowledge about the true ability of the respondents.

In our adaptation of IRT, an item in IRT can be identified with a problem in AI (e.g., an Atari 2600 game), and an individual (or subject) can be identified with an AI method, technique or system. Recently, there has been more understanding about how IRT should be applied to classification at the level of instances in [Prudêncio *et al.*, 2015; Martínez-Plumed *et al.*, 2016].

3 Data

The Arcade Learning Environment (ALE) was introduced by [Bellemare *et al.*, 2015], after compiling a good number of games for the Atari 2600, a popular console of the late 1970s and most of the 1980s. The simplicity of the games from today's perspective and the use of a visual input of 210×160 RGB dots at 60Hz makes the benchmark sufficiently rich (but still simple) for the AI algorithms of today.

The results of [Mnih *et al.*, 2015] for ALE being superhuman for many games have popularised ALE as a benchmark for learning systems that improve from experience. The number of platforms, techniques and papers that use ALE today is such that the results on this Atari 2600 ALE are analysed when talking about progress in AI[1].

We have performed a search to find the papers that have used ALE since its introduction. We will focus on those techniques that are not allowed to look ahead using a simulator (this is common in search-based approaches [Naddaf, 2010; Lipovetzky *et al.*, 2015; Shleyfman *et al.*, 2016]). Instead, we will use the results obtained with truly learning approaches (most, but not necessarily all, using reinforcement techniques, usually in conjunction with deep learning). In this category, we are flexible about whether the results include human assistance (such as human start examples or learning by demonstration). For instance, we consider the results for the "noop" and "humanstarts" settings.

[1] http://www.milesbrundage.com/blog-posts/my-ai-forecasts-past-present-and-future-main-post

Game	# Actions	Difficulty	Discrimination
Alien	18	-	-
Amidar	10	2.49232336	78.0941338
Assault	7	-0.77078409	1.8293886
Asterix	9	1.17327222	74.4803265
Asteroids	14	-	-
Atlantis	4	-1.10992006	2.6870913
Bank Heist	18	1.27306997	24.8152159
Battle Zone	18	2.83531991	2.4685478
Beam Rider	9	1.40680369	1.1584455
Bowling	6	-	-
Boxing	18	-2.11368712	1.4113013
Breakout	4	-0.44196066	4.0757191
Centipede	18	-4.93797512	-0.4883670
Chopper Command	18	2.83899376	2.4495428
Crazy Climber	9	-0.91089795	4.1155110
Demon Attack	6	-0.68064430	3.3891392
Double Dunk	18	-0.35997097	1.8784440
Enduro	9	0.85847901	2.1029635
Fishing Derby	18	1.28989165	3.1236767
Freeway	3	0.60365230	1.1637861
Frostbite	18	2.74588098	147.1717478
Gopher	8	-0.67401643	23.2702166
Gravitar	18	-	-
H.E.R.O	18	9.14254376	0.4492266
Ice Hockey	18	2.50984497	2.0125731
James Bond	18	-0.32240713	3.2068925
Kangaroo	18	0.87031456	1.1254110
Krull	18	-	-
Kung-Fu Master	14	-0.22134822	2.0462940
Montezuma's Revenge	18	-	-
Ms. Pacman	9	-	-
Name This Game	6	-0.08162203	3.7408238
Pong	3	-0.04440702	34.1502733
Private Eye	18	-	-
Q*Bert	6	1.39864132	1.1884510
River Raid	18	1.66434969	3.6145164
Road Runner	18	-0.67393443	22.9443401
Robotank	18	-6.66322664	0.2756640
Seaquest	18	2.87005804	146.3441990
Space Invaders	6	0.16420283	2.8756194
Star Gunner	18	-0.32522627	5.0075491
Tennis	18	10.48605210	-0.1116351
Time Pilot	10	0.60796743	1.3446705
Tutankham	8	1.98175005	0.6102680
Up and Down	6	-0.39948025	1.5542507
Venture	18	-4.95755096	-0.5808805
Video Pinball	9	3.16049027	-0.2957503
Wizard of Wor	10	0.50861249	1.9847390
Zaxxon	18	0.87547205	2.7558438

Table 1: Games from the Arcade Learning Environment (ALE) analysed in this paper jointly with the estimated IRT parameters. Alien, Asteroids, Bowling, Gravitar, Montezuma, Ms. Pacman, Private Eye and Krull were not finally included in the IRT analysis as they have no variance after they were binarised according to being above or below human performance.

