Preliminary Results of a Systematic Review: Quality Assessment of Conversational Agents (Chatbots) for People with Disabilities or Special Needs

Maria Laura de Filippis, Stefano Federici, Maria Laura Mele, Simone Borsci, Marco Bracalenti, Giancarlo Gaudino, Antonello Cocco, Massimo Amendola, and Emilio Simonetti

1 Department of Philosophy, Social and Human Sciences and Education, University of Perugia, Perugia, Italy
2 Department of Cognitive Psychology and Ergonomics, Faculty of BMS, University of Twente, Enschede, The Netherlands
3 Department of Surgery and Cancer, Faculty of Medicine, NIHR London IVD, Imperial College, London, UK
4 Design Research Group, School of Creative Arts, Hertfordshire University, Hatfield, UK
5 DGTCSI-ISCTI, Directorate General for Management and Information and Communications Technology, Superior Institute of Communication and Information Technologies, Ministry of Economic Development, Rome, Italy
6 Department of Public Service, Prime Minister’s Office, Rome, Italy

Abstract. People with disabilities or special needs can benefit from AI-based conversational agents, which are used in competence training and well-being management. Assessment of the quality of interactions with these chatbots is key to being able to reduce dissatisfaction with them and to understand their potential long-term benefits. This will in turn help to increase adherence to their use, thereby improving the quality of life of the large population of end-users that they are able to serve. We systematically reviewed the literature on methods of assessing the perceived quality of interactions with chatbots, and identified only 15 of 192 papers on this topic that included people with disabilities or special needs in their assessments. The results also highlighted the lack of a shared theoretical framework for assessing the perceived quality of interactions with chatbots. Systematic procedures based on reliable and valid methodologies continue to be needed in this field. The current lack of reliable tools and systematic methods for assessing chatbots for people with disabilities and special needs is concerning, and may lead to unreliable systems entering the market with disruptive consequences for users. Three major conclusions can be drawn from this systematic analysis: (i) researchers should adopt consolidated and comparable methodologies to rule out risks in use; (ii) the constructs of satisfaction and acceptability are different, and should be measured separately;
(iii) dedicated tools and methods for assessing the quality of interaction with chatbots should be developed and used to enable the generation of comparable evidence.

**Keywords:** Chatbots · Conversational agents · People with disability · People with special needs · Usability · Quality of interaction

1 Introduction

Chatbots are intelligent conversational software agents which can interact with people using natural language text-based dialogue [1]. They are extensively used to support interpersonal services, decision making, and training in various domains [2–5]. There is a broad consensus on the effectiveness of these AI agents, particularly in the field of health, where they can promote recovery, adherence to treatment, and training [6, 7] for the preparation of different competencies and the maintenance of well-being [3, 8, 9]. In view of this, an evaluation of the perceived quality of engagement with chatbots is key to being able to reduce dissatisfaction, facilitate their possible long-term benefits, increase loyalty and thus improve the quality of life of the large population of end-users that they are able to serve. Chatbots are interaction systems, and irrespective of their domain of application, their output in terms of the quality of interaction should be planned and measured in conjunction with their users, rather than by applying a system-centric approach [1]. A recent review by Abd-Alrazaq and colleagues [6] found that in the field of mental health, researchers typically only test chatbots in a randomized controlled trial. The efficiency of interaction is seldom assessed, and is generally done by looking at non-standardized aspects of interaction and qualitative measurements that do not require comparisons to be made. This unreliable method of testing the quality of interaction of these devices or applications through a wide and varied range of variables is endemic in all fields that use chatbots, and makes it difficult to compare the results of these studies [1, 10, 11]. While some qualitative guidelines and tools have emerged [1, 12], it is still hard to find agreement on which factors should be tested. As argued by Park and Humphry [13], the implementation of these innovative systems should be based on a common framework for assessing the perceived interaction quality, in order to prevent chatbots from being regarded by their end-users as merely another source of social alienation, and being discarded in the same way as any other unreliable assistive technology [14, 15]. A common framework and guidelines on how to determine the perceived quality of chatbot interaction are therefore required. From a systems perspective, a subjective experience of consistency arises from the interaction between the user and the program in specific conditions and contexts. Subjective experience cannot be measured merely by believing that the optimal performance of the system as perceived by the user is the same as a good user experience [16]. The need to quantify the objective and subjective dimensions of experience in a reliable and comparable manner is a lesson that has been learned by those in the field of human-computer interaction, but has yet to be learned in the field of chatbots, as outlined by Lewis [17] and Bendig and colleagues [18]. Chatbot developers are forced to rely on the umbrella framework provided by the International
Organization for Standards (ISO) 9241-11 [19] for assessing usability, and ISO 9241-210 [20] for assessing user experience (UX), due to the absence of a common assessment framework that specifies comparable evaluation criteria. These two ISO standards define the key factors of interaction quality: (i) effectiveness, efficiency and satisfaction in a specific context of use (ISO 9241-11); and (ii) control (where possible) of expectations over time concerning use, satisfaction, perceived level of acceptability, trust, usefulness and all those factors that ultimately push users to adopt and keep using a tool (ISO 9241-210). Although these standards have not yet been updated to meet the specific needs of chatbots and conversational agents, the two aspects of usability and UX are essential to the perceived quality of interaction [21]. Until a framework has been developed and broad consensus reached on assessment criteria, practitioners may benefit from the assessment of chatbots against these ISO standards, as they allow for an evaluation of the interactive output of these applications. This paper examines how aspects of perceived interaction quality are assessed in studies of AI-based agents that support people with disabilities or special needs. Our systematic literature review was conducted in accordance with the PRISMA reporting checklist.

