CHAPTER 9

WHAT EYE MOVEMENTS AND FACIAL EXPRESSIONS TELL US ABOUT USER-FRIENDLINESS: TESTING A TOOL FOR COMMUNICATORS AND JOURNALISTS

Jenny Lindholm, Klas Backholm and Joachim Högväg

ABSTRACT

Technical solutions can be important when key communicators take on the task of making sense of social media flows during crises. However, to provide situation awareness during high-stress assignments, usability problems must be identified and corrected. In usability *studies, where researchers investigate the user-friendliness of a product, several types of data gathering methods can be combined. Methods may include subjective (surveys and observations) and psychophysiological (e.g. skin conductance and eye tracking) data collection. This chapter mainly focuses on how the latter type can provide detailed clues about user-friendliness. Results from two studies are summarised. The tool tested is intended to help communicators and journalists with monitoring and handling social media content during times of crises.*

Keywords: Laboratory testing; psychophysiology; situation awareness; social media; usability testing; biometric measurement

INTRODUCTION AND TEST METHODS

Social media can be of assistance during crisis communication when individuals, organisations, the mass media and government agencies need to seek or provide information and engage in dialogue. Research on the use of social media in crisis situations shows that both the public (Reuter & Spielhofer, 2017) and key communicators (Haataja, Laajalahti, & Hyvärinen, 2016) increasingly turn to social media during times of crisis.[1] Yet, social media provide several challenges and obstacles for communication (Spence, 2016). A recent example is Hurricane Harvey that hit Texas with severe flooding in August 2017. Misinformation was spread on social media, as well as by news agencies, for example, through outdated photos, a hoax in the form of support information for victims and fake images of sharks swimming up the freeways (Qiu, 2017). Arguably, mass media and authorities no longer have control over information production, resulting in higher stakes for credibility. Hence, a central challenge to both journalists and other key communicators is to develop relevant strategies to monitor and verify the enormous flow of information.

Technical solutions can simplify this task. However, especially during high-stress assignments, the usability of such solutions, for example, monitoring tools, must be excellent. With rising stress levels, the ability to multi-task is lowered and the focus remains on the main task at hand. In a crisis situation, it is important that as much as possible of the user's cognitive abilities are available for critical decision-making and maintaining situation awareness. Situation awareness can be defined as 'being aware of what is happening around you and what that information means to you now and in the future' (Endsley & Jones, 2016, p. 13). This means that a complex or poorly designed tool can occupy the user's cognitive abilities to an unnecessary degree and cause added stress, which is a detriment to maintaining situation awareness.

In this chapter, the term usability is understood as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use (ISO 9241, 2010). In other words, a tool such as a computer program, a web page or a physical product with good usability helps the user reach their goals, and does so while causing the least possible amount of cognitive load. Reaching a state of cognitive overload, partly caused by usability problems, might lead to dangerous errors (Mendl, 1999). Using a tool should make the user feel positive and content – what is often referred to as a good user experience. Achieving this means that preplanning and research before designing such a product are important for finding out *who* the users are, *what* their goals are and *how* they want to go about reaching them. In the RESCUE project, the research on user goals has been done in the mapping phase. For more information on this, as well as a more extensive discussion on usability and the features of the RESCUE tool, please refer to Chapter 8 in this volume.

This chapter focuses on a specific part of usability testing, mainly tool prototype testing in a laboratory setting. A laboratory enables the researchers to measure, for example, the users' stress levels using psychophysiological measurements, in a simulated crisis situation. Current best practices for usability testing in laboratories are discussed and illustrated through two tests conducted within the RESCUE project.

LABORATORY TESTING OF USABILITY

To determine if a product's usability goals are achieved, its performance can be measured through predefined tasks. Such tests can be conducted in a laboratory setting or through field studies and there is an extensive discussion on the distinction between these methods in the human-computer interaction (HCI) literature (see, e.g. Kjeldskov & Skov, 2014). Traditionally, task performance tests, where the user is asked to perform a certain task and rated on for example task completion and task time, are often conducted in a laboratory setting (Dumas & Redish, 1999; Rubin, 1994), since such a setting provides a more controlled environment (Duh, Tan, & Chen, 2006). Laboratory studies are conducted with precise instruments and a high replicability. In contrast, field studies take place in 'the real world', providing rich and grounded data, as well as ecological validity (Kjeldskov & Skov, 2014). Interestingly, a study conducted by Tullis, Fleischman, McNulty, Cianchette, and Bergel (2002) compared laboratory tests to field usability testing of websites and found the behaviour of test participants to be strikingly similar in both settings. When completing the tasks, the users encountered similar problems and devoted a similar amount of time to solve them. Hence, different environments do not necessarily lead to different kinds of behaviour. Nevertheless, some measurements are better suited for a laboratory setting, such as direct observations to capture certain elements or behaviour (Tullis et al., 2002), as well as psychophysiological measurements (Cacioppo, Tassinary, & Berntson, 2007). These kinds of measurements are the main focus of this chapter.

