The Advent of Speech Based NLP QA Systems: A Refined Usability Testing Model

Diarmuid Lane, Robin Renwick(✉), John McAvoy, and Philip O'Reilly

University College Cork, Cork, Ireland
{113456738,117223923}@umail.ucc.ie,
{J.McAvoy,Philip.OReilly}@ucc.ie

Abstract. This paper outlines a refined usability testing model, developed for industry specific comparative usability testing of two natural language processor (NLP) based question answering (QA) systems. The systems operate over differing modalities, one through text and the other speech. The revised model combines two existing usability testing frameworks, the System Usability Metric (SUM) and the System Usability Scale (SUS). It also integrates the context-specific determined target user - financial managers working within the financial services industry. The model's metric weightings are determined through key informant interviews. The presented model will be the working framework from which a series of comparative usability tests will be carried out within the target organisation.

Keywords: Natural language processor · Question Answering System · Usability

1 Introduction

As we move further into the 21st century, interaction with automated natural language processor (NLP) based question answering (QA) systems will become commonplace (Levy 2016; Panetta 2017). The trend may be seen emerging through commercially available *zeitgeist* tools such as Amazon's Alexa, Google Assistant, Apple's Siri and IBM's Watson. QA system development may be seen through the seemingly unstoppable adoption of text based 'helper' systems proliferating the internet and mobile applications - colloquially known as 'chat-bots' (López et al. 2017). QA systems are a subfield of natural language processing research, focused on the location and retrieval of specific data points relating to user questions posed in natural language. NLP powered QA systems act as human-computer interfaces; delivering accurate and efficient responses to queries through the modes of text or speech (Lopez et al. 2011). QA systems are becoming increasingly commonplace in contemporary workplaces, carrying out a range of activities - from administrative workplace tasks through to complex human resource functions (Knight 2016). It is estimated that by 2020 over 50% of large enterprises will have internally deployed 'chat-bots' to carry out, or augment, business functions. It is envisioned this will lead to more fluid interactions with Information Systems (IS), whilst correspondingly reducing levels of human-error

A. Marcus and W. Wang (Eds.): HCII 2019, LNCS 11586, pp. 152–163, 2019.
https://doi.org/10.1007/978-3-030-23535-2_11

(Goasduff 2018). It was noted as early as 2001 that speech based interactions would become an accepted mode of communication between human and computer (Hone and Graham 2001), while recent research has indicated a significant upturn in the market penetration of bi-directional speech-based interactions (Moore et al. 2016; Moore 2017a, 2017b; Simonite 2016). However, the question remains which mode of communication is preferred, or more effective for businesses implementation – speech or text.

The design of QA interactions between human and computer will determine the mode and interface through which we engage. Core to the concept of QA interaction design is the measure of usability. Usability is a complex term, with myriad interpretations found in varied, and sometimes disparate, fields: design science; design engineering; information science; human computer interface (HCI) design; user interface (UI) design; and user experience (UX) design (Green and Pearson 2006). Research has highlighted the need for 'usable' NLP based QA systems (Hausawi and Mayron 2013; Ferrucci et al. 2009). However, there remains a paucity of research within the context of modality comparison. With this in mind, this paper proposes a usability testing model to assess NLP QA systems, incorporating specific nuances associated with both speech and text based interfaces. The model is developed by reviewing and extending literature from a number of disciplines, paying particular attention to IS publications, user experience and/or usability literature, and the field of HCI design. The model is refined in the development stage through interaction with key industry informants. This paper describes existing usability testing models, details how these are viewed as inadequate for the specific context of comparative usability testing between speech and text, and outlines the development of a new model within a context led scenario.

2 Usability Literature Review

A definition of usability is set forth by the International Organisation of Standards, through ISO 9241-11. The international standard details usability as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO 1998, p. 1). A number of studies base testing metrics on this definition, or encapsulate its importance in defining a testing model (Green and Pearson 2006; Hone and Graham 2001; Möller et al. 2008; Rohrer et al. 2016; Sauro and Kindlund 2005). Criticism has been leveled at the ISO 9241-11 definition from various angles, most notably from within the usability testing profession. Usability is the characteristic of a product that inherently makes it 'usable', but the term may mean different things to different people, and is dependent on contextual, behavioural, and/or situational aspects (Quesenbury 2003; Rohrer et al. 2016). The core criticism aimed at the ISO definition is that specific users will view concepts such as effectiveness and efficiency differently. The lack of flexibility in the ISO definition has previously been noted, as has its suitability as a working base from which to start (Quesenbury 2003). Incorporating project specifics along with the target user into the process of building a revised understanding ensures that any devised model can move away from generality towards specificity.

