Validating the Creature Believability Scale for Videogames

Nuno Barreto(✉), Rui Craveirinha, and Licinio Roque

CISUC, Department of Informatics Engineering,
University of Coimbra, 3004 516 Coimbra, Portugal
nbarreto@dei.uc.pt

Abstract. We present the validation of a scale to assess creature believability in videogames. We define Creatures as all zoomorphic entities not qualifying as fundamentally human-like, whether possessing or not anthropomorphic features. The scale, derived from a previous research, contains 26 items in 4 dimensions – Biological/Social Plausability, Relationship with the Environment, Adaptation, and Expression. The results of a Confirmatory Factor Analysis, using 19 subjects, originated a model with 4 factors, a CFI of 0.795, and RMSEA of 0.111. While not a good fit, it is very close to a mediocre fit, which is a potentially promising result. Further validation is needed with more subjects in the future.

Keywords: Believability scale · Creature design · Evaluation
Videogames

1 Introduction

Studies have shown believable actors are perceived to be more engaging than non-believable ones [4,11,21,24,25]. However, believability, as a construct, is still in its infancy [24]. Works by Loyall [15], Bates [4] and Mateas [17], transcribe their believability criteria from Disney's rules of thumb listed in"Illusion of Life: Disney Animation" [23], and are, in turn, used as basis for several agent architectures and algorithms (e.g. Warpefelt [25], Parenthöen et al. [18]). These criteria are used at face value, without any validation to assess how they measure believability, or if they do so at all.

Assessing believability has been discussed in Togelius et al. [24], arguing for subjective methods, such as self-reports used by Arrabales et al. [2]. This work adjusted a scale used to evaluate cognitive functions in avatars, and allowed assessing a form of believability. However, its range over other types of believability is reduced, since agents are judged as avatars for players, not simulated living beings.

This paper's contribution is the revision of the believability scale presented by Barreto, Craveirinha and Roque [3], providing a first attempt at validation through Confirmatory Factor Analysis.

E. Clua et al. (Eds.): ICEC 2018, LNCS 11112, pp. 198–204, 2018.
https://doi.org/10.1007/978-3-319-99426-0_17

We can foresee two applications for such a scale, at the moment: to aid the analysis and comparison of fauna, in game worlds, across presentation, autonomy and interactivity (even between different games), and as a set of heuristics to use during game development. This would allow game designers the means to evaluate how believable their creatures are, and consequently, the potential to improve their game's believability and engagement.

The paper's structure is as follows: Sect. 2 describes methodology. Section 3 presents Results and Scale Revision, where we perform a revision of the scale's underlying theoretical model within its Confirmatory Factor Analysis. Section 4 details conclusions.

2 Methodology

Aiming to provide an initial framework for the study of believability in creatures, a first believability scale was proposed in Barreto, Craveirinha and Roque [3]. After establishing a formal definition for creatures, an initial iteration was created with 46 statements. After a first round of Exploratory Factor Analysis, 26 statements achieved an acceptable loading value and thus made it into the existing model, grouped into 4 dimensions. Note however, there was no confirmation of the hypothesized model. As suggested in Spector [22], we now provide an additional validation and revision phase of the scale, through Confirmatory Factor Analysis.

As suggested in Togelius et al. [24], we designed an online self-report[1] where ten randomly sorted video clips (40 to 60 s long), from various videogame sources, were shown, each with at least one creature engaged in a specific activity. This duration allowed numerous, and time consuming (i.e. learning) creature activities to be demonstrated. After each clip, subjects answered a fail-safe question [13, 16], and scored each item in the scale.

We chose games that were recent, to reduce bias created by technological limitations, and that had creatures with extensive on-screen presence, to increase perceivable activity. The following creature/game pairs were selected: Radstags (Fallout 4 [6]), Wolves and Tigers (Life of Black Tiger [1]), Chop (Grand Theft Auto V [20]), D-Horse (Metal Gear Solid V: The Phantom Pain [14]), Aliens (No Man's Sky [10]), Sea Vulture (Risen [19]), Wolves (The Elder Scrolls V: Skyrim [5]), Deer (theHunter [8]), Trico (The Last Guardian [9]), Antilopes (The Witcher III: The Wild Hunt [7]).

Our survey was deployed online, with 19 users participating (12 Male and 7 Female). The sample is admittedly short, but we think that for the purpose of a first validation attempt, it can provide meaningful insight regarding the model. The average age was 31.5 ± 11.9. Distribution of level of education was: 11% with Highschool degree, 37% with Bachelor's degree, 47% with Master's degree, 5% with Doctorate degree. Academic background included Science, Technology, Engineering and Math (68%), Humanities (Literature, Social Sciences, etc.)