PSYCHOPHYSIOLOGICAL MEASUREMENTS FOR STRESS DETECTION

In the field of HCI, the use of psychophysiological methods is both effective and increasingly popular (e.g. Jacucci, Fairclough, & Solovey, 2015; Silva, Fairclough, Holzinger, Jacob, & Tan, 2015). The term psychophysiological means using measurements of physiology to interpret the states of the mind. Cowley et al. (2016) describe this process as aiming at extracting quantitative indices of essentially qualitative, cognitive or affective states. The indices of emotional states can be either internal – and derive from the autonomic nervous system and the central nervous system, such as SC or EEG, or external – such as ocular signals or recordings of video or audio (Cowley et al., 2016). This is why choosing the suitable measurement is an important question that requires working knowledge of the different psychophysiological reactions.

In this chapter, we give a brief introduction to some of the most common measurements of psychophysiology. Moreover, this chapter discusses the use of combined measurements and mixed methods. A more detailed discussion of available methods with a specific focus on HCI can be found in the work of Cowley et al. (2016) or in the *Handbook of Psychophysiology* by Cacioppo et al. (2007). The specific focus of this chapter is on measurements that can help identify negative emotions and stress. Identifying stressful situations is an important aim in the usability testing of tools to be used in crisis situations.

Electrodermal Activity

Electrodermal activity (EDA) represents changes in the electrical properties of the skin (Dawson, Schell, & Filion, 2007). SC is one method to measure this activity. SC is helpful in identifying arousal (Boucsein & Backs, 2000), and is seen as a reliable indicator of stress (Healey & Picard, 2005; Lunn & Harper, 2010). Measuring SC is usually done by attaching two finger electrodes to the participant's fingers.

SC is regulated by the sympathetic nervous system and can be measured as fluctuations caused by the activation of sweat glands. Such changes are often subtle and imperceptible to the individual but can provide invaluable information into how the individual perceives or responds to stimuli, such as when trying out a new product. This means that SC is an indicator of cognitive and emotional activity (Dawson et al., 2007). Using SC in laboratory settings, as well as in HCI studies, is a well-established method (Cowley et al., 2016).

Eye Tracking

The eyes provide information on the central nervous system and cognition and activity (Cowley et al., 2016). Eye tracking can be used to gain information on internal cognitive states by providing information on the user's visual attention. Eye tracking is a valuable way of evaluating a computer interface (Crowe & Narayanan, 2000). Knowing where test persons look when they use an interface is key to understanding the cognitive processes of the user. Knowing this can highlight mismatches in the user's mental model of how they think the system should work, where they assume information and functionality should be placed, compared with how the interface is actually designed. Thus, during a system prototyping stage, good and bad design choices can be evaluated with eye tracking (Goldberg & Kotval, 1999).

During task performance in a usability test, eye movements can be either fixational, and provide data on location, duration and frequency or saccadic, in the form of amplitude or average speed. During fixations, the eye movement is stopped as the brain obtains clear visual information. After this, the fixation is released and the eye moves in a saccade to a different spot to stop and fixate again. The eye has to stop moving when obtaining visual information as the image during saccades becomes blurry (Uttal & Smith, 1968).

Facial Expressions

The use of facial expressions as an indicator of emotional behaviour is receiving increased attention in the HCI literature. There is a relationship between user's emotions and perceived usability problems (Branco, Firth, Encarnação, & Bonato, 2005). Moreover, facial expressions can be detected with a user camera, making it an unobtrusive measurement of emotional reactions. Facial expressions are detected through action units that correspond to independent movement of face muscles, creating unique and meaningful expressions, such as anger provoked by a product feature that doesn't work well. These expressions can be measured quantitatively (McDuff, El Kaliouby, & Picard, 2015).

