A New Approach to Measure User Experience with Voice-Controlled Intelligent Assistants: A Pilot Study

Félix Le Pailleur(✉), Bo Huang, Pierre-Majorique Léger, and Sylvain Sénécal

HEC Montréal, Montréal, Canada
{felix.le-pailleur,bo.huang,pierre-majorique.leger,
sylvain.senecal}@hec.ca

Abstract. Voice-controlled intelligent assistants use a conversational user interface (CUI), a system that relies on natural language processing and artificial intelligence to have verbal interactions with end-users. In this research, we propose a multi-method approach to assess user experience with a smart voice assistant through triangulation of psychometric and psychophysiological measures. The approach aims to develop a richer understanding of what the users experience during the interaction, which could provide new insights to researchers and developers in the field of voice assistant. We apply this new approach in a pilot study, and we show that each method captures a part of emotional variance during the interaction. Results suggest that emotional valence is better captured with psychometric measures, whereas arousal is better detected with psychophysiological measures.

Keywords: Human-Computer Interactions · Conversational user interface · User experience · Vocal assistant · Arousal · Valence · Emotion

1 Introduction

Voice assistants (e.g., Alexa, Google Assistant, Siri) are voice-controlled devices that allow consumers to use their voice to make queries such as listening to music, accessing the latest news, or answering general questions. In the U.S., it is estimated that conversational user interface (CUI) users will surpass 123 million by 2021, which represents an increase of 44% since 2017 (Petrock 2019). In addition, a recent study shows that Amazon has sold 75 million dollars' worth of smart speakers around the globe in 2018, a growth of 600% over the last year (Tung 2018).

Although voice assistants have become omnipresent in our phones, vehicles, and homes, to date, academic research that aims at developing methods to study these increasingly popular technologies is still lacking (Nass 2005; Sciuto et al. 2018; Lopatovska and Oropeza 2018; Lopatovska and Williams 2018; Jiang 2015). In fact, not all traditional methods for evaluating the user experience appears to be suited to the context of interaction with intelligent voice assistants. For instance, the "Think Aloud" method

M. Kurosu (Ed.): HCII 2020, LNCS 12182, pp. 197–208, 2020.
https://doi.org/10.1007/978-3-030-49062-1_13

(Fonteyn et al. 1993) where the researcher asks the participant to verbalize what he or she is doing and thinking while performing a task does not apply in this context since the participant is already using his/her voice to interact with the device.

Therefore, the goal of this paper is to propose a new approach to evaluate user experience during vocal interactions with voice assistants. Specifically, we propose to bonify self-reported measures used *before* and *after* the task with psychophysiological measures (i.e., electrodermal activity and micro facial expressions) to investigate the automatic and non-conscience reaction *during* the interaction. To test the feasibility of our new approach, we conducted a laboratory experiment in which participants (N = 11) were instructed to interact with Alexa. To elicit emotional reactions from participants, we designed a set of tasks likely to generate a wide range of discrete emotions.

The article is structured as follows. We first review existing research using self-reported measures in the context of voice assistant, then we discuss related work on psychophysiological measurement in the HCI literature. Next, we explain our research methodology as well as summarize the results and their interpretations in the discussion.

2 Current Research on Voice Assistants Using Self-reported Measures

Past research on user interaction with voice assistants has been using both qualitative and quantitative research methods such as questionnaires, diaries, and interviews.

Questionnaires are a widely used tool since they allow researchers to manage a large amount of data from participants quickly and inexpensively (De Singly 2016). There are several forms in which questionnaires can be presented. For example, using Likert scales, questionnaires can be quickly presented to participants before or after completing a task without hindering the flow of the experiment. In a study conducted by Jiang et al. (2015), participants were asked to complete a sequence of 10 tasks using the vocal assistant Cortana on a smartphone, and a questionnaire was used to assess frustration, success, effort, and reuse intentions. For every task, the participant only had to answer a questionnaire regarding their experience using a standard 5-point Likert scale, the most commonly used question model for measuring affective variables (Brown 2000; Burns and Grove 2005).

