Analysis of the Protocols Used to Assess Virtual Players in Multi-player Computer Games

Cindy Even[1][2], Anne-Gwenn Bosser[1], and Cédric Buche[1]

[1] Lab-STICC, ENIB, CERV, 25 Rue Claude Chappe, 29280 Plouzané, France,
{even,bosser,buche}@enib.fr
[2] Virtualys, 41 Rue Yves Collet, 29200 Brest, France

Abstract. Recently, the development of believable agents has gained a lot of interest and many solutions have been proposed by the research community to implement such bots. However, in order to make advances in this field, a generic and rigorous evaluation that would allow the comparison of new systems against existing ones is needed. This paper provides a summary of the existing believability assessments. Seven features characterising the protocols are identified. After a comprehensive analysis, recommendations and prospects for improvement are provided.

Keywords: Believability, Evaluation, Benchmarking, Turing test, Non player characters, Virtual players, Bots, Multi-player games

1 Introduction

Computer games can be populated by human players represented by their avatar as well as computer-controlled players, also known as Non-Player Characters (NPCs) or bots. They may have different roles in the game such as acting as traders, providing services, quests or clues to the human players. In multi-player games, a special type of bot - that we call “virtual player” - can be used instead of human players. Their role is to play the game as a human player would. They are necessary for players who want to practice before facing human opponents, players who do not have the possibility to connect with other players, or to fill in spots on a server when there are too few human players. The popularity of a video game (and therefore its commercial success) is linked to the quality of these bots. For example, an unbeatable bot would be frustrating to play against while a predictable one would be boring. Indeed, according to Livingstone [1], modern video games do not require unbeatable Artificial Intelligence (AI) but believable AI. Also, recent experimental results [2] show that believable bots increase users' enjoyment. Different approaches have been adopted for the development of believable bots, such as systems based on connectionist models [3, 4], production systems [5, 6] or probabilistic models [7–9] - to mention just a few. Generally, the proposed systems are not assessed, and when they are, the results obtained can not be compared as different protocols have been used. However, in order to make advances in this field, many authors [10, 11, 8] pointed out the need of a generic and rigorous evaluation that would allow the comparison of new systems against existing ones. The evaluation of AI in games research has been identified as one of the main challenges in game AI research [12]. In this paper, we review evaluation techniques for assessing the believability of virtual players and we provide a comprehensive analysis of the evaluation features. We conclude by suggesting prospects for improvement.

2 Assessing Believability

Authors have worked on criteria-based assessment methods [13], [14] where the believability of bots is ranked by the amount of criteria they meet. Even though such lists can be interesting, it can be difficult to take all the items into account during the assessment [15]. These lists are rather intended to provide a roadmap for the design of human-like bots. However, the notion of believability being highly personal, subjective assessments are a more common approach to measuring believability.

A way of evaluating AI is to organise competitions. According to Togelius [16], the advantage of competitions is that they provide fair, transparent and reusable benchmarks. In recent years we have seen the emergence of competition oriented toward the implementation of human-like (or believable) opponents such as the 2K Botprize competition [17] or the Turing Test track of the Mario AI Championship [18]. The BotPrize is particularly interesting as it has evolved significantly over the years. It is a variant of the Turing test [19] which uses the "Deathmatch" game-type mode of the video game *Unreal Tournament 2004 (UT2004)*. For the first two editions [17], each human judge played against a human confederate and a bot. At the end of each round, the judges were asked to evaluate the two opponents on a rating scale and to record their observations. For the next edition, a new protocol was implemented [20], in order to make the judging process part of the game. A weapon in the game had two firing modes that could be used to tag an opponent as being human or bot. Both bots and humans were equipped with the judging gun and could vote. This modification to the system introduced a bias in the evaluation process as the game-play was adversely affected. Whilst players previously had to move quickly in order to not present an easy target, in the new competition players are tempted to stop and observe their opponents to make a judgement [21]. Furthermore, judges may be inclined to attempt to communicate through movements and shooting patterns [6]. This kind of behaviour would not naturally occur in normal game-play. For the last edition[1], the novelty was the addition of a third-person believability assessment (i.e. the judges observe the game).