There are three core criticisms of the ISO based definition (Green and Pearson 2006; Quesenbury 2003). The first is that the model is too focused on well-defined tasks and goals; ignoring the more intangible elements of user experience. The second problem is that the definition foregrounds efficiency and effectiveness, without catering for products where these are of lesser concern; where pleasure and engagement are more relative to any measure of usability. The final criticism focuses on the use of the word 'satisfaction', viewed as too narrow a testing term (Quesenbury 2003). Similar sentiments have been discussed within HCI design and research, where "'usability' is a construct conceived by the community to denote a desired quality of interactive systems and products" (Tractinsky 2018, p. 131). The issues associated with the usability construct are thought to have stemmed from the fact that usability incorporates both subjective and objective dimensions, broadening the distance between the usability construct and its measures (Hone and Graham 2001; Sauro and Lewis 2009; Tractinsky 2018). The context specific nature of usability increases the ambiguity associated with the construct and its measurement (Brooke 2013). In practice this implies a disjunct between how usability is defined, and how it is actually tested.

A model for usability may be drawn in which the construct is understood by its perceived components in a reflective manner, as opposed to a formative model which first understands components and then attempts to define how they drive a specific frame (Tractinsky 2018). The distinction between the two methods is important, as concerns regarding how project specific usability and testing contexts should inform the development of any new usability testing model are incorporated. However, it should be remembered that existing frameworks offer overarching guidance, or reference, from which more exacting project specific models may be developed (Quesenbury 2003).

2.1 ISO 9241-11 Standard Based Frameworks

Green and Pearson (2006) outline an ISO 9241-11 standard based usability testing model, designed for the user-based testing of an e-commerce website. The focus is on the user, with any potential pitfalls encountered while using the system viewed as the measure of whether the product is 'usable' or not. Understanding usability in this context means that it extends "…beyond the issues of ease of use, ease of learning, and navigation…with additional pressure for the design to be intuitive" (Green and Pearson 2006, p. 66). Alternatively, a usability testing model for web-based information systems based on the ISO 9241-11 standard examines the correlation between web-based systems 'quality' and overall usability (Oztekin et al. 2009). Causation is created between the overall 'quality' of a system, or interface, and its measured level of usability.

The single, standardised and summated usability metric (SUM) is based on the ISO 9241-11, and the American National Standards Institute (ANSI) dimensions of usability: effectiveness, efficiency, and satisfaction (Sauro and Kindlund 2005). From these dimensions four usability metrics are derived - time on task, number of errors, task completion, and average satisfaction. The metrics have been selected as base standards for the testing of usability. The four measures show significant correlation; a weighted average of the standardised values convey the maximum amount of information in one combined score (Sauro and Kindlund 2005). One key feature of SUM is the equal weighing denoted to the three dimensions of usability (efficiency, effectiveness, and satisfaction); offering an unbiased overall 'systems usability' score. SUM has been successfully extended for a multitude of purposes: measuring the usability of mobile applications (Avouris et al. 2008), through to the testing of educational software (Lado et al. 2006).

The Practical Usability Rating by Experts (PURE) framework was developed in conjunction with industry representatives at Intel Security (Rohrer et al. 2016). PURE is a framework that includes ratings as given by teams of trained evaluators, testing validity against independent studies completed by specific usability experts. The comparisons offer guidance on the verifiability and accuracy of the testing procedure, as well as the governing framework. PURE derives an understanding of usability from the ISO 9241-11 standard, acknowledging interaction between three standard components of user experience while simultaneously stating its core focus as usability. By integrating knowledgeable evaluators drawn from the target firm, the PURE model fosters a co-opt design process, in which prospective end users engage in the development, testing, and refinement process. This ensures that the target user's understanding of 'usability' is integrated into the framework; a method suggested as being integral to a successful and accurate derivation of a testing model (Quesenbury 2003; Tractinsky 2018). The overall success of the framework is built on the specified outline of what the interface, or product, is going to be used for. Once a specification is detailed, it is possible to outline project specific metrics which will be used within the testing framework.