[1] https://goo.gl/forms/VUbhu4lrvavk3W2X2.

(26%) and Arts (Illustration, Music, etc.) (5%). Average contact with media (videogames, movies, tv) per week was 20 h for 32% of users, 20 h to 40 h for 42%, 40 h to 60 h for 16%, and >60 h for 11%.

3 Results and Scale Revision

Each <subject, videogame> tuple was considered a separate answer, resulting in 190 entries (19 subjects × 10 clips), since each clip had its own context, creature(s) and activities. The data underwent a Confirmatory Factor Analysis (CFA) using a Maximum-Likelihood Path Analysis to test the scale's theoretical model's goodness-of-fit. Model data was compared using χ^2, $\frac{\chi^2}{df}$, CFI and RMSEA indexes.

A first structured equation model was considered (model 4F), transcribing the original scale [3]: its four dimensions were converted to latent variables, each with a covariance link to the remainder ones. These dimensions were also linked to their respective items, measured variables. The indexes obtained are listed in Table 1.

Observing the values, we first conclude the null model, as expected, provides a bad fit. Specifically, $\frac{\chi^2}{df}$ is above 5 (9.768), CFI is below 0.8 (0.000) and RMSEA is above 0.1 (0.215); all values lie within a bad fit on their respective ranges. Similarly, the original model (4F), also seems a bad fit. While its $\frac{\chi^2}{df}$ value was 4.594 (below 5 yet above 2), the others, CFI (0.630) and RMSEA (0.138) were above 0.6 and 0.1 respectively, suggesting a bad fit.

Table 1. Goodness-of-fit indexes for alternative models for the scale. (n = 190)

	Step	$\chi^2(df)$	$\frac{\chi^2}{df}$	CFI	RMSEA
	Null model	3174.497 (325)	9.768	0.000	0.215
1F	1 Common-Factor	1082.293 (291)	3.719	0.722	0.120
4F	4 Factors (w/o adj.)	1345.917 (293)	4.594	0.630	0.138
4FW	4 Factors (w/ adj.)	852.059 (264)	3.125	0.795	0.111

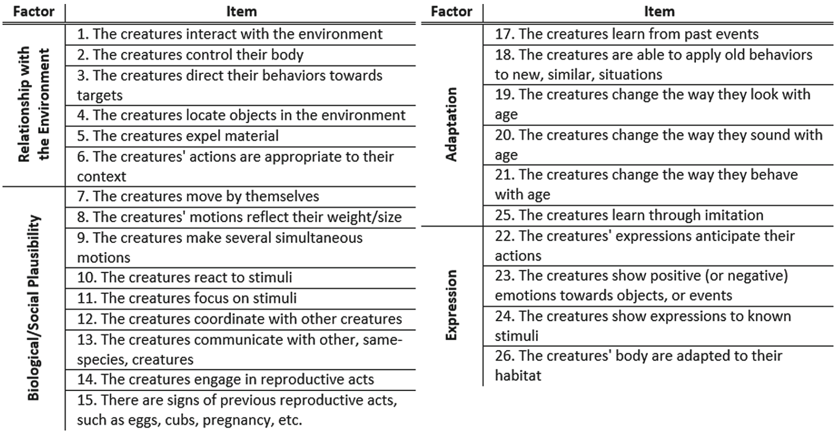
Since the original model seemed a bad fit for the data, we devised alternatives. One alternative involved a model (1F) which, unlike the previous, hypothesizes only 1-Common Factor, or dimension (Believability), to explain variances. Surprisingly, results are slightly better than 4F. Although $\frac{\chi^2}{df}$ (3.719) is also below 5 but above 2, its CFI (0.722), while still below, is closer to 0.8, the threshold for a mediocre fit (according to Hooper, Coughlan and Mullen [12]). Its RMSEA (0.120), while near 0.1, is still above it and a bad fit.

We also used hints from the estimates' modification indexes, formulating model 4FW as follows:

- We added covariances between residuals.
- Item 25 (“The creatures learn through imitation”) was changed to reflect Adaption.
- While several items had a loading value inferior to 0.4, the conventional cut threshold, we opted to remove only item 16 (“The creatures’ same-stimuli reactions change over time”), because of the item’s sentence confusing wording.
- Data outliers were removed according to their Mahalanobis Distance p1 and p2 values (when both equaled zero).
- Removed the covariances between the latent variable pairs <“Relation with the Environment”,“Adaptation”> and <“Adaptation”,“Biological/Social Plausibility”> as their estimated value was nearly zero.