Overall, we integrate about 40 techniques from about a dozen papers [Mnih *et al.*, 2013; Mnih *et al.*, 2015; Furelos Blanco, 2015; Gruslys *et al.*, 2017; He *et al.*, 2016; Nair *et al.*, 2015; O'Donoghue *et al.*, 2017; Pritzel *et al.*, 2017; Salimans *et al.*, 2017; Schaul *et al.*, 2015; Talvitie and Bowling, 2015; Van Hasselt *et al.*, 2016; Wang *et al.*, 2015]. We discarded some papers because they did not include results for the 49 games that are most common in many papers. As some results (especially DQN) are reported repeatedly for some papers, we removed all results with a correlation higher than 0.99. In other cases, the results for the same technique with different parameters were kept.

For the ease of comparison, we use normalised scores (where 0 is like random, and 100 is like human). Finally, in order to apply binary IRT, we consider success when the normalised score is above 100.

Table 1 shows the games we used for the analysis. Alien, Asteroids, Bowling, Gravitar, Montezuma, Ms. Pacman and Private Eye were below human performance for all techniques, meaning that variance was 0 and hence inappropriate for IRT. Krull was always above human performance. The data and code can be found in http://users.dsic.upv.es/~fmartinez/IRT/AtariIRT.html

Discretising results by being above or below human performance may look very anthropocentric. This is true, but we have to remember that these games were designed for humans. Indeed, this human-based normalisation can be helpful to see those games for which the result of AI techniques can show a negative discrimination (best techniques do worse), which could tell us something about these games.

4 Models, item characteristic curves and parameter analysis

A 2-parameter IRT logistic model (2PL) is learned for each Atari 2600 game, fitting the probability of correct response for all techniques according to their abilities. We used a 2-parameter model instead of a 3-parameter model including the guessing parameter because this parameter (which tells us how likely the examinees are to obtain the correct answer by guessing) is not considered in this analysis under the assumption that the probability of a random-guessing technique being better than human is equal to zero.

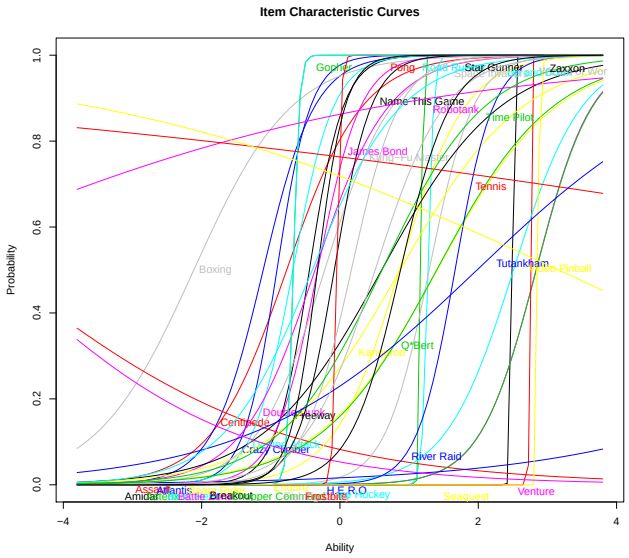

Figure 2: All ICC plots

We adopt MLE to estimate all the model parameters of all instances and the classifier abilities simultaneously, as usual in IRT. In particular, for generating the IRT models, we used the `ltm` R package[2], which implements the previously mentioned Birnbaum's method. The model parameters characterise the instance difficulty and discrimination power (see

[2] https://cran.r-project.org/web/packages/ltm/

Table 1 and Figure 2 for all parameters and curves). The ability of a technique is also estimated (see Figure 3) by the MLE method in the IRT package under different contexts (levels of instance difficulty).

