

Impact of Performance and Subjective Appraisal of Performance on the Assessment of Technical Systems

Matthias Haase¹(✉), Martin Krippel², Mathias Wahl¹,
Swantje Ferchow¹, and Jörg Frommer¹

¹ Department of Psychosomatic Medicine and Psychotherapy, Medical Faculty,
Otto-von-Guericke University Magdeburg, Magdeburg, Germany
matthias.haase@med.ovgu.de

² Institute for Psychology, Department Methodology, Psychodiagnostics and Evaluation
Research, Otto-von-Guericke University Magdeburg, Magdeburg, Germany

Abstract. Technical systems of the future are companion systems. These systems should be individualized, adaptive and accommodating. In order to create this technology, detailed evidence about users' behavior is needed. In this study, user-specific factors (user performance, subjective appraisal of user performance and user characteristics) are examined in terms of their impact on user satisfaction. In the WOZ experiment "last minute", 130 subjects interacted with a simulated speech-controlled technical system and had to complete a specific task. Over the course of the experiment, the subjects had to cope with different challenging situations. Using bivariate and point-biserial correlations, significant correlations for age and NEO-FFI personality dimension extraversion on a user's assessment of the simulated system were determined. Consequently, the postulated model could not be empirically proven, but provides important information for future studies .

Keywords: Companion system · User characteristics · Personality traits · Wizard of oz experiment · User experience · User satisfaction

1 Introduction

"Technical systems of the future are companion systems - cognitive technical systems, with their functionality completely individually adapted to each user" [1]. To gain this functionality, systems should be able to adapt to the individual abilities, preferences and needs of their users. This level of adaptiveness requires a recognition of users' situative contexts and their particular conditions. Creating such so-called cognitive technical systems is more than just a process of technical realization. It is also necessary to analyze the behavior of users when interacting with technical systems [2], the ascriptions users make to the system [3] and user characteristics that influence their interactions with technical systems [4].

Research efforts regarding user satisfaction, which is an important aspect of companion technology, are also relevant to other approaches, like usability or user experience (UX). In the present empirical study, we analyzed the impact of user-specific variables (user performance, subjective appraisal of performance and user characteristics) on satisfaction with a simulated speech-based system. First, we will give an overview of the relevance of user satisfaction in the context of interaction with technical systems.

1.1 Satisfaction

User satisfaction surveys are an established and reliable method of the quantitatively assessing technical systems [5, 6]. Taking user satisfaction into account not only serves to systematically eliminate undesired weaknesses or faults, as seen in the development of applications for mobile devices [7, 8], it also represents an essential element of several theoretical models and constructs, like the Information System (IS) Success Model [e.g. 9] or User Experience (UX) [10]. Indeed, although the IS Success Model according to DeLone and McLean [11] focused primarily on the efficiency and economy of information systems, user satisfaction still has an important function. DeLone and McClean [11] postulated that the success of an information system depends on system quality, information, quality, use, user satisfaction, individual impact and organizational impact. During the revision of the IS Success Model [9], the revision of the category “use” to “intention to use” represents a further step towards a user-oriented perspective. Intention to use is affected by user satisfaction, among other variables. According to some authors, the term usability has started to become replaced by the concept of UX [e.g. 12]. Accordingly, user satisfaction has come into the spotlight because it is part of one of the most significant influencing factors of UX.

Whether in UX or the IS Success Model, the evaluation of user satisfaction generally occurs with the help of questionnaires. The frequently-used End User Computing Satisfaction (EUCS) questionnaire can be viewed as exemplary for the IS Success Model; it consists of the scales content, accuracy, format, ease of use and timeliness.

In contrast, the AttrakDiff [13] focuses on different aspects like pragmatic or hedonic quality. However, in most assessment systems, or rather in the methods they apply, variables that influence the assessment process, like user characteristics, performance or subjective appraisal of performance [8], are not considered. This is astonishing, at least for UX, where several reports verify influences of the technical system as well as influences of users [e.g. 14].