The study of facial expressions as emotion detectors is based on a long history of empirical research (e.g. Allport, 1924; Ekman, 1973, 1994). Universal relationships between facial muscle movements and particular emotions have been studied extensively starting from the seminal work of Ekman and Keltner (1970) (for more recent work, see e.g. Chen et al., 2016; Jack, Sun, Delis, Garrod, & Schyns, 2016). However, a study by Gendron, Roberson, van der Vyer, and Feldman Barrett (2014) questions its universality and shows that emotion perception is also dependent on cultural and conceptual contexts. Using computer vision algorithms to identify landmarks of the face, such as the eyebrows, the nose and the corner of the mouth, makes it possible to more easily identify facial expressions. Software, using deep learning algorithms (McDuff et al., 2016) can identify pixels in the key regions of the face, and classify facial expressions that are mapped to emotions such as anger, joy or sadness.

TESTING A TOOL FOR SOCIAL MEDIA INFORMATION GATHERING

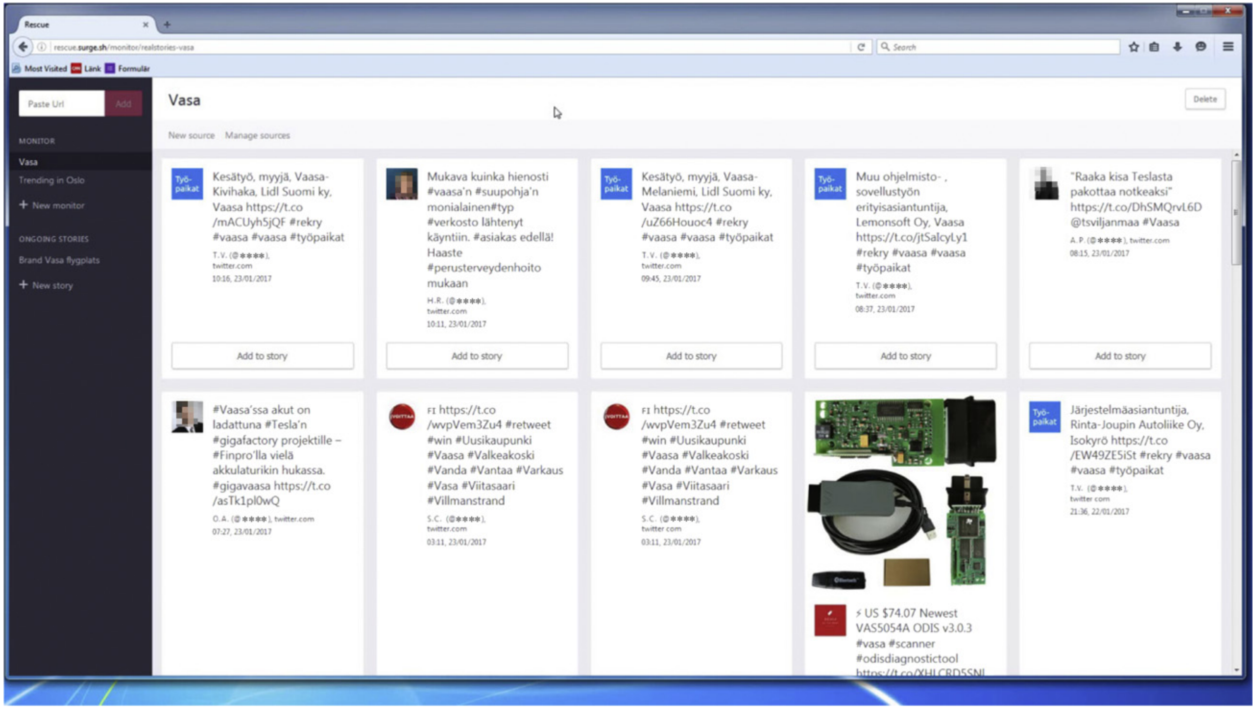
In this part of the chapter, we illustrate how psychophysiological testing can be used in a laboratory setting to test a tool intended for gathering social media information. The results stem from two usability studies conducted within the RESCUE project. The main focus is on how psychophysiological measurements were included in each test and for what purpose. Moreover, this chapter discusses how different types of psychophysiological data, in combination with subjective data, provide information about usability. The setup of the studies, as well as a results discussion in relation to emergency communication, can be found in Chapter 8.