Third person assessment was also used in [22]. Videos were recorded where an expert player played against bots and human players with different levels. After watching videos, judges were asked to evaluate human-likeness on a 7-point

[1] Human-Like Bots Competition, presented at the IEEE CIG conference by Raúl Arrabales : `http://www.slideshare.net/array2001/arrabales-bot-prize2014v2`

Likert scale. A similar approach was used in [5] and [8] but with a different First Person Shooter (FPS) game (Quake II). The protocol's characteristics of these player believability assessments can be found in Table 1 along with relevant references regarding *player believability* (the belief that a character is controlled by a human player [23, 24]), and *character believability* (the belief that the character itself is real [24]) assessments.

As we can see from the descriptions below, the protocols used in the past for the assessment of virtual player's believability have characteristics that vary significantly. The process of judging the behaviours of a bot is, by nature, a subjective process [10, 11, 1] as it depends on the perceptions of the people playing or watching the game. Having no obvious physical attributes or features that can be measured, the only solution for measuring the believability of bots that can be considered is the use of a questionnaire [10]. In some cases, the players fill the questionnaire after playing the game for a few minutes, in other cases they vote during the game. The judgement can be done by the players or by observers, and different types of questionnaires are used such as ranking or comparison. In the next section we propose to analyse the characteristics of the protocols collected in Table 1.

3 Assessment's characteristics analysis

3.1 Application

The application used for the evaluation process can be pre-existing or developed specially for the test. The implementation of a sample game can be necessary when no open-source games are available [27] but it needs to be well-thought-out in order to not introduce bias unintentionally.

There are many advantages when choosing a pre-existing video game. According to Tencé et al. [23], the game needs to be a multi-player game (indeed, the role of virtual players is to be played against) offering a lot of interaction between the players. Action, role playing, adventure and sport games meet these criteria. Adventure and sport games tend to be difficult to modify and in particular, they rarely offer the possibility to add customised bots. The main draw-back of role playing games is that they rely in large part on communication and natural language which is not what we intend to evaluate here. Similarly, in order to not impact the assessment, all "chat" options should be disabled [17]. Action games, especially FPSs, are often a good choice. For the BotPrize contest, Hingston [17] chose UT2004 because it is affordable, readily available, customizable, bots and humans can play together and do not need to be collocated, and it is easy to interface a bot to the game. Togelius et al. [24] argued that FPS are not suitable for believability assessments as players encounter their opponents for only a few seconds and in the middle of a chaotic situation. For this reason they preferred to use the single player game *Infinite Mario Bros* which does not

Table 1: Comparison of the existing experiments

Reference	**Application**	**1^{st} or 3^{rd} person assessment**		**Duration**	**No. of judges**	**Judges' level**			**Information given**	**Subjective assessment type**				**How**
						novice	**medium**	**expert**		**binary**	**comparison**	**scale**	**comments**	
[5]	Quake II Deathmatch	3^{rd}	16 x 1 video candidate's view	3 min	8	✓		✓	A	✓		✓ 1 to 10		n/a
[10]	Simulation of a bar	3^{rd}	2 simulations global view	as long as needed	13	✓	✓		B		✓ 2 choices	✓ 1 to 5		pen & paper
[11]	Pong game	3^{rd}	video global view	n/a	n/a	n/a			A		✓ 4 choices		✓	n/a
[8]	Quake II Deathmatch	3^{rd}	15 x 3 videos 1^{st} person view	20 sec	20	✓	✓	✓	A			✓ 1 to 5	✓	on-line
[25]	CoPeFoot	1^{st}		n/a	48	✓		✓	C	✓				pen & paper
[17] (BotPrize v1)	UT2004 Deathmatch	1^{st}		10 min	5		✓	✓	A			✓ 1 to 5	✓	n/a
[20] (BotPrize v2)	UT2004 Deathmatch	1^{st}		n/a	7		✓	✓	A	✓				in-game
[4]	UT2004 Deathmatch	1^{st}		15 min	3		✓		n/a	✓				in-game
		3^{rd}	10 x 1 video 3^{rd} person view	1 min	12		✓		n/a	✓				crowdsourcing platform
[22]	UT2004 Capture The Flag	3^{rd}	1 x 4 videos 1^{st} person view	n/a	10	n/a			n/a			✓ 1 to 7		n/a
[18]	Infinite Mario Bros	3^{rd}	2 videos global view	1 min	73	n/a			n/a		✓ 4 choices			on-line
[26]	Everyday life of the Darug people	3^{rd}	14 x 2 videos 1^{st} person view	n/a	43	n/a			B		✓ 3 choices	✓ 1 to 5		on-line