1.2 Performance and Subjective Appraisal of Performance

As mentioned before, UX considers several factors related to both the product (e.g. technical systems) and the user (e.g. user characteristics). Winter and colleagues [14] extracted in a literature review 21 important factors influencing UX (e.g. adaptability, efficiency, originality, timeliness, transparency, identity, intuitive operation, usefulness, trust). This large number of variables is not surprising, because a common definition of UX and its factors does not exist. Usually, UX is defined as the “relationship between the product and the user” [15, S.27], “interaction between a user and a product, including the degree to which all our senses are gratified” [16, S. 57] or “primarily evaluative feeling (good-bad) while interacting with a product” [12, S. 12]. All of these descriptions focus primarily on interaction as well as well-being and disposition on the user side. However, Tullis and Albert [10] define UX as follows: “there are two main aspects of the user experience: performance and satisfaction” [10, S.44]. Satisfaction is defined similarly to the introduced concepts but performance is described as “all about what the user actually does in interacting with the product. It includes measuring the degree to

which users can accomplish a task or set of tasks successfully” [10, S.44]. In this sense, performance is based on measurements like the amount of effort, the number of mistakes or failed attempts as well as the time that was required. Thus, performance tends to be recognized in a work-related context [e.g. 17, 18]. Including user performance, the authors’ [10] perspective on UX is quite different from the general view that UX is characterized mainly by the performance of the system (e.g. adaptability, efficiency, originality). When it comes to the subjective appraisal of performance, user’s ability self-concept seems to be an important factor. Humans actively try to explain their own performance, which leads to very differential causal attributions [19]. Hence, users may refer to good or bad performance as accidental, as caused by external conditions or as a reflection of their own competence or deficits. These are exclusively cognitive processes. Szalma and Hancock [8] postulate that user characteristics influence their subjective appraisal of their performance. In turn, this appraisal of performance can influence satisfaction with a technical system. Therefore, measuring disposition or performance alone seems to be insufficient.

1.3 User Characteristics

Besides the already mentioned connection between subjective appraisal and performance, user characteristics are also associated with user performance.

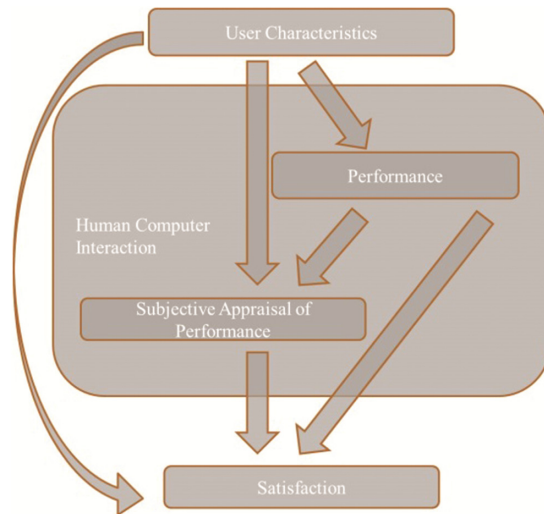


Fig. 1. Schematic illustration of user-specific factors (user characteristics, performance, subjective appraisal of performance) that influence user satisfaction.

Motowidlo and van Scooter [20] report an influence of experience and personality on performance as well. In addition, Rösner and colleagues [2] as well as Haase und colleagues [4] were able to empirically demonstrate the impact of user characteristics on performance with a technical system. More precisely, these empirical studies looked at differences

in user performance in terms of the user characteristics age, experience with computer systems as well as the NEO-FFI personality dimensions [Costa] neuroticism and agreeableness over the course of time [2, 4]. Participants with higher performance on average were younger, more experienced with technical systems and showed lower levels of neuroticism as well as higher levels of agreeableness. Due to a lack of previous empirical research, this study analyzes the interconnections between user characteristics, performance (realized as dialog success) and subjective appraisal of performance and how these influence user satisfaction, and the assessment of technical systems (Fig. 1).

