Weiss *et al. Computational Cognitive Science* (2015) 1:6
DOI 10.1186/s40469-015-0006-9

Computational Cognitive Science
a SpringerOpen Journal

REVIEW Open Access

Evaluating embodied conversational agents in multimodal interfaces

Benjamin Weiss[1*], Ina Wechsung[1], Christine Kühnel[1] and Sebastian Möller[1]

Abstract

Based on cross-disciplinary approaches to Embodied Conversational Agents, evaluation methods for such human-computer interfaces are structured and presented. An introductory systematisation of evaluation topics from a conversational perspective is followed by an explanation of social-psychological phenomena studied in interaction with Embodied Conversational Agents, and how these can be used for evaluation purposes. Major evaluation concepts and appropriate assessment instruments – established and new ones – are presented, including questionnaires, annotations and log-files. An exemplary evaluation and guidelines provide hands-on information on planning and preparing such endeavours.

Keywords: User experience; Usability; Human-computer interaction; Evaluation methods

Introduction

Speech is the most common way for humans to communicate, and thus it has been tempting to use speech as a modality to interact with machines for a long time. However, whereas human speech communication is normally multimodal, as it makes use of the auditory and the visual senses, speech as an interaction modality towards machines was mostly limited to the acoustic channel in the past. With the advent of Embodied Conversational Agents (ECAs) as a means to convey machine output, and with advanced possibilities to recognize speech not only from the acoustic signal, but also from lip and other facial information, which is conveyed visually and captured through cameras, this situation has drastically changed. Today, interactions between humans and machines can also take place in the form of a "dialogue" between a human and an ECA, where both interaction partners convey linguistic as well as extra-linguistic information through both the acoustic and the visual channel.

However, despite these abilities, the spoken interaction between humans and ECAs is still a very unusual way of interacting with machines. The reasons for the observable reluctance to using ECAs in multimodal human-machine interaction are multifold: Humans automatically convey extra-linguistic information such as information about their mental state, emotions, or social relationships, when communicating audio-visually via speech. Machines are commonly not able to extract and interpret these types of information in a reliable way. In turn, the abilities to generate such information on the machine side are still limited, and humans cannot always interpret machine-generated extra-linguistic information in the correct way. This mismatch is frequently at the origin of communication problems.

As a matter of fact, it has to be decided for each application anew whether the use of an ECA is beneficial in a multimodal human-machine interface, and whether it will enhance high quality, usability and finally acceptance on the part of the human user. Such a decision can only be based on a thorough evaluation of the ECA in the respective application scenario.

For evaluations, it is important to distinguish the concept of performance from the concept of quality. Whereas performance indicates whether a unit is able to carry out the function which it has been designed for, quality is the result of a human perception and judgement process, in which the human compares what s/he perceives against some type of internal reference, which has been built on the basis of experiences, expectations, and functional or social demands (Le Callet et al. 2012). Thus, quality is much more difficult to measure than performance.

*Correspondence: benjamin.weiss@tu-berlin.de
[1] Quality and Usability Lab, Deutsche Telekom Innovation Laboratories, Technische Universität Berlin, Ernst-Reuter-Platz 7, Sekr. TEL 18, 10587 Berlin, Germany

Depending on the application domain, different performance and quality aspects may be important. For example, in a tutoring system it may be desirable to engage the user in a conversation which leads to a better understanding of the concepts to be tutored. Here, it may be desirable that the ECA is able to convey feedback about the correctness of user answers in a fine-grained way. In a game, the interaction between human and ECA may just need to be entertaining, thus joy of use may be a decisive factor. In contrast, in an information retrieval task, it is essential that the information is correctly exchanged between user and system, thus the efficiency of the exchange may be highly relevant. Finally, in a care-taking situation where the ECA is part of a care-taking robot, social aspects of the interaction (such as conveying empathy, provoking emotions, etc.) may be important, in addition to the assurance of correct care-taking functions.

The examples show that there will be no universal method to evaluate ECAs which are integrated into multimodal interfaces in all circumstances. Thus, appropriate methods have to be designed for the purpose at hand, and for each application scenario. It is essential that the used methods provide valid and reliable results, in that they address the correct target with minimum uncertainty of the measurement result. This article aims at providing the basis for selecting evaluation methods which fulfil these criteria.

The Review starts with theories which are useful for describing interactions between humans and ECAs, as well as phenomena which have been observed in such interactions. Section "Evaluation and assessment methods" then reviews common evaluation and assessment methods. Definitions and specific assessment instruments of quality aspects are presented in Section "Aspects of evaluation". These quality aspects or concepts have been organized in terms of a taxonomy originally developed for multimodal human-machine interaction at large, and tailored to human-ECA interactions. Finally, we provide guidelines for practically conducting experiments and studies, along with one exemplary evaluation of an ECA in order to illustrate some basic principles. More details can be found in the related literature listed in the Conclusion.

Review

Theoretical foundations of evaluation

According to Scherer (2004), nonverbal information typically accompanies spoken interaction and can

- *substitute* (e.g., nodding, shrugging),
- *amplify* (e.g., simultaneously nod and saying "yes"),
- *contradict* (e.g. unconsciously, or to signal irony), or
- *modify* (relativising an utterance with a smile or gesture)

verbal information on a semantic level.

Consequently, numerous scientists are dedicated to producing such nonverbal signals for ECAs in a natural and situationally adequate way, as well as to recognize and process such signals originating from the users. Furthermore, usability studies and psychological or linguistic experiments are conducted which address the effect of such technological capabilities on the user and the conversation. In the following, major evaluation topics are presented, originating from communication theories and from psychological effects observed in face-to-face conversation.

Communication theories

One example of research on multimodal human-computer interaction (HCI) is the identification of systematic variation in users' interaction style between systems providing only spoken interaction and those providing also the visual channel (Oviatt 2008). For example, if users are aware of the capability to use nonverbal information, spoken content is likely to exhibit linguistic underspecification, which is solved with the support of nonverbal information and world or context knowledge.

This multidimensionality of information conveyed in face-to-face conversation is not covered by information theory (Shannon and Weaver 1949). One of the problems in applying this theory to face-to-face conversation is the assumption of a common code, which works for technical transmission of information, for which this theory was developed, especially for reduction of redundant information, but not in the case of HCI, as users are usually not aware of all capabilities (i.e. code) of that particular system. Also, the assumption of a purely symbolic and intentional communication neglects the continuous negotiation going on between interlocutors, also called *grounding* (Clark 1996), as well as unintentionally produced signals, which are useful to recognize the meaning and intention of a speaker and are therefore typically considered by the conversational partner.

Other communication theories do take into account multiple aspects of a message: According to Schulz and Thun (1981) a multimodal message comprises information about the sender (e.g. mood, age), the relationship (e.g. attitude, status), and the demand on something (perlocutionary act) apart from the factual information. Of course, the question arises, whether such aspects of a message are necessary for the interaction with a system, and whether users are motivated by an ECA to expect proper processing of them.

According to Krauss and Fussell (1996), communication theories relevant for social psychology can be separated into four classes: *Encoder/decoder* models like information theory (Shannon and Weaver 1949), *Intentionalist* models focusing on the task of understanding the

speaker's intention instead of just the utterance meaning (e.g. speech act theory by Searle (1969)), models centred on the process of *Perspective-Taking* as an important aspect of conversation (e.g., models of feedback or the client-centred counselling strategy by Rogers (1951)). Last, *Dialogic* models treat conversation as socially dependent co-joint work of the interlocutors.

According to dialogic models, the main goal of an interaction is to accomplish a shared world view. This emphasizes topics of interaction control, meta-communication and discourse to prevent and solve communication problems and to establish grounding. For the application of user–ECA interaction, any usage beyond mere controlling services or devices will exhibit aspects of such social and co-joint activity. In fact, there is not much motivation to build ECAs for mere controlling tasks. However, the degree of human-likeness or naturalness a user assumes the ECA can provide represents a problematic issue. Users are heterogeneous in this respect and not only interact differently, dependent on expert knowledge and personality, but also evaluate the conversational quality differently due to diverging internal references.

Arguing on the basis of a dialogic approach (Clark 1996), the *grounding problem* is defined for ECAs in Brennan (1998): In addition to obtaining a shared world view in human conversation, one major goal of collaboratively working with an ECA is the negotiation about its verbal and nonverbal capabilities. This grounding process is currently not properly supported in interfaces based on natural spoken language, especially concerning signals of positive evidence and coordination of acceptance of system messages. Therefore, the interaction with spoken dialogue systems, embodied or not, is still inferior compared to approaches of direct manipulation, which intrinsically provide grounding mechanisms (e.g., visualization of capabilities due to labelled buttons in Graphical User Interfaces, GUIs), despite the theoretical advantages of ECAs and spoken dialogue systems providing "natural interaction" (cf. Section "Computers as social actors").

Still, the four approaches presented above do not necessarily differ concerning the topics studied, but they do provide unique views and theoretical assumptions. For example, back-channelling phenomena are investigated in at least three approaches, but they are not a likely topic for intentionalist models. Following major research areas tackled by these communication theories, topics for evaluation can be identified:

1. *Smoothness and success of interaction related to nonverbal interaction management* ("turn-taking" and "back-channel"): E.g.: Is the ECA capable to produce and understand (multimodal) turn-taking/back-channel signals and does this improve subjective or instrumental measures of interaction quality? One example for ECAs is the study of Pardo et al. (2010), which is concerned with instrumentally and subjectively measured effects of an ECA providing a nonverbal communication strategy compared to a spoken dialogue system without embodiment. From a usability point of view, smoothness of interaction represents an aspect of the concept ease-of-use (Section "Ease-of-use"), and success is related to usability (Section "Usability and user experience"). For more information on instrumental measures of smoothness, please refer to Section "Interaction parameters".
2. *Naturalness of interaction, as expected by the user.* This can concern either the ECA's production or understanding capabilities and it can be focused either on the verbal part (social and contextual adequacy; user utterance understanding) or the aspect of multimodality (adequacy of interplay between verbal and nonverbal information; correctness of interpretation of relevant multimodal signals). One HCI example is the reduction of user frustration due to signalling affective states in problematic situations (Hone 2006). Affective responses are covered in Section "Joy-of-use" about joy-of-use.
3. *Functions of nonverbal signals can be tested by synthesizing these with an ECA*, comparable to studies with human interlocutors. Such functions can be linguistically defined (e.g., concerning the type of information or the conversational structure), but they can also refer to the three non-factual aspects of a message (Schulz and Thun 1981), e.g. the special cases of politeness (relationship, demand), status (relationship), or affective state (self-revelation) (Op den Akker and Bruijnes 2012). Testing the usability of such synthesized signals would initially require function tests with participants. However, the effect on users can be evaluated for both concepts, ease-of-use and joy-of-use (Sections "Ease-of-use", "Joy-of-use"). Thus, the goal of evaluation should be clearly defined.
4. Related to smoothness and naturalness, but not unique to ECAs, are *special communicative processes* like negotiations, repairs, grounding terms, adjacency-pairs, alignment and giveness in the course of a conversation. Typically, manual or semi-automatic annotations are used for such analysis (cf. Section "Interaction parameters").

One central aspect of the dialogic approach is the inherent social nature of conversation. As relevant implication, evaluation based only on usability approaches is limited and likely to fail, as social processes have

to be taken into account. In the following, the effects of persona, social facilitation, and visual/auditory attributions are presented to raise awareness and provide explanations for potential issues in human-ECA interaction.

Computers as social actors

According to Duffy (2003) robots need to have anthropomorphic, i.e. human-like qualities, in order to be capable of meaningful social interactions. If there is a human-computer interface exhibiting speech or also visual human features, human social behaviour, especially nonverbal signals, seem to be often triggered automatically (Sproull et al. 1996; Vogeley and Bente 2010). This might also be true for non-anthropomorphic interfaces to some extent (Reeves and Nass 1996), but a rich amount of empirical results has confirmed this effect for anthropomorphic and speech interfaces.

Persona effect: Concerning usability, ECAs are assumed to possess an inherent benefit over other interaction paradigms in HCI, as they provide "natural" and thus "intuitive" interaction as learned by humans throughout their whole lifetime, or to phrase it differently: ECAs could relieve users of the need to learn technical details of new technological services and interfaces (Takeuchi and Naito 1995).