2 Methods

A systematic review was carried out of journal articles investigating the relationship between chatbots and people with disabilities or special needs over the last 10 years. To determine whether and how the quality of interaction with chatbots was evaluated in line with ISO standards of usability (ISO 9241-11) and UX (ISO 9241-210), this review sought to answer the following research questions:

R1. How are the key factors of usability (efficiency, effectiveness, and satisfaction) measured and reported in evaluations of chatbots for people with disabilities or special needs?

R2. How are factors relating to UX measured and reported in assessments of chatbots?

We included in our review studies that: (i) referred to chatbots or conversational interfaces/agents for people with disabilities or special needs in the title, abstract, keywords or main text; (ii) included empirical findings and discussions of theories (or frameworks) of factors that could contribute to the perceived quality of interaction with chatbots, with a focus on people with various types of disability.

We excluded records that did not include at least one group of end-users with a disability in either the testing or the design of the interaction, and studies that focused on: (i) testing emotion recognition during the interaction exchange, or assessing applications for detecting the development of disability conditions or disease; (ii) chatbots supporting people with alcoholism, anxiety, depression or traumatic disorders; (iii) the assessment of end-user compliance with clinical treatment, or assessment of the clinical effectiveness of using AI agents as an alternative to standard (or other) forms of care without considering the interaction exchange with the chatbot; and (iv) the ethical and legal implications of interacting with AI-based digital tools.

Records were retrieved from Scopus and the Web of Science using the Boolean
operators (AND/OR) to combine the following keywords: chatbot*, conversational agent*, special needs, disability*. We searched only for English language articles.

3 Results

A total of 147 items were retrieved from Scopus and Web of Science. A further 53 records were added based on a previous review of chatbots in mental health in [6]. After removing eight duplicates, a scan of the remaining 192 records by title and abstract was performed by two authors (MLDF, SB). Articles that defined their scope as including the assessment of interactions between chatbots and conversational agents and people with various types of intellectual disabilities or special needs were retained. The full text of 68 records was then scanned to look for articles mentioning methods and factors for assessing the interactions of people with disabilities or special needs with chatbots. The final list consisted of 15 documents [3, 8, 9, 22–33], 80% of which had already been discussed in previous work by Abd-Alrazaq et al. [6] for different purposes.

Of the 15 records that matched our criteria, 80% examined AI agents in terms of supporting people with autism and (mild to severe) mental disabilities, while the other 20% focused on the testing of applications to support the general health or training of people with a wide range of disabilities. The main goal of 66.6% of the applications was to support health and rehabilitation, while the remaining studies focused on solutions to support learning and training for people with disabilities. In terms of their approach to assessment, 46.7% of the studies used surveys or questionnaires, 26.7% applied a quasi-experimental procedure, and the remaining 26.7% tested chatbots using randomized controlled trials (i.e., testing the use of the agent versus standard practice with a between design) that assessed several aspects relating to the quality of the interaction. Factors relating to usability (i.e., effectiveness, efficiency, and satisfaction) were partly assessed, with 80% of the studies reporting measures of effectiveness, 26.7% measures of efficiency and 20% measures of satisfaction. In terms of UX, acceptability was the most frequently reported measure (26.7% of the cases) while a few other factors (e.g., engagement, safety, helpfulness, etc.) were measured using various approaches.

