Design Systems and the Need for a Usability Method: Assessing the Fitness of Components and Interaction Patterns in Design Systems Using Atmosphere Methodology

P. Johansson, S. Mardh

***Abstract*—**The present study proposes a usability test method, Atmosphere, to assess the fitness of components and interaction patterns of design systems. The method covers the user's perception of the components of the system, the efficiency of the logic of the interaction patterns, perceived ease of use as well as the user's understanding of the intended outcome of interactions. These aspects are assessed by combining measures of first impression, visual affordance and expectancy. The method was applied to a design system developed for the design of an electronic health record system. The study was conducted involving 15 healthcare personnel. It could be concluded that the Atmosphere method provides tangible data that enable human-computer interaction practitioners to analyze and categorize components and patterns based on perceived usability, success rate of identifying interactive components and success rate of understanding components and interaction patterns intended outcome.

***Keywords*—**Atomic design, Atmosphere methodology, design system, expectancy testing, first impression testing, usability testing, visual affordance testing.

I. Introduction

IN developing systems and products in HCI (Human-Computer Interaction) organizations, it has become increasingly common for designers to use, or develop, design systems. A design system typically comprises reusable components that can be combined into design patterns. Applying a design system enables teams to adopt a coherent design and create a unified user experience across their systems or products. Originated from [1], design patterns have become a fundamental concept for HCI professionals in design activities for systems and products [2]-[4]. Even though modular software solutions have existed since the 1960s [5], design systems sprung to life in the early 2010s with Google's Material Design [6] which combined atomic design developed by Frost [7] with best practices for pattern libraries. Today, Material Design is one of the most widely known design systems. Google [6] describes it as "Material Design is a visual language that synthesizes the classic principles of good design with the innovation of technology and science". Material Design is a best-practice design language made public and the guidelines are updated and revised frequently by Google since 2014. Many design systems are built on the structure of atomic design, seeing design as divided into atoms, molecules and organisms [7].

Atomic design is a metaphor from the domain of chemistry, where different fragments can be used to create many different things. Translated into the HCI domain this helps us consider user interfaces as both complete as a whole and a collection of building blocks at the same time. Development of pattern-based design requires a holistic view and centralized approach of all UI components that comprise the user interface [7]. Even though design systems have existed for more than a demi-decade, research on usability testing for design systems is scarce. Although there are a large number of usability methods for assessing systems and products, no formal usability testing method has been presented to cope with the validation of components and interaction patterns of design systems.

Usability assessment methods evolved in the early 1980s from traditional ergonomics & human factor techniques and include a wide variety of methods [8]. Design solutions and concepts based upon design systems are often tested through expert evaluations or scenario-based testing methods to measure success rates or time to completion of tasks. However, expert evaluations are prone to be biased and tend to uncover different sets of problems from user testing, called the "Evaluator Effect" [9]-[11]. A common way of capturing the users' thoughts during usability tests is think-aloud protocols, as we cannot directly observe what the user is thinking [12]. There are several different think-aloud techniques, Traditional think-aloud, Speech-communication and Coaching [13]-[15]. Boren and Ramey [15] proposed the speech-communication think-aloud protocol as a less disruptive alternative to technique developed by Ericsson and Simon [13], which focuses on only providing insistent probes.

Complications arise when products and systems are being evaluated on a system level without considering evaluating the components of the solution. This would be equivalent to trying to make any system or product perform better by examining the system as a whole instead of tweaking the components of the system that really affect performance. Scenario-based testing can tell you whether a user can complete tasks, the test administrator can measure the progress in different ways and later propose design alterations. However, scenario-based

Patrik Johansson is with Zoezi, Linkoping, 58331 Sweden (corresponding author, phone: +46-706190790; e-mail: patrik@zoezi.se).

Selina Mardh is with Cambio Group, Linkoping, 58331 Sweden (e-mail: selinda.mardh@cambio.se).

testing will not provide sufficient insights into if and how design system components are intuitive and provide clues of usage/affordance to the user. This might have varying consequences depending on the type of system or application at hand, for example websites that are relatively uncomplicated to update and modify. However, when it comes to very complex products, for example security-critical ultra-large scale electronic health record (EHR) systems or supersonic fighter aircrafts, design issues might be extremely costly and require huge efforts to resolve. These are examples of systems where unintuitive or non-logic design is safety-critical and in some cases might result in loss of life. We argue that evaluating the components and interaction patterns comprising a design system is equally critical to a design as testing the ready-made system/product.