The two usability tests have received ethical approval. During and after the tests, all data were collected, stored and analysed according to ethical guidelines. The tool tested was developed by the Norwegian company Bengler, following a mapping study of user needs (Backholm et al., 2017), similar available tools, and current best practices for monitoring and validating social media content. The results from the mapping study emphasise that automatic tool features need to be combined with several manual functions, and presented with a clear visualisation. Such a tool would provide better situation awareness during crisis work, by avoiding the problem of too much automatisation, creating a situation where the user is unaware of what the system is doing, 'an out of the loop situation' (Endsley & Kiris, 1995, p. 381).

In the tool, one of the main ideas is to work with different so called events. This is a specified field of interest, for instance, an unfolding disaster. An event is created by setting up search criteria and including content from several social media platforms, such as tweets or Instagram images, as well as information from different websites or RSS-feeds. All the information is displayed in a feed; furthermore, the user is able to save highly important information from the feed to a specific page. The saved content can be visualised in different ways, for example, on a map with geographical information or through an evaluation scale of importance and trustworthiness. The users can also work in a monitor mode, simply monitoring different search criteria that do not yet belong to a specific story. However, the content in monitor mode can easily be added to a new or existing story (Figure 1).

The tool was tested on journalists with working experience of at least one year, who were also familiar with the most common social media

Figure 1: A Close-to-Working Tool Prototype that Was Used During the Second Usability Test.



platforms and actively used social media either at work or at home. Seventeen journalists participated in the first test, and 15 of these chose to participate in the second test.[2] A more detailed description of the sample is available in Chapter 8.

USABILITY TEST 1: A WIREFRAME PROTOTYPE

The first prototype was a semi-interactive wireframe prototype with limited functionality. During the test, the user could click on predefined areas and navigate the tool through static images. The main focus of the first test was to see if users understood 'the big picture', investigating whether the features included fit the mental models that journalists use when gathering, monitoring and validating information. The test is described in detail in an article by Backholm et al. (2017). In this chapter, we focus on the test measurements of psychophysiology.

The laboratory part of the test was conducted in a room furnished as an office, where the test person could use a stationary computer to try out the prototype. One researcher guided the participant through the prototype giving out 34 predefined tasks. The tasks were divided into two main categories. Either the participant was asked about the meaning of different features without trying them out, 'What do you think location means?', or the participant could click on a feature and answer questions such as 'Click on location, what happened? Was this what you expected?' The task completion was graded in three steps by the researchers using a pass, struggle or fail scale. The concurrent think-aloud method (Nielsen, 1993) was used during the test, which means that participants were asked to voice their opinions during task completion. Another researcher observed the test from the control room and had access to the gaze patterns of the participant. A recording of the screen shown to the participant during the test, and corresponding eye tracking data were collected. No other psychophysiological measurements were used during the first usability test. After the test, participants were interviewed about their main impressions of the prototype, and filled out a questionnaire.

Key Conclusions from the First Usability Test

First, a key conclusion from the first usability test is that when analysing a rudimentary prototype, eye tracking can provide valuable insight into the cognitive processes of the user who is trying to understand the prototype. This means that the eyes move to the position where the test person expects the information to be, according to their mental model of the system. Moreover, eye tracking data were collected during the test to be used for verification of the researchers' observations of test participants' performance during tasks. This information was used by the researcher situated in the control room, who followed the gaze patterns of the participant. Combining eye tracking with subjective measurements, such as observations and questionnaires, is one way of externally validating the information and to validate the consensus between different observers (inter-rater observations).

Second, the wireframe prototype was still visually sparse and functionally far from a usable version. Hence, extensive use of psychophysiological measurements was unnecessary at this stage. No other psychophysiological measurements than eye tracking were beneficial for the test since the main focus was to understand key concepts of a more fundamental nature. Such research questions for early prototype testing are often better answered by using a more communicative approach than by including psychophysiological measurements.

A third conclusion from the first test is that since a concurrent think-aloud method was applied, talking would affect psychophysiology, making such measurements untrustworthy. It would be difficult to control if an emotional reaction stems from the discussion, or from the interface. Thus, a conclusion is that, when combining several types of measurements in a usability study, psychophysiological data collection should be kept separate from communication based methods (e.g. think-aloud).

Based on these three conclusions, the use of psychophysiological measurements in rudimentary prototype testing has limited benefits and suitability. Thus, we want to emphasise the use of appropriate measurements to assist in answering the research questions.