Similarly, diaries have also been used frequently as a method for qualitative research because it provides access to users' subjective impressions and more importantly, reflections on their interactions. This technique is advantageous since studies have shown that the presence of a stranger, e.g., researcher, might affect the way a user will interact with a voice assistant since it is mainly used in a private or comfortable context (e.g., home, with friends or alone) (Easwara Moorthy and Vu 2015). Hence, diaries offer a suitable alternative or an addition for qualitative research that might be affected by the presence of a researcher in a laboratory (Nicholl 2010). Researchers have used this method in a variety of contexts to user experience after the use of a voice assistant. For instance, Lopatovska and Williams (2018) used a diary log in studying user personification of Alexa. The study data were collected primarily through a structured online diary, which participants were asked to complete once a day for four days. The diary was also the primary method in Lau et al. (2018)'s study on users' privacy concerns when interacting with voice assistant. Through the analysis of the diary logs, they found that many non-users did not see the utility of smart speakers or did not trust speaker companies. Other studies went further. They found innovative ways to conduct data collections to understand how Alexa was used in participant's households with multiple members on a long period in a more natural way without having to report their interactions in a diary (Sciuto et al. 2018; Lopatovska and Oropeza 2018). For example, a recent study by Porcheron et al. (2018) used a Conditional Voice Recorder (CVR), a device that is activated when Alexa is turned on, to record the interaction. That way, it is possible to record multiple interactions with the voice assistant and family members in a natural context of use.

As a common tool in the HCI literature, interviews are often used as a complementary method in conjunction with the above-discussed methods. For example, in order to study user sharing practices of voice assistants, Garg and Moreno (2019) used semi-structured interviews in addition to diary logs. In a similar vein, in-depth interviews were conducted to have a better understanding of the collected conversational logs with voice assistants in investigating of Alexa's in-homer usage pattern (Sciuto et al. 2018).

Finally, observations are the only traditional methods providing a way to record user behavior during the interaction directly. For instance, in a recent study examining user interaction with voice assistants in public spaces, the area around Alexa was observed at different times of the day and different days for one week, totalling 5.5 observation hours and 132 persons observed (Lopatovska and Oropeza 2018). However, observation provides little insight on how the user feels emotionally and cognitively during the interaction.

3 Psychophysiological Measures in HCI

As presented above, most studies used qualitative or quantitative methods, mostly relying on self-reported measures. Although they provide extensive and informative results on user interaction with voice assistants, these methods alone may suffer from not precisely measuring what the user really experienced at the moment of the interaction. Researchers are calling for multi-method approaches that consider what the users really experience and perceive (Vom Brocke et al. 2020). For instance, it is possible that these results mainly "assess the user's reflection on the interaction, but not the interaction itself" (Georges et al. 2017, p. 91). Therefore, we posit that what users have really experienced might be different from their subjective evaluation of their experience.

Research in Human-Computer Interaction (HCI) has used psychophysiological measures as a viable indicator of cognitive and emotional states such as cognitive effort or frustration (Rowe et al. 1998; De Guinea et al. 2013, 2014; Giroux-Huppé et al. 2019; Beauchesne et al. 2019; Lourties et al. 2018; Agourram et al. 2019; Maunier et al. 2018). The literature has shown that user's emotional and cognitive states can also be inferred using psychophysiological signals, such as electrodermal activity (EDA), heart rate, eye-tracking, and facial expressions (see Riedl and Léger 2016 and Riedl et al. 2020).

By using self-reported measures only, researchers can face various cognitive biases such as social desirability (de Guinea et al. 2014). For example, psychologists suggest that the presence of a stranger (e.g., researcher) can change the way one will interact and, in our case, use a voice assistant to respond in the most socially desirable way (Piedmont

2014, p. 6036–6037). For example, by asking participants their likelihood to use a voice assistant in multiple environments (e.g., alone at home, in the metro or at work), Easwara and Vu (2015) found that the social context in which the interaction occurs, influence the information transmitted to the vocal assistant. Hence, psychophysiological measurement tools can contribute to overcoming bias coming from self-reported measures or observations (Xiong and Zuo 2020).

Thus, in the context of assessing the experience of users while they are interacting with a voice assistant, psychophysiological tools are an interesting add-on because they make it possible to complement traditional means of measurements (e.g., questionnaires, interviews), but especially to bring a precision on a specific emotional state, in time, to which a user cannot remember (Lourties et al. 2018). For example, it might be difficult for a participant, in the context of evaluating an intelligent voice assistant, to remember how he/she felt at a particular moment of the interaction (e.g., when he/she felt frustrated after the CUI gave an irrelevant answer to his/her question).