4 Discussion

The results suggest that the main focus of studies of chatbots for people with disabilities or special needs is the effectiveness of such apps compared with standard practice, in terms of supporting adherence to treatment. The results can be summarized in accordance with our research questions as follows:

R1. A total of 80% of the studies [3, 8, 9, 23, 25, 27, 30–33] tested the effectiveness of chatbots according to the ISO standard [19], i.e., the ability of the app to perform correctly, allowing the users to achieve their goals. Only 26.7% of the studies [9, 25, 26, 32] also investigated efficiency, by measuring performance in terms of
time or factors relating to the resources invested by participants to achieve their goals. Only 20% [9, 22, 23] referred to an intention to gather data on user satisfaction in a structured way, and only one study [23] used a validated scale (e.g., the System Usability Scale, or user metrics of UX [34]). In another, practitioners adapted a standardized questionnaire without clarifying the changes to the items [22], and a qualitative scale was used in a further study [9].

R2. Acceptability was identified as an assessment factor in 26.7% of the studies [9, 22, 24, 25]. Despite the popularity of the technology acceptance model [35, 36], acceptability was measured in a variety of ways (e.g., lack of complaints [25]) or treated as a measure of satisfaction [24]. A total of 53% of the studies used various factors to assess the quality of interaction, such as the overall experience, safety, acceptability, engagement, intention to use, ease of use, helpfulness, enjoyment, and appearance. Most used non-standardized questionnaires to assess the quality of interaction. Even when a factor such as safety was identified as a reasonable form of quality control, in compliance with ISO standards for medical devices [37] and risk analysis [38], the method of its measurement in these studies was questionable, i.e., assessing a product to be safe based on a lack of adverse events [9].

5 Conclusion

The results of the present study suggest that informal and untested measures of quality are often employed when it comes to evaluating user interactions with AI agents. This is particularly relevant in the domain of health and well-being, where researchers set out to measure the clinical validity of tools intended to support people with disabilities or special needs. The risk is that shortcomings in these methods could significantly compromise the quality of chatbot usage, ultimately leading to the abandonment of applications that could otherwise have a positive impact on their end-users. Three major findings can be identified from this systematic analysis. (i) Researchers tend to consider a lack of complaints as an indirect measure of the safety and acceptability of tools. However, safety and acceptability should be assessed with consolidated and comparable methodologies to rule out risks in use [37–39]. (ii) Satisfaction, intended as a usability metric, is a different construct from acceptability, and these two constructs should be measured separately with available standardized questionnaires [39, 40]. (iii) Although dedicated tools and methods for assessing the quality of interaction with chatbots are lacking, reliable methods and measures to assess interaction are available [17, 19, 21, 37], and these should be adopted and used to enable the generation of comparable evidence regarding the quality of conversational agents.
References