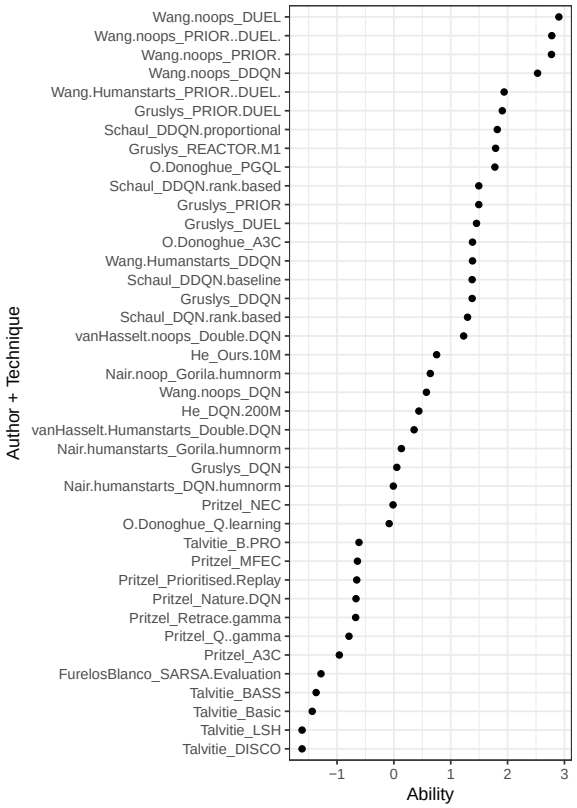

Figure 3: IRT-estimated abilities for the AI techniques included in this study.

The item parameter that is easiest to understand is difficulty. Because of the MLE estimation method, the value is not equal but well correlated to the percentage of techniques which score is above 100. Intuitively, easy games are solved by almost all techniques, and difficulty games are those that are only solved by very able techniques. Figure 5 shows the ICCs of those 6 most difficult Atari 2600 games with positive discrimination. However, in Figure 6 we can also observe that the difficulty value for "Tennis" game is second to none (10.5), but its discrimination is negative. Is this game particularly difficult or is it just a useless game since most able techniques do worse than those less able ones? What it is clear is that, in order to determine which games are more informative or useful for the analysis of new AI algorithms, we also need to look at the discrimination parameter.

In this case, the discrimination parameter (slope) measures the capability of a game to differentiate between techniques. Therefore, when applying IRT to evaluate techniques, the slope of an instance can be used to indicate if the game is useful to distinguish between strong or weak techniques for a problem. Figure 4 shows those ICCs of the most discriminating games. From the 41 games analysed, 37 had positive discrimination values so that the probability of correct responses (score above 100) is positively related to the estimated ability of the techniques. However, negative discrimination values were observed for 4 games (Figure 6). The latter means that these games are most frequently solved (score above 100) by the weakest techniques. These cases are anomalous in IRT (usually referred to as abstruse or idiosyncratic items) and, therefore, these games should be considered with extreme care for the analysis of new AI algorithms.

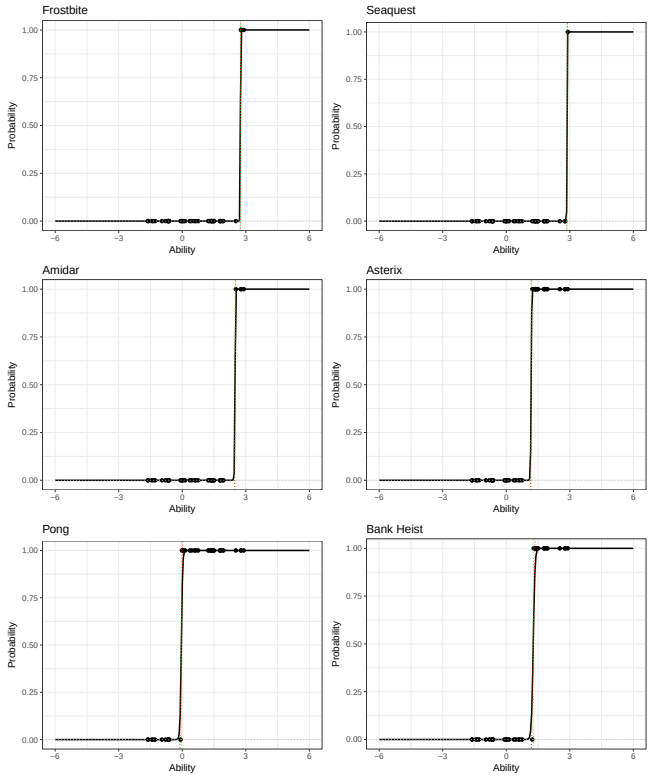

Figure 4: ICCs of the 6 most discriminating Atari 2600 games. Technique abilities are also included in the ICCs, plotted at $y = 1$ if their score is above 100, and at $y = 0$ otherwise. Difficulty and discriminating values are represented, respectively, with green and red dashed lines.

5 Discussion

The results above show that some games are not really very useful for the analysis of new AI algorithms. This is not referring to those games that were discarded because all techniques are above or below human performance, but to those games that have very low or even negative discriminating power. These games will not very useful to detect if new techniques are really improving overall for the whole benchmark. Our recommendation, especially for hyperparameter tuning, is to focus on those games that are discriminating, and also use those whose difficulty is close to the estimated ability of the technique, resembling adaptive testing.