This benefit of establishing a "natural", i.e. social situation, could be empirically observed, as the existence of anthropomorphic human interfaces, even those just accompanying traditional interfaces like web-sites, can result in a higher quality (more positive subjective evaluation) of the service or interface and/or in higher performance (efficiency and effectiveness measured as scores and time-to-complete). This *persona effect*, however, is highly debated, as it seems to be dependent on task, system, and situation (Dehn and van Mulken 2000; Foster 2007; Yee et al. 2007). Also, there is the question, whether the persona effect resembles an effect of mere presence like the social facilitation effect (see below) or if it is a result of a more natural and intuitive interaction.

At least, compared to pure spoken dialogue systems, ECAs can facilitate interaction (Dohen 2009), as human processing benefits from multimodal signals in decreased load and higher neural activity (Stein et al. 2009). So, if not engaged in multiple tasks, an ECA might be less demanding to interact with, if the nonverbal signals are produced and recognized well.

Despite this assumed and also observed persona effect, the design decisions and evaluation concepts concerning ECAs are not to be taken easily: As mentioned already, natural interaction can only have its positive effect, if it is supported well by the ECA, and visually human-like ECAs may fuel too high expectations for the dialogic and nonverbal capabilities. Although a natural interaction is aimed at and users do show signs of social phenomena, some users may like to avoid such social interaction for certain tasks in favour of a paradigm of direct manipulation; just like some customers prefer anonymous online shopping or self-service to personal service.

Therefore, the users' perspective on what the task is, what expectations are on the interaction, and how the ECA is perceived, are important to understand evaluation results. The social situation created by an ECA can also increase attention and even arousal on the user's side (see below), which led to the idea to provide an ECA as highly salient and interesting single interface for various services, e.g. domestic applications or guide for visitors in public spaces, although attention can of course also result in distraction from the relevant task (Takeuchi and Naito 1995). Therefore, an additional evaluation topic is:

5. *Does the mere presence of a (particular) ECA improve the quality of the interaction for the given domain and task?* In order to answer this question, a direct comparison would have been carried out, conducting either an empirical experiment or a field test with at least two conditions (cf. Section "Experiments and field tests").

In line with the persona effect, the presence of an ECA can, for example, result in increased entertainment in a game application (Koda and Maes 1996) and more nonverbal signals are sent in interaction compared to other interfaces (e.g., more eye contact interpreted as increased attention, cf. Takeuchi and Naito (1995)). Also, observations of social effects like politeness, impression management and social facilitation indicate that ECAs can indeed establish a social situation resulting in the automatic appearance of typical phenomena of social psychology (Sproull et al. 1996).

Social facilitation: With *social facilitation*, performance of a person is enhanced in the presence of other people for tasks of low complexity, while performance decreases for difficult, complex tasks. The latter is known as *social inhibition*. This has also been observed for interaction with ECAs (Sproull et al. 1996) and robots (Riether et al. 2012; Wechsung et al. 2012a). One explanation of this effect is an increase in attention and arousal due to the social situation (Guerin 1993). Social facilitation might be more likely for ECAs with gender different to the one of the user (Karacora et al. 2012).

Treating an ECA as a social actor might be used by designers and developers in order to increase attention

(humans are aware of other social actors) and engagement (social interaction as motivation and reward for interacting with the interface), but they have to be aware of the possibility of social inhibition. A sixth topic of evaluation is thus:

6. *Differences in internal states and interactive behaviour due to the presence of an ECA.* Studying this research area typically involves empirical experiments (Section "Experiments and field tests").

Attributions There is a rich canon on human attractiveness and on its relevance for attributions. Although there are positive effects of attractiveness also for non-anthropomorphic interfaces (e.g., "what is beautiful is usable" Tractinsky et al. (2000)), attractive virtual humans might be easier to build than non-human attractive interfaces. ECAs can be built to evoke *attributions* of intelligence and competence, and increase attention. On the downside, stereotypes might apply, in relation to haircut, facial geometry, age, gender, skin colour; even clothes and speaking style have to be defined and cannot stay "neutral". However, this might result in the attribution of a personality (Nass et al. 2001; Sproull et al. 1996; Marakas et al. 2000). Therefore, simple design decisions may result in incongruence with the user's expectations for the service or impression of the brand. Therefore, a last evaluation topic can be named:

7. *Studying attributions and attitude formation due to characteristics of an ECA.* For this topic, the particular situation, defined by user, domain, task and application have to be taken into account. As for Question 6, this question is typically addressed by conducting empirical experiments (Section "Experiments and field tests").

Uncanny valley: One particular risk when aiming at human-like interfaces is the *uncanny valley* effect: This effect describes an increase of the perceived familiarity with increasing resemblance to human appearance only until a certain level of human-likeness is reached. However, small divergence from human-likeness results to an uneasy feeling. At that level, familiarity drops and increases again if the human-likeness is perfect (Mori 1970). The uncanny valley effect can even be observed in neural activity (Saygin et al. 2012). The authors conclude that the uncanny valley effect may be based on perceptual mismatch: For very human-like robots equally human-like movements are expected. However, if those expectations are not met, the uncanny valley effect may occur. This dependency of the effect on the expectation of (and reference as) a real human is the reason to present it along with attributions in this subsection.

Whereas the evaluation topics 1–4 are related to special aspects of an ECA or robot, topics 5–7 are about using an ECA or robot at all. The subsequent difference lies in the possible evaluation methods, as answering questions within topics 5–7 requires a comparison with a non-embodied condition. In the following section, we briefly describe evaluation methods with and without real users and present typical assessment methods, before describing specific instruments to assess concepts of interest in Section "Aspects of evaluation".

Evaluation and assessment methods

According to Preece et al. (1994) evaluation methods can be distinguished using four different criteria.

1. *The question addressed with the evaluation;* i.e.:
   - to see if the system is good enough
   - to compare two alternatives and to see if one system is better than another one
   - to get the system closer to the real world
   - to see how well the system is working in the real world or
   - to see if the system complies to certain standards
2. *The particular stage in the engineering cycle;* formative, process-oriented evaluation can be distinguished from summative, goal-oriented evaluation
3. *The user involvement;* with user-centred, empirical methods on the one side, and expert-centred, analytical-formal methods one the other side
4. *The type of data collected* (qualitative or quantitative data)

In addition to the last criterion, direct and indirect measurements can be distinguished in the context of usability and quality evaluation (Wechsung et al. 2012b). Direct measurements are assessed directly from the user. Indirect measurements refer to manually annotated or instrumentally obtained data (i.e. automatically logged interaction data or physiological parameters).

Established usability evaluation methods are already presented elsewhere (e.g. Noor (2004)). Therefore, we focus on methods explicitly used with ECAs or adapted to them. Also, evaluation data obtained from experiments are covered here in more detail. These include questionnaire and interview data, as well as data obtained indirectly, i.e. by recording log-files, recording or annotating user behaviour, etc. For the case of indirect measures, the examples of eye-tracking data and interaction parameters are presented. Particular questionnaires are referred to in the appropriate paragraphs of Section "Aspects of evaluation", where the quality aspect assessed by each questionnaire is described.

Cognitive Walkthrough

The *Cognitive Walkthrough* is a task-based, expert-centred, analytical method (Holzinger 2005) based on exploratory learning and problem solving theory (Wharton et al. 1994). It takes into account that users often learn the interface by exploring it, instead of reading the manual.

Experts, usually designers or psychologist, analyse the functionalities of the interface based on a description of the system, a description of the tasks the end user will carry out, a list of actions necessary to perform the tasks, and a description of user and usage context. Critical information is recorded by the experts using a standardized protocol. The procedure itself involves the following five steps (Wharton et al. 1994):

1. Definition of inputs for the walkthrough (e.g. identifying the users, defining the tasks to evaluate, describing the interface in detail)
2. Calling in the experts
3. Analysing the action sequences for each task
4. Protocolling critical information
5. Revising the interface

The experts analyse the interface based on the following four questions: (1) *Will the user try to achieve the right effect?* (2) *Will the user notice that the correct action is available?* (3) *Will the user associate the correct action with the effect to be achieved?* (4) *If the correct action is performed, will the user see that progress is being made toward solution of the task?*

In the context of the evaluation of ECAs the procedure of the Cognitive Walkthrough, which was initially developed for graphical user interfaces, may be used unchanged. However, regarding the four central evaluation questions Ruttkay and Op den Akker (2004) suggest adapting and expanding the question (2) and question (4) as follows: (2) *Will the user be aware of what they can do to achieve a certain task: to talk to several (all) ECAS on the screen, to use gestures, gaze and head movement as those are perceived by the ECAS? Will users notice (when) they need to give an answer? Will they notice whom they may address (who is listening)?* (4) *Will there be a feedback to acknowledge the (natural, may be multi-modal) answer given by the user? Does the feedback indicate for the user if his/her action was correct?* (In case of natural communication, this distinction means if something 'syntactically correct' was said and thus properly parsed, or something was said which is (one of the) expected answers in such a situation.)

As for almost all formative-analytical methods, the biggest advantage of the Cognitive Walkthrough is that end users as well as an implemented system are not necessary. Disadvantages are the quite low level of information, as only the ease of learning is investigated (Wharton et al. 1994). Moreover, a Cognitive Walkthrough might be very time consuming for complex systems. Note that, the Cognitive Walkthrough is strictly task-based and will only be able to evaluate the ease-of-use of an interface rather than its joy-of-use.

Heuristic evaluation

Heuristic Evaluation is a method of the so-called Discount Usability Engineering, a resource conserving, pragmatic approach (Nielsen and Molich 1990), aiming to overcome the argument that usability evaluation is too expensive, too difficult and too time consuming. In a Heuristic Evaluation, several experts check whether the interface complies with certain usability principles (heuristics). To ensure an independent, unbiased judgement of every evaluator, they do not communicate to find an aggregated judgement until each of them investigated the interface on his/her own (Nielsen 1994). The result of a Heuristic Evaluation is a list of usability problems and the respective explanations. Additionally, problems might be judged according to their frequency and pertinence.

According to Nielsen, three to five experts will find 60–70 % of the problems, with no improvements for more than ten evaluators (Nielsen 1994). However, this statement has repeatedly caused disputes; research provides support (Virzi 1992) as well as contrary findings (Spool and Schroeder 2001; Woolrych and Cockton 2001). In a series of studies, Molich and colleagues come to the conclusion that 4–6 experts are sufficient, but only for one iteration cycle and only with real experts (Molich and Dumas 2008).

The Heuristic Evaluation is a cheap method and quick to apply, and it can be conducted throughout the whole development cycle (Holzinger 2005). The probably most widespread heuristics are the ten guidelines proposed by Nielsen (1994). Moraes and Silveira (2006) updated and adapted this set first to the evaluation of animated agents in general and later to the evaluation of pedagogical agents in particular (Moraes and Silveira 2009).

Model-based evaluation

For *Model-Based Evaluation* on a very general level, two different approaches can be distinguished. The first approach has its origin in cognitive psychology and focuses on the cognitive process while interacting with an interface. The other approach is rooted in the engineering domains and is focusing on the prediction of user behaviour patterns. Within both approaches, user models are employed for the predictions.

Methods of the first approach are usually addressing low-level parameters like task execution time, memory processes or cognitive load and are largely bottom-up

oriented. Starting point to define user models are theories and findings from cognitive psychology. Examples are the methods GOMS (Goals, Operator, Methods, Selection rules) (Card et al. 1983), the Cognitive Complexity Theory (CCT) (Kieras and Polson 1985), or ACT-R (Adaptive Control of Thought–Rational) by Anderson and his group (e.g. Anderson and Lebiere (1998)).

As all these cognitive models are well-grounded in theory, they provide useful insights in user behaviour. Although cognitive modelling is an active research field, so far it has not been received particularly well by usability practitioners and only rarely finds its way into non-academic evaluations (Engelbrecht et al. 2009; Kieras 2003). Reasons are their often high complexity (Kieras 2003) and possibly the aforementioned low level of the information possible to gain with cognitive modelling.

An engineering-based, statistically-driven approach attempts to provide more high level information, e.g. if the user is "satisfied" with the system, and therefore rather utilizes top-down strategies. Here, user models are usually defined based on real user data and are not necessarily linked to cognitive theories (Engelbrecht et al. 2009). Most of these methods and algorithms were developed for spoken dialogue systems, with PARADISE (Paradigm for Dialogue System Evaluation) (Walker et al. 1997) likely being the most widespread one. PARADISE uses linear regression to predict user satisfaction based on interaction parameters such as task success or task duration.