2.2 Usability Metrics

Employing effective and quantifiable usability metrics reveal important numerically based information pertaining to user experience (Tullis and Albert 2013). There are three distinct categories of measurement in the ISO standard of usability: effectiveness (the act of completing a given task); efficiency (the level of effort needed to compete a given task); and satisfaction (the extent to which a user enjoyed the experience of performing a given task). Within these three categories lies a host of sub-metrics which may be applied by usability practitioners in specific usability study scenarios:

- *Task success* metrics are concerned with whether participants are able to complete a given task associated with an interface, product, or system. When testing for task success, identifying and testing completion rates removes certain aspects of ambiguity.
- *Task time* is concerned with the length of time needed to complete a specific task. In order to test efficiency, a method is set forth by the National Institute of Standards and Technology (NIST): efficiency is based on a combination of the task success metric, and the task time metric (NIST 2001).
- *Learnability* refers to how easy it is for users to complete specific tasks associated with an interface, product, or system the first time they encounter the design (Nielsen 2003). Due to the nature of the design presented in this research, the concept of learnability allows comparisons to be made of two modes of interaction - speech and text.

- *Self-reported metrics* are critical to gaining an understanding of users' perceptions about an interface, product, or system. These metrics give the tester a generalised insight into how a user 'feels' about a system. They also allow a tester to evaluate any differences that may exist between how a user perceives a system, and how the system actually is (Bangor et al. 2008; Hone and Graham 2001; Silvervarg et al. 2016).

Objective usability metrics provide an accurate measure of a systems performance, quantified using usability performance indicators. Self-reported metrics are employed in order to quantify a user's subjective opinion of a system's usability. An adequate model must be employed in order to accurately measure this subjective opinion. A review of such models is necessary prior to integrating subjective metrics into a context specific usability framework.

System Usability Scale (SUS)
The System Usability Scale (SUS) was developed as a quick and effective usability survey scale, developed to assess users' subjective perception of usability with respect to a given system, or design (Brooke 1996). SUS has become a leading model in the usability industry, having been successfully extended to assess interface technologies, novel hardware platforms, and voice response systems (Bangor et al. 2008). The original SUS questionnaire was composed of 10 statements, scored on a Likert scale - measuring responses to answers ranging from 'strongly agree' to 'strongly disagree'. The questionnaire alternates between positive and negative questions concerning the usability of the system being tested. A percentile score is then derived based on the provided responses (Brooke 1996).

There have been some criticisms of SUS, with researchers noting that erroneous answers appear at various stages of the testing response procedure. Respondents have been found to mistakenly agree with negative questions, leading to unreliable questionnaire scores (Sauro and Lewis 2011). Due to the continued reporting of user error, a number of semantic changes have been made to the questionnaire since it first appeared; roughly 90% of tests since its formation have been found to use a revised version (Bangor et al. 2008). When SUS was originally developed it was intended to assess users perceived usability, yielding a single usability score (Brooke 1996). As SUS was designed prior to the proliferation of speech based systems we see today, a review of speech based usability testing is required before the development of a context specific usability testing framework.

2.3 Speech Based Usability Testing

Since the emergence of Siri in 2011, speech based interfaces have become increasingly commonplace in society (Moore 2017a, b). However, overall adoption metrics, usability testing frameworks, and/or specific metrics for evaluating mode of interaction have not been developed. A subjective assessment model for speech-based interfaces has appeared, with the emergence of the SASSI model (Hone and Graham 2001). Work has also been done predicting the quality and usability of spoken dialogue interfaces (Möller et al. 2008). However, it is still not understood if speech based interaction is genuinely demanded, preferred, or if the proposed mode of interaction is capable of providing the type and level of usability demanded from the end user. For mainstream adoption to occur two needs must be met: “the need to align the visual, vocal and behavioural affordances of the system, and second, the need to overcome the huge mismatch between the capabilities and expectations of a human being and the features and benefits offered by even the most advanced autonomous social agent” (Moore 2017a, p. 10).