1. Radziwill, N.M., Benton, M.C.: Evaluating quality of chatbots and intelligent conversational agents. arXiv preprint arXiv:1704.04579 (2017)
2. Ammari, T., Kaye, J., Tsai, J.Y., Bentley, F.: Music, search, and IoT: how people (really) use voice assistants. ACM Trans. Comput.-Hum. Interact. 26 (2019). https://doi.org/10.1145/3311956
3. Beaudry, J., Consigli, A., Clark, C., Robinson, K.J.: Getting ready for adult healthcare: designing a chatbot to coach adolescents with special health needs through the transitions of care. J. Pediatr. Nurs. 49, 85–91 (2019). https://doi.org/10.1016/j.pedin.2019.09.004
4. Costa, S., Brunete, A., Bae, B.C., Mavridis, N.: Emotional storytelling using virtual and robotic agents. Int. J. Hum. Robot. 15 (2018). https://doi.org/10.1142/S0219843618500068
5. Dmello, S., Graesser, A.: AutoTutor and affective AutoTutor: learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Trans. Interact. Intell. Syst. 2 (2012). https://doi.org/10.1145/2395123.2395128
6. Abd-Alrazaq, A.A., Alajlani, M., Alalwan, A.A., Bewick, B.M., Gardner, P., Househ, M.: An overview of the features of chatbots in mental health: a scoping review. Int. J. Med. Inf. 132 (2019). https://doi.org/10.1016/j.ijmedinf.2019.103978
7. Fadhil, A., Wang, Y., Reiterer, H.: Assistive conversational agent for health coaching: a validation study. Methods Inf. Med. 58, 009–023 (2019)
8. Burke, S.L., et al.: Using virtual interactive training agents (ViTA) with adults with autism and other developmental disabilities. J. Autism Dev. Disord. 48(3), 905–912 (2017). https://doi.org/10.1007/s10803-017-3374-z
9. Ellis, T., Latham, N.K., DeAngelis, T.R., Thomas, C.A., Saint-Hilaire, M., Bickmore, T.W.: Feasibility of a virtual exercise coach to promote walking in community-dwelling persons with Parkinson disease. Am. J. Phys. Med. Rehabil. 92, 472–485 (2013). https://doi.org/10.1097/PHM.0b013e31828cd466
10. Balaji, D., Borsci, S.: Assessing user satisfaction with information chatbots: a preliminary investigation. University of Twente, University of Twente repository (2019)
11. Tariverdiyeva, G., Borsci, S.: Chatbots’ perceived usability in information retrieval tasks: an exploratory analysis. University of Twente, University of Twente repository (2019)
12. IBM http://conversational-ux.mybluemix.net/design/conversational-ux/practices/
13. Park, S., Humphry, J.: Exclusion by design: intersections of social, digital and data exclusion. Inf. Commun. Soc. 22, 934–953 (2019)
14. Federici, S., Borsci, S.: Providing assistive technology in Italy: the perceived delivery process quality as affecting abandonment. Disabil. Rehabil. Assist. Technol. 11, 22–31 (2016). https://doi.org/10.3109/17483107.2014.930191
15. Scherer, M.J., Federici, S.: Why people use and don’t use technologies: introduction to the special issue on assistive technologies for cognition/cognitive support technologies. NeuroRehabilitation 37, 315–319 (2015). https://doi.org/10.3233/NRE-151264
16. Bevan, N.: Measuring usability as quality of use. Softw. Qual. J. 4, 115–130 (1995). https://doi.org/10.1007/BF00402715
17. Lewis, J.R.: Usability: lessons learned... and yet to be learned. Int. J. Hum.-Comput. Interact. 30, 663–684 (2014). https://doi.org/10.1080/10447318.2014.930311
18. Bendig, E., Erb, B., Schulze-Thuesing, L., Baumeister, H.: The next generation: chatbots in clinical psychology and psychotherapy to foster mental health – a scoping review. Verhaltenstherapie (2019). https://doi.org/10.1159/000501812
19. ISO: ISO 9241-11:2018 Ergonomic Requirements for Office Work with Visual Display Terminals – Part 11: Guidance on Usability. CEN, Brussels (2018)
20. ISO: ISO 9241-210:2010 Ergonomics of Human-System Interaction – Part 210: Human-Centred Design for Interactive Systems. CEN, Brussels (2010)

21. Borsci, S., Federici, S., Malizia, A., De Filippis, M.L.: Shaking the usability tree: why usability is not a dead end, and a constructive way forward. Behav. Inform. Technol. 38, 519–532 (2019). https://doi.org/10.1080/0144929x.2018.1541255

22. Ali, M.R., et al.: A virtual conversational agent for teens with autism: experimental results and design lessons. arXiv preprint arXiv:1811.03046 (2018)

23. Cameron, G., et al.: Assessing the usability of a chatbot for mental health care. In: Bodrunova, S.S., et al. (eds.) INSCI 2018. LNCS, vol. 11551, pp. 121–132. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-17705-8_11

24. Konstantinidis, E.I., Hitoglou-Antoniadou, M., Luneski, A., Bamidis, P.D., Nikolaidou, M.M.: Using affective avatars and rich multimedia content for education of children with autism. In: 2nd International Conference on PERvasive Technologies Related to Assistive Environments (PETRA 2009), pp. 1–6 (2009). https://doi.org/10.1145/1579114.1579172