An interesting analysis would be to study the average correlation for all techniques (0.2746862) and games

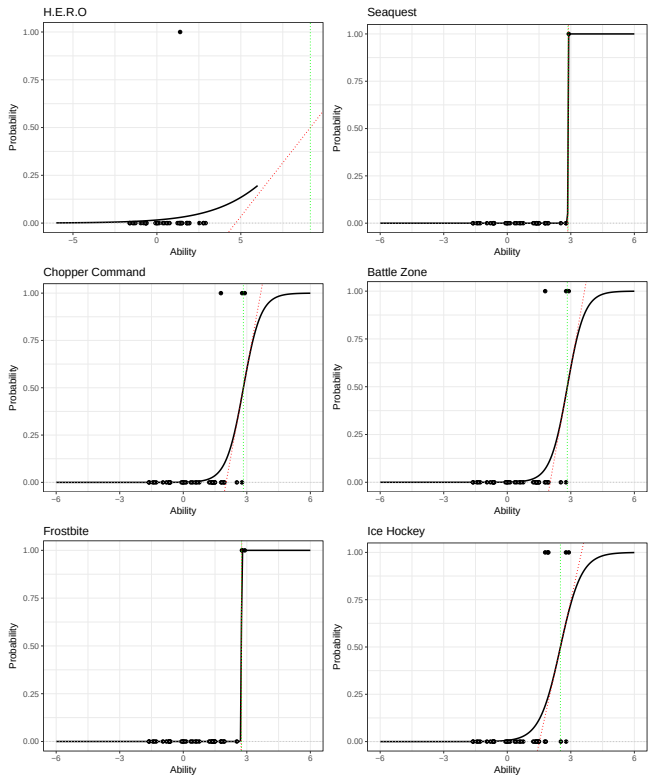

Figure 5: ICCs (with positive slope) of the 6 most difficult Atari 2600 games. Technique abilities are also included in the ICCs, plotted at $y = 1$ if their score is above 100, and at $y = 0$ otherwise. Difficulty and discrimination values are represented, respectively, with green and red dashed lines.

(0.3218037) in order to analyse whether there is a general factor, and also whether this is increasing with time, focusing on whether AI techniques are creating more general techniques that perform better overall and not at the cost of performing very badly at a small subset of problems, in line with the analysis of generality in AI [Hernandez-Orallo, 2016b].

As future work, we would like to use quantitative IRT models to consider the scores for the games without the discretisation (above/below human performance). This could produce further insights in the understanding of the benchmark.

References

[Bellemare *et al.*, 2015] Marc Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015. 1, 2

[Birnbaum, 1968] A. Birnbaum. *Statistical Theories of Mental Test Scores*, chapter Some Latent Trait Models and Their Use in Inferring an Examinees Ability. Addison-Wesley, Reading, MA., 1968. 2

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

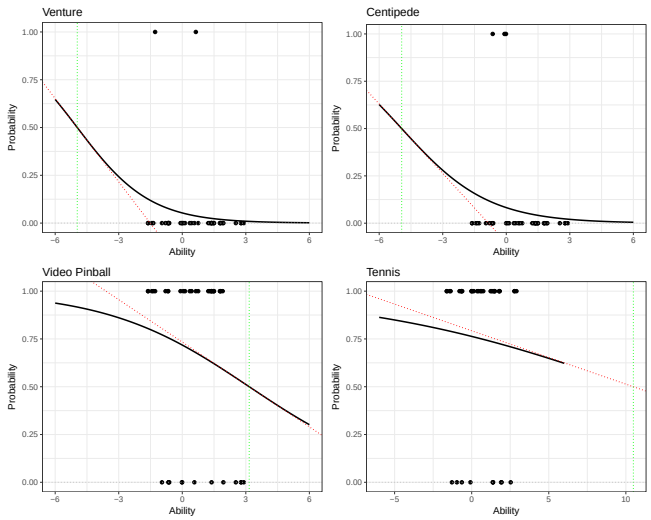

Figure 6: Examples of ICCs of those Atari Games with negative discrimination. Technique abilities are also included in the ICCs, plotted at $y = 1$ if their score is above 100, and at $y = 0$ otherwise. Difficulty and discrimination values are represented, respectively, with green and red dashed lines.