How users react at the moment of interacting with a device comes from unconscious and automatic mechanisms (De Guinea et al. 2013). The most accurate way to assess how they felt at one particular moment is with the psychophysiological response to the stimuli rather than their perception of what motivates their reaction (Dijksterhuis and Smith 2005).

In this research, we contribute to the literature on human interaction with voice assistants by proposing a multi-method approach to study user experience with a voice assistant by combining both psychological and psychophysiological measures, which could provide insights to researchers and developers in the field of intelligent assistants Specifically, this study leverages electrodermal activity and micro facial expressions based on Ekman's universal facial expressions (Ekman 1997) (happy, sad, angry, surprised, scared, disgusted) and emotional valence (positive-negative) in studying user experience with intelligent assistants. In the next section, we show how psychophysiological measures can offer interesting additional information to conventional self-reported measures.

3.1 Arousal

Arousal is an emotional state related to psychophysiological activity, which is linearly manifested from "calm" to "aroused" (Deng and Poole 2010; Russell 2003). Being aroused by a specific stimulus results typically in a feeling of alertness, readiness, or mobility (e.g. body movement, deep breath) (Boucsein 2012). This emotional state can be measured with Electrodermal Activity (EDA), which can assess the changes in the skin conductance response (SCR) from the nervous system functions (Braithwaite et al. 2013; Dawson et al. 2000; Bethel 2007). It is an easy to use and reliable psychophysiological measure that has been widely used in NeuroIS research (Léger et al. 2014; Brocke et al. 2013; Giroux-Huppé et al. 2019; Lamontagne et al. 2020). Arousal can also be measured perceptually by using the self-reported measure such as the Self-Assessment manikin rating (SAM), in which users report their perceived emotional state for a specific stimulus, such as excited, wide-awake, neutral, dull, calm (Bradley and Lang 1994).

However, the main advantage of using a psychophysiological measure to assess arousal is that it is not invasive, requires no overt behaviour to be recorded, and offers an ecologically valid portrait of the user's arousal, at any time during an experiment

(Dirican and Göktürk 2011). For instance, in a study on child-robot interaction, Leite et al. (2013) measured user's arousal through skin conductance and found that such a method is valuable and reliable for capturing interaction with social robots. Also, it can be used to complement and validate traditional survey methods (e.g. questionnaires).

Moreover, in a study measuring the effects of time pressure and accuracy using a computer mouse, participants were asked to paint rectangles with a decreasing time limit. Heiden et al. (2005) found that there was a significant difference in electrodermal data between task difficulty levels. Finally, in a study providing a systematic assessment of IS construct validity, de Guinea et al. (2013) found that the convergent validity of arousal was evidenced by the significant correlation between the SAM scale and the electrodermal data.

3.2 Valence

Emotional valence refers to the emotional response, with negative emotions (e.g., fear, anger, sadness) on one side of the spectrum and positive emotions (e.g., joy, surprise) on the other, to a specific stimulus (Lane et al. (1999). Valence can easily be measured perceptually with self-reported measure (e.g., SAM Scale) as the intensity of positive emotions minus the intensity of negative emotions expressed within a range from -1 to 1 (Bradley and Lang 1994). Another way to measure valence is by interpreting facial expressions, which are expressed by the micro-movements of facial muscles (e.g. frowning when angry) (Ekman 1993). It used to be that the only way to interpret facial expressions was via a trained observer who would observe and note changes in facial expressions based on the Facial Action Coding System (FACS) by Ekman and Friesen (1997).

Today, this time-consuming method is replaced with automatic facial analysis tools (AFA), which can automatically recognize the small changes in facial action units (e.g. raising a brow, chin raise, jaw drop, etc.) and interpret data based on the FACS (Cohn and Kanade 2007, Ekman 1997).