We propose a usability test method, *Atmosphere,* to assess the fitness of components and interaction patterns in design systems. The *Atmosphere methodology* is an eclectic choice of tests to assess the essential aspects of a design system. We suggest that a method for testing a design system developed for a specific context should cover the following three areas: first impression, visual affordance and expectancy. The combination of these aspects will cover the user's perception of the components of the system as well as the efficiency of the logic of the interaction patterns of the design system.

A. First Impression

For HCI practitioners, first impression testing has become a widespread method for measuring the users' first impressions of a design that is new to the user. First impression testing is a relatively fast method to be performed on simple designs to gain early insights. It has an easy protocol to develop and follow [16]. First impression testing indicates what information and impressions stay with the user after viewing a design for only a few seconds. This typically tells you what the user thinks about overall design structure, aesthetics and recognizable content. Research in the HCI field indicates that users' perception of the usability of a system is affected by aesthetic factors [17]. Tractinsky [18] has shown that there are strong correlations between user's satisfaction from using the system and their perceptions of its aesthetics and usability. Arguments have been made that aesthetics should be greatly considered during the design phase to cope with the behavioral and psychological responses toward products [19], [20]. Liu et al. [21] identified that a user's first moment at a new website was critical to the loyalty of the user, and whether users decided to stay or leave. Guo [22] suggests that users' first impressions are only marginally influenced by the difference in perceived usability, but significantly affected by the differences in aesthetics, and designers should prioritize aesthetic design in early design stages. Grishin and Gillian [23] highlight that designers could be provided design advantages from a deeper and more profound understanding about how and when aesthetics affect user responses towards products.

There are several ways of testing users' first impressions; the most common method is the 5 second test (5ST). Gronier [16] argues that 5 seconds seems to be sufficient time for the test participant to have a first impression of the interface and the quality of the design components. This study suggests that first impression testing is implemented in a methodology to assess the fitness of design system components.

B. Visual Affordance

Affordance is a concept made popular in the HCI community through Donald Norman's book The Psychology of Everyday Things [24]. Norman's [24] definition of affordance reads "the term affordance refers to the perceived and actual properties of the thing, primarily those fundamental properties that determine just how the thing could possibly be used.". A high level of affordance could heighten the user experience by informing users of possible and appropriate actions without instructions. Affordance is especially valuable to help novice users achieve fundamental understanding, or to compel users in exploring systems without fear of making mistakes leading to unwanted results. Previous research has argued that visual affordance is closely linked to action [25]. It has been shown that representations for actions, elicited by an object's visual affordance, serve to inform users of the most afforded action, even when these objects and affordances are irrelevant to current goals [26]. Hence, users can be supported or invited to certain types of actions based on the visual affordance of objects. It is desirable in most contexts to have strong visual affordance in the system since the system might then be perceived as intuitive and easy to learn. This study proposes visual affordance testing as an important aspect of usability testing of design systems as a way of finding out which components are perceived as interactive and which are not perceived as interactive.

C. Expectancy

The main benefit of using expectation measures in usability practices is to help prioritize the usability issues [27], [28]. Albert and Dixon [27] argue that expectancy is based on a subject's previous experiences, prompting that in order to get valuable results for testing the test participants should not be completely novice nor extreme experts. In order to get reliable results, data should be obtained regarding the test participants' experience using similar systems and how relatable they are to the domain of the system at hand.