Comparing fit indexes with the alternative models, it is clear 4FW yields better results. Its $\frac{\chi^2}{df}$ is the highest at 3.2, within the mediocre fit range. The CFI (0.795) is close to 0.8 and it can be argued this borderlines a mediocre fit. The RMSEA (0.111) is around 0.1. However, while this model fares better than previous ones, all of them are bad fits or borderline mediocre fits. Therefore, additional experiments need to be conducted to expand these result; with this in mind, we propose model 4FW for a revision proposal, as it had the best results. The original model (4F) and the one we considered (4FW) are illustrated in Fig. 1, and Table 2 lists the revised scale.

Table 2. The revised believability scale obtained after the CFA.



Factor	Item
Relationship with the Environment	1. The creatures interact with the environment
	2. The creatures control their body
	3. The creatures direct their behaviors towards targets
	4. The creatures locate objects in the environment
	5. The creatures expel material
	6. The creatures' actions are appropriate to their context
Biological/Social Plausibility	7. The creatures move by themselves
	8. The creatures' motions reflect their weight/size
	9. The creatures make several simultaneous motions
	10. The creatures react to stimuli
	11. The creatures focus on stimuli
	12. The creatures coordinate with other creatures
	13. The creatures communicate with other, same-species, creatures
	14. The creatures engage in reproductive acts
	15. There are signs of previous reproductive acts, such as eggs, cubs, pregnancy, etc.
Adaptation	17. The creatures learn from past events
	18. The creatures are able to apply old behaviors to new, similar, situations
	19. The creatures change the way they look with age
	20. The creatures change the way they sound with age
	21. The creatures change the way they behave with age
	25. The creatures learn through imitation
Expression	22. The creatures' expressions anticipate their actions
	23. The creatures show positive (or negative) emotions towards objects, or events
	24. The creatures show expressions to known stimuli
	26. The creatures' body are adapted to their habitat

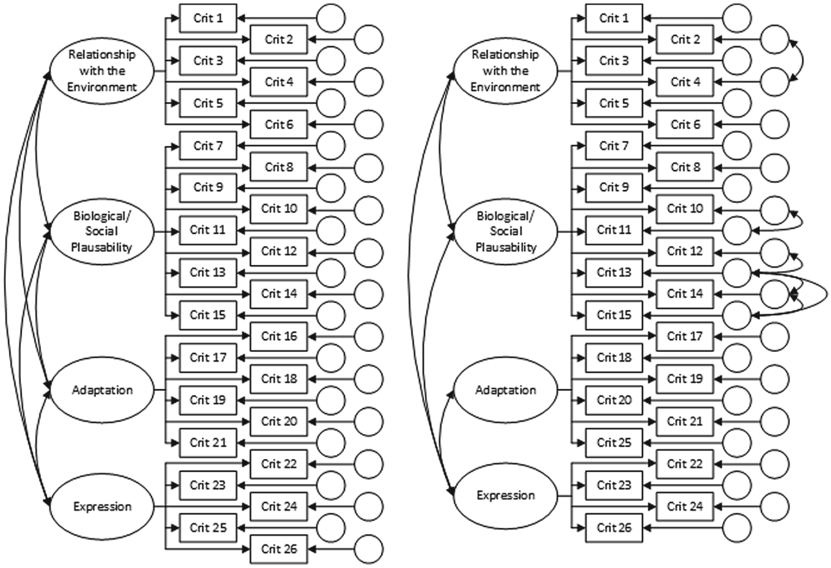


Fig. 1. Structured equation models of model 4F (left) and model 4FW (right). Ovals are the scale's construct's dimensions, rectangles are items and circles are residuals; single-headed arrows explain variance while the remaining show covariances.

4 Conclusion

We attempted to validate a Creature Believability Scale. After modeling the scale as a Structural Equation, we performed a Confirmatory Factor Analysis, leading to adjustments using three different approaches. From those, we chose one with an improved fit, taking into account the scale's future utility. With the selected approach, we revised the scale. After revision, 1 item was removed from the original 26-item scale, while one other switched dimensions.

The items' statements could also be revised by interviewing several subjects and assessing if they former are similarly interpreted across the latter. Revising the test setup altogether is also an option; for instance, using actual gameplay in lieu of video clips. Backtracking towards the scale's initial steps would allow reanalyzing the scale through an Exploratory Factor Analysis, though using a different, and larger, sample.