A model-based approach explicitly tailored to multimodal system is PROMISE (Beringer et al. 1997), an extension of Walker's PARADISE. However, studies applying PROMISE are scarce, possibly because some of the parameters are relatively ill defined (e.g. the way of interaction), and it is not specified how they should be assessed. Just recently a well-defined set of interaction parameters for multimodal interaction was proposed (Kühnel 2012), yielding reasonable prediction performance ($>$ 50 % accuracy) for user judgements in general (see Section "Interaction parameters"). However, most data analysed with respect to interaction parameters was not collected with experiments including an ECA. For those experiments with an ECA, significant correlations were found for the concepts of helpfulness and cognitive demand (cf. Section "Aspects of evaluation"). Using this approach, Pardo et al. (2010) compared voice-only interaction to an ECA, and observed significant differences suggesting a "rougher" interaction for the voice-only version.

Experiments and field tests

Experimental evaluation investigates specific aspects of the interaction behaviour under controlled conditions (Sturm 2005). In the simplest experimental design, one hypothesis is formulated and two experimental conditions, differing only regarding the manipulated factor to be investigated (independent variable), are set-up (Dix et al. 2003). All differences occurring in the measured variables (dependent variable) are attributed to the manipulations of the independent variable (Dix et al. 2003). Experiments allow for collecting high quality data as interfering variables are controlled and/or eliminated. Experiments provide, if carried out carefully, causal inference. Thus, experiments are essential to establish and verify theories (Gerrig and Zimbardo 2007); accordingly, experiments are a useful method for evaluation of ECAs. However, with experiments being strongly controlled, user behaviour might be rather unnatural (Sturm 2005). In the worst case, results can be an experimental artefact. Another drawback is the high amount of resources required to set up and conduct a proper experiment.

When evaluating HCI, field tests aim at ecologically valid conditions by embedding the actual usage in an authentic daily situation. This affects not only the environment, but all context factors like user motivation and references. Although field tests can thus prevent the issue of (laboratory) experiment being unnatural, field tests are usually much more costly in terms of effort and time. In contrast to experiments, which reduce or control potential influence factors, results from field tests are typically much harder to interpret because of unknown or uncontrolled variables, and do not allow for causal inferences (Bernsen and Dybkjær 2009; Dix et al. 2003).

In HCI, both methods are sometimes difficult to distinguish, as, e.g., field tests can be carried out in different conditions like experiments. The actual choice will depend on the evaluation/research question (e.g. with topics from cognitive psychology most likely studied with experiments, and topics about acceptability most likely studied with field tests) and on the domain: Although in general, all topics in Section "Theoretical foundations of evaluation" can be studied with experiments, topics 5 and 6 concerning the effect of an ECA on the system's quality and on users' internal states and behaviour may be more difficult to validly study for an ECA companion than a learning tutor, as the companion is expected to be embedded in daily life routine.

In the following paragraphs, we will go into detail for two major categories of dependent variables, questionnaires and indirect measures (eye-tracking data and interaction parameters).

Interviews and questionnaires

Questionnaires and interviews are indispensable to measure the users' judgements of the system (Holzinger 2005), as interaction data will not necessarily reflect the users' perceptions (Naumann and Wechsung 2008). Interviews

and questionnaires are often used to assess user satisfaction, emotions or attitudes towards the system. If reliable questionnaires or interviews are available, they are relatively cheap and easy to employ. Whereas interviews typically aim at assessing qualitative data, questionnaires allow for collecting quantitative data, and for statistical analysis.

Standardized, well-validated questionnaires tailored to ECAs are rather rare; consequently, self-made questionnaires or questionnaires developed for unimodal systems are often employed. Both approaches are problematic: Self-constructed questionnaires are usually not properly validated (Larsen 2003) and questionnaires developed for unimodal GUI-based systems may not provide valid and reliable results for ECAs. Consequently, the constructs measured are quite diverse and the results are hardly comparable. Moreover, without a proper validation, it is uncertain if a questionnaire actually measures the constructs which it was intended to measure (Larsen 2003).

Notable exceptions are the Agent Persona Instrument (API) (Baylor and Ryu 2003) and the Attitude Towards Tutoring Agent Scale (ATTAS) (Adcock and Eck 2005). Both are relatively well-known and psychometrically tested questionnaires. However, both are limited to the evaluation of pedagogical agents. A questionnaire intended to be applicable to a wide range of ECAs has just recently been proposed with the Conversational Agents Scale (CAS) (Wechsung et al. 2013). The CAS is based on the dimensions of the theoretical framework by Ruttkay and colleagues (Ruttkay et al. 2004). The original dimensions are: satisfaction, engagement, helpfulness, naturalness and believability, trust, perceived task difficulty, likeability and entertainment. According to (Ruttkay et al. 2004) satisfaction and engagement are high-level constructs, which comprise several of the other dimensions. For example, engagement may include likeability and trust; hence, these meta-aspects can be said to derive from the sub-aspects. Consequently, the questionnaire comprises the sub-aspects only, which are helpfulness, naturalness and believability, trust, perceived task difficulty, likeability and entertainment. Note that, to date only the German version of the CAS is validated. A preliminary validation of an English translation is presented in Section "An experiment testing two different talking heads". In Section "Aspects of evaluation" additional questionnaires are presented to assess specific quality aspects.

Although interaction data does not necessarily reflect user perception, such data can be of interest for evaluation. For HCI, instrumental and manually annotated data is usually used to describe the course of interaction, either because it represents a genuine evaluation question (e.g. the relationship between measured and user perceived interaction duration). In this case, such data can complement questionnaire data, e.g. for topics concerning smoothness or naturalness of interaction (topics 1 and 2); nonverbal and verbal functions (topics 3 and 4) or user's internal states (topic 6 from Section "Theoretical foundations of evaluation").

Eye-tracking

One popular physiological method to assess data instrumentally is eye-tracking. Recorded data can be used to monitor and record user behaviour. The exact purpose of using eye-tracking should be noted: As a method of usability testing, information can be obtained about which items of a GUI are salient or not. However, often used "heatmaps" can be very misleading due to the aggregation of data.

In research, eye-tracking is for example used to study visual search strategies, to measure aspects of affect, like interest, fatigue, or to assess arousal using fixation times, blinking rate or pupil dilation as measures. The book by Holmqvist et al. (2011) is advised as a good start for the interested reader. Still, properly assessing physiological measures requires a lot of experience, e.g. to deal with practical problems with lighting, reflections etc., especially in mobile settings and under realistic usage circumstances.

Interaction parameters

The parametrisation of individual interactions on the basis of data extracted from manually annotated or automatically logged (test) user interactions commonly accompanies experimental evaluation. So-called interaction parameters quantify the flow of the interaction, the behaviour of the user and the system, and the performance of the devices involved in the interaction. These parameters provide useful information for system development, optimization and maintenance. The parametrisation of interaction is thus complementary to quality judgements assessed via questionnaires – by addressing the system's performance from a system developer's point-of-view.

For more than two decades of experience with spoken dialogue systems, researchers and developers have defined, used, and evaluated interaction parameters for the named purposes, summarized for example in Möller (2005). With the emergence of multimodal systems, this approach has been stipulated for this new domain as well (Dybkjær et al. 2004). Several annotation schemes for multimodal interaction have been published (Gibbon et al. 2000; López-Cózar et al. 2005), but researchers build "their own corpora, codification and annotation schemes" mostly "ad hoc" (López-Cózar et al. 2005, p. 121). Recently, a dialog acts annotation scheme was standardized for human face-to-face conversation. These dialog acts are multimodal per definition, as they

are considered the conceptual or perceptual origin of human (humanlike) multimodal signals (ISO 24617-2 2012). This scheme could be directly used for ECAs as well, as long as the multimodal system does not offer additional interaction capabilities except for the ECA (e.g. a touch-screen).

The benefit of a nonverbal communication strategy with an ECA-based interface have been analysed using interaction parameters and questionnaires together (Pardo et al. 2010). While the results assessed via interaction parameters and questionnaires pointed in the same direction only the interaction parameters yielded significant differences. Later, similar results were found comparing a user-adaptive to a non-adaptive ECA (Weiss et al. 2013), suggesting to not only rely on questionnaires.

In Kühnel (2012) a first set of interaction parameters for multimodal interaction has been proposed, focusing mainly on spoken, gestural and GUI-based input as well as spoken and graphical output. For the evaluation of the interaction with ECAs or ECA-based system no such set of parameters exists. In the following the adaptation of multimodal interaction parameters to ECAs will be discussed and some exemplary parameters will be proposed.

One basic concept common to all interaction parameters is that they can only be measured based on an interaction between a system and at least one user – for example during a laboratory or field test. The interaction is thus influenced by system and user characteristics and behaviour. As these aspects are interrelated they can usually not be separated. Consequently, interaction parameters will reflect not only the interaction but also the characteristics of the interlocutor. The requirements for recording interaction parameters are therefore similar to the ones applicable for experimental evaluation (cf. Section "Best practice and example evaluation" on best practice).

If the system is available as a glass-box, some system-side parameters can be extracted from log-data. The same is true for selected user-side parameters, possibly with an amount of uncertainty. These parameters can be used for adaptive systems and to monitor the interaction online. But for many parameters a manual transcription and annotation of recorded video and audio files is indispensable, thus making a laboratory test setting necessary.

Most interaction parameters which have been proposed for spoken dialogue systems (Möller 2005) can be directly transferred to the context of interaction with ECAs – at least for those systems where speech input and output plays a major role. For other parameters, the definition has to be adapted. Some parameters, such as speech input-related metrics, have to be mirrored for every input modality. Depending on the accompanying modalities, other parameters might have to be added – for example parameters known from graphical user interfaces. And there are new parameters inherent to the interaction with ECAs which should be considered (see below).

Most parameters are annotated on a turn-by-turn basis and later summed up or averaged over the number of turns for the whole dialogue or sub-dialogues. It is thus necessary to precisely define the beginning and end of a turn in a multimodal interaction. For an interaction at least one pair of user and system turns is necessary; this has been named "exchange" in Fraser (1997). In Fig. 1 one complete exchange is depicted, including associated time-related parameters.

Most of the concepts related to the overall dialogue and the communication (e.g. dialogue duration, system and user turn duration (Gibbon et al. 2000), system response delay (Price et al. 1992), can be transferred to multimodal interaction unmodified – system response delay, for example, has been used before as a parameter for multimodal systems to measure dialogue efficiency (Foster et al. 2009). Their measurement is based on the definition of user and system turns as described above.

Compared to an interaction with a spoken dialogue system, the interaction with an ECA, and thus every single turn, is potentially far more complex. This is mainly due to two aspects: ECAs mirroring human-human conversation should include and understand feedback mechanisms. And even if this is not the case, input and output via multiple modalities adds complexity.

On the output side, that is the behaviour and output of the ECA, interaction parameters should be logged automatically. But for a useful evaluation the user's behaviour and input has to be taken into consideration. If the ECA applies feedback or turn taking mechanisms, for example, does the user perceive those and react accordingly? A possible measure is the delay between a turn taking action from the ECA and the taking over by the user.

Whenever the user's reaction to visual features is analysed it has to be ensured that the user was actually looking at the ECA in the crucial moment. This can be achieved automatically with an eye tracking system or by annotation of video data.

A human-like ECA might induce higher expectations in the user. Confronted with a talking head or even a fully bodied ECA, which uses gestures and mimics to convey meaning, the user might expect the system to understand the same modalities. Thus it makes sense to annotate how often the user tried to interact via modalities not offered by the system, such as pointing to an object while only spoken input is possible.

Measures such as task success are only meaningful for task-oriented interactions. To compute metrics for task success the goal of the interaction has to be known. In a laboratory setting this can be achieved

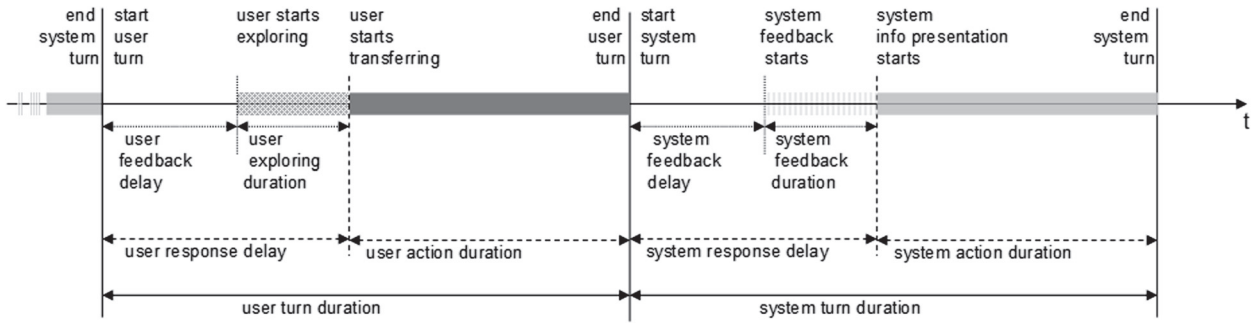

Fig. 1 Time-related interaction parameters

by defining explicit tasks to be fulfilled by the participants of the study. But even when defining explicit tasks it is possible that the participant does not understand the task correctly or accidentally skips parts of the task, resulting in deviations from the original interaction path.