A well-studied method for expectancy is the SEQ (Single Ease Question), in which the test participant is asked how difficult tasks were to perform after they conducted them [29]. However, Albert and Dixon [27] argue that SEQ needs to be complemented with a pre-stated expectancy from the test participant in order to compare the expectancy with the actual result. The analysis for an expectancy test focuses on the differences between the expectations and actual experience, referred to as disconfirmation. The disconfirmation can be either positive, when the experience exceeds the expectations, or negative, when the experience is below the expectation [30]. The expectancy disconfirmation theory suggests that customers judge their satisfaction in relation to an already established level of expectation, and marketing research

studies have found that a primary driver for satisfaction is a customer's expectancy [31], [32]. Studies have been performed to measure expectancy for tasks often spanning several interactions in order to reach a task goal or complete an assignment. However, no research studies have been found where expectancy has been measured for every interaction as proposed by the *Atmosphere methodology* in this study. The theoretical difference is that in the *Atmosphere methodology,* every interaction is considered a task, while in other studies a task may consist of several interactions before reaching the goal. This study suggests that expectancy is a key factor in discovering users' understanding of the components and patterns of which the design system comprise.

II. Method

A. Study Design

The present study assessed a design system developed for an EHR system. The design system builds on the atomic design structure. The study was performed during four weeks including 15 participants. Each test occasion lasted approximately 50 minutes. To include as many as possible of the components and interaction patterns that the evaluated design system comprises in a way that did not overcomplicate the application, two versions of the application had to be created, the first version tested on seven participants and the second version tested on eight participants. The participants were given the same test details and protocol, in the same order, evaluating the same aspects of the design system. The participants were all daily users of the system and well familiar with the current, implemented version of the system and its functions. The test occasion was recorded after written informed consent by each of the participants. The study followed the ethical principles and guidelines regarding research, recordings and informed consent by the World Medical Association Declaration of Helsinki [33].

B. Participants

15 participants were recruited by internal customer-company contacts where end-users were asked to be part of establishing a new generation of the EHR system at hand. Amongst the participants, six different user roles in healthcare that interact with the system on a daily basis were represented. Age ranged from 26-64. To be eligible to participate in the study, the participants had to be part of the intended end-user group (healthcare personnel), and working in the relevant context (full-time or part-time). They also had to have experience working with the EHR system at hand on a daily basis, Swedish-speaking and be competent to provide consent. The participant group was limited to 15 participants as studies have shown that most of the usability issues are found within that number [34]-[36].

C. Material

The user tests were performed on a desktop-built application based on the design system. The application was a one-page view incorporating components and patterns from the design system. A laptop with a resolution of 1920x1080 was used. A recording software was installed on the laptop. The recording software recorded on-screen activities and sound.

A moderator script was developed for the test administrator to make sure that all the participants received the exact same information and in the same order. A pre-test questionnaire regarding demographics, such as gender, age and role was administered. It also had questions on work related experience of the system at hand as well as technical experience. A post-test questionnaire, i.e., the System Usability Scale, SUS [37] was used to gain insight into the users' overall perception of the usability of the system. SUS is a validated method for measuring and calculating usability [38]-[40] and has become an industry standard when it comes to measuring usability.

In order to capture the test participants' thoughts on each component they wanted to interact with, we used the speech-communication think-aloud protocol presented by [14]. While practicing the speech-communication think-aloud protocol, we deliberately used variations of the protocol during different phases of the test process, *Intervention* and *Instruction* described by [12].

D. Procedure

The tests were performed in a meeting room at the clinic in the hospital where the test participants worked. The test was performed with one participant at a time. The set-up was exactly the same for all participants. A formal greeting and presentation were done by the test administrator at the beginning of the test. All test participants received the same information, read directly from the moderator script. The information included the purpose of the test, how the test would be conducted and instructions to the think-aloud protocol. The tasks were administered in the same order for all participants.

Before the test began, test participants were asked to fill in the pre-test questionnaire and a written informed consent regarding their voluntary participation in the study as well as the approval of recording the screen and sound. After the test participant signed the consent form the test administrator started the recording.