2 Method

2.1 Research Questions

After examining the aforementioned findings from previous research, we generated and examined the following research questions: What impact do user characteristics and performance have on participants' subjective appraisal of performance? What impact do user characteristics, performance and subjective appraisal of performance have on the assessment of the simulated system? For statistical analysis, bivariate correlation (Spearman's rho for ordinal variables) and point-biserial correlation (rpb coefficient when one variable is dichotomous) has been used. If significant correlations are shown, multiple linear regression models will be used to explore the effect of user characteristics, performance and subjective appraisal of performance on the AttrakDiff subscales.

2.2 Sample

Basically, the sample was differentiated with regard to age and educational level. Participants were between 18 and 29 or over 60 years old and were equal distributed into a "lower educational level" (secondary school or modern secondary school certificate, apprenticeship as the highest educational/occupational qualification) and a "higher educational level" (general matriculation standard, studies at a university or a university of applied sciences). Altogether, we recruited 135 participants; three participants did not fill out the psychometric questionnaires and at two experiments we had recording problems. The final sample comprised 130 participants, one of which could not be properly assigned to a level of education.

2.3 WoZ Experiment Last Minute

We developed a Wizard of Oz experiment (WoZ) where participants had to interact with a simulated speech-controlled cognitive technical system [21]. The speech output runs via a text-to-speech system (TTS). The participants only got the information that they "are talking to a prototype of a computer program designed to assist users in dealing with everyday tasks. What is unique about this program is that it adapts itself individually to its users. For this purpose, you will be run through some tasks and test situations in the course of this session." [21, S. 19]. The interaction with the simulated system was

similar to an exploration task, because participants had to first test the skills and constraints of the system. Then, they received the task of packing luggage for a vacation with the help of the system. They could choose clothes and other pieces of luggage out of 12 categories (e.g. tops, trousers, shoes, accessories). It is suggested that they will be going on a summer vacation. Participants collect their luggage with the aid of the system. This stage is called baseline (BSL). After finishing the eighth category, the system gives the information that the luggage has exceeded the weight limit. Participants have to remove items before they can add new items. This limitation wasn't announced, and every participant got this information at the same point during the experiment. This stage is called the weight limit barrier (WLB). After the tenth category, participants are faced with another challenge situation. The system specifies that it has received delayed weather information and the participants will be going to a place with winter weather. This is called the weather information barrier (WIB). At this point, participants need to adapt their task-solving strategy under increased time pressure. Subsequent to this challenge situation, a randomized sample of participants receives an intervention focused on general psychotherapeutic factors (resource activation, problem actualization, accomplishment and clarification) [22]. After completing all categories, participants get the chance to unpack some items and replace them with more appropriate items (revision stage (RES)). In cooperation with the simulated system, participants had to solve a mundane task (packing luggage for a holiday trip) that necessitates planning, problem solving and strategy change. In the experiment, participants need to handle a large amount of interacting variables. WIB, especially, presents a complex set of problems, according to Funke [23]. This means that the problems are complex, enmeshed, dynamic and nontransparent and therefore not so easy to resolve.

2.4 Independent Variables

User Characteristics: We collected data regarding sociobiographic variables such as age, level of education and aspects like experience with computer devices. Participants also completed different questionnaires such as the NEO-FFI (regarding Big Five Personality Traits) [24].

Performance (Dialog Success): Performance was assessed by measuring the dialog success, which is operationalized as systems' reactions to users' verbal expressions [25]. "During the experiment, all contributions of speech output of the computer system, including their exact times, were logged. Afterwards, those outputs were chosen which represented a reaction of participants' interaction-contributions (e.g. phrases, single words or longer silence). This allowed a categorization of interaction-contributions of participants without regards to contents of transcripts: system is able to process a contribution (positive logs) and system is not able to process a contribution (negative logs). ... The positive logs represent all contributions of participants which could be processed by the system" [4]. Negative logs can be the result of a synonym failure (e.g. participant said shorts instead of jeans) or utterances that are not implemented [21].