Performance measures should be taken into account as they are good indicators for system developers and might be useful for the interpretation of some dialogue issues. A low performance of one or more recognizers might explain an above-average dialogue duration.

For ECAs the synchrony of speech and lip movement and also the correct alignment of gestures, mimicry and speech are of high importance. This can be measured by the lag of time (LT) between corresponding modalities or by the overall number of times corresponding output modalities have been asynchronous (calculated based on a pre-defined threshold). For spoken output, different methods to assess TTS quality have been proposed but they have not been applied widely (Möller 2005).

Further evaluation methods

In the above sections established evaluation methods are presented; note that this overview is not exhaustive as primarily methods were considered, which have been modified for and applied to the evaluation of ECAs. Additional evaluation methods are for example *Review-Based Evaluation* and the *Think Aloud* approach. For the Review-Based Evaluation, existing experimental findings and principles are employed to provide a judgement. Relevant literature has to be analysed in order to approve or disapprove the design of the interface (Dix et al. 1993). Review-Based Evaluation is faster and more economical than conducting an experiment oneself. But wrong conclusions might be drawn if the selection of the considered studies is not done with the required prudence. Studies in which the Review-Based Evaluation is applied in the context of the evaluation of ECAs are rather rare.

Another prominent usability evaluation method is Think Aloud. Participants are encouraged to 'think aloud', i.e. to verbalize, and externalize their thoughts (Holzinger 2005). This might be done during the interaction or after the interaction as retrospective Think Aloud. For the latter, the user is confronted with video recordings of the test session and is asked to comment on them. Although retrospective Think Aloud is less intrusive than online Think Aloud, it might possibly be affected by memory biases. The Think Aloud method can be used for free exploration of the interface as well as for conducting concrete tasks. The main advantages of the Think Aloud method are the low effort required for the preparation of the test and the rich amount of information which can potentially be gathered. The disadvantages are the often 'unnatural' situation due to the constant verbalization (Lin et al. 1997); often the experimenter has to repeatedly advise the participants to actually think aloud in order to keep the participants talking (Ericsson and Simon 1980). Additional problems are the systematic biases due to social desirability. Moreover, for interfaces offering speech input such as ECAs, the non-retrospective version of this method is inappropriate, as Think Aloud and speaking to the system simultaneously is not possible (Lewis 2012).

Aspects of evaluation

Some requirements and issues to consider when n ECAs are presented in Ruttkay and Pelachaud (2004). For a start, decisions on user characteristics and evaluation criteria should be made explicit (Ruttkay et al. 2004). Therefore, we concentrate on user characteristics not described there, first of all mood and needs.

User characteristics

Before evaluation studies are conducted, the targeted user group and their characteristics need to be defined. In the following paragraphs relevant individual characteristics are explained and respective measurement instruments

are presented either to confirm the target group or to cluster users.

Abilities

Abilities refer to perceptual-cognitive abilities and knowledge as well as to motor and physical capabilities. A variety of methods can be found to assess perceptual-cognitive abilities, including clinical intelligence tests like the Wechsler Adult Intelligence Scale (WAIS) (Wechsler 2008). Such tests typically include sub-scales; for instance, the WAIS IV addresses perceptual abilities with the Perceptual Reasoning Index and memory capabilities with the Working Memory Index.

There are validated instruments for the assessment of HCI-related abilities and knowledge (Smith et al. 2007; Van Vliet et al. 1994). The former instrument measures computer experience, the latter is a questionnaire for assessing computer literacy. For motor capabilities the AMPS (Assessment of Motor and Process Skills) (Fisher 1996) can be employed.

Personality

Personality includes personality variables, like psychological personality traits, e.g. the so-called Big Five (Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism), and attitudes. To assess personality variables, psychometric questionnaires are available: They can be measured with the NEO-FFI (Costa and McCrae 1992) or with the briefer Big Five Inventory (John et al. 1991). Both questionnaires are also available in a short version (Rammstedt and John 2007).

In this category belongs also the attitude towards computers (Richter et al. 2000) or technical affinity in general (Karrer et al. 2009; Weiss et al. 2012).

Demographics

For age, it is often assumed that not the chronological age "per se" causes differences, but rather characteristics like a smaller degree of previous experience (Chalmers 2003), the age-related decrease of cognitive abilities (Wolters et al. 2010) and motor impairments (Carmichael 1999). Demographics are usually easy to assess with simple, self-constructed questionnaires presented before, or after the test.

Mood

Moods are the affective quality of experiences, constantly experienced but often only sporadically reflected consciously (Morris 2005; Silvia and Warburton 2006). Compared to emotions, moods are lacking objects, are psychologically diffuse, relatively long lasting, and are structured simply; moreover, they are not differentiated by patterns of appraisal (Silvia and Warburton 2006).

Research suggests that people in a good mood are, compared to people in a bad mood, more likely to employ less elaborate information processing strategies like heuristics (Bless et al. 1996), for an overview see (Schwarz and Clore 2003). Thus, their information processing might lack logical consistency. For evaluative judgements this means, that evaluation is probably more context-driven than content-driven (Bless et al. 1996). More specifically in terms of usability studies, people in a positive mood might give less exact evaluation ratings, as positive mood is associated with less attention given to details and less information being considered. Moreover, they may be more influenced by the contextual factors e.g. the setting and scenario. Additionally, memory recall is mood congruent: good moods make recall of positive experiences more likely than bad moods and vice versa (Kahneman 2003a).

A lightweight instrument to measure mood is the so-called Faces Scale (Andrews and Whitey 1976). This non-verbal scale shows seven faces ranging from "very sad" to "very happy". Participants are asked to indicate the face, which matches their current mood best. Another questionnaire is the Brief Mood Introspection Scale (BMIS) (Mayer and Gaschke 1988). The BMIS consists of 16 items (plus one global item) and measures mood on four sub-scales (Pleasant-Unpleasant Mood, Arousal-Calm Mood, Positive-Tired Mood, Negative-Relaxed Mood).

Needs

According to Hassenzahl and colleagues (Hassenzahl et al. 2010), the main motivation to use interactive technologies is the fulfilment of psychological needs. The most salient needs in the context of HCI have been identified as the needs for stimulation, relatedness, competence and popularity. For example, a very bored person may ask "stupid" questions to a spoken dialogue system in order to fulfil the need for stimulation. It should be noted that psychological needs do not match biological-physiological needs such as hunger or thirst. However, like biological needs psychological needs are assumed to be largely invariant across human beings (Sheldon et al. 2002). Of course, the level of fulfilment and deprivation of each need changes constantly. A questionnaire to assess the level of experienced need fulfilment can be found in Hassenzahl et al. (2010). It is an adapted version of a questionnaire, originally developed by Sheldon and colleagues in the context of personality psychology (Sheldon et al. 2002).

Usability concepts

In this paragraph, the judgemental processes leading to the formation of quality, as well as quality aspects or concepts relevant for the usage of and interaction

with ECAs are presented. The structure is based on the third layer of the taxonomy for multimodal systems (Wechsung et al. 2012b). Where appropriate, evaluation dimensions from (Ruttkay et al. 2004) are associated (cf. Fig. 2).

Judgemental process

Research indicates that judgement and decision-making processes involve two systems, the cognitive-rational and the emotive-intuitive system (Epstein 1994; Kahneman 2003b).

The cognitive-rational systems is, compared to the emotive system, more analytic, logical, abstract, active, controlled, rule-based and slower (Kahneman 2003b; Lee et al. 2009); it is the deliberate mode of judgements (Kahneman 2003b). The emotive-intuitive system, on the other hand, is characterized by automatic, associative effortless and often emotionally charged operations (Kahneman 2003b); it is the automatic mode of judgements. These automatic, intuitive judgements of the emotive system are monitored by the cognitive system and may be corrected or overridden (Kahneman 2003b). However, the monitoring is rather loose, as the analytical conscious processing in the cognitive system requires mental resources and thus induces cognitive load (Kahneman and Frederick 2002). Hence, the judgements of the emotive-intuitive system determine preferences unless the cognitive system intervenes (Kahneman 2003b) and in every action or thought the emotional system is, at least unconsciously, engaged (Picard 1997).

However, for a long time the only emotion considered in HCI was frustration, and how to prevent it. Only during the last decade have emotions and affect become a major research topic in HCI. In line with the findings reported above, it is argued that every interaction with technological systems involves a wide range of emotions (Brave and Nass 2007). In Hassenzahl et al. (2010) the close relationship between experience, emotions, and affect – proposed earlier (McCarthy and Wright 2004) – is emphasized. Their position is, according to (Hassenzahl et al. 2010), strongly influenced by the work of Dewey, who describes emotions as the "qualities of experiences". Thus, a positive experience is linked to a positive emotion and vice versa. As evidence for the assumptions above it could be shown that apparent affective responses towards a target are used as information and therefore influence evaluative judgements (Schwarz and Clore 2003), especially when the judgements refer to preference or "likeability" judgements (cf. Section "Mood").

Accordingly, evaluative judgements of a system are not solely based on the system's attributes, but on the feelings, the user has towards the system and the mode of judgement used when forming the judgements. For the case of ECAs, likeability as one concept (Ruttkay et al. 2004) resembles the attitude towards the ECA on the basis of joy-of-use and ease-of-use (cf. Section "Usability and user experience").

Hedonic and pragmatic qualities

Traditionally HCI focused on enhancing the efficiency and effectiveness of the system (Preece et al. 1994). The major concern was to prevent negative emotions (Hassenzahl and Tractinsky 2006). With the paradigm shift from instrumental to non-functional qualities (cf. Section "Usability and user experience") concepts of positive or hedonic psychology were adapted and transferred to HCI. Hedonic Psychology focuses on concepts like enjoyment, pleasantness, but also unpleasantness rather than on attention and memory (Kahneman 2003a) – two key topics that have been

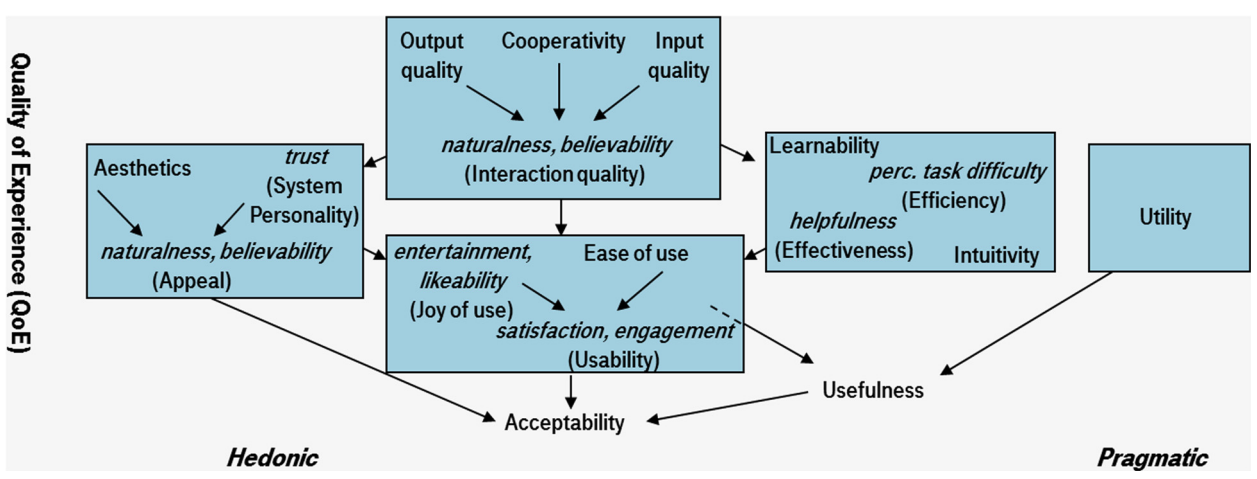

Fig. 2 Quality aspects from (Wechsung et al. 2012b). Addressed concepts from (Ruttkay et al. 2004) are italic and lowercase, with corresponding original aspects in brackets

the focus of psychological research. Analogous to this development, HCI research also moved away from the "classical" cognitive information processing paradigm (Kaptelinin et al. 2003) towards concepts like Affective Computing (Picard 1997) and Emotional Design (Norman 2004). Nowadays the aim is not only to prevent rage attacks as a result of a crashing computer, but to facilitate positive emotions while interacting with an interactive system (Hassenzahl and Tractinsky 2006).