USABILITY TEST 2: A CLOSE-TO-WORKING PROTOTYPE

A second usability test of a close-to-working prototype with a reactive interface was conducted. Central results from the first usability test were used in the process to transform the static first prototype into an interactive one (see the previous book chapter by Backholm et al., 2017). In the second prototype, live data were used for a more realistic user experience of key functions in the tool. Since the aim of the RESCUE project is to develop a tool for crisis situations, the second usability test focused on a simulated crisis situation. The simulation, a fire at the regional airport, was based on a pre-made dataset of social media content. The simulation unfolded in the background during the prototype test, meaning that the feed was updated with new content at specific time intervals.

The second test was also conducted in a laboratory setting, due to the fact that the second prototype was still not suitable for field deployment (e.g. some functionality was still missing). The test subject was seated in front of a computer, in a room furnished as an office. One researcher was seated next to the participant, while another observed the test from the control room. In the control room, the researcher saw a mirror of the computer screen, as well as the participant's eye movements represented by a red dot, superimposed on the mirrored screen. To be able to identify usability problems, the participants were asked to complete 36 tasks in a predefined order. The tasks were either specific requests or questions where the participant could interact with the tool, such as 'Please add a twitter hashtag to the search', or 'Where do you think the presented information originates from?' Other tasks concerned the meaning of specific functions, such as 'What do you think the meaning of selected is?' Tasks focused on tool functionality, and the test participant was thus not depending on reaching specific phases of the unfolding crisis simulation to carry out tasks. Both researchers graded the task completion as either a 'pass' (no problems), 'struggle' (some problems, but completed the task) or 'fail' (did not complete the task).

In comparison to the first usability test where a concurrent think-aloud method was used, during this test the participants were first asked to complete the tasks without speaking – and then, if necessary, give instant feedback (i.e. retrospective think-aloud). One reason for this test design modification was that it more resembles a naturalistic usage setting. Another was due to the incorporation of psychophysiological measurements. Talking affects both SC levels and facial expressions detections (Johnson & Campos, 1967) and has a negative impact on task performance (van den Haak, de Jong, & Schellens, 2003). The psychophysiological measurements used in this study were SC, facial recognition and eye tracking. SC and facial expressions were used to identify stress levels and related emotional responses. Eye movements provided non-verbal information about the test person's actions during task completion. After the test, participants were interviewed about prototype features and completed a survey.

Key Conclusions from the Second Usability Test

The conclusions from the second usability study can be summarised into seven key points. First, since one of the main aims of the second usability test was to analyse the functionality of the tool, the participant was asked to complete tasks in a structured manner. The task completion was carried out without concurrent verbalisation or think-aloud. However, since the users are quiet during the task, psychophysiological measurements and/or post-interviews provide valuable information on how the users react during the different tasks.

Second, some usability problems could be further analysed and understood using eye tracking metrics. One example was the task to add a URL, which caused problems for all users. The eye tracking analysis showed that the participant had indeed noticed the add URL text field in the tool during the test, but when asked to add a URL, they struggled or failed to find this functionality. This is a good example of how eye tracking fixations are no proof that the correct mental models have been made, meaning that the user may have seen the function but not made the correct connections to the functionality of the tool. This finding also emphasises the importance of correct placement of functions. Hence, the eye tracking analysis can provide valuable information on where the user assumes the functionality should be placed, by looking at their gaze patterns when they try to complete tasks.

A third conclusion is that SC measurement was beneficial in situations when the participant had an emotional reaction. Such a reaction can be caused by both positive or negative emotions. Therefore, data should be analysed by looking at the whole data set and by identifying SC arousal peaks together with, for example, eye tracking data, to pinpoint what the person was doing that might have elicited this response and to label whether the response was positive or negative.

Fourth, we analysed the SC data using computer algorithms (PhysiOBS; Liapis, Karousos, Katsanos, & Xenos, 2014). The analysis was used both to compare the stress level during different tasks and to validate the laboratory setting. If the participants would not have been experiencing any stress during the test, the results would have been less valid. Using computer algorithms is a highly accurate addition to manual and subjective analysis (Liapis et al., 2014; Liapis, Katsanos, Sotiropoulos, Xenos, & Karousos, 2015).

We can further conclude that facial expressions provide a more detailed understanding of the emotional state of the participant. This means that positive or negative arousal can be identified, as well as what kind of positive or negative emotion the participant is experiencing (e.g. engagement, sadness, joy or disgust). For example, in the second usability test, the combination of facial expressions and the task completion grading into pass, struggle or fail showed that participants tended to express positive emotions after failing a task. Using joy in a situation like this can be understood as a healthy coping mechanism used to alleviate stress – and the observer gradings of tool performance were important to avoid a misinterpretation of collected facial expressions.