▨ Character believability assessment.

A Judges are told that there is a mix of bots and humans.

B Judges know the nature of each entity.

C Judges are given no information.

meet the criteria of being a multi-player game.

3.2 1^{st} or 3^{rd} person assessment

Believability assessment may consider both first person and third person reports. In first person assessment, the judge has two simultaneous roles: to play the game, and to judge opponents. On the other hand, in third person assessment, the judge is only a spectator observing the game being played.

In [5, 18, 4] the authors argued that assessing believability from a first-person perspective might be distracting since the judge has to pay attention both to the game experience and to the behaviour of the other players for the evaluation. Livingstone answered in his paper [1] with : "*in game development the aim is to satisfy the needs of the players of a game and not those of watchers*". However, even if computer games are primarily designed for the players, video game spectating has recently become a popular activity [28, 29]. In Cheung et al. paper [28], the authors report that there are some spectators that actually prefer to watch professionals playing rather than playing the game themselves.

First person assessment is possible only with applications that can be played by at least two players simultaneously. The third person assessment however, can be used with any application. When performing a third person assessment, judges are asked to give their judgement after watching a video of the game previously recorded. To reduce the subjectivity and the guesswork, Gorman et al. [8] suggested to show more than one video to the judges in order for them to have a basis for comparison. They also pointed out the risk of introducing a bias when selecting videos for the assessment. The person in charge of the selection might pick parts of the video that could influence the responses.

When recording the videos, different points of view can be used. In some cases the application does not offer many possibilities. The Pong game for example, can only be played with a global view, representing the tennis table and the two paddles. In other video games such as FPSs, it is possible to choose between the first and third person view. Therefore, videos can be recorded from the confederate's or the candidate's first or third person view.

Confederate's 1^{st} or 3^{rd} person view The confederate's 1^{st} person view is most commonly used for assessing the believability of bots. This might be due to the fact that it is easily recorded during game play, particularly during a first person assessment. These points of view allow us to capture the game as if the judges were in-play. The main drawback of these points of view is that a considerable portion of recording can not be used. Indeed, all the moments when the confederate is in the environment without facing the candidate are useless and need to be cut from the video.

Candidate's 1^{st} person view When using the candidate's first person view, the judges have less resources to evaluate the entity : for instance, they can not see its movements.

Candidate's 3^{rd} person view This solution has never been used in our knowledge. Yet it could be especially interesting since it would capture both the perception and the actions of the candidates. This could allow a better understanding of the decisions made by the candidate. Moreover, it would not require cuts in the recording as even the time when the candidates are alone in the environment could be used for the judgement, which would be time saving and would reduce the risk of introducing the aforementioned bias when selecting videos for the assessment.

3.3 Duration

The duration of video and game play varies greatly from one experience to another, going from 20 seconds to as long as the judge desires. It might depend on the nature of the game but most of the time, the choice of the experiment's duration relies on the organisers' opinion [24, 18] and is never justified. In their experiment, Soni et al. [2] tried to examine the role of predictability by using two different bots during their assessment. Unfortunately, the subjects did not notice any difference between the two bots. The authors hypothesised that the experiment was too short and that longer sessions could give the judges enough time to make a distinction. The observation of Paritosh et al. [30] regarding the Loebner competition[2] (the first formal instantiation of a Turing Test) is similar. They argue that the test is too short (only few minutes) to allow any depth in the judgement. Even if it is important to allow enough time for the judges to make a judgement, the assessment can not be too long as it can induce inattention or mistakes due to judges' boredom or fatigue [31].