It is suggested to differentiate between pragmatic qualities and hedonic qualities (Hassenzahl et al. 2000). While pragmatic qualities refer to functional aspects of a system and are closely related to the classical concept of usability, hedonic qualities cover the interface's non-instrumental aspects (Hassenzahl 2003). Pragmatic qualities were found to be a hygiene factor removing barriers hindering the fulfilment of the users' needs (Hassenzahl et al. 2010). Hence, a system's n qualities enable need fulfilment, but are themselves not a source of a positive experience. Hedonic qualities are associated with a system's ability to evoke pleasure and the psychological well-being of the user (Hassenzahl 2003). They are motivators and reflect the products capability to create a positive experience (Hassenzahl et al. 2010).

Ease-of-use

Ease-of-use is closely related to pragmatic qualities. Key aspects are the traditional usability measures described by the International Organization for Standardization (ISO) in the ISO 9241-11 standard (ISO 9421-11 1998). Namely, these are effectiveness, which is the accuracy and completeness of goal attainment with respect to user and context of use, and efficiency, i.e. the required effort and resources related to the accuracy and completeness (cf. topics 1 and 5 from Section "Communication theories"). In Ruttkay et al. (2004) helpfulness is proposed as an aspect relevant for the evaluation, as it covers supportive behaviour for effective interaction with an ECA. Another dimension, naturalness and believability, actually corresponds to two different quality aspects, namely naturalness of its embodiment (the hedonic aspect of appeal) and interaction quality. An ECA is considered as believable, if it is naturalness is "meeting the expectations" (cf. Ruttkay et al. (2004)).

Although efficiency is often measured via perceived task duration, it may also refer to the mental effort. This means, the perceived mental effort is considered as a resource. In the context of the evaluation of ECAs, perceived task difficulty is suggested as an additional concept, which is closely related to the perceived mental effort (Ruttkay et al. 2004): Perceived mental effort is often used as an indicator of task difficulty. In addition to the aspects presented above, learnability also determines a system's ease-of-use (Dix et al. 2003). Learnability describes how well a new user can effectively interact with a system and maximize performance. On a lower level, learnability includes predictability, synthesizability, familiarity, generalizability and consistency (Dix et al. 2003).

Concerning ECAs, intuitivity is often given as a major reason to provide embodiment (Ruttkay et al. 2004). Therefore it is a relevant quality aspect for the topic 5 (cf. Section "Computers as social actors").

The vast majority of standardized usability questionnaires cover these constructs. Examples are the QUIS (Shannon and Weaver 1987), the SUS (Brooke 1996), the IsoMetrics (Gediga et al. 1999), the AttrakDiff (Hassenzahl et al. 2003) and the SASSI (Hone and Graham 2000). It has to be noted that the questionnaires' subscales are not necessarily named efficiency, effectiveness, and learnability. The SASSI sub-scale Speed is strongly related to efficiency, the scale Pragmatic Qualities on the AttrakDiff refers to both, efficiency and effectiveness.

Besides questionnaires, expert-oriented procedures such as the Cognitive Walkthrough (Wharton et al. 1994) are regularly used for the evaluation of ease-of-use. Please note, that the expert-oriented procedures do not involve users. Therefore, they do not assess the quality as perceived by the user, although they may provide an estimation of the user's perceptions.

Although there is a wide range of methods available for assessing a system's ease of use, few of them are so far particularly tailored to ECAs. Using established questionnaires like the Software Usability Measurement Inventory (SUMI) (Kirakowski and Corbett 1993) and the Questionnaire for User Interaction Satisfaction (QUIS) (Shannon and Weaver 1987) might be problematic as they were developed for unimodal graphical user interfaces like websites. However, the scale measuring Pragmatic Qualities of the AttrakDiff may provide meaningful results (cf. Section "Hedonic and pragmatic qualities"). A possible explanation is that the AttrakDiff measures on a relatively high level appropriate for a variety of different interfaces. The SUS questionnaire also offers rather generic questions, which might be adaptable to ECAs (Brooke 1996). However, there are instrumental means to assess the cognitive load. In experimental conditions, providing a secondary task (dual task paradigm) can be used to increase the load so that indirect measures show differences in cognitive load more clearly (cf. Stevens et al. (2013) for details, especially for a method suited for ECAs).

Joy-of-use

Joy-of-use is the positive feeling a user has when using technical or interactive systems (Schleicher and Trösterer 2009) and is associated with hedonic qualities. It is related to the concept of entertainment and likeability, the

attitude of a user towards the ECA (Ruttkay et al. 2004) (cf. topic 2 from Section "Communication theories").

There is evidence for the assumptions that positive experiences during interactions are related to the fulfilment of human needs (Hassenzahl et al. 2010). Moreover, a link between need fulfilment and a system's hedonic, non-functional qualities is suggested under the precondition that the experience is attributed to the system and not to the context (e.g. the person can attribute a positive experience with a system to the device or ECA itself or to the situation or both). Other aspects of joy-of-use are aesthetics, system personality and stimulation. Aesthetics covers the "pleasure attained from sensory perception" (Hekkert 2006). The system's personality includes system factors like voice (e.g. gender of the voice), the wording of the voice prompts, or its visual appearance. Personality is particularly important in the context of the evaluation of (Ruttkay et al. 2004) concept of trust. The ECA's appeal subsumes aesthetics and personality and is related to naturalness and believability (Ruttkay et al. 2004) (cf. topic 7 from Section "Computers as social actors").

Stimulation describes a system's ability to enable personal development respectively the proliferation of knowledge and the development of skills, e.g. by providing innovative and/or exciting features (Hassenzahl 2003).

A variety of methods is available to measure joy-of use and related aspects but before deciding on a measurement method it has to be defined which aspect should be assessed. The questionnaire proposed in Lavie and Tractinsky (2004) is suitable for measuring the visual aesthetics, but not for aesthetics perceived via other modalities. The AttrakDiff (Hassenzahl et al. 2003) measures hedonic qualities on a higher level and is not limited to unimodal interfaces. For measuring hedonic qualities during the interaction the Joy-Of-Use-Button (Schleicher and Trösterer 2009) and psycho-physiological parameters are available options, the latter being the most resource-intensive method (Schleicher 2009).

Another well validated and widely used instrument is the Self-Assessment Manikin (Bradley and Lang 1994), which measures the arousal, pleasure and dominance linked to affective reactions on three non-verbal scales.

If the aim is to measure specific emotions, LemTool (Huisman and van Hout 2008) or PrEmo (Desmet 2004) may be used. However, both tools are so far only validated for specific application areas: Lemtool for websites and PrEmo for non-interactive products.

Although a wide range of methods assessing hedonic, affective qualities are nowadays available, a recent review (Bargas-Avila and Hornbæk 2011) indicates that questionnaires, more specifically the hedonic qualities sub-scales of Hassenzahl's AttrakDiff and the Self-Assessment Manikin are by far the most popular instrument.

Please note that for evaluations of affective qualities care has to be taken, when deciding whether the measurements will take place during or after the interaction. Apart from the general memory biases (e.g. consistency bias, change bias, stereotypical bias (for an overview see Schacter (2001)) several memory biases are documented regarding the retrospective assessment of emotional experiences. It was shown that retrospective reports of affective experiences are mainly based on the moment of the peak of the affect intensity and on the moment of the ending of the experience (Kahneman et al. 1993). This so-called peak-end rule could also applied in the context of interface evaluation (Hassenzahl and Sandweg 2004; Hassenzahl and Ullrich 2007). Accordingly, retrospective questionnaires reflect the remembered affective experience, but might give only little information on the specific aspects, which lead to the global evaluation (Wechsung et al. 2012c).

Usability and user experience

The ISO 9241-11 standard (ISO 9421-11 1998) defines 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".

This definition sets the focus on ergonomics and hence on the task-oriented, functional aspects of human machine interaction, which were the major themes in the early years of HCI. Satisfaction is often regarded as an affective result of high efficiency and effectiveness.

According to (Hassenzahl and Tractinsky 2006) aspects like "fun" and "experience" were already presented during the late 1980s. However, the authors also point out that it took a number of years until these ideas were adopted by the HCI community, with the new concept User eXperience (UX) becoming increasingly popular only during the last decade. The origins of the term UX probably lie in the work of Donald Norman at Apple Computers (Norman et al. 1995).

Although the term UX became omnipresent, the concept itself was neither being well defined nor well understood (Law et al. 2008). The lack of a shared view on UX (and indeed the subsequent need for one) became obvious, when many companies just exchanged the label usability with the label user experience, but kept on doing the same task-centred usability testing and engineering they did before (Hassenzahl 2008). In academia on the other hand, lots of research was conducted aiming to define UX and its preconditions. To date, relevant literature offers numerous models, theories and definition of user experience, joy-of-use, hedonic qualities or emotional design (Desmet and Hekkert 2007; Forlizzi

and Battarbee 2004; Hassenzahl et al. 2000; Jordan 2000; McCarthy and Wright 2004; Norman 2004). A survey among researchers and practitioners showed how heterogeneous the views on UX are (Law et al. 2009); however, the surveys' authors were able to deduce the following shared understanding: UX can be described,

> "as a dynamic, context-dependent and subjective [concept], which stems from a broad range of potential benefits the users may derive from a product".

A similarly broad definition is given in ISO 9421-210 (2010). Here, UX is defined as

> "a person's perceptions and responses that result from the use or anticipated use of a product, system or service".

Following the arguments above, appropriate assessing methods for usability need to measure both, joy-of use and ease-of-use. This concept of usability is difficult to distinguish from UX, which we do not consider here any further. Although several questionnaires measure ease-of-use, only a few include joy-of-use. An affect scale is included in the SUMI (Kirakowski and Corbett 1993). The Post-Study System Usability Questionnaire (PSSUQ) and the Computer System Usability Questionnaire (CSUQ) (cf. Lewis (1995)) additionally measure frustration. The AttrakDiff's Attractiveness scale, measuring pragmatic as well as hedonic qualities, is probably closest to the presented concept of usability (Hassenzahl et al. 2003). As usability – as understood in this perspective – comprises pragmatic and hedonic aspects, it is represented by the meta-concepts satisfaction and engagement (cf. Ruttkay et al. (2004)), with the relevance of each of them depending on the particular domain and application (cf. topic 6 from Section "Computers as social actors").

Apart from the questionnaires presented above, other suitable methods to assess joy-of use and ease-of-use include qualitative approaches like the Repertory Grid Technique (Kelly 1955), the Valence Method proposed by (Burmester et al. 2010) and the UX Curve (Kujala et al. 2011).

Utility and usefulness

Utility has been defined as the functionality or capability of the system (Grudin 1992), related to the tasks the user wants to accomplish with the system. Thus, usability and utility are separate concepts. This also means that an interface may have zero usability but still a high utility. For example, a software program may offer all the functions a user needs, but the interface is so bad, that the user is unable to access this function. The other way around, a highly usable interface may have no utility. Usefulness comprises both usability and utility. Accordingly, a system is useful, if it provides the functionality to support the tasks, the user wants to accomplish, and if the user can employ these function in an easy (e.g. efficient, effective) and pleasurable manner.

It is noteworthy, that the distinction between usability, usefulness, and utility is often fuzzy (Landauer 1995) and sometimes those terms are used synonymously. For example, usefulness can be defined as the

> "degree to which a given system matches the tasks it is intended to support" (Lindgaard 1994).

This definition basically equals the definition of utility presented above. Thus, it is often not clear what was measured, when results regarding one of the three concepts are reported, thus such results are difficult to interpret.

Moreover, methods measuring utility and usefulness, as understood in the first paragraph of this section, are rather rare. Particularly utility is difficult to assess in classical lab tests as typically only tasks are selected that are supported by the product (Cordes 2001). Domain relevant tasks the system is not capable of are not presented. If users are told that the given tasks may not be solvable, the users tended to terminate the task earlier and terminate more tasks than users receiving the exactly same instruction, except for the hint that the tasks may not be solvable (Cordes 2001). Thus, it is likely that the user assumes that the utility to fulfil the tasks is provided by the interface, if not stated otherwise in the instructions.