25. Lahiri, U., Bekele, E., Doehrmann, E., Warren, Z., Sarkar, N.: Design of a virtual reality based adaptive response technology for children with autism. IEEE Trans. Neural Syst. Rehabil. Eng. 21, 55–64 (2013). https://doi.org/10.1109/TNSRE.2012.2218618

26. Ly, K.H., Ly, A.-M., Andersson, G.: A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. Internet Interv. 10, 39–46 (2017). https://doi.org/10.1016/j.invent.2017.10.002

27. Milne, M., Luerssen, M.H., Lewis, T.W., Leibbrandt, R.E., Powers, D.M.W.: Development of a virtual agent based social tutor for children with autism spectrum disorders. In: International Joint Conference on Neural Networks (IJCNN 2010), pp. 1–9 (2010). https://doi.org/10.1109/IJCNN.2010.5596584

28. Razavi, S.Z., Ali, M.R., Smith, T.H., Schubert, L.K., Hoque, M.: The LISSA virtual human and ASD teens: an overview of initial experiments. In: Traum, D., Swartout, W., Khooshabeh, P., Kopp, S., Scherer, S., Leuski, A. (eds.) IVA 2016. LNCS (LNAI), vol. 10011, pp. 460–463. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47665-0_55

29. Smith, M.J., et al.: Job offers to individuals with severe mental illness after participation in virtual reality job interview training. Psychiatr. Serv. 66, 1173–1179 (2015). https://doi.org/10.1176/appi.ps.201400504

30. Smith, M.J., et al.: Virtual reality job interview training for individuals with psychiatric disabilities. J. Nerv. Mental Dis. 202, 659–667 (2014). https://doi.org/10.1097/NMD.0000000000000187

31. Tanaka, H., Negoro, H., Iwasaka, H., Nakamura, S.: Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. PLoS ONE 12, e0182151 (2017). https://doi.org/10.1371/journal.pone.0182151

32. Wargnier, P., Benveniste, S., Jouvelot, P., Rigaud, A.-S.: Usability assessment of interaction management support in Louise, an ECA-based user interface for elders with cognitive impairment. Technol. Disabil. 30, 105–126 (2018). https://doi.org/10.3233/TAD-180189

33. Smith, M.J., et al.: Virtual reality job interview training in adults with autism spectrum disorder. J. Autism Dev. Disord. 44(10), 2450–2463 (2014). https://doi.org/10.1007/s10803-014-2113-y

34. Borsci, S., Federici, S., Bacci, S., Gnaldi, M., Bartolucci, F.: Assessing user satisfaction in the era of user experience: comparison of the SUS, UMUX and UMUX-LITE as a function of product experience. Int. J. Hum.-Comput. Interact. 31, 484–495 (2015). https://doi.org/10.1080/10447318.2015.1064648

35. Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D.: User acceptance of information technology: toward a unified view. MIS Q.: Manag. Inf. Syst. 27, 425–478 (2003)
36. Federici, S., Tiberio, L., Scherer, M.J.: Ambient assistive technology for people with dementia: an answer to the epidemiologic transition. In: Combs, D. (ed.) New Research on Assistive Technologies: Uses and Limitations, pp. 1–30. Nova Publishers, New York (2014). https://doi.org/10.13140/2.1.3461.4405

37. IEC: IEC 62366-1:2015 Medical Devices – Part 1: Application of Usability Engineering to Medical Devices. CEN, Brussels (2015)

38. ISO: ISO 14971:2007 Medical Devices – Application of Risk Management to Medical Devices. CEN, Brussels (2007)

39. Borsci, S., Federici, S., Mele, M.L., Conti, M.: Short scales of satisfaction assessment: a proxy to involve disabled users in the usability testing of websites. In: Kurosu, M. (ed.) HCI 2015. LNCS, vol. 9171, pp. 35–42. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21006-3_4

40. Borsci, S., Buckle, P., Walne, S.: Is the lite version of the usability metric for user experience (UMUX-LITE) a reliable tool to support rapid assessment of new healthcare technology? Appl. Ergon. 84, 103007 (2020)