3.4 Number of judges

The assessment being of subjective nature, it seems important to collect a significantly large number of judgements in order to cancel out the biases introduced by that type of assessment [32]. The use of on-line surveys eases the collection and treatment of results. For their experiment, Llargues Asensio et al. [4] used a crowd-sourcing platform for mobile devices that allows to conduct a video-based poll experiment where the users can vote at the end of each video clip.

3.5 Judges' and confederates' expertise

The level of the judges is sometimes taken into account for the experiment. As it has been noticed by Mac Namee [10], the experience of players in video games can introduce a difference between the subjects. In general, for an experienced

[2] `http://www.loebner.net/Prizef/loebner-prize.html`

player it will be quicker and easier to recognise a bot than for a novice player. For example, in Laird et al.'s paper [5], only the expert player made no mistake in differentiating between bots and humans. Novice players might not fully know the rules of the game or the available actions which could make the whole experience too confusing and they would not be capable to sensibly evaluate the players' behaviours [1].

Another interesting element that has been taken into account in [5, 22, 18] is the level of the confederates. They have a major role in the assessment as their behaviours directly influence the judges' evaluation. For example, a high-performing expert-player confederate could easily be mistaken for a bot by non-expert players [6]. On the contrary, novice-players confederates who are still learning how to play the game and how to use the controls might be mistaken with a weak bot by expert players. Confederates should be provided with sufficient time for gaining control over the game rules and commands before starting the evaluation. Hingston [17] avoided these potential problems by choosing confederates who were all of a reasonable level of experience, i.e. neither expert nor novice.

3.6 Information given to the judges

As we can see in Table 1, judges can be given different information before starting the experiment. Most of the time, they are informed that they will see a combination of bots and human players (A in Table 1). In other cases, (B) they know the nature of the entity they are evaluating. Finally, (C) judges are not informed as to the purpose of the experiment. For instance, in [25], judges were invited to play a football video game, where all the players had a number. After a given time, the game was paused and they were given a table and the following instructions: "Cross the box corresponding to the two players controlled by humans in the simulation, if and only if you are confident in your answer. If in doubt, write nothing". The analysis of the results revealed that judges were considerably better at distinguishing bots from human players after the first attempt.

In two other experiments, half of the participants were informed that the other character in the game would be controlled by another person, while the other half were informed that it would be controlled by a computer (AI). In fact, for all the participants, the character were controlled by a computer in [33], and by a human in [34]. In the first experiment, the participants who played against the character that they believed to be human-controlled, reported stronger experiences of presence, flow, and enjoyment. And in the second experiment, the participants exhibited greater physiological arousal and reported greater presence and likeability when the character was introduced as being human controlled rather than computer controlled. These results demonstrate that the information given to the judges can significantly alter their judgement.

3.7 Subjective assessment types

When assessing players' believability in a game, players are asked to give their opinion [24]. Their answer can have the form of a free response or of forced data retrieved through questionnaires.

Free response answers can contain much richer information but they are also much harder to analyse appropriately. Sometimes judges have the opportunity to give a free response in the form of comments [17]. These comments can be useful for identifying areas for improvement for the bots implementation but are generally not used for evaluation.
On the other hand, by using a questionnaire, subjects are constrained to choose between some specific items, yielding data that is easier to analyse. Different types of forced questionnaires can be identified [24] :

- *Binary* : Subjects can answer by *Yes* or *No* to a simple question (e.g. *is this player a bot?*, or, *is this bot believable?*).
- *Scale* : Judges are asked to rate the humanness of the players' behaviour or to choose an answer within a list (e.g. [8] *1: Human, 2: Probably Human, 3: Don't Know, 4: Probably Artificial, 5: Artificial*).
- *Comparison* : Subjects are asked to compare two or more players (e.g. *did player A or B act more like a human player?*).