E. Tests

First Impression

The test session started off with a description of the first impression test (5ST), where the test participants were informed that the application would be presented to them for 5 seconds. The application was displayed to the test participants for 5 seconds, timed by the test administrator, whereafter the screen was turned off. The test participants' input was divided into two sections, first, the test participants were asked to describe what they observed and what they could remember in their own words. Second, when the test participant had given all the input they could, the test administrator would ask participants to specifically describe their opinions on the following areas:

- Information
- Purpose of application

- Structure, layout, design

Visual Affordance and Expectancy

After the first impression test had been performed, the next phase of the test began, focusing on evaluating visual affordance and expectancy. The test participants were presented with a live test environment of the same application as in the first impression test. They were instructed to identify all interactive components in the application, one at a time, and provide a pre-stated expectancy of the interaction before interacting with the component. To test their expectation of a component, the think-aloud protocol variation *Intervention* was used after the test participants had identified a component they thought was interactive, but before the test participants interacted with the component. The intervention was used for all components about 2-10 seconds after the test participants identified it. After each identified component, the test administrator intervened; asking the test participants to pre-state what they expected would happen if they interacted with the component. The approach was to understand what the test participants thought would be the outcome of interacting with a component before actually interacting with it.

The think-aloud protocol variation *Instruction* was used after the test participants interacted with a component. After the test participants interacted with the component, they were presented with a SEQ about their experience of the interaction outcome. The test participants were asked to evaluate the usability of the outcome, e.g., if the solution was different from what they expected, if it was better or worse than they expected and in what way. This instruction was used for all components about 5-30 seconds after the test participants interacted with it. The procedure was repeated for each interaction in the application until the test participants verbally confirmed that they could not identify any more interactive components. Assistance was only provided if the test participants tried to interact with components before pre-stating their expectancy. After the test had been conducted, the test participants were given the SUS as a post-test questionnaire.

Data Collection and Analysis

Data from the *first impression test* were gathered and analyzed using a combination of open word choice and open-ended preference explanation. In the first section where test participants were asked to describe what they observed and what they could remember, this study used *open word choice*. In the analysis phase, the words of the test participants from the open word choice were categorized into positive, neutral or negative expressions. The words were then categorized to find patterns in the results. As open word choice does not give insights into *why* the test participants feel in a certain way, we also performed an open-ended preference explanation as a complementary data-gathering method. The test participants were asked to explain and expand on the why of a design. The data from the open-ended preference explanation were analyzed using a formative approach with the following metrics: information, perceived purpose and structure, layout and design.

Data from the *visual affordance test* were gathered through video and audio recordings. The analysis focused success rate, i.e., proportion of correctly identified interactive components. In counting the found components, a percentage of identification could be made.

Data from the *expectancy* test were gathered through video and audio recordings, where the analysis focused on the *understanding rate of interaction results*. We wanted to see how many of the interactions occurred as the test participant had expected, as well as calculate to what extent they described them in the same way as the interaction occurred. We also analyzed how many of the interactions the user's expectancy differed from the interaction outcome. The outcome was analyzed and interpreted as either positive or negative depending on the test participants' feedback. After the test had been conducted, the test participants were given a SUS as a post-test questionnaire.

III. Results

The results section displays results from two aspects; first, the methodological aspect which includes results concerning how well the methodology worked, i.e., the efficiency of the instructions, as well as whether the participants correctly perform the different parts of the test. Second, the results from the tests included are presented.

A. First Impression

Results on the Methodological Aspects

In the first impression test, it was concluded that all of the test participants could readily understand and act upon the instructions given. There were no incidents of a test participant misunderstanding the instructions or performing actions contradictory to instructions. Using the combination of open word choice and open-ended preference explanation provided the possibility to categorize assessments in a way that provided data about the test participant's first impression of the system, as well as understanding why they felt as they did. Data also included the test participants' perception on what they were supposed to do as well as their thoughts on structure, layout and design.

Results from the Tests

In the open word choice, the comments were categorized as either positive, neutral or negative.

TABLE I
Results from First Impression Testing, Depicting Data with Open Word Choice

Positive	Neutral	Negative
(1) Clear	(3) Looks like today	(1) Missing features
(2) Comprehensive		(3) Too much information
(1) Nice		(5) Messy
(1) Good		(1) Unusual
		(1) Dated
(1) Understandable		
Σ = 6	Σ = 3	Σ = 11

The comments from the open-ended preference explanations were clustered into core categories according to content.