“By using the experimental values, it is possible to generate a ‘log quotient’ ($Quotient_{logs} = \frac{N_{positiv\ logs}}{N_{all\ logs}}$) for each participant. The log quotient permits for an intra- and inter-individual comparison of different time stages during the experiment. A high value indicates that a participant succeeded more in adaption to the conditions” [4]. For the following analyses, participants’ performance was considered over the course of the entire experiment (BSL, WLB, WIB and RES).

Subjective Appraisal of Performance: At the end of the experiment, participants were asked for their satisfaction with the results. Participants’ statements were evaluated and divided into five categories (5. satisfied, 4. relatively satisfied, 3. neither, 2. relatively unsatisfied and 1. unsatisfied). This categorization allowed for statistical analysis.

Rating of the System (User Satisfaction): After finishing the experimental part, participants rated the simulated system with the standardized AttrakDiff [13], which means that they assessed its hedonic and pragmatically quality with the help of 28 pairs of adjectives (e.g., simple or complicated) over four subscales (pragmatic quality (PQ), hedonic quality identification (HQI) and stimulation (HQS) and the overall appeal or attraction (ATT)). A high value on a subscale means users’ requirements are met for this specific area.

3 Results

3.1 Subjective Appraisal of Performance

The bivariate correlation between user characteristics (age, gender, experiences with technical systems and NEO-FFI personality dimensions neuroticism and agreeableness), dialog success and subjective appraisal of performance revealed just a statistically significant correlation between subjective appraisal of performance and experience with technical system (Spearman’s $\rho = 0.244$, $p < 0.007$). Mentionable dialog success and subjective appraisal of performance did not correlate (Spearman’s $\rho = 0.000$, $p < 0.998$). The NEO-FFI personality dimension neuroticism and subjective appraisal of performance showed a trend toward negative correlation (Spearman’s $\rho = -0.161$, $p < 0.08$).

The findings support the view that participants with more technical experience (Fig. 2) and lower scores in NEO-FFI subscale neuroticism (Fig. 3) stated that they were more satisfied with their performance.

3.2 Assessment of a Simulated Technical System

First, a bivariate correlation between independent variables dialog success, subjective appraisal of performance and user characteristics as well as the AttrakDiff scales was examined. Age and the AttrakDiff scale HQS showing a correlation ($r_{pb} = 0.180$, $p < 0.041$) and the NEO-FFI extraversion and also the AttrakDiff scale HQS showing a negative correlation (Spearman’s $\rho = -0.180$, $p < 0.040$).

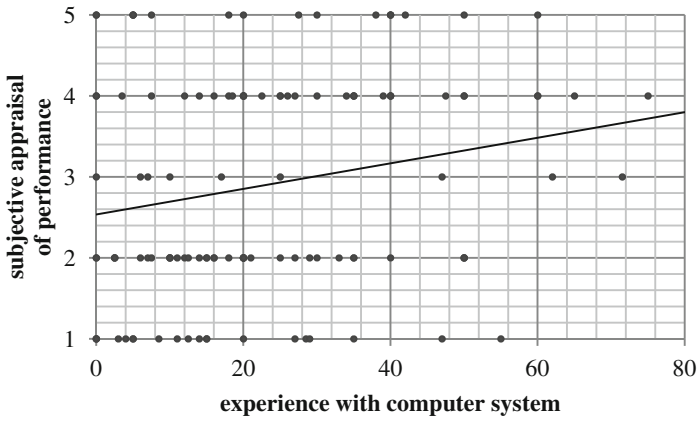


Fig. 2. Experience with computer systems (hours working with computer systems per week) and appraisal of performance (1 = not satisfied, ..., 5 = satisfied).