With ranking questionnaires, it is not possible to analyse the interpretation of the rating categories across subjects [35]. To minimise the subjective notion of scaling and allow a fairer comparison between the subjects' answers, comparison and boolean questions can be used [24]. But as mentioned by Hingston [17], a binary choice might have the effect of forcing the subjects to "toss a coin" if they are unable to choose an answer. In an effort to reduce subjectivity, in [10, 1] subjects were not asked to rate believability, instead, they were asked to compare two players and say which was more believable or acted more like a human player. The choice items may be presented in different ways, for instance, the subjects can choose between 2 solutions (*player A* or *player B*). They can also be offered more options such as *there is no difference*, or *both equally* and *none of them*, following the 4 alternative forced choice (4-AFC) protocol proposed by Yannakakis and Hallam [36].

4 Discussion

When studying the protocols used in the past to assess virtual players' believability, we identified some characteristics that varied significantly from one assessment to another, giving results that can not be correlated.

Application First of all, different types of games were used such as FPS, sport or platform games. The main criterion when choosing the game is that it needs to be a multi-player game where one can face virtual players. The second criterion, which restricts significantly the range of games that can be considered, is that it has to be possible to interface a bot.

1st or 3rd person assessment Even when the types of games used in the assessments were similar, judges had different roles. They were either part of the game (first person assessment), with the ability to interact with the candidates but also with the risk of modifying the game-play. Or they were spectators (third person assessment), assessing a game in which they were not involved. The recent interest for game spectating can be an additional argument in favour of this choice. For this type of assessment, the judges watch videos of the game. These videos can be recorded using different points of view. The most commonly used is the confederate's first person view but a solution that seems to have potential and needs to be tested is the candidate's third person point of view.

Duration The duration of the assessment is another characteristic that can vary significantly. Judges might give a random answer if they do not have enough time to evaluate a bot. In order to avoid this situation it seems important to define a minimum assessment duration.

Number of judges As the notion of believability is very subjective, it is important to collect a large number of judgements. The use of an on-line questionnaire or crowd-sourcing platform seems unavoidable as they can allow for the collection of more data that would give more accurate results. In order for the protocol to be rigorous, a minimum number of participants must be defined.

Judges' and confederates' expertise The judges' and confederates' level of experience is sometimes taken into account. In general, we recommend training novices before involving them in the roles of judge or confederate as they need to know the rules, the commands and to have experimented with the game. Otherwise, confederates could easily be mistaken with weak bots and judges could be too confused to be able to make a judgement. It would be interesting to study the influence of the judges' level on the results when the number of judges is high.

Information given to the judges As we saw in 3.6, recent experiments have shown the influence of the information given to the judges on their judgement. This part of the assessment protocol needs to be carefully designed in order to avoid introducing a bias. When conducting a first person assessment, the game-play might be modified if the judges know the aim of the assessment. The only way to avoid this is to keep the question secret and to ask the player only at the end of the game, whether he thought he was playing against a human player or a bot. Of course, the player could be asked only once. During a third person assessment, the best solution seems to be keeping the nature of the candidate secret and telling the judges that they would see a mix of bots and human players, so that they have no prejudices.

Subjective assessment types Finally, different types of questionnaire have been used (binary, scale or comparison) to collect the judges' opinions, giving data that can not be compared from one assessment to another. Regardless of the type of questionnaire, the question(s) as well as the offered solutions will have to be adapted according to the type of assessment (first or third person) and the information previously given to the judges.

5 Conclusion & Future Work

Virtual players play a major role in the success of video games. A new challenge is to develop believable bots that could blend in among human players. Over the years, different approaches have been used for the implementation of such bots. However most of the time, these bots were either not evaluated, or they were evaluated using different protocols. Yet, in order to make improvements in the development of believable bots, a generic and rigorous evaluation needs to be set up, that would allow the comparison between new systems and existing ones. According to Clark et al. [37], "*standardised tests are an effective and practical assessment of many aspects of machine intelligence, and should be part of any comprehensive measure of AI progress*". Although the evaluation of bots' performance can be performed through objective measures (comparing score or time spent to complete a level), the evaluation of bots' believability is complex due to its subjective aspect.