Usability research of speech based interfaces designed specifically for the elderly has unearthed complex discourse surrounding themes such as privacy, dependency, loss of control, and the affordances of speech based interfaces to tackle psychological and societal issues such as loneliness (Portet et al. 2013). Themes such as these are beyond the bounds of this study, but are worth considering when detailing how speech based interfaces may impact procedures and processes at any level of human computer interaction. Specific issues have also been raised with respect to the interactional abilities of speech interfaces. Differences in use of vocabulary and grammatical structure have been noted as being key determinants in scores for usability, with speech based interfaces responding differently to accents, regional dialects, and sentence syntax - sometimes to the frustration of users or detriment of the system (Portet et al. 2013; Vacher et al. 2015).

Researchers have investigated the functionality and usability of competing proprietary speech based user interface systems (López et al. 2017). Four of the most widely available speech based interfaces were included in the study. The systems were evaluated with respect to a defined task list, offering results against specific operational modes: shopping and buying assistant; care assistant; travel and entertainment assistant; and administrative assistant (López et al. 2017). Similar to the aforementioned SUS model, a 5-point Likert scale was used to assess the ‘naturality’ of interactions, while ‘correctness’ was measured with respect to task success. The research concluded that improvements could be made by all interface systems and that further study is needed to understand effectiveness and usability of the available proprietary systems (López et al. 2017; Torres et al. 2017). With this in mind, the research presented in this paper acknowledges the exacting requirements for testing two natural language processor based QA systems operating under two alternative modalities. A refined model is required that encompasses existing models, as discussed, as well as the project specific context that the research addresses.

3 Initial Usability Testing Model

The model presented in this paper is designed to evaluate the usability of two distinct forms of question answering systems, one interfaced through speech, another through text. Similar to the previously outlined models, the foundation definition of usability is that set forth by ISO 9241-11. The paper incorporates the three main dimensions of usability (efficiency, effectiveness, and satisfaction); chosen as base constructs for the usability testing framework. The three dimensions have initially been allocated equal weighting, provisioning for a summated Question Answering System Usability Score. This scoring system is viewed as similar to that of Sauro and Kindlund (2005). The model has been designed to test both objective and subjective indicators of usability, assessing performance and perceived usability of the two alternate QA systems. The objective measures are determined by testing carried out similar to ISO based frameworks, and SUM. Scores are determined through specific usability indicators. The subjective score is determined by metrics as put forward by SUS. The research proposes that alternative score weightings are required for increased accuracy and project specific context. These are determined through interviews with key informants from within the financial services industry.

Internal usability metrics have been selected to measure and quantify dimensions of usability (see Fig. 1). ‘Time-on-Task’ has been chosen to measure the effort required by a user to complete a given task, referred to as ‘Task Proficiency’. It determines the usability dimension of ‘Efficiency’. ‘Number of Errors’, and ‘Completion’ were chosen to measure the dimension of ‘Effectiveness’. The above metrics are based on system performance, and are classed as objective measures (Tullis and Albert 2013). The dimension of ‘Satisfaction’ is associated with the users’ perspective while using a system. An adapted System Usability Scale (SUS) will be used to assess the users’ perceived usability of the system, and their overall satisfaction level. SUS will act as a measure of users’ subjective view of the system; identifying, analysing, and quantifying the perceived usability associated with the QA system from the users’ perspective (Lewis and Sauro 2009). Through the use of both subjective and objective usability metrics, the model builds on previous usability testing models; offering an unbiased and comprehensive view of overall system usability.

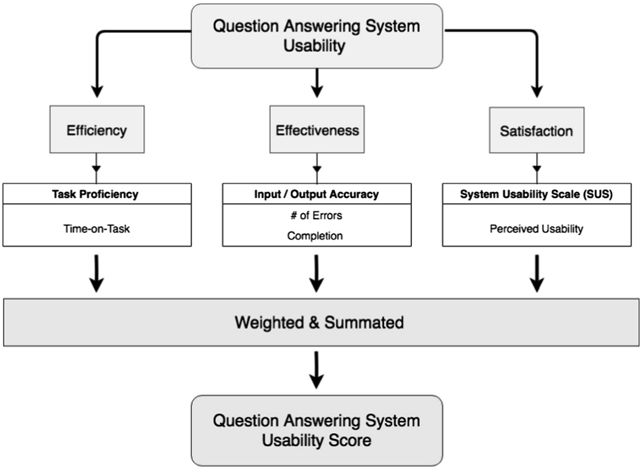

Fig. 1. Proposed usability model

4 Key Informant Interviews

In order to refine the testing model, qualitative interviews were conducted with professionals from the financial services industry, viewed as key informants (KI). The use of key informants as an approach for data collection is beneficial when there is a lack of underlying theoretical understanding (Babbie 1998). This technique has been previously applied in a number of scenarios:

- The study of inter-organisational research when there is a lack of archived data to support such a study (Kumar et al. 1993).
- Developing e-retailing theory when there was a lack of foundational understanding of online retailing (Cowles et al. 2002).
- Using key informants to analyse a groups attitudes and beliefs towards software project reviews (McAvoy 2006).

The user's perception of what is important to them with respect to the usability of question answering systems has not been documented, so key informants were selected to gather this information. Four key informants were chosen from four financial services organisations. The positions held by the key informants are as follows: Chief Operating Officer - Head of Innovation EMEA; Financial Modeling Analyst; Equity Trader; Business Intelligence Analyst. Interviews were conducted on a semi-structured (Myers and Newman 2007; Schultze and Avital 2011) basis. Within their roles, the individuals perform tasks similar to those intended to be completed by the proposed QA system; querying large documents and retrieving answers based on analysis of specified data and data points.

The key informants agreed that the usability dimensions (efficiency, effectiveness, and satisfaction) should have varied weightings based on the specific functional demands associated with the financial services industry. The weighting adjustments suggested by the KIs highlight the specific use case context of QA systems, contrasting strongly with similar usability models devised for alternate contexts. All four interviewees placed significant emphasis on 'Effectiveness', ranking it as the most important dimension of usability. The interviewees stated that correct answers were essential in order for them to "trust" a QA system. Each interviewee stated that QA 'Effectiveness', be it in the context of speech or text is "vital". 'Efficiency' was weighted as the second most important dimension, with one interviewee stating that a QA system must "meet real-world demands of immediate information access". 'Satisfaction' was ranked as the least important usability dimension in all but one case. Consensus formed around the idea of satisfaction being derived only if the system was both efficient, and effective. Satisfaction was perceived to be of least importance; one interviewee stating: "I would be far more concerned with system performance, rather than satisfaction".

5 Refined Usability Weightings from Key Informant Insights

The refined weightings as given by the key informants (see Table 1) highlight the industry specific nuances associated with both speech and text based question answering systems. Consensus formed around 'Effectiveness' taking precedence over both ‘Efficiency’ and ‘Satisfaction’. Based on interviews, ‘Effectiveness’ accounts for 47.5% of the overall Question Answering System Usability Score. ‘Efficiency’ is designated an overall contribution of 32.5%, while ‘Satisfaction’ contributes the remaining 20%. This is reflective of the importance of effectiveness and efficiency as regards the specific QA use case within any given financial services organisation. The layout, and metrics, associated with the model remain unchanged, and are shown in Table 1.

Table 1. Usability dimension weighting.

Usability dimension	Initial weighting	Refined weighting
Efficiency	33%	32.5%
Effectiveness	33%	47.5%
Satisfaction	33%	20%

By combining objective and subjective metrics along with specific refined usability weightings, the question answering usability scoring model attempts to reduce the disjunct between the construct of usability, and its measures (Hone and Graham 2001; Tractinsky 2018). The dimensions of usability have been weighted to incorporate specific nuances associated with the proposed use case, based on the experience of Key Informants; adapting the question answering usability score to more accurately score industry specific QA systems.

6 Conclusion

The research presented in this paper proposes a context specific question answering system usability testing model. It is designed to assess the usability of two alternate natural language processing based question answering systems, operating through alternate modes of communication - speech and text. The weightings associated with the model have been refined through interviews with key informants, applicable to the specific use case context of financial services professionals. The disparity between the original SUM weightings and the Question Answering System Usability Score weightings reflects the view that a ‘one size fits all’ approach is not always applicable, especially in the case of speech and text. This is evident when testing the usability of natural language processing based question answering technologies within the financial services industry. In order to assess the validity of the model, two natural language processing based question answering proofs of concepts are in development, and a usability testing process will be completed at a future date. The testing procedure will assess a text-based user interface, using IBM Watson Assistant, alongside a speech-based user interface developed using Amazon’s Alexa. Comparative analysis will be completed to understand which interface is preferred by the proposed target user, and why this is the case. It is imagined that relative question answering usability scores will reflect user preference. Both proofs of concepts will be tested against the proposed research model in a financial services organisation.