This technology allows us to accurately detect facial expressions in real-time by distinguishing between a set of discrete emotions such as angry, happy, disgusted, sad, scared, surprised. For example, Danner et al. (2014) used this technology to examine participants' facial reactions when tasting orange juice samples to compare implicit measures from the tool with explicit measures from the questionnaire. They found that the software was accurate to report changes in the participant's micro facial expressions between the different samples. Zaman and Shrimpton-Smith (2006) found that, compared to a user questionnaire, data captured by facial micro-expressions is more effective in measuring instant emotions and fun of use. Also, their results suggest that questionnaire data was instead a reflection of the outcome of a task, than a genuine self-reflection of how the user felt when accomplishing the task. Similarly, in a recent study, Lourties et al. (2018) explored the convergent validity of self-reported measures with psychophysiological measures. Their results suggest that the experience lived by a participant is not the same as it is reported. Users self-evaluate their emotional valence more accurately at the end than at the beginning of a task, while they evaluate their arousal more accurately only at the beginning of a task.

To the best of our knowledge, no studies have yet used automatic facial analysis in conjunction with the precise triangulation of electrodermal activity to study user experience with a voice assistant. The proposed triangulated method could provide new insights for this learning or evaluation context using voice only.

4 Method

To test the feasibility of using psychophysiological measures in conjunction with psychometric measures to evaluate user experience with voice assistant, we conducted a pilot laboratory experiment where participants were invited to actively interact with Alexa through Amazon's (Amazon Inc, Seattle, WA) Echo Dot (3rd generation) device by completing a series of tasks. A total of 11 subjects participated in the experiment (4 males, 7 females, mean age = 24; sd = 5.48) and received a $20 gift card as compensation. This project was approved by the IRB of our institution.

4.1 Participants and Design

Since this is a feasibility study, and we wanted to generate as much as variance in the data, we designed a within-subject experiment where each participant was instructed to perform a sequence of interactions. The experiment has one factor with two conditions: impossible tasks (i.e., queries that Alexa was unable to complete) and possible tasks (i.e., queries that Alexa was able to complete) in order to induce negative emotions such as frustration. Participants were randomly assigned to two different sets of tasks wherein one condition, they completed possible tasks before impossible tasks and in the other condition, we reversed the sequence. During the experiment, participants completed a set of 8 interactions in total.

4.2 Procedure and Measures

Participants were informed that they would have to complete a total of 8 tasks. The goal of each task was explained under the form of pictograms on a tablet.

Participants completed a short questionnaire after each interaction as well as a final questionnaire at the end of the study, followed by a brief interview. To measure user perceptions, the 5-point Self-Assessment Manikin (SAM) scale (Bradley and Lang 1994) was used. The tool allows to directly measure a person's perceived emotional reaction to a stimulus, such as valence and arousal. Respectively, the scales range from sad (1) happy (5) and calm (1) to excited (5).

For the psychophysiological arousal measure, we collected EDA with a Biopac MP-160 (Biopac, Goleta, USA) device with pre-gel sensors placed on the palm of the participant's non-dominant hand to capture changes in skin conductivity.

Electrodermal measures were standardized using as a reference a baseline captured on each participant before the experiment. The baseline consists of measuring the normal electrodermal activity unique to each participant, so that variations from the baseline can be compared. Also, results were rescaled from −1 to 1 for analysis purpose.

Finally, psychophysiological emotional valence was captured via micro facial expressions with the software FaceReader (Noldus, Wageningen, Netherlands). This non-obtrusive method can detect up to six emotions: happy, sad, angry, surprised, scared, and disgusted. Valence value was calculated by subtracting the value of the "happy" emotion and the value of the highest negative emotion (Noldus, FaceReader).

Since the objective of this study is to investigate user experience at the moment of interaction with a voice assistant, only psychophysiological measures that were captured at the moment of listening to Alexa's answers were retained for analysis. It is the participant's reactions to the response given by the voice assistant that interests us.

4.3 Material and Apparatus

The apparatus was installed in a quiet room with a mirror window, to reduce noise or external stimulation to make sure there was no interruption and that our psychophysiological data would be good quality (see Fig. 1 for a detailed setup).