An additional conclusion from the study is that in our test sample there was a large variety of how overt the participant's emotional expressions were. Being overt means that some participants used a high amount of facial expressions during the test, while others faces were more neutral (covert) throughout the different tasks. Using several measurements assure a better understanding of participants that express their emotions on a varying degree of overtness.

Finally, combining three types of psychophysiological data (eye tracking, SC and facial expressions) together with subjective data (observations, interviews and questionnaires) provides rich usability data where the parts complement each other. One example is how facial expressions and interviews can be used to verify SC reactions. This means that we get a holistic understanding of whether the arousal is due to positive or negative emotions and what kind of emotion this is (e.g. anger or frustration). We can also pinpoint the specific feature of the interface causing this. Understanding the context and the person's reactions also provide us with information on whether the usability problem is about practical issues, such as when a function is hard to notice visually (not seeing it) or is not intuitively positioned (eluding the correct functionality) – or if the problem rather reflects more general discrepancies between the participant's own mental models of how to carry out a specific task and the tool's functionality.[3]

DISCUSSION AND RECOMMENDATIONS

Monitoring social media flows is especially challenging during a crisis situation, where misinformation and rumours are easily spread. Developing technical support solutions with good usability can help key communicators maintain situation awareness in a stressful situation. Laboratory testing is a useful way to identify potential usability problems in such technical support tools.

This final part of the chapter provides some more general lessons learned and practical advice for how to design a relevant psychophysiological usability test for a laboratory setting. We sum up several layers of pros and cons with combining different methods. First, we present recommendations regarding the validity of laboratory testing, and second, we discuss some challenges and limitations with this approach.

Recommendations for the Use of Psychophysiological Measurements in Usability Prototype Testing

Understanding the Choice of Methods in Relation to Prototype Level

- Using psychophysiological data is more suitable for high fidelity prototypes. A low fidelity prototype often lacks the content and functional fidelity that in a final product would affect, for example, gaze patterns or stress levels. Psychophysiological measurements are therefore less valuable for low fidelity prototype testing, due to the nature of the research questions to be answered at that stage of the development process. Subjective measurements such as think-aloud are more suited for early prototype testing.
- Securing relevant data by using several types of psychophysiological measurements
  We recommend a mixed methods approach, using both subjective and psychophysiological measurements when investigating high fidelity prototypes. This is important since there are individual

differences, for instance in the form of overt and covert expression of emotions. This means that some people have very expressive faces suitable for emotional analysis, while for people with less visible facial expressions other methods such as SC are a better measurement of emotional response. In general, collecting several types of data provides a more robust result in studies with small sample sizes.

- Combining the think-aloud method and psychophysiological data collection
  The problem with using a concurrent think-aloud method in combination with psychophysiological measurements is that, for example, skin response and facial responses are affected by speech. Think-aloud does not, to the same extent, affect eye tracking metrics, especially when data are analysed qualitatively (i.e. as a tool for observation validation). We recommend carefully considering which type of think-aloud approach to use (concurrent or retrospective) in relation to the psychophysiological methods included and the level of prototype complexity in the test.

- Methodology consideration for large data sets
  A central challenge is how to cope with the vast data amounts that psychophysiological measurements result in, especially when combining several measurements types. This places demands on the analysis software used, such as in the synchronisation of data between different measurement types (i.e. eye-tracking and skin response). Thus, we emphasise the importance of a clear and structured methodology as well as an understanding of the limitations of different software programmes.

- Understanding the challenges when using psychophysiological measurements with interactive stimuli
  Using interactive test material in high fidelity prototypes creates a test environment that is more complex and harder to control. The interactive nature gives the user more options on how to use the prototype, and creates a more realistic test situation. However, this means that what the user is experiencing is not constant between participants. Predefined tasks and strict adherence to test protocol are crucial to minimising the differences in situations between test persons.

- Understanding the choice of method in relation to generalisation and tool design issues
  The use of a laboratory setting for prototype testing has its limitations, due to the fact that the situation is never completely natural. This is a specific challenge when testing a tool to be used in a crisis situation, since such a situation is nearly impossible to simulate in a lab. However, conducting usability tests during real life crises is also challenging, since a crisis is unexpected and sudden. Furthermore, when it comes to prototype testing, the main focus should be on first pinpointing and correcting usability problems, regardless of situational context.