Usefulness is partly covered in the PSSUQ as well as in the CSUQ. Those questionnaires are identical, except that PSSUQ was developed for use after a usability test, and is thus addressing specific tasks, whereas the CSUQ asks about the general system and is suitable for surveys (Lewis 1995). Pedagogical ECAs, however, belong to a specific task domain. For these, usefulness is best evaluated by assessing training success with indirect measures. Note that utility and usefulness can best be assessed in field studies with real test users in natural settings, as most applications for ECAs aim at utilizing social aspects.

Acceptability

Acceptability has been described as how readily a user will actually use the system (Wechsung et al. 2012b). Quality aspects alone are not sufficient to explain whether a service will be accepted or not (Larsen 2003; Möller 2005) and thus, it is not included as quality aspect in Fig. 2. In fact, acceptability is determined by a complex interaction of multiple aspects (Larsen 2003), with the influencing factors (e.g. user, context, system) assumed to be highly relevant (Möller 2005); especially service factors like the costs of the system and, depending on the system, privacy and security issues are of major concern.

Best practice and example evaluation

Hands-on issues

The main goal of a usability evaluation is to identify possible usage problems beforehand. This section gives hands-on issues to consider, based on a recommendation (VDE-ITG-Richtlinie 2011). Information on treating participants and their data can typically be found in ethics regulations of national research councils.

One important prerequisite is that the internal structure of the system is unknown to the user. Thus, discrepancies between the assumptions and expectations of the user and the developer can be identified. This entails that the evaluation should – at a certain point – be carried out without active involvement of the developer since the influence on test procedure and/or participant can have a strong impact on results.

A multitude of usability problems arises from intransparent functions and badly designed (e.g. unnatural) visual or verbal information. A first option to uncover flaws of this kind is to let the current implementation be reviewed without bias by a colleague. This "internal evaluation" quickly reaches its limits as colleagues are often too familiar with the capabilities and logic of the ECA in question and thus do not evaluate impartially.

User tests are more costly than evaluations done by experts. At the same time it is astounding which unforeseen flaws can be uncovered by a naive user. Therefore, user tests are expedient especially for new and unusual products. Test users should correspond to the expected main user group concerning age, gender, educational background, experience with technology etc.

The quality of usability tests depends on the selection of tasks the user has to fulfil. The tasks should be precisely worded and representative for later usage. The task description should neither evoke unrealistic scenarios nor hint implicitly at the favourable approach.

The investigator should be recognizably "neutral". If this is not the case, it is possible that by suggestive questions or hints the test user is guided towards the correct solution. Also, the participant might rate the product positively believing that this is socially desired.

Furthermore, user tests (but not expert evaluations) are only of limited informational value if the function or interface is not fully functional yet. This is due to the fact that users have often difficulties imagining interaction steps which are not implemented yet, or else expect those to function perfectly. In this case the tasks should be chosen such that only already implemented functions are necessary.

User tests with application- or system-versions which are functionally not stable are of limited relevance as a user will either include system crashes in her judgement (even when asked not to do so) or else doubt the seriousness of the evaluation.

Usability tests of prototypes for which only minor adjustments are still possible should be avoided. If users realize that their suggestions for improvements have not been adopted they often react with an increased rejection.

Guideline to setting up an evaluation study

When planning an evaluation it is recommendable to answer the following questions (cf. Sonntag et al. 2004) one after the other and as thoroughly as possible:

- purpose: what shall be achieved by the evaluation?
- object: what is the test object (system)?
- time: in which development stage shall the system be evaluated?
- method: which method is suitable to reach the relevant goals?
- participants: who will be test users and how will they be found?
- help: which kind of help shall be offered to the test user and to what extent does the experimenter offer support?
- function: which functions are under test and what are the suitable tasks to evaluate them?
  - At which point is a task successfully finished?
  - Under which conditions can a task be abandoned?
- results: which data will be assessed and how? How will the data be analysed?
- setup: what kind of equipment is necessary for the evaluation (additional hard- or software, such as cameras or microphones)?
  - Where and when shall the study be carried out?
  - How long shall each run take?
  - Who will be conducting the evaluation?
  - How many participants are needed and how shall they be compensated?

Unfortunately, no answers are given in Sonntag et al. 2004. We tried to provide a starting point to answer some of these questions, i.e. on *purpose* (cf. the four evaluation topics in Section "Communication theories" and the four psychological effects in Section "Computers as social actors"), on *method* (cf. Section "Evaluation and assessment methods" on methods and Section "Usability concepts" on concepts to assess), and on *participants* (cf. Section "User characteristics" on user characteristics). For the remaining questions, it is advised to examine the methods applied in studies similar to the one planned.

An experiment testing two different talking heads

This experiment is basically a replication of an existing German one (Weiss et al. 2010) conducted for Australian English speaking participants (approved by the UWS Human Research Ethics Committee). The aim was to validate an English version of a German questionnaire to assess the perceptual quality of different talking heads. The main concepts of the questionnaire are Ease-of-Use and Joy-of-Use. In order to induce variance in ratings, we tested four different versions consisting of two different visual models and two different voices. All versions were presented to 18 participants (4 men, 14 women) as metaphors of a spoken-dialogue system controlling a smart living room. The age of the participants ranged between 18 and 48 years (M = 24.61, SD = 6.95). For their attendance they received either money or course credit. The participants were seated in front of a table inside a laboratory room which is designed for audio and video experiments. There were two screens (19") placed side by side in from of the participant. The metaphor was displayed on one. On the other screen the participants received visual information simulating feedback from an answering machine and an electronic program guide according to the task.

The participants interacted via headphones with the metaphor using free speech (instead of a fully functional system, the speech recognition and part of the interaction control was replaced by a human operator, called Wizard-of-Oz paradigm). They were asked to complete a domain-specific scenario consisting of seven different tasks with each of the four metaphors (head and voice combinations). These tasks were grouped in an answering machine scenario consisting of three tasks and an electronic program guide scenario consisting of four tasks.

The dialogue flow was controlled: the tasks were written on separate cards and offered to the participants in a predefined order. Every participant had to carry out both scenarios once with each metaphor. To avoid boredom the tasks were altered slightly in expression and content while the level of difficulty of each task remained constant. The order of scenarios was varied between participants.

After each scenario, a brief quality assessment was conducted. After both scenarios, participants had to fill in an English version of an early version of the CAS questionnaire (Wechsung et al. 2013) based on the complete interaction with one metaphor version.

The most interesting result is the evaluation of the English CAS version. Based on Cronbach's alpha, the translation into English can be confirmed in general, but there are differences to the German version. First, some of the items are better removed to obtain higher α values. Second, for some subscales, α is too low:

- Helpfulness (α=.78; .79 without "meaningful"): *impractical–practical, helpful–not helpful, destructive–constructive, useless–useful, meaningful–meaningless*
- Naturalness (α=.90): *real–not real, human like–artificial, unnatural–natural, unrealistic–realistic*

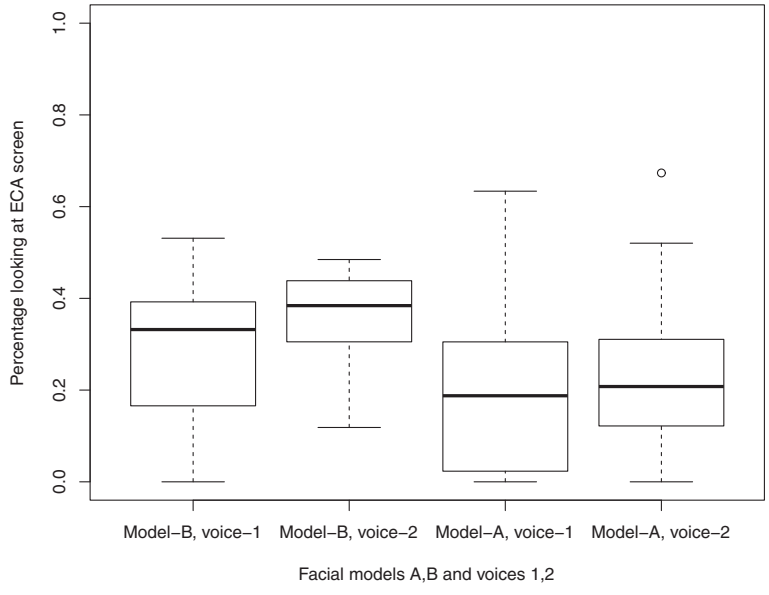

Fig. 3 Percentage of time looking at the ECA screen

- Trust (α=.61; .76 without "candid"):
  candid–uncandid (try: "sincere–insincere" (Wechsung et al. 2013)), *honest–dishonest, non credible–credible, trustworthy–untrustworthy*
- Perceived Task Difficulty (α=.75; .76 without "exhausting"):
  exhausting–not exhausting ("taxing" provided as alternative (Wechsung et al. 2013)), *undemanding–demanding, non laborious–laborious* ("effortful" provided as alternative in Wechsung et al. (2013)), *easy–difficult*
- Likeability (α=.73; due to an error we did not include "I was pleased", i.e. "likable" in Wechsung et al. (2013)):
  pleasant–unpleasant, friendly–unfriendly, likeable–not likeable (or agreeable–not agreeable (Wechsung et al. 2013))
- Entertainment (α=.68; .79 without "conventional"):
  diversified – monotonous (not monotonic (Wechsung et al. 2013)), *lame–enthralling, unconventional–conventional, boring–entertaining*

Apart from this first step towards a validated English version of CAS, this experiment also exemplifies one possible usage of eye-tracking data. Data from 16 participants was analysed. As two separate screens were used, it was of interest to obtain information on the amount of time a participant looked at the ECA screen, measured as time the eyes were tracked, which was less than 50 %. The assumption that missing data was most widely originated by the users looking at the second screen or reading the task instruction could be confirmed during the experiment.

Results show a significantly lower amount of time looking at the ECA screen for one particular facial model, but no effect for the two voices used (cf. Fig. 3). As there was no initial hypothesis, final interviews with the participants were analysed to find a reason for this difference in eye-tracking data. The results suggest that those participants familiar with model A concentrated more on other visual information, i.e. the second screen or the task descriptions. For the domain of a smart living room, the results motivate to study the development of user experience with ECAs over time more closely. The question raised is, whether initial evaluations are still valid after familiarization. However, this will require field tests instead of laboratory experiments (cf. Ring et al. (2015) for an example of a one week ECA evaluation for elderly users).

Conclusion

We presented different motivations and topics of evaluations, as well as methods and assessment instruments to conduct evaluations. Nevertheless, evaluating multimodal interactive systems, especially ECAs, is a broad field that can hardly be covered by a single article. Therefore, we concentrated on evaluative concepts and those methods related to usability. Highly recommended for further reading are the comprehensive book by Bernsen and Dybkjær (2009), the edited volume by Ruttkay and Pelachaud (2004), including not only conceptual, but also exemplary studies (Hone 2006; Poppe et al. 2014; Stevens et al. 2013) as examples for more fundamental experiments.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors drafted the manuscript and contributed to the taxonomy (displayed in Fig. 2). BW and IW were responsible for the Theoretical Foundations of Evaluation and analysed the data for the questionnaire in Best Practice. IW, BW, and CK contributed to the Evaluation and Assessment Methods, CK, SM, and BW worked in particular on log-data, annotation, and defined interaction parameters. IW was responsible for the Aspects of Evaluation. All authors read and approved the final manuscript.