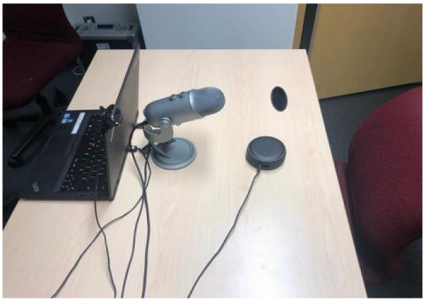

Fig. 1. Experimental setup

Our experimental setup was composed of an Alexa device, a microphone, mounted with a camera, and a digital tablet was installed. During the experiment, participants were interacting with the device. Facial expressions during the experiment were captured using a Logitech camera (Newark, USA), and recorded with the software Media Recorder (Noldus, Wageningen, Netherlands). The software Observer XT (Noldus, Wageningen, Netherlands) and CubeHX (Montréal, Canada) was used to precisely and temporally synchronize all psychophysiological measurements, in line with the guidelines proposed by Léger and colleagues (Léger et al. 2014, 2019; Courtemanche et al. 2018). Statistics were performed using the Statistical Analysis System 9.4 (SAS Inst., U.S.A.).

5 Results

To analyze the data, we first performed several linear mixed-effects regressions where each of the measures was entered as a dependent variable (see Table 1 for detailed results). For self-reported measures, namely the valence and arousal, we found that participants reported significantly more positive valence in the possible tasks, compared to impossible tasks (t (76) = −3.77, p < .001), which was expected. This suggests that participants felt more positive emotions than negative emotions when having successful interactions with the voice assistant. However, arousal did not show a significant difference (t (76) = 0.54, p = .59, NS) between the two task sets.

Table 1. Summary of results: means standard deviation and linear regression

	Possible tasks	Impossible tasks	Estimate	Std. error	t-value	p-value
Valence (self-reported)	3.65 (0.96)	3 (0.83)	−0.65	0.18	−3.77	p < .001
Arousal (self-reported)	2.45 (0.96)	2.36 (0.93)	−.09	0.17	−0.54	p = .59
Arousal (Psychophysiological)	−0.01 (0.33)	0.07 (0.30)	0.08	0.01	7.46	p < .0001
Valence (Psychophysiological)	0.03 (0.35)	0.004 (0.31)	0.01	0.01	−0.94	p = 0.35

Note: Standard deviations are reported in parentheses.

For psychophysiological measures, arousal results suggest that impossible tasks generate much higher EDA than possible tasks (t (2638) = 7.46, p < .0001). This means that participants experienced a much higher aroused emotional state when they were having difficulties during their interactions. However, in terms of the valance, we did not find a significant difference between possible and impossible tasks (t (1776) = −0.94, p = .35, NS). The following table presents the descriptive statistics and regression results.

In order to understand the relationship between the two self-reported measures and the psychophysiological measures, we conducted two additional linear mixed-effects regression analyses. The results showed that the self-reported arousal is positively correlated with psychophysiological arousal (t (2638) = 3.82, p < .0001). However, surprisingly, our analysis revealed that self-reported valence was negatively correlated with psychophysiological valence (t (1776) = −5,09 ρ < .0001).

6 Discussion

Our main contribution with this methodological paper is through the triangulation of psychological and psychophysiological measures since, to the best of our knowledge, this study is the first to compare results from both psychophysiological and self-reported measures in the context of user interaction with a voice assistant. Specifically, we found that for arousal, results from EDA showed a significant difference between possible tasks and impossible tasks (but the self-reported measure did not capture such difference). In contrast, for valence, the self-reported measure was more effective than the automatic facial analysis (AFA) in detecting variance in valence. Since previous studies mainly used self-reported measures in studying user interaction with voice assistant, our study contributes by showing the benefit of a multimethod approach in this context, as each method captures a distinct emotional dimension. This suggests that during interaction with a voice assistant, what users experienced might not be exactly the same as reported by themselves. We note that this finding is in line with previous research that combining both methods in studying similar emotional states (i.e., arousal and valence) (Lourties et al. 2018).

Also, the results suggest that the self-perceived arousal was consistent with the psychophysiological responses measured with electrodermal activity when combing both task sets, as they showed a significant positive correlation. These results support previous findings in HCI research using EDA and extend these findings in user interaction context with voice assistants. For example, De Guinea et al. (2013) found that the convergent validity of arousal was evidenced by the significant correlation between the SAM scale measure and the electrodermal measure. Such correlation was evidenced in the current research as well.

Moreover, our results indicate that the emotions inferred from the user's facial expressions by AFA during the interaction complement the self-perceived emotional valence reported by the users. However, we note that there is a discrepancy between valence inferred based on AFA and the reported by questionnaire. For example, they are negatively correlated in general when combining both tasks. To investigate this surprising result, we conducted further observation analysis by analyzing the video recordings of our participants performing the tasks. We found a tendency of several participants smiling when they were not able to complete an impossible task, but a smile emanating from frustration rather than joy, which would be aligned with self-reported valence results.