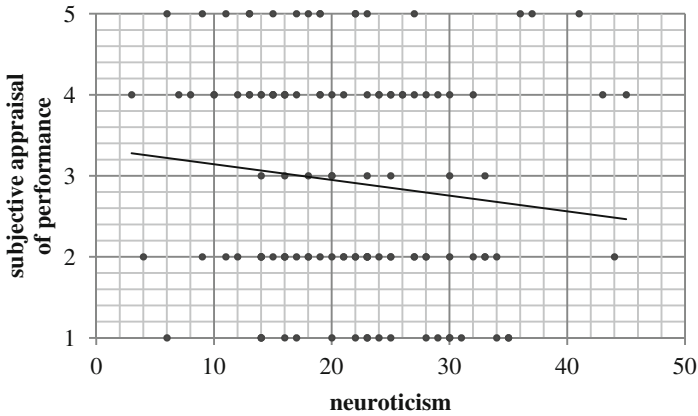


Fig. 3. NEO-FFI subscale neuroticism (high value = high level of neuroticism, low value = low level of neuroticism) and subjective appraisal of performance (1 = not satisfied, ..., 5 = satisfied)

The descriptive account showed that older participants (Fig. 4) as well as participants with lower scores in NEO-FFI subscale extraversion (Fig. 5) rated higher quality of stimulation (HQS). Due to the small number of significant correlations no linear regression models are executed.

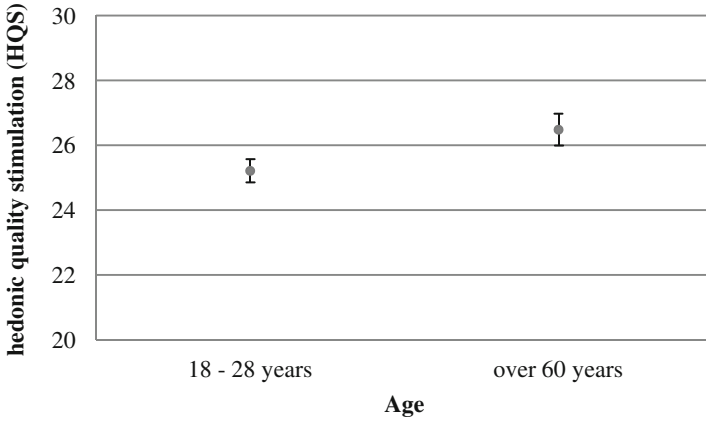


Fig. 4. Age (dichotomous, younger vs. older participants) and AttrakDiff subscale HQS (high value = high quality of stimulation, low value = low quality of stimulation).

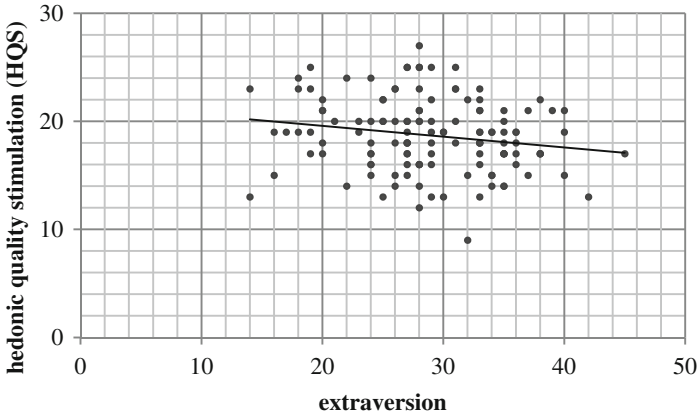


Fig. 5. NEO-FFI subscale extraversion (high value = high level of extraversion, low value = low level of extraversion) and AttrakDiff subscale HQS (high value = high quality of stimulation, low value = low quality of stimulation).