References

Avouris, N., Fiotakis, G., Raptis, D.: On measuring usability of mobile applications. In: International Workshop, p. 38 (2008)

Babbie, E.R.: The Practice of Social Research. Wadsworth Pub., Belmont (1998)

Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. Int. J. Hum. Comput. Interact. **24**(6), 574–594 (2008)

Brooke, J.: SUS: a retrospective. J. Usability Stud. **8**(2), 29–40 (2013)

Brooke, J.: SUS-a quick and dirty usability scale. Usability Eval. Ind. **189**(194), 4–7 (1996)

Cowles, D.L., Kiecker, P., Little, M.W.: Using key informant insights as a foundation for e-retailing theory development. J. Bus. Res. **55**(8), 629–636 (2002)

Ferrucci, D., et al.: Towards the open advancement of question answering systems. IBM Research Report, IBM, Armonk, NY (2009)

Goasduff, L.: Chatbots will appeal to modern workers, smarter with Gartner (2018). https://www.gartner.com/smarterwithgartner/chatbots-will-appeal-to-modern-workers/. Accessed 6 Mar 2018

Green, D., Pearson, J.M.: Development of a web site usability instrument based on ISO 9241-11. J. Comput. Inf. Syst. **47**(1), 66–72 (2006)

Hausawi, Y.M., Mayron, L.M.: Towards usable and secure natural language processing systems. In: Stephanidis, C. (ed.) HCI 2013. CCIS, vol. 373, pp. 109–113. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39473-7_23

Hone, K.S., Graham, R.: Subjective assessment of speech-system interface usability. In: Seventh European Conference on Speech Communication and Technology, Aalborg, Denmark (2001)

International Standards Organisation (ISO): Ergonomic Requirements for Office Work with Visual Display Terminals (VDT)s-Part II Guidance on Usability, ISO/IEC 9241-11 (1998)

Knight, W.: The HR person at your next may actually be a bot. MIT Technology Review (2016). https://www.technologyreview.com/s/602068/the-hr-person-at-your-next-job-may-actually-be-a-bot/. Accessed 6 Mar 2018

Kumar, N., Stern, L.W., Anderson, J.C.: Conducting interorganizational research using key informants. Acad. Manag. J. **36**(6), 1633–1651 (1993)

Lado, M.J., Méndez, A.J., Roselló, E.G., Dacosta, J.G., Pérez-Schofield, J.B.G., Cota, M.P.: R-interface: an alternative GUI for MATLAB. Comput. Appl. Eng. Educ. **14**(4), 313–320 (2006)

Levy, H.P.: Gartner's top 10 strategic predictions for 2017 and beyond: surviving the storm winds of digital disruption, smarter with gartner (2016). https://www.gartner.com/smarterwithgartner/gartner-predicts-a-virtual-world-of-exponential-change/. Accessed 2 Feb 2018

Lewis, J.R., Sauro, J.: The factor structure of the system usability scale. In: Kurosu, M. (ed.) HCD 2009. LNCS, vol. 5619, pp. 94–103. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02806-9_12

López, G., Quesada, L., Guerrero, L.A.: Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of speech-based natural user interfaces. In: Nunes, I. (ed.) AHFE 2017. AISC, vol. 592, pp. 241–250. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60366-7_23

Lopez, V., Uren, V., Sabou, M., Motta, E.: Is question answering fit for the semantic web?: a survey. Semant. Web **2**(2), 125–155 (2011)

McAvoy, J.: Evaluating the evaluations: preconceptions of project post-mortems. Electron. J. Inf. Syst. Eval. **9**(2), 65–72 (2006)

Möller, S., Engelbrecht, K.P., Schleicher, R.: Predicting the quality and usability of spoken dialogue services. Speech Commun. **50**(8–9), 730–744 (2008)