In this paper we analysed the protocols previously used to assess the believability of virtual players. We identified seven features that characterise the assessments and which vary significantly from one to another. When designing a new protocol, these features need to be chosen carefully in order to not introduce a bias into the evaluation. After an in-depth analysis of these protocols, we gave recommendations for the features that are well established. In order for the protocol to be rigorous and reusable, other features still need further study and testing to be determined.

References

1. Livingstone, D.: Turing's test and believable AI in games. Computers in Entertainment **4**(1) (jan 2006) 6
2. Soni, B., Hingston, P.: Bots trained to play like a human are more fun. In: IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). (jun 2008) 363–369
3. van Hoorn, N., Togelius, J., Wierstra, D., Schmidhuber, J.: Robust player imitation using multiobjective evolution. In: 2009 IEEE Congress on Evolutionary Computation, IEEE (may 2009) 652–659
4. Llargues Asensio, J.M., Peralta, J., Arrabales, R., Bedia, M.G., Cortez, P., Peña, A.L.: Artificial Intelligence approaches for the generation and assessment of believable human-like behaviour in virtual characters. Expert Systems with Applications **41**(16) (2014) 7281–7290
5. Laird, J.E., Duchi, J.C.: Creating Human-Like Synthetic Characters with Multiple Skill Levels: A Case Study Using the Soar Quakebot. Papers from 2001 AAAI Spring Symposium, Artificial Intelligence and Interactive Entertainment I (2001) 54–58

6. Polceanu, M.: Mirrorbot: Using human-inspired mirroring behavior to pass a turing test. In: IEEE Conference on Computational Intelligence in Games (CIG'13), IEEE (2013) 1–8
7. Le Hy, R., Arrigoni, A., Bessière, P., Lebeltel, O.: Teaching Bayesian behaviours to video game characters. Robotics and Autonomous Systems **47**(2-3) (jun 2004) 177–185
8. Gorman, B., Thurau, C., Bauckhage, C., Humphrys, M.: Believability Testing and Bayesian Imitation in Interactive Computer Games. From Animals to Animats 9 **1** (2006) 655–666
9. Tencé, F., Gaubert, L., Soler, J., De Loor, P., Buche, C.: CHAMELEON: Online Learning for Believable Behaviors Based on Humans Imitation in Computer Games. Computer Animation and Virtual Worlds (CAVW) **24**(5) (2013) 477–496
10. Mac Namee, B.: Proactive Persistent Agents: Using Situational Intelligence to Create Support Characters in Character-Centric Computer Games. PhD thesis, University of Dublin, Trinity College (aug 2004)
11. McGlinchey, S., Livingstone, D.: What believability testing can tell us. In: Proceedings of the International Conference on Computer Games: Artificial Intelligence, Design, and Education. (2004) 273–277
12. Lucas, S.M., Mateas, M., Preuss, M., Spronck, P., Togelius, J.: Artificial and Computational Intelligence in Games (Dagstuhl Seminar 12191). Dagstuhl Reports **2**(5) (2012) 43–70
13. Hinkkanen, T., Kurhila, J., Pasanen, T.A.: Framework for evaluating believability of non-player characters in games. In: AI and Machine Consciousness. (2008)
14. Arrabales, R., Ledezma, A., Sanchis, A.: ConsScale: A Pragmatic Scale for Measuring the Level of Consciousness in Artificial Agents. Journal of Consciousness Studies **17**(3-1) (2010) 131–164
15. Arrabales, R., Ledezma, A., Sanchis, A. In: ConsScale FPS: Cognitive Integration for Improved Believability in Computer Game Bots. Springer Berlin Heidelberg, Berlin, Heidelberg (2012) 193–214
16. Togelius, J.: How to Run a Successful Game-Based AI Competition. IEEE Transactions on Computational Intelligence and AI in Games **8**(1) (mar 2016) 95–100
17. Hingston, P.: A Turing Test for Computer Game Bots. IEEE Transactions on Computational Intelligence and AI in Games **1**(3) (sep 2009) 169–186
18. Shaker, N., Togelius, J., Yannakakis, G.N., Poovanna, L., Ethiraj, V.S., Johansson, S.J., Reynolds, R.G., Heether, L.K., Schumann, T., Gallagher, M.: The turing test track of the 2012 Mario AI Championship: Entries and evaluation. In: IEEE Conference on Computational Intelligence in Games (CIG'13), IEEE (2013) 1–8
19. Turing, A.M.: Computing machinery and intelligence. Mind **59**(236) (1950) 433–460
20. Hingston, P.: A new design for a Turing Test for Bots. In: Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games, IEEE (aug 2010) 345–350
21. Thawonmas, R., Murakami, S., Sato, T.: Believable judge bot that learns to select tactics and judge opponents. In: IEEE Conference on Computational Intelligence and Games (CIG'11). (2011) 345–349
22. Acampora, G., Loia, V., Vitiello, A.: Improving game bot behaviours through timed emotional intelligence. Knowledge-Based Systems **34** (2012) 97–113
23. Tencé, F., Buche, C., De Loor, P., Marc, O.: The challenge of believability in video games: Definitions, agents models and imitation learning. In Mao, W., Vermeersch, L., eds.: 2nd Asian Conference on Simulation and AI in Computer Games (GAMEON-ASIA'10), Eurosis (2010) 38–45