4 Conclusion

The further development of technical systems will primarily rely on characteristics such as availability, functionality, and adaptability to individual preferences, weaknesses and needs. User satisfaction represents a central aspect not only of the development of so-called companion technologies but also of other models (e.g. UX and the IS Success Model [11, 12]). Previous studies have not sufficiently considered user-specific factors that influence satisfaction with technical systems [8]. On the basis of an interdisciplinary review of the literature and of previous empirical findings [2, 4], a predominantly user-oriented model of satisfaction was developed. Both Rösner and Haase and colleagues

[2, 4] were able to find statistically significant effects of sociobiographical factors such as age and experience with computer systems as well as the NEO-FFI personality dimensions of neuroticism and agreeableness on performance (dialog success).

The present study first investigated whether user characteristics and performance influence users' appraisals of their own performance. However, statistically significant correlation could only be found with regard to experience with computer systems. Significant trends could only be found at the Big Five personality dimension neuroticism. At first glance, this is surprising, since with regard to gender, for example, Dickhäuser and Stiensmeier-Pelster [26] note that women tend to attribute failure in their work with technical systems to their own deficits, while men attribute failure to the technical system itself. This study was not able to confirm these results. With regard to satisfaction with the technical system, measured on the basis of participants' evaluations of the simulated technical system (AttrakDiff), only statistically significant correlations be found with regard to the age and Big-Five personality dimension extraversion (NEO-FFI). Nevertheless, the results of this study should not lead to a rejection of the model presented here.

Limitations: The first “last minute” experimental setup had some methodological vulnerabilities. For example, self-report (e.g. affective state, satisfaction) and assessing the simulated system in different stages of the experiment were missing. It also had been neglected to request participants' motivation and locus of control with regard to computer systems. In a revision of the “last minute” experiment these aspects have been taken into account [27].

Acknowledgements. The presented study is performed in the framework of the Transregional Collaborative Research Centre SFB/TRR 62A Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG). The responsibility for the content of this paper remains with the authors.

References

1. Wendemuth, A., Biundo, S.: A companion technology for cognitive technical systems. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) COST 2102. LNCS, vol. 7403, pp. 89–103. Springer, Heidelberg (2012)
2. Rösner, D., Haase, M., Bauer, T., Günther, S., Krüger, J., Frommer, J.: Desiderata for the design of companion systems. *KI - Künstliche Intelligenz* **30**(1), 53–61 (2016)
3. Krüger, J., Wahl, M., Frommer, J.: Making the system a relational partner: users' ascriptions in individualization-focused interactions with companion-systems. In: Berntzen, L.: Böhm, S. (eds.) Proceedings of the 8th International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC 2015), pp. 48–54. IARIA XPS Press (2015)
4. Haase, M., Krippel, M., Ferchow, S., Otto, M., Frommer, J.: Influence of user characteristics on coping with stress. In: Human-Computer Interaction. LNCS. Springer, Berlin (2016). Submission ID: 611
5. Bailey, J.E., Pearson, S.W.: Development of a tool for measuring and analyzing computer user satisfaction. *Manage. Sci.* **29**(5), 530–545 (1983)