As a future research avenue, researchers have found a way to overcome this kind of situation by focusing on a new set of emotions called epistemic. For example, D'Mello and Calvo (2013) report in their E-learning study with students that "boredom," "confusion," "curiosity," "happiness," and "frustration" where the most common affective states felt during learning and reading situations. In particular, the affective state of "confusion" might be interesting to test in our context since there can be much discrepancy between what the participant expects to get as an answer and the actual answer given by the intelligent voice assistant since speech recognition is not yet optimal. We are currently running a new study where we are considering the affective states "boredom," "confusion," and "curiosity."

Our experience is limited by the fact that it took place in a user experience laboratory. Thus, the user experience may have been slightly different than if it had taken place in a more natural setting. Future research could extend the current study to other real-life settings such as home and office where interaction with voice assistant is more frequent. In addition, our experiment only measured EDA and facial expressions, while many other tools and measurements suggested by the literature still need to be tested in our specific study context. Hence, it would be interesting for future research to consider a more natural set up and to add more psychophysiological tools. Also, rarely do voice assistant users use their device without performing other tasks at the same time. The main advantage of this tool is that it allows the user to perform a vocal command when he can perform something else simultaneously (e.g. walking, driving or listening to television). In our opinion, the idea of adding pupillometry to measure cognitive load (Sirois and Brisson 2014; Léger et al. 2018) in a multi-tasking context using a vocal assistant would be an excellent contribution to the research in HCI.

References

Agourram, H., Alvarez, J., Sénécal, S., Lachize, S., Gagné, J., Léger, P.-M.: The relationship between technology self-efficacy beliefs and user satisfaction – user experience perspective. In: Kurosu, M. (ed.) HCII 2019. LNCS, vol. 11568, pp. 389–397. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22636-7_29

Beauchesne, A., et al.: User-centered gestures for mobile phones: exploring a method to evaluate user gestures for UX designers. In: Marcus, A., Wang, W. (eds.) HCII 2019. LNCS, vol. 11584, pp. 121–133. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23541-3_10

Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry **25**(1), 49–59 (1994)

Braithwaite, J.J., Watson, D.G., Jones, R., Rowe, M.: A guide for analyzing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. Psychophysiology **49**(1), 1017–1034 (2013)

Brocke, J.V., Riedl, R., Léger, P.M.: Application strategies for neuroscience in information systems design science research. J. Comput. Inf. Syst. **53**(3), 1–13 (2013)

Brown, J.D.: What issues affect Likert-scale questionnaire formats. Shiken JALT Test. Eval. SIG Newslett. **4**(1), 27–30 (2000)

Burns, N., Grove, S.K.: The Practice of Nursing Research: Conduct, Critique and Utilization. Elsevier, Amsterdam (2005)

Bethel, C.L., Salomon, K., Murphy, R.R., Burke, J.L.: Survey of psychophysiology measurements applied to human-robot interaction. In: 16th IEEE International Symposium on Robot & Human Interactive Communication (2007)

Nass, C.: Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship. MIT Press, Cambridge (2005)

Cohn, J.F., Kanade, T.: Use of automated facial image analysis for measurement of emotion expression. In: Handbook of Emotion Elicitation and Assessment, pp. 222–238 (2007)

Courtemanche, F., Léger, P.-M., Dufresne, A., Fredette, M., Labonté-LeMoyne, É., Sénécal, S.: Physiological heatmaps: a tool for visualizing users' emotional reactions. Multimedia Tools Appl. **77**(9), 11547–11574 (2018)

Danner, L., Sidorkina, L., Joechl, M., Duerrschmid, K.: Make a face! Implicit and explicit measurement of facial expressions elicited by orange juices using face reading technology. Food Qual. Prefer. **32**, 167–172 (2014)

De Guinea, A.O., Titah, R., Leger, P.M.: Explicit and implicit antecedents of users' behavioral beliefs in information systems: a neuropsychological investigation. J. Manag. Inf. Syst. **30**(4), 179–210 (2014)

Guinea, D., Ortiz, A., Titah, R., Léger, P.-M.: Measure for measure: a two study multi-trait multi-method investigation of construct validity in IS research. Comput. Hum. Behav. **29**(3), 833–844 (2013)