24. Togelius, J., Yannakakis, G.N., Karakovskiy, S., Shaker, N.: Assessing Believability. In Hingston, P., ed.: Believable Bots: Can Computers Play Like People? Springer Berlin Heidelberg (2012) 215–230
25. Bossard, C., Benard, R., De Loor, P., Kermarrec, G., Tisseau, J.: An exploratory evaluation of virtual football player's believability. In: Proceedings of 11th Virtual Reality International Conference (VRIC'09). (2009) 171–172
26. Bogdanovych, A., Trescak, T., Simoff, S.: What makes virtual agents believable? Connection Science (2016)
27. Bernacchia, M., Hoshino, J.: AI platform for supporting believable combat in role-playing games. In: Proceedings of the 19th Game Programming Workshop in Japan. (2014) 139–144
28. Cheung, G., Huang, J.: Starcraft from the stands: understanding the game spectator. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2011) 763–772
29. Kaytoue, M., Silva, A., Cerf, L.: Watch me playing, i am a professional: a first study on video game live streaming. Proceedings of the 21st international conference companion on World Wide Web (June 2009) (2012) 1181–1188
30. Paritosh, P., Marcus, G.: Toward a Comprehension Challenge, Using Crowdsourcing as a Tool. AI Magazine **37**(1) (2016) 23–30
31. Brace, I.: Questionnaire design: How to plan, structure and write survey material for effective market research. Kogan Page Publishers (2008)
32. Hyman, H.H., Center, N.O.R.: Interviewing in social research. A research project of the National Opinion Research Center. University of Chicago Press (1954)
33. Weibel, D., Wissmath, B., Habegger, S., Steiner, Y., Groner, R.: Playing online games against computer- vs. human-controlled opponents: Effects on presence, flow, and enjoyment. Computers in Human Behavior **24**(5) (sep 2008) 2274–2291
34. Lim, S., Reeves, B.: Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. International Journal of Human Computer Studies **68**(1-2) (2010) 57–68
35. Friedman, H.H., Amoo, T.: Rating The Rating Scales. The Journal of Marketing Management **9**(3) (1999) 114–123
36. Yannakakis, G.N., Hallam, J.: Real-time game adaptation for optimizing player satisfaction. IEEE Transactions on Computational Intelligence and AI in Games **1**(2) (2009) 121–133
37. Clark, P., Etzioni, O.: My Computer is an Honor Student but how Intelligent is it? Standardized Tests as a Measure of AI. AI Magazine **37**(1) (2016) 5–12