6. Hassenzahl, M.: The thing and I: understanding the relationship between user and product. In: Blythe, M., Overbeeke, C., Monk, A.F., Wright, P.C. (eds.) *Funology: From Usability to Enjoyment*, pp. 31–42. Kluwer Academic Publishers, Dordrecht (2003)
7. Lin, C., He, Y., Pedrinaci, C., Domingue, J.: Feature LDA: a supervised topic model for automatic detection of web API documentations from the web. In: Cudré-Mauroux, P., Hefflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (eds.) *ISWC 2012, Part I. LNCS*, vol. 7649, pp. 328–343. Springer, Heidelberg (2012)
8. Szalma, J.L., Hancock, P.: Task loading and stress in human-computer interaction: theoretical frameworks and mitigation strategies. In: Sears, A., Jacko, J.A. (eds.) *Human Factors and Ergonomics. The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*, pp. 115–132. Lawrence Erlbaum Associates, New York (2007)
9. DeLone, W.H., McLean, E.R.: Information system success: the quest for the dependent variable. *Inf. Syst. Res.* **3**(1), 60–95 (1992)
10. Tullis, T., Albert, B.: *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Elsevier, Amsterdam (2013)
11. DeLone, W.H., McLean, E.R.: The DeLone and McLean model of information systems success: a ten-year update. *J. Manag. Inf. Syst.* **19**(4), 9–30 (2003)
12. Hassenzahl, M.: User experience (UX): toward an experiential perspective on product quality. In: Brangier, É., Michel, G., Bastien, J.M.C., Carbonell, N. (eds.) *IHM 2008 Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine*, pp. 11–15. ACM, New York (2008)
13. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Szwillus, G., Ziegler, J. (eds.) *Berichte des German Chapter of the ACM. Mensch & Computer 2003*, pp. 187–196. Vieweg + Teubner Verlag, Wiesbaden (2003)
14. Winter, D., Schrepp, M., Thomaschewski, J.: Faktoren der User Experience: Systematische Übersicht über produktrelevante UX-Qualitätsaspekte. In: Fischer, H., Endmann, A., Krökel, M. (eds.) *Mensch und Computer 2015 - Usability Professionals*, pp. 33–41. De Gruyter, Berlin (2015)
15. McNamara, N., Kirakowski, J.: Functionality, usability, and user experience. *Interactions*, **13**(6), 26–28 (2006)
16. Desmet, P., Hekkert, P.: Framework of product experience. *Int. J. Des.* **1**(1), 57–66 (2007)
17. LePine, J.A., van Dyne, L.: Voice and cooperative behavior as contrasting forms of contextual performance: evidence of differential relationships with big five personality characteristics and cognitive ability. *J. Appl. Psychol.* **86**(2), 326–336 (2001)
18. Borman, W.C., Motowidlo, S.J.: Task performance and contextual performance: the meaning for personnel selection research. *Hum. Perform.* **10**(2), 99–109 (1997)
19. Stiensmeier-Pelster, J., Schöne, C.: Fähigkeitsselbstkonzept. In: *Handbuch der Psychologie, Handbuch der Pädagogischen Psychologie*, vol. 10, pp. 62–73. Hogrefe, Göttingen (2008)
20. Motowidlo, S.J., van Scotter, J.R.: Evidence that task performance should be distinguished from contextual performance. *J. Appl. Psychol.* **79**(4), 475–480 (1994)
21. Frommer, J., Rösner, D., Haase, M., Lange, J., Friesen, R., Otto, M.: *Project A3 - Detection and Avoidance of Failures in Dialogues*. Pabst Science Publisher, Lengerich (2012)
22. Grawe, K.: Grundriß einer allgemeinen psychotherapie. *Psychotherapeut* **40**, 130–145 (1995)
23. Funke, J.: Complex problem solving: a case for complex cognition? *Cogn. Process.* **11**(2), 133–142 (2010)

24. McCrae, R.R., Costa, P.T.: A contemplated revision of the NEO five-factor inventory. *Pers. Individ. Differ.* **36**(3), 587–596 (2004)
25. Rösner, D., Frommer, J., Friesen, R., Haase, M., Lange, J., Otto, M.: LAST MINUTE: a multimodal corpus of speech-based user-companion interactions. In: Workshop Abstracts, LREC 2012, Istanbul, Turkey, pp. 2559–2566. ELRA, Istanbul (2012)
26. Dickhäuser, O., Stiensmeier-Pelster, J.: Erlernte hilflosigkeit am computer? geschlechtsunterschiede in computerspezifischen attributionen. *Psychologie in Erziehung und Unterricht* **27**, 486–496 (2002)
27. Ferchow, S., Haase, M., Krüger, J., Vogel, M., Wahl, M., Frommer, J.: Speech matters – psychological aspects of artificial versus anthropomorphic system voices in user-companion interaction. In: *Human-Computer Interaction*. LNCS. Springer, Berlin (2016). Submission ID: 1110