De Singly, F.: Le questionnaire, 4th edn. Armand Colin, Paris (2016)

Dirican, A.C., Göktürk, M.: Psychophysiological measures of human cognitive states applied in human-computer interaction. Procedia Comput. Sci. **3**, 1361–1367 (2011)

Easwara Moorthy, A., Vu, K.-P.L.: Privacy concerns for use of voice activated personal assistant in the public space. Int. J. Hum.-Comput. Interact. **31**(4), 307–335 (2015)

Ekman, P.: Facial expression and emotion. Am. Psychol. **48**(4), 384 (1993)

Ekman, P., Keltner, D.: Universal facial expressions of emotion. In: Segerstrale, U., Molnar, P. (eds.) Nonverbal Communication: Where Nature Meets Culture, pp. 27–46. Lawrence Erlbaum Associates, Inc., Mahwah (1997)

Fonteyn, M.E., Kuipers, B., Grobe, S.J.: A description of think aloud method and protocol analysis. Qual. Health Res. **3**(4), 430–441 (1993)

Georges, V., Courtemanche, F., Sénécal, S., Léger, P.M., Nacke, L., Pourchon, R.: The adoption of psychophysiological measures as an evaluation tool in UX. In: International Conference on HCI in Business, Government, and Organizations, pp. 90–98. Springer, Cham, July 2017

Giroux-Huppé, C., Sénécal, S., Fredette, M., Chen, S.L., Demolin, B., Léger, P.-M.: Identifying psychophysiological pain points in the online user journey: the case of online grocery. In: Marcus, A., Wang, W. (eds.) HCII 2019. LNCS, vol. 11586, pp. 459–473. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23535-2_34

Jiang, J., et al.: Automatic online evaluation of intelligent assistants. In: Proceedings of the 24th International Conference on World Wide Web - WWW 2015. Presented at the 24th International Conference (2015). https://doi.org/10.1145/2736277.2741669

Lamontagne, C., et al.: User test: how many users are needed to find the psychophysiological pain points in a journey map? In: Ahram, T., Taiar, R., Colson, S., Choplin, A. (eds.) IHIET 2019. AISC, vol. 1018, pp. 136–142. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-25629-6_22

Lane, R.D., Chua, P.M., Dolan, R.J.: Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. Neuropsychologia **37**(9), 989–997 (1999)

Lau, J., Zimmerman, B., Schaub, F.: Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. In: Proceedings of the ACM on Human-Computer Interaction, vol. 2, no. CSCW, pp. 1–311 (2018)

Léger, P.-M., Courtemanche, F., Fredette, M., Sénécal, S.: A cloud-based lab management and analytics software for triangulated human-centered research. In: Davis, F.D., Riedl, R., vom Brocke, J., Léger, P.-M., Randolph, A.B. (eds.) Information Systems and Neuroscience. LNISO, vol. 29, pp. 93–99. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-01087-4_11

Léger, P.M.,Davis, F.D., Cronan, T.P., Perret, J.:Neuropsychophysiological correlates of cognitive absorption in an enactive training context. Comput. Hum. Behav. **34**, 273–283 (2014)

Léger, P.-M., Charland, P., Sénécal, S., Cyr, S.: Predicting properties of cognitive pupillometry in human–computer interaction: a preliminary investigation. In: Davis, F.D., Riedl, R., vom Brocke, J., Léger, P.-M., Randolph, A.B. (eds.) Information Systems and Neuroscience. LNISO, vol. 25, pp. 121–127. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-67431-5_14

Lopatovska, I., Oropeza, H.: User interactions with "Alexa" in public academic space. Proc. Assoc. Inf. Sci. Technol. **55**(1), 309–318 (2018). https://doi.org/10.1002/pra2.2018.14505501034

Lopatovska, I., Williams, H.: Personification of the Amazon Alexa: BFF or a mindless companion. In: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, pp. 265–268. ACM, March 2018

Lourties, S., Léger, P.-M., Sénécal, S., Fredette, M., Chen, S.L.: Testing the convergent validity of continuous self-perceived measurement systems: an exploratory study. In: Nah, F.F.-H., Xiao, B.S. (eds.) HCIBGO 2018. LNCS, vol. 10923, pp. 132–144. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91716-0_11