Received: 29 May 2015 Accepted: 10 August 2015
Published online: 18 August 2015

References

Adcock, AB, & Eck, RNV (2005). Reliability and factor structure of the attitude toward tutoring agent scale (ATTAS). *Journal of Interactive Learning Research*, *16*(2), 195–217.
Anderson, JR, & Lebiere, C (1998). *The Atomic Components of Thought*. Hillsdale: Lawrence Erlbaum Associates.
Andrews, FM, & Whitey, SB (1976). *Social Indicators of Well-being. Americans Perception of Life Quality*. New York: Plenum Press.
Bargas-Avila, JA, & Hornbæk, K (2011). Old wine in new bottles or novel challenges? A critical analysis of empirical studies of user experience, In *Proc. ACM Conf. on Human Factors in Computing Systems (CHI)* (pp. 2689–2698).
Baylor, A, & Ryu, J (2003). The API (agent persona instrument) for assessing pedagogical agent persona, In *Proc. World Conference on Educational Multimedia, Hypermedia and Telecommunications (EDMEDIA)* (pp. 448–451).
Beringer, N, Kartal, U, Louka, K, Schiel, F, Türk, U (1997). PROMISE: A procedure for multimodal interactive system evaluation, In *Proc. Workshop on Multimodal Resources and Multimodal Systems Evaluation* (pp. 77–80).
Bernsen, NO, & Dybkjær, L (2009). *Multimodal Usability*. Human-Computer Interaction Series. London: Springer.
Bless, H, Clore, GL, Schwarz, N, Golisano, V, Rabe, C, Wölk, M (1996). Mood and the use of scripts: Does a happy mood really lead to mindlessness? *Journal of Personality and Social Psychology*, *71*(4), 665–679.
Bradley, MM, & Lang, PJ (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59.
Brave, S, & Nass, C (2007). Emotion in human-computer interaction. In A Sears & J Jacko (Eds.), *The Human-Computer Interaction Handbook. Fundamentals, Evolving Technologies and Emerging Applications*. 2nd edn: Lawrence Erlbaum.
Brennan, SA (1998). The grounding problem in conversations with and through computers. In SR Fussell & J Kreuz (Eds.), *Social and Cognitive Psychological Approaches to Interpersonal Communication*. Hillsdale: Erlbaum.
Brooke, J (1996). SUS: a "quick and dirty" usability scale. In PW Jordan, B Thomas, BA Weerdmeester, AL McClelland (Eds.), *Usability Evaluation in Industry*. London: Taylor and Francis.
Burmester, M, Mast, M, Jäger, K, Homans, H (2010). Valence method for formative evaluation of user experience, In *Proc. ACM Conf. on Designing Interactive Systems (DIS)* (pp. 364–367).
Card, SK, Moran, TP, Newell, A (1983). *The Psychology of Human-Computer Interaction*. London: Lawrence Erlbaum Associates.

Carmichael, A (1999). *Style Guide for the Design of Interactive Television Services for Elderly Viewers*. Winchester, UK: Independent Television Commission.

Chalmers, PA (2003). The role of cognitive theory in human-computer interface. *Computers in Human Behavior*, *19*(5), 593–607.

Clark, HH (1996). *Using Language*: Cambridge University Press.

Cordes, RE (2001). Task-selection bias: A case for user-defined tasks. *International Journal of Human Computer Interaction*, *13*(4), 411–419.

Costa, PTJ, & McCrae, RR (1992). *NEO PI-R Professional Manual*. Odessa: Psychological Assessment Resources, Inc.

Dehn, DM, & van Mulken, S (2000). The impact of animated interface agents: a review of empirical research. *International Journal of Human-Computer Studies*, *52*, 1–22.

Desmet, PMA (2004). From disgust to desire: How products elicit emotions. In DC Hekkert & J McDonagh van Erp (Eds.), *Proc. Int. Conf on Design and Emotion*.

Desmet, PMA, & Hekkert, P (2007). Framework of product experiences. *International Journal of Design*, *1*, 57–66.

Dix, A, Finlay, J, Abowd, G, Beale, R (1993). *Human-Computer Interaction*: Prentice Hall.

Dix, A, Finlay, J, Abowd, G, Beale, R (2003). *Human-Computer Interaction*, 3rd edn: Prentice Hall.

Dohen, M (2009). Speech through the ear, the eye, the mouth and the hand. In A Esposito, A Hussain, M Marinaro (Eds.), *Multimodal Signals: Cognitive and Algorithmic Issues*. Berlin: Springer.

Duffy, BR (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*, 177–190.

Dybkjær, L, Bernsen, NO, Minker, W (2004). Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, *43*, 33–54.

Engelbrecht, KP, Quade, M, Möller, S (2009). Analysis of a new simulation approach to dialogue system evaluation. *Speech Communication*, *51*(12), 1234–1252.

Epstein, S (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, *49*, 709–724.

Ericsson, K, & Simon, H (1980). Verbal reports as data. *Psychological Review*, *87*, 215–251.

Fisher, AG (1996). *Assessment of Motor and Process Skills (AMPS). Vol. 2: User Manual*, 5th edn. Fort Collins: Three Star Press.

Forlizzi, J, & Battarbee, K (2004). Understanding experience, In *Proc. ACM Conf. on Designing Interactive Systems (DIS)* (pp. 261–268).

Foster, MA (2007). Enhancing human-computer interaction with embodied conversational agents, In *Proc. Int. Conf. on Universal Access in Human-computer Interaction: Ambient Interaction* (pp. 828–837).

Foster, MA, Giuliani, M, Knoll, A (2009). Comparing objective and subjective measures of usability in a human-robot dialogue system, In *Proc. of Annual Meeting of the ACL Joint with the Int. Conf. on Natural Language Processing* (pp. 879–887).

Fraser, N (1997). Assessment of interactive systems. In D Gibbon, R Moore, R Winski (Eds.), *Handbook on Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.

Gediga, G, Hamborg, KC, Düntsch, I (1999). The IsoMetrics usability inventory: An operationalisation of ISO 9241-10. *Behaviour and Information Technology*, *18*, 151–164.

Gerrig, RJ, & Zimbardo, PG (Eds.) (2007). *Psychology and Life*, 18 edn. Essex: Pearson.

Gibbon, D, Mertins, I, Moore, R (Eds.) (2000). *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Norwell: Kluwer Academic Publishers.

Grudin, J (1992). Utility and usability: Research issues and development contexts. *Interacting with Computers*, *4*, 209–217.

Guerin, B (1993). *Social Facilitation*: Cambridge University Press.

Hassenzahl, M (2003). The thing and I: Understanding the relationship between user and product. In MA Blythe, K Overbeeke, AF Monk, PC Wright (Eds.), *Funology. From Usability to Enjoyment*. Dordrecht: Kluwer.

Hassenzahl, M (2008). User experience (UX): Towards an experiential perspective on product quality, In *Proc. Int. Conf. of the Association Francophone d'Interaction Homme-Machine.*

Hassenzahl, M, & Sandweg, N (2004). From mental effort to perceived usability: Transforming experiences into summary assessments, In *Proc. ACM Conf. on Human Factors in Computing Systems (CHI)* (pp. 1283–1286).

Hassenzahl, M, & Tractinsky, N (2006). User experience – a research agenda. *Behaviour and Information Technology*, *25*, 91–97.

Hassenzahl, M, & Ullrich, D (2007). To do or not to do: Differences in user experience and retrospective judgments depending on the presence or absence of instrumental goals. *Interacting with Computers*, *19*, 429–437.

Hassenzahl, M, Burmester, M, Koller, F (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität, In *Proc. Mensch and Computer. Interaktion in Bewegung* (pp. 187–196).

Hassenzahl, M, Platz, A, Burmester, M, Lehner, K (2000). Hedonic and ergonomic quality aspects determine a software's appeal, In *Proc. ACM Conf. on Human Factors in Computing Systems (CHI)* (pp. 201–208).

Hassenzahl, M, Diefenbach, S, Göritz, A (2010). Needs, affect, and interactive products - facets of user experience. *Interacting with Computers*, *22*, 353–362.

Hekkert, P (2006). Design aesthetics: Principles of pleasure in product design. *Psychology Science*, *48*, 157–172.

Holmqvist, K, Nyström, M, Andersson, R, Dewhurst, R, Jarodzka, H, van de Weijer J (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures*: Oxford University Press.

Holzinger, A (2005). Usability engineering for software developers. *Communications of the ACM*, *48*(1), 71–74.

Hone, K (2006). Animated agents to reduce user frustration: the effects of varying agent characteristics. *Interacting with Computers*, *18*(2), 227–245.

Hone, KS, & Graham, R (2000). Towards a tool for the subjective assessment of speech system interfaces (sassi). *Natural Language Engineering*, *6*, 287–303.

Huisman, G, & van Hout, M (2008). The development of a graphical emotion measurement instrument using caricatured expressions: the LEMtool, In *Proc. Int. Workshop Human-Computer Interaction* (pp. 5–8).

ISO 24617-2 (2012). *Language resource management – Semantic annotation framework (SemAF), Part 2: Dialogue acts*. Geneva: ISO.

ISO 9421-11 (1998). Ergonomic requirements for office work with visual display terminals (VDTs), Part 11: Guidance on usability specification and measures.

ISO 9421-210 (2010). Ergonomics of human system interaction, Part 210: Human-centred design for interactive systems.

John, OP, Donahue, EM, Kentle, RL (1991). *The Big Five Inventory–Versions 4a and 54*. Berkeley: University of California.

Jordan, PW (2000). *Designing Pleasurable Products*. London: Taylor and Francis.

Kahneman, D (2003a). Objective happiness. In D Kahneman, E Diener, N Schwarz (Eds.), *Well-being: Foundations of Hedonic Psychology*. New York: Russell Sage.

Kahneman, D (2003b). A psychological perspective on economics. *American Economic Review*, *93*(2), 162–168.

Kahneman, D, & Frederick, S (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T Gilovich, D Griffin, D Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press.

Kahneman, D, Fredrickson, BL, Schreiber, CA, Redelmeier, DA (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, *4*, 401–405.

Kaptelinin, V, Nardi, BA, Bödker, S, Carroll, JM, Hollan, JD, Hutchins, E, Winograd, T (2003). Post-cognitivist hci: second-wave theories, In *Proc. ACM Conf. on Human Factors in Computing Systems (CHI). Extended Abstracts*.

Karacora, B, Dehghani, M, Krämer-Mertens, N, Gratch, J (2012). The influence of virtual agents' gender and rapport on enhancing math performance, In *Proc: COGSCI* (pp. 563–568).

Karrer, K, Glaser, C, Clemens, C, Bruder, C (2009). Technikaffinität erfassen – der Fragebogen TA-EG, In *Proc. 8. Berliner Werkstatt Mensch-Maschine-Systeme* (pp. 196–201).

Kelly, GA (1955). *The Psychology of Personal Constructs*. New York: Norton.

Kieras, DE (2003). Model-based evaluation. In JA Jacko & A Sears (Eds.), *The Human-computer Interaction Handbook*. Mahwah: Lawrence Erlbaum Associates.

Kieras, DE, & Polson, PG (1985). An approach to the formal analysis of user complexity. *International Journal of Man-Machine Studies*, *22*, 365–394.

Kirakowski, J, & Corbett, M (1993). SUMI – the software usability measurement inventory. *British Journal of Educational Technology*, *24*(3), 210–212.

Koda, T, & Maes, P (1996). Agents with faces: the effect of personification, In *Proc. IEEE International Workshop on Robot and Human Communication* (pp. 189–194).

Krauss, RM, & Fussell, SR (1996). Social psychological models of interpersonal communication. In ET Higgins & A Kruglanski (Eds.), *Social Psychology: A Handbook of Basic Principles*. New York: Guilford.

Kühnel, C (2012). *Quantifying Quality Aspects of Multimodal Interactive Systems*. T-Labs Series in Telecommunication Services. Berlin: Springer.

Kujala, S, Roto, V, Väänänen-Vainio-Mattila, K, Karapanos, E, Sinnelä, A (2011). UX curve: A method for evaluating long-term user experience. *Interacting with Computers*, *23*(5), 473–483.

Landauer, TK (1995). *The Trouble with Computers: Usefulness, Usability, and Productivity*. Cambridge, USA: MIT Press.

Larsen, LB (2003). Assessment of spoken dialogue system usability – what are we really measuring? In *Proc: EUROSPEECH* (pp. 1945–1948).

Lavie, T, & Tractinsky, N (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies*, *60*(3), 269–298.

Law, E, Roto, V, Vermeeren, A, Korte, J, Hassenzahl, M (2008). Towards a shared definition of user experience, In *Proc. ACM Conf. on Human Factors in Computing Systems (CHI)* (pp. 2395–2398).

Law, E, Roto, V, Hassenzahl, M, Vermeeren, A, Korte, J (2009). Understanding, scoping and defining user experience: a survey approach, In *Proc. ACM Conf. on Human Factors in Computing Systems (CHI)* (pp. 719–728).

Le Callet, P, Möller, S, Perkis, A (2012). *Qualinet White Paper on Definitions of Quality of Experience, Version 1.1*. Lausanne, Switzerland: European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003).

Lee, L, Amir, O, Ariely, D (2009). In search of homo economicus: Cognitive noise and the role of emotion in preference consistency. *Journal of Consumer Research*, *36*, 173–187.

Lewis, JR (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*(1), 57–78.

Lewis, JR (2012). Usability testing. In G Salvendy (Ed.), *Handbook of Human Factors and Ergonomics*. 4th edn (pp. 1267–1312). New York: John Wiley.

Lin, HX, Choong, YY, Salvendy, G (1997). A proposed index of usability: a method for comparing the relative usability of different software systems. *Behaviour and Information Technology*, *16*, 267–278.