Moore, R.K., Li, H., Liao, S.H.: Progress and prospects for spoken language technology: what ordinary people think. In: INTERSPEECH, San Francisco, California, pp. 3007–3011 (2016)

Moore, R.K.: Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In: Jokinen, K., Wilcock, G. (eds.) Dialogues with Social Robots. LNEE, vol. 999, pp. 281–291. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-2585-3_22

Moore, R.K.: A needs-driven cognitive architecture for future 'intelligent' communicative agents. In: Proceedings of EU Cognition "Cognitive Robot Architectures", vol. 1855, pp. 50–51 (2017b)

Myers, M.D., Newman, M.: The qualitative interview in IS research: examining the craft. Inf. Org. **17**(1), 2–26 (2007)

National Institute of Standards and Technology (NIST). ANSI/INCITS 354-2001: Common Industry Format (CIF) for usability test reports (2001). https://www.irit.fr/~Philippe.Truillet/ens/ens/upssitech/3ASRI/ihm/outils/ANSI_NCITS_354.pdf

Nielsen, J.: Usability 101: introduction to usability. Nielsen Norman Group (2003). https://www.nngroup.com/articles/usability-101-introduction-to-usability. Accessed 23 Feb 2018

Oztekin, A., Nikov, A., Zaim, S.: UWIS: An assessment methodology for usability of web-based information systems. J. Syst. Softw. **82**(12), 2038–2050 (2009)

Panetta, K.: Gartner top strategic predictions for 2018 and beyond, smarter with Gartner (2017). https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2018-and-beyond/. Accessed 4 Apr 2018

Portet, F., Vacher, M., Golanski, C., Roux, C., Meillon, B.: Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. Pers. Ubiquitous Comput. **17**(1), 127–144 (2013)

Quesenbury, W.: Dimensions of usability: defining the conversation, driving the process. In: UPA 2003 Conference, Scottsdale, AZ, pp. 23–27 (2003)

Rohrer, C.P., Boyle, F., Wendt, J., Cole, S., Sauro, J.: Practical usability rating by experts (PURE): a pragmatic approach for scoring product usability. In: CHI 2016 Extended Abstracts, San Jose, CA (2016)

Sauro, J., Kindlund, E.: A method to standardize usability metrics into a single score. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Portland, Oregan, pp. 401–409 (2005)

Sauro, J., Lewis, J.R.: Correlations among prototypical usability metrics: evidence for the construct of usability. In: Proceedings of the SIGCHI conference on human factors in computing systems, Boston, Massachusetts, pp. 1609–1618. ACM (2009)

Sauro, J., Lewis, J.R.: When designing usability questionnaires, does it hurt to be positive? In: Proceedings of the Conference in Human Factors in Computing Systems, CHI 2011, Vancouver, BC, pp. 2215–2224. ACM (2011)

Schultze, U., Avital, M.: Designing interviews to generate rich data for information systems research. Inf. Org. **21**(1), 1–16 (2011)

Silvervarg, A., et al.: Perceived usability and cognitive demand of secondary tasks in spoken versus visual-manual automotive interaction. In: INTERSPEECH, San Francisco, CA, pp. 1171–1175 (2016)

Simonite, T.: Google thinks you're ready to converse with computers. MIT Technology Review (2016). https://www.technologyreview.com/s/601530/google-thinks-youre-ready-to-converse-with-computers/. Accessed 13 Feb 2018

Torres, J., Vaca, C., Abad, C.L.: What ignites a reply?: Characterizing conversations in microblogs. In: Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, Austin, Texas, pp. 149–156. ACM (2017)

Tractinsky, N.: The usability construct: a dead end? Hum. Comput. Interact. **33**(2), 131–177 (2018)

Tullis, T., Albert, W.: Measuring the user experience: collecting, analyzing, and presenting usability metrics, 2nd edn. Newnes, Waltham (2013)

Vacher, M., Caffiau, S., Portet, F., Meillon, B., Roux, C., Elias, E., Lecouteux, B., Chahuara, P.: Evaluation of a context-aware voice interface for ambient assisted living: qualitative user study vs. quantitative system evaluation. ACM Trans. Access. Comput. (TACCESS) **7**(2), 5 (2015)