Maunier, B., et al.: Keep calm and read the instructions: factors for successful user equipment setup. In: Nah, F.F.-H., Xiao, B.S. (eds.) HCIBGO 2018. LNCS, vol. 10923, pp. 372–381. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91716-0_29

Dawson, M.E., Schell, A.M., Filion, D.L.: The electrodermal system. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (eds.) Handbook of Psychophysiology, vol. 2. Cambridge University Press, Cambridge (2000)

Nicholl, H.: Diaries as a method of data collection in research. Paediatr. Care **22**(7), 16–20 (2010). https://doi.org/10.7748/paed2010.09.22.7.16.c7948

Noldus FaceReader methodology. https://info.noldus.com/free-white-paper-on-facereader-methodology

Petrock, V.: Voice Assistant Use Reaches Critical Mass. Retrieved from e-Marketer database, 15 August 2019

Piedmont, R.L.: Social desirability bias. In: Encyclopedia of Quality of Life and Well-Being Research, pp. 6036–6037 (2014). https://doi.org/10.1007/978-94-007-0753-5_2746

Riedl, R., Fischer, T., Léger, P.-M., Davis, F.: A decade of NeuroIS research: progress, challenges, and future directions. Data Base Adv. Inf. Syst. **51** (2020, in press)

Rowe, D.W., Sibert, J., Irwin, D.: Heart rate variability: indicator of user state as an aid to human-computer interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 480–487. ACM Press/Addison-Wesley Publishing Co, January 1998

Sciuto, A., Saini, A., Forlizzi, J., Hong, J.I.: Hey Alexa, what's up? In: Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS 2018. Presented at the 2018 (2018). https://doi.org/10.1145/3196709.3196772

Sirois, S., Brisson, J.: Pupillometry. Wiley Interdisc. Rev. Cogn. Sci. **5**(6), 679–692 (2014)

Tung, L.: Amazon: We sold tens of millions of Echo devices in 2018, and Alexa has now 70 000 skills. Retrieved from ZDnet database, 20 December 2018

Vom Brocke, J., Hevner, A., Léger, P.M., Walla, P., Riedl, R.: Advancing a NeuroIS research agenda with four areas of societal contributions. Eur. J. Inf. Syst. **29**, 9–24 (2020)

Xiong, J., Zuo, M.: What does existing NeuroIS research focus on? Inf. Syst. **89**, 101462 (2020). https://doi.org/10.1016/j.is.2019.101462

Zaman, B., Shrimpton-Smith, T.: The FaceReader: measuring instant fun of use. In: Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles, Chicago, pp. 457–460. ACM, October 2006

D'Mello, S., Calvo, R.A.: Beyond the basic emotions: what should affective computing compute? In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, pp. 2287–2294 (2013)

Heiden, M., Lyskov, E., Djupsjöbacka, M., Hellström, F., Crenshaw, A.G.: Effects of time pressure and precision demands during computer mouse work on muscle oxygenation and position sense. Eur. J. Appl. Physiol. **94**(1–2), 97–106 (2005)

Leite, I., Henriques, R., Martinho, C., Paiva, A.: Sensors in the wild: exploring electrodermal activity in child-robot interaction. In: 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 41–48. IEEE, March 2013

Boucsein, W.: Electrodermal Activity, 2nd edn. Springer, Boston (2012). https://doi.org/10.1007/978-1-4614-1126-0

Russell, J.A.: Core affect and the psychological construction of emotion. Psychol. Rev. **110**(1), 145 (2003)

Deng, L., Poole, M.S.: Affect in web interfaces: a study of the impacts of web page visual complexity and order. MIS Q, 711–730 (2010)

Dijksterhuis, A., Smith, P.K., Van Baaren, R.B., Wigboldus, D.H.: The unconscious consumer: effects of environment on consumer behavior. J. Consum. Psychol. **15**(3), 193–202 (2005)

Garg, R., Moreno, C.: Exploring everyday sharing practices of smart speakers. In: IUI Workshops, January 2019

Porcheron, M., Fischer, J.E., Reeves, S., Sharples, S.: Voice interfaces in everyday life. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–12 April 2018

Riedl, R., Léger, P.M.: Fundamentals of NeuroIS. Studies in Neuroscience, Psychology and Behavioral Economics. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-45091-8