Lindgaard, G (1994). *Usability Testing and System Evaluation: A Guide for Designing Useful Computer Systems*. London: Chapman and Hall.

López-Cózar, López-Cózar Delgado, R, Araki, M (2005). *Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment*. Chinchester: Wiley.

Marakas, GM, Johnson, RD, Palmer, JW (2000). A theoretical model of differential social attributions toward computing technology: when the metaphor becomes the model. *International Journal of Human-Computer Studies*, *52*, 719–750.

Mayer, JD, & Gaschke, YN (1988). The experience and meta-experience of mood. *Journal of Personality and Social Psychology*, *55*, 102–111.

McCarthy, J, & Wright, P (2004). *Technology as Experience*: MIT Press.

Molich, R, & Dumas, JS (2008). Comparative usability evaluation (cue-4). *Behaviour and Information Technology*, *27*(3), 263–281.

Möller, S (2005). *Quality of Telephone-based Spoken Dialogue Systems*. New York: Springer.

Moraes, MC, & Silveira, MS (2006). How am i? guidelines for animated interface agents evaluation, In *Proc. IEEE/WIC/ACM Intern. Conf. on Intelligent Agent Technology (IAT)* (pp. 200–2003).

Moraes, MC, & Silveira, MS (2009). Design guidelines for animated pedagogical agents, In *Proc. IFIP World Conf. on Computers in Education (WCCE)* (pp. 1–10).

Mori, M (1970). Bukimi no tani (the uncanny valley). *Energy*, *7*, 33–35.

Morris, WN (2005). *Mood: The Frame of Mind*. New York: Springer.

Nass, C, Isbister, K, Lee, EJ (2001). Truth is beauty: Researching conversational agents. In J Cassell (Ed.), *Embodied Conversational Agents*. Cambridge, Massachusetts: MIT Press.

Naumann, A, & Wechsung, I (2008). Developing usability methods for multimodal systems: The use of subjective and objective measures, In *Proc. Int. Workshop on Meaningful Measures: Valid Useful User Experience Measurement (VUUM)* (pp. 8–12).

Nielsen, J (1994). Heuristic evaluation. In J Nielsen & RL Mack (Eds.), *Usability Inspection Methods*. New Work: John Wiley and Sons.

Nielsen, J, & Molich, R (1990). Heuristic evaluation of user interfaces, In *Proc. ACM Conf. on Human Factors in Computing Systems (CHI)* (pp. 249–256).

Noor, C (2004). Empirical evaluation methodology for embodied conversational agents: On conducting evaluation studies. In Z Ruttkay & C Pelachaud (Eds.), *From Brows to Trust: Evaluating Embodied Conversational Agents*. Dordrecht: Kluwer.

Norman, D (2004). *Emotional Design: Why We Love (or Hate) Everyday Things*. New York: Basic Books.

Norman, D, Miller, J, Henderson, A (1995). What you see, some of what's in the future, and how we go about doing it: HI at Apple Computer, Inc., In *Proc. ACM Conf. on Human Factors in Computing Systems (CHI)* (p. 155).

Op den Akker, R, & Bruijnes, M (2012). Computational models of social and emotional turn-taking for embodied conversational agents: a review. COMMIT deliverable. http://doc.utwente.nl/80451/1/Akker12computational.pdf.

Oviatt, S (2008). Multimodal interfaces. In J Jacko & A Sears (Eds.), *The Human-Computer Interaction Handbook. LNCS 5440*. Mahwah: Lawrence Erlbaum and Associates.

Pardo, D, Mencia, BL, Trapote, AH (2010). Non-verbal communication strategies to improve robustness in dialog systems: a comparative study. *Journal of Multimodal User Interfaces*, *3*, 285–297.

Picard, R (1997). *Affective Computing*. Cambridge, USA: MIT Press.

Poppe, R, Böck, R, Bonin, F, Campbell, N, de Kok, I, Traum, D (2014). The special issue: From multimodal analysis to real-time interactions with virtual agents. *Journal of Multimodal User Interfaces*, *8*, 1–3.

Preece, J, Rogers, Y, Sharp, H, Benyon, D, Holland, S, Carey, T (1994). *Human-Computer Interaction*. Wokingham: Addison-Wesley.

Price, P, Hirschman, L, Shriberg, E, Wade, E (1992). Subject-based evaluation measures for interactive spoken language systems, In *Proc. DARPA Workshop* (pp. 281–292).

Rammstedt, B, & John, OP (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. *Journal of Research in Personality*, *41*, 203–212.

Reeves, B, & Nass, C (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*: Cambridge University Press.

Richter, T, Naumann, J, Groeben, N (2000). Attitudes toward the computer: Construct validation of an instrument with scales differentiated by content. *Computers in Human Behavior*, *16*, 473–491.

Riether, N, Hegel, F, Wrede, B, Horstmann, G (2012). Social facilitation with social robots? In *Proc. ACM/IEEE Int. Conf. on Human-Robot Interaction* (pp. 41–48).

Ring, L, Shi, L, Totzke, K, Bickmore, T (2015). Social support agents for older adults: longitudinal affective computing in the home. *Journal of Multimodal User Interfaces*, *9*, 79–88.

Rogers, C (1951). *Client-Centered Therapy*. Cambridge, Massachusetts: The Riverside Press.

Ruttkay, Z, & Op den Akker, R (2004). Affordances and cognitive walkthrough for analyzing human-virtual human interaction. In A Esposito, NG Bourbakis, N Avouris, I Hatzilygeroudis (Eds.), *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction. LNCS 5042*. Heidelberg: Springer.

Ruttkay, Z, & Pelachaud, C (2004). *From Brows to Trust: Evaluating Embodied Conversational Agents*. Dordrecht: Kluwer.

Ruttkay, Z, Dormann, C, Noot, H (2004). Ecas on a common ground – a framework for design and evaluation. In Z Ruttkay & C Pelachaud (Eds.), *From Brows to Trust: Evaluating Embodied Conversational Agents*. Dordrecht: Kluwer.

Saygin, AP, Chaminade, T, Ishiguro, H, Driver, J, Frith, C (2012). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive Affective Neuroscience*, *7*(4), 413–422.

Schacter, DL (2001). *The Seven Sins of Memory: How the Mind Forgets and Remembers*. Boston: Houghton Mifflin.

Scherer, K (2004). The functions of non-verbal signs in conversation. In R St. Clair & H Giles (Eds.), *The Social and Psychological Contexts of Language*. Hillsdale: Erlbaum.

Schleicher, R (2009). *Emotionen und Peripherphysiologie*: Pabst Science Publishers.

Schleicher, R, & Trösterer, S (2009). Der Joy Of Use Button, In *Proc. Mensch and Computer* (pp. 419–422).

Schulz, V, & Thun, F (1981). *Miteinander Reden: Störungen und Klärungen. Psychologie der Zwischenmenschlichen Kommunikation*. Reinbek: Rowohlt.
Schwarz, N, & Clore, GL (2003). Mood as information: 20 years later. *Psychological Inquiry*, *14*, 296–303.
Searle, J (1969). *Speech Acts: An Essay in the Philosophy of Language*: Cambridge University Press.
Shannon, C, & Weaver, W (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.
Shannon, C, & Weaver, W (1987). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Reading: Addison-Wesley.
Sheldon, KM, Elliot, AJ, Kim, Y, Kasser, T (2002). What is satisfying about satisfying events? testing 10 candidate psychological needs. *Journal of Personality and Social Psychology*, *80*, 325–339.
Silvia, PJ, & Warburton, JB (2006). Positive and negative affect: Bridging stages and traits, In *Comprehensive Handbook of Personality* (pp. 268–284). Hoboken: John Wiley and Sons.
Smith, B, Caputi, P, Rawstorne, PR (2007). The development of a measure of subjective computer experience. *Computers in Human Behavior*, *23*, 127–145.
Sonntag, D, Jacobs, O, Weihrauch, C (2004). Usability guidelines for use case applications. Technical report, Deutsches Forschungsinstitut für Künstliche Intelligenz. Theseus Report CTC WP4, Task 4.1, MS3.
Spool, J, & Schroeder, W (2001). Test web sites: five users is nowhere near enough, In *Proc. ACM Conf. on Human Factors in Computing Systems (CHI)* (pp. 285–286).
Sproull, L, Subramani, M, Kiesler, S, Walker, JH, Waters, K (1996). When the interface is a face. *Human-Computer Interaction*, *11*, 97–124.
Stein, BE, Stanford, TR, Ramachandran, R, Perrault, TJJ, Rowland, BA (2009). Challenges in quantifying multisensory integration: Alternative criteria, models, and inverse effectiveness. *Experimental Brain Research*, *198*, 113–126.
Stevens, C, Gibert, G, Leung, Y, Zhang, Z (2013). Evaluating a synthetic talking head using a dual task: Modality effects on speech understanding and cognitive load. *International Journal in Human-Computer Studies*, *71*, 440–454.
Sturm, J (2005). On the usability of multimodal interaction for mobile access to information services. PhD thesis, Radboud University Nijmegen.
Takeuchi, A, & Naito, T (1995). Situated facial displays: Towards social interaction, In *Proc. Conf. on Human Factors in Computing Systems* (pp. 450–455).
Tractinsky, N, Katz, AS, Ikar, D (2000). What is beautiful is usable. *Interacting with Computers*, *13*, 127–145.
Van Vliet, PJA, Kletke, MG, Chakraborty, G (1994). The measurement of computer literacy - a comparison of self-appraisal and objective tests. *International Journal of Human-Computer Studies*, *40*(5), 835–857.
VDE-ITG-Richtlinie (2011). Messung und Bewertung der Usability von Kommunikationsendeinrichtungen. Technical report, Informationstechnische Gesellschaft im VDE.
Virzi, RA (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, *34*, 457–468.
Vogeley, K, & Bente, G (2010). "Artificial humans": Psychology and neuroscience perspectives on embodiment and nonverbal communication. *Neural Networks*, *23*(8–9), 1077–1090.
Walker, MA, Litman, D, Kamm, CA, Abella, A (1997). PARADISE: A general framework for evaluating spoken dialogue agents, In *Proc. of the 35th Annual Meeting of the Association of Computational Linguistics* (pp. 271–280).
Wechsler, D (2008). *Adult Intelligence Scale (WAIS-IV)*, 4th edn: Pearson.
Wechsung, I, Weiss, B, Ehrenbrink, P (2013). Development and validation of the conversational agents scale (CAS), In *Proc. Interspeech, Lyon* (pp. 1106–1110).
Wechsung, I, P., E, Schleicher, R, Möller, S (2012a). Investigating the social facilitation effect in human-robot-interaction, In *Proc. Int. Wksh on Spoken Dialogue Systems Technology (IWSDS)* (pp. 1–10).
Wechsung, I, Engelbrecht, KP, Kühnel, C, Möller, S, Weiss, B (2012b). Measuring the quality of service and quality of experience of multimodal human-machine interaction. *Journal on Multimodal User Interfaces*, *6*(1), 73–85.
Wechsung, I, Jepsen, K, Burkhardt, F, Köhler, A, Schleicher, R (2012c). View from a distance: comparing online and retrospective UX-evaluations, In *Proc. Int. Conf on Human-Computer Interaction with Mobile Devices and Services Companion (MobileHCI)* (pp. 113–118).
Weiss, B, Willkomm, S, Möller (2013). Evaluating an adaptive dialog system for the public, In *Proc. Interspeech* (pp. 2034–2038).
Weiss, B, Wechsung, I, Marquardt, S (2012). Assessing ict user groups, In *Proc. ACM NordiCHI* (pp. 1–9).
Weiss, B, Kühnel, C, Wechsung, I, Fagel, S, Möller, S (2010). Quality of talking heads in different interaction and media contexts. *Speech Communication*, *52*(6), 481–492.
Wharton, C, Rieman, J, Lewis, C, Polson, P (1994). The cognitive walkthrough method: A practitioner's guide. In J Nielsen & RL Mack (Eds.), *Usability Inspection Methods* (pp. 105–140). New York: John Wiley and Sons.
Wolters, M, Engelbrecht, KP, Gödde, F, Möller, S, Naumann, A, Schleicher, R (2010). Making it easier for older people to talk to smart homes: Using help prompts to shape users' speech. *Universal Access in the Information Society*, *9*(4), 311–325.
Woolrych, A, & Cockton, G (2001). Why and when five test users aren't enough, In *Proc. of HCI* (pp. 105–108).
Yee, N, Bailenson, JN, Rickertsen, K (2007). A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces, In *Proc. Conf. on Human Factors in Computing Systems* (pp. 1–10).