When developing a tool to be used during high-stress work, the tool itself should be intuitive to use, and not contribute to additional cognitive load or emotional stress caused by poor usability. In conclusion, psychophysiological measurements are helpful when trying to identify and understand the emotional and cognitive states of the user. We emphasise the combination of different psychophysiological measurements, as well as including subjective observations and ratings, to get a holistic understanding of the user experience, considering for example individual differences in emotional transparency. We also emphasise that laboratory tests should be combined with field tests of a more or less final version of a product, if possible.

NOTES

1. In this chapter, we refer somewhat interchangeably to crises, emergencies, disasters and catastrophes. The distinction between the events can be either the scale of the disruption or the cause of the event. However, such events cause a threat to societal values and demand some sort of response from different actors.

2. Dropouts were due to relocation and sick leave.

3. A more detailed discussion on these so-called lower- and higher-level problems is found in Chapter 8, by Backholm et al. (2017).

ACKNOWLEDGEMENTS

This work was supported by the Norwegian Research Council under Grant 233975/H20; the Högskolestiftelsen i Österbotten Foundation.

REFERENCES

Allport, F. H. (1924). *Social psychology*. Boston, MA: Houghton Mifflin.

Backholm, K., Ausserhofer, J., Frey, E., Larsen, A. G., Hornmoen, H., Högväg, J., & Reimerth, G. (2017). Crises, rumours and reposts: Journalists' social media content gathering and verification practices in breaking news situations. *Media and Communication*, *5*(2), 67–76. doi:10.17645/mac.v5i2.878

Boucsein, W., & Backs, R. W. (2000). Engineering psychophysiology as a discipline: Historical and theoretical aspects. In W. Boucsein & R. W. Backs (Eds.), *Engineering psychophysiology. Issues and applications* (pp. 3–30). London: Lawrence Erlbaum.

Branco, P., Firth, P., Encarnação, L. M., & Bonato, P. (2005). Faces of emotion in human-computer interaction. In *Proceedings of CHI'05 extended abstracts on human factors in computing systems*. New York: Association for Computer Machinery (pp. 1236–1239).

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. (Eds.). (2007). *Handbook of psychophysiology*. Cambridge: Cambridge University Press.

Chen, C., Crivelli, C., Garrod, O., Fernandez-Dols, J. M., Schyns, P., & Jack, R. (2016). Facial expressions of pain and pleasure are highly distinct. *Journal of Vision*, *16*(12), 210. doi:10.1167/16.12.210

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L. ... Ravaja, N. (2016). The psychophysiology primer: A guide to methods and a broad review with a focus on human–computer interaction. *Foundations and Trends® in Human–Computer Interaction*, *9*(3–4), 151–308. doi:10.1561/1100000065

Crowe, E. C., & Narayanan, N. H. (2000). Comparing interfaces based on what users watch and do. In *Proceedings of eye tracking research and* *applications symposium 2000*. New York, NY: Association for Computing Machinery (pp. 29–36).

Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 200–223). Cambridge: Cambridge University Press.

Duh, H. B. L., Tan, G. C., & Chen, V. H. H. (2006). Usability evaluation for mobile device: A comparison of laboratory and field tests. In *Proceedings of the 8th conference on human-computer interaction with mobile devices and services*. New York, NY: Association for Computing Machinery (pp. 181–186).

Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing*. Portland, Ore.: Intellect books.

Ekman, P. (1973). Cross-cultural studies of facial expression. In P. Ekman (Ed.), *Darwin and facial expression: A century of research in review* (pp. 169–222). New York, NY: Academic Press.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin*, *115*(2), 268–287. doi:10.1037//0033-2909.115.2.268

Ekman, P., & Keltner, D. (1970). Universal facial expressions of emotion. *California Mental Health Research Digest*, *8*(4), 151–158.

Endsley, M. R., & Jones, D. G. (2016). *Designing for situation awareness: An approach to user-centered design*. New York: CRC Press.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, *37*(2), 381–394. doi:10.1518/001872095779064555

Gendron, M., Roberson, D., van der Vyer, J. M., & Feldman Barrett, L. (2014). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion*, *14*(2), 251–262. doi:10.1037/a0036052

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: Methods and constructs. *International Journal of*

Industrial Ergonomics, *24*(6), 631–645. doi:10.1016/s0169-8141(98)00068-7

Haataja, M., Laajalahti, A., & Hyvärinen, J. (2016). Expert views on current and future use of social media among crisis and emergency management organizations: Incentives and barriers. *Human Technology*, *12*(2), 135–164. doi:10.17011/ht/urn.201611174653

Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 156–166. doi:10.1109/tits.2005.848368

ISO 9241. (2010). *Ergonomics of human-system interaction—Part 210: Human-centered design for interactive systems*. Geneva: International Standard Organization.