This research tries to provide a complementary insight into creature design. The existing literature on believability is either focused on humans or narratives, either centering on behaviors or expressions. This Creature Believability Scale provides a tool to assess believability of non-humanoid creatures, but also unifies several perspectives on how to convey believability.

The results show the scale, in its current form, needs further research. The model was deemed a borderline mediocre fit, so additional studies need to be conducted to understand why it is so. In the future, we will conduct a new

Confirmatory Factor Analysis phase with additional subjects (est. 300 to 500 required), to provide further validation.

We believe the revised scale is a step towards guidelines that improve design practices, allowing designers to evaluate their work's believability. While not ideal, we believe it is promising, and a first step towards the goal set for the research.

Acknowledgements. This work was supported by the FCT Ph.D. Grant SFRH/BD/100080/2014.

References

1. 1Games: Life of black tiger (2017)
2. Arrabales, R., Ledezma, A., Sanchis, A.: ConsScale FPS: cognitive integration for improved believability in computer Game Bots. In: Hingston, P. (ed.) Believable Bots: Can Computers Play Like People?, pp. 193–214. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-32323-2_8
3. Barreto, N., Craveirinha, R., Roque, L.: Designing a creature believability scale for videogames. In: Munekata, N., Kunita, I., Hoshino, J. (eds.) ICEC 2017. LNCS, vol. 10507, pp. 257–269. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66715-7_28
4. Bates, J.: The role of emotion in believable agents. Commun. ACM **37**, 122–125 (1994)
5. Bethesda Game Studios: The Elder Scrolls V: Skyrim. [DVD-ROM] (2011)
6. Bethesda Game Studios: Fallout 4. [DVD-ROM] (2015)
7. CD Projekt RED: The Witcher 3: Wild Hunt. [DVD-ROM] (2015)
8. Games, E.: theHunter (2005)
9. genDESIGN and SIE Japan Studio: The Last Guardian. [Blu-Ray] (2016)
10. Hello Games: No man's sky. [Blu-Ray] (2016)
11. Hingston, P.: Believable Bots: Can Computers Play Like People?. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-32323-2
12. Hooper, D., Coughlan, J., Mullen, M.R.: Structural equation modeling: guidelines for determining model fit. Electron. J. Bus. Res. Methods **6**(1), 53–60 (2007)
13. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2008, pp. 453–456. ACM, New York (2008)
14. Kojima Productions: Metal Gear Solid V: The Phantom Pain. [Blu-Ray] (2015)
15. Loyall, A.B.: Believable agents: building interactive personalities. Ph.D. thesis, Carnegie Mellon University Pittsburgh (1997)
16. Mason, W., Suri, S.: Conducting behavioral research on Amazon's Mechanical Turk. Behav. Res. Methods **44**(1), 1–23 (2012)
17. Mateas, M.: An oz-centric review of interactive drama and believable agents. Technical report, School of Computer Science Carnegie Mellon University, July 1997
18. Parenthöen, M., Tisseau, J., Morineau, T.: Believable decision for virtual actors. In: 2002 IEEE International Conference on Systems, Man and Cybernetics (2002)
19. Piranha Bytes: Risen. [DVD-ROM] (2009)
20. Rockstar North: Grand Theft Auto V. [Blu-Ray] (2013)

21. Rosenkind, M., Winstanley, G., Blake, A.: Adapting bottom-up, emergent behaviour for character-based AI in games. In: Bramer, M., Petridis, M. (eds.) Research and Development in Intelligent Systems XXIX, pp. 333–346. Springer, Heidelberg (2012). https://doi.org/10.1007/978-1-4471-4739-8_26
22. Spector, P.E.: Summated Rating Scale Construction: An Introduction. Sage, Thousand Oaks (1992)
23. Thomas, F., Johnston, O.: The Illusion of Life: Disney Animation. Hyperion, Santa Clara (1997)
24. Togelius, J., Yannakakis, G.N., Karakovskiy, S., Shaker, N.: Assessing believability. In: Hingston, P. (ed.) Believable Bots: Can Computers Play Like People?, pp. 215–230. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-32323-2_9
25. Warpefelt, H.: The non-player character - exploring the believability of NPC presentation and behavior. Ph.D. thesis, Stockholm University, Faculty of Social Sciences, Department of Computer and Systems Sciences, Stockholm, Sweden (2016)