[De Ayala, 2009] Rafael Jaime De Ayala. *Theory and practice of item response theory*. Guilford Publications, 2009. 1

[Embretson and Reise, 2000] S. E. Embretson and S. P. Reise. *Item response theory for psychologists*. L. Erlbaum, 2000. 1

[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. 1

[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. 1

[Furelos Blanco, 2015] Daniel Furelos Blanco. Learning and generalization in atari games. 2015. 3

[Gruslys *et al.*, 2017] Audrunas Gruslys, Mohammad Gheshlaghi Azar, Marc G Bellemare, and Remi Munos. The reactor: A sample-efficient actor-critic architecture. *arXiv preprint arXiv:1704.04651*, 2017. 3

[He *et al.*, 2016] Frank S He, Yang Liu, Alexander G Schwing, and Jian Peng. Learning to play in a day: Faster deep reinforcement learning by optimality tightening. *arXiv preprint arXiv:1611.01606*, 2016. 3

[Hernández-Orallo *et al.*, 2017] José Hernández-Orallo, Marco Baroni, Jordi Bieger, Nader Chmait, David L Dowe, Katja Hofmann, Fernando Martínez-Plumed, Claes Strannegård, and Kristinn R Thórisson. A new ai evaluation cosmos: Ready to play the game? *AI Magazine*, 2017. 1

[Hernández-Orallo, 2016a] J. Hernández-Orallo. Evaluation in artificial intelligence: From task-oriented to ability-oriented measurement. *Artificial Intelligence Reviews*, .(.):., 2016. 1

[Hernandez-Orallo, 2016b] Jose Hernandez-Orallo. Is spearmans law of diminishing returns (slodr) meaningful for artificial agents? In *ECAI 2016: 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands-Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*, volume 285, page 471. IOS Press, 2016. 5

[Hernández-Orallo, 2017] J. Hernández-Orallo. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press, 2017. 1

[Lipovetzky *et al.*, 2015] Nir Lipovetzky, Miquel Ramirez, and Hector Geffner. Classical planning with simulators: results on the atari video games. In *International Conference on Automated Planning and Scheduling, Proceedings of the 7th Workshop on Heuristics and Search for Domain-independent Planning (HSDIP)*, 2015. 2

[Martínez-Plumed *et al.*, 2016] Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Making sense of item response theory in machine learning. In *European Conference on Artificial Intelligence, ECAI*, pages 1140–1148, 2016. 2

[Mnih *et al.*, 2013] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. 3

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. 1, 2, 3

[Naddaf, 2010] Yavar Naddaf. Game-independent ai agents for playing atari 2600 console games. *MSc, Dep. of Computing Science, University of Alberta*, 2010. 2

[Nair *et al.*, 2015] Arun Nair, Praveen Srinivasan, Sam Blackwell, Cagdas Alcicek, Rory Fearon, Alessandro De Maria, Vedavyas Panneershelvam, Mustafa Suleyman, Charles Beattie, Stig Petersen, et al. Massively parallel methods for deep reinforcement learning. *arXiv preprint arXiv:1507.04296*, 2015. 3

[O'Donoghue *et al.*, 2017] Brendan O'Donoghue, Remi Munos, Koray Kavukcuoglu, and Volodymyr Mnih. Combining policy gradient and q-learning. *International Conference on Learning Representation*, 2017. 3

[Pritzel *et al.*, 2017] Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. *arXiv preprint arXiv:1703.01988*, 2017. 3

[Prudêncio *et al.*, 2015] Ricardo BC Prudêncio, José Hernández-Orallo, and Adolfo Martínez-Usó. Analysis of instance hardness in machine learning using item response theory. In *Second International Workshop on Learning over Multiple Contexts in ECML 2015. Porto, Portugal, 11 September 2015*, 2015. 2

[Salimans *et al.*, 2017] Tim Salimans, Jonathan Ho, Xi Chen, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017. 3

[Schaul *et al.*, 2015] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015. 3

[Shleyfman *et al.*, 2016] Alexander Shleyfman, Alexander Tuisov, and Carmel Domshlak. Blind search for atari-like online planning revisited. *Heuristics and Search for Domain-independent Planning (HSDIP)*, page 85, 2016. 2

[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. 1

[Talvitie and Bowling, 2015] Erik Talvitie and Michael Bowling. Pairwise relative offset features for atari 2600 games. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. Citeseer, 2015. 3

[Van Hasselt *et al.*, 2016] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI*, pages 2094–2100, 2016. 3

[Wang *et al.*, 2015] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. Dueling network architectures for deep reinforcement learning. *arXiv preprint arXiv:1511.06581*, 2015. 3