Jack, R. E., Sun, W., Delis, I., Garrod, O. G., & Schyns, P. G. (2016). Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General*, *145*(6), 708–730. doi:10.1037/xge0000162

Jacucci, G., Fairclough, S., & Solovey, E. T. (2015). Physiological computing. *Computer*, *48*(10), 12–16. doi:10.1109/mc.2015.291

Johnson, H. J., & Campos, J. J. (1967). The effect of cognitive tasks and verbalization instructions on heart rate and skin conductance. *Psychophysiology*, *4*(2), 143–150. doi:10.1111/j.1469-8986.1967.tb02751.x

Kjeldskov, J., & Skov, M. B. (2014). Was it worth the hassle? Ten years of mobile HCI research discussions on lab and field evaluations. In *Proceedings of the 16th international conference on human-computer interaction with mobile devices & services*. New York: Association for Computing Machinery (pp. 43–52).

Liapis, A., Karousos, N., Katsanos, C., & Xenos, M. (2014). Evaluating user's emotional experience in HCI: The physiOBS approach. In M. Kurosu (Ed.), *Proceeding of the international conference on human-computer interaction*. Cham: Springer (pp. 758–767).

Liapis, A., Katsanos, C., Sotiropoulos, D., Xenos, M., & Karousos, N. (2015). Recognizing emotions in human computer interaction: Studying stress using skin conductance. In J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque, & M. Winckler (Eds.), *Proceedings of human-computer interaction*. Cham: Springer (pp. 255–7262).

Lunn, D., & Harper, S. (2010). Using galvanic skin response measures to identify areas of frustration for older web 2.0 users. In *Proceedings of the 2010 international cross disciplinary conference on web accessibility (W4A)*. New York, NY: Association for Computing Machinery (pp. 34–43).

McDuff, D., El Kaliouby, R., & Picard, R. W. (2015). Crowdsourcing facial responses to online videos. In *Proceedings of the affective computing and intelligent interaction (ACII), 2015 international conference*. IEEE (pp. 512–518).

McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., & Kaliouby, R. E. (2016). AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. New York: ACM (pp. 3723–3726).

Mendl, M. (1999). Performing under pressure: Stress and cognitive function. *Applied Animal Behaviour Science*, *65*(3), 221–244. doi:10.1016/s0168-1591(99)00088-x

Nielsen, J. (1993). *Usability engineering*. Boston, NY: Academic Press.

Qiu, L. (2017, August 28). *A shark in the street, and other Hurricane Harvey misinformation you shouldn't believe*. Retrieved from https://www.nytimes.com/2017/08/28/us/politics/shark-hurricane-harvey-rumors.html?mcubz=3

Reuter, C., & Spielhofer, T. (2017). Towards social resilience: A quantitative and qualitative survey on citizens' perception of social media in emergencies in Europe. *Technological Forecasting and Social Change*, *121*, 168–180. doi:10.1016/j.techfore.2016.07.038

Rubin, J. (1994). *Handbook of usability testing*. Hoboken, NJ: Wiley.

Silva, H. P. D., Fairclough, S., Holzinger, A., Jacob, R., & Tan, D. (2015). Introduction to the special issue on physiological computing for human-computer interaction. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *21*(6), 29. doi:10.1145/2688203

Spence, P. R. (2016). Introduction to the special section on social media and the crisis lifecycle. *Computers in Human Behavior*, *54*(1), 587–588. doi:10.1016/j.chb.2015.08.044

Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., & Bergel, M. (2002). An empirical comparison of lab and remote usability testing of web sites. In *Proceedings of the usability professionals association conference*. Orlando, Fla. http://doi.org/10.1.1.457.3080

Uttal, W. R., & Smith, P. (1968). Recognition of alphabetic characters during voluntary eye movements. *Attention, Perception, & Psychophysics*, *3*(4), 257–264. doi:10.3758/bf03212741

van den Haak, M., de Jong, M., & Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, *22*(5), 339–351. doi:10.1080/0044929031000