Information: There were 17 comments made on information. According to content, the comments were subordinated into three core categories, namely "Good information" (6 comments), "A lot of information" (6 comments) and "A little bit messy" (5 comments). The comments in the core category "Good information" concerned information content, i.e., the content was perceived as to the point (3 comments) and easy to overview (3 comments). In the core category "A lot of information", the comments only concerned the aspect of there being a lot of information (the answers had no negative nor positive coloring). Three aspects of information were brought up in the core category "A little bit messy", namely that it was hard to know where to focus (2 comments), that the design was a bit busy (2 comments) and that the information made you lose focus (1 comment).

Purpose of the application: There were 21 comments made on Perceived purpose. The comments were subordinated into four core categories, namely "Attain information" (12 comments), "Document" (5 comments), "Support" (3 comments) and "Act" (1 comment). Comments on "Attain information" concerned aspects of overview (3 comments), finding (5 comments) and reading (4 comments) information. The core category "Document" had comments on documenting (4 comments) and filling in information (1 comment). The core category "Support" was made up of comments on getting help on sorting (1 comment) and moving to other types of information (2 comments). The comment on the core category "Act" concerned booking (1 comment).

Structure, layout and design: 58 comments were made in relation to the question on structure, layout and design. The comments were clustered according to content and subordinated into four core categories; "Structure" (29 comments), "Layout" (nine comments), "Design" (13 comments) and "Perceived emotion" (seven comments). The core category "Structure" (29 comments) had comments on perceived clear-cut structure (13 comments), six comments concerned that it was hard to understand the structure, five comments concerned center of attention and five comments stated that it was easy to understand the structure. "Layout" had nine comments all addressing work space (five comments) and columns (four comments). The comments on "Design" concerned font size (three comments), color (seven comments) and shape (10 comments). For the core category "Perceived emotion" (seven comments), five were positive and two negative.

B. Visual Affordance and Expectancy

Results on the Methodological Aspects

In the visual affordance and expectancy test, it was concluded that all of the test participants could readily understand and act upon the instructions given. There were a few incidents where the test participants interacted with components before giving a pre-stated expectancy. This happened due to eagerness rather than misunderstanding the instructions. Quantitative data about the overall perceived usability were collected using the SUS questionnaire [37].

Results from the Tests

For visual affordance, the data were collected through observing all video recordings and counting all components that the test participants identified, cross checking it against a spreadsheet with all components in the test environment. The mean percentage for rate of identification was 64,62%, with a range from 37,14% to 82,86%. In the analysis of the data, we were able to detect a rate of error identification at 1,86%, where test participants identified graphical elements as interactive components, or tried to use keyboard navigation that was not available.

Data were collected in the same way for expectancy to calculate the rate of understanding, by analyzing what pre-stated expectancy the test participant gave before interacting with a component. The mean percentage for rate of understanding was 69,94%, with a range from 48,15% to 91,30%.

We were able to observe a positive trend in how the test participants experienced the interaction outcomes. Of all interactions there was a 55,06% positive experience of the interactions and only 10,13% negative. The negative experience is representative to only a few components and interaction patterns that the target group interacted with. For this part we also had missing data at 34,81% due to no input or input that was not possible to interpret. The data from the SUS questionnaire were collected and benchmarked, showing a mean SUS results value of 75,3.

IV. Discussion

The aim of this study was to develop a usability method for evaluating the fitness of components and interaction patterns in a design system. Usability testing is key in releasing products or services to avoid usability issues. Although *Atmosphere methodology* is a usability method, it targets a new area of testing as there is no test method targeting specifically design systems in the field today. The main contribution of Atmosphere methodology is the ability to perform standardized testing of design systems in order to assure that components and interaction patterns are easy to understand and helps the user interact with the product as intended. The proposed usability method *Atmosphere* was simple to implement and easy for participants to understand. There were no stated or observed disturbances on the test participants' side regarding the use of think-aloud variations *Intervention* and *Instruction*. In this study the participant group was set to 15 participants as studies have shown that most of the usability issues are found within that number [34]-[36]. This gives us a high level of certainty to identify a large percentage of existing usability problems.

The results indicate that prominent results can be achieved by following the presented protocol, enabling HCI practitioners to gain valuable insights into perceived usability, affordance and expectancy of a design system's components and interaction patterns. Through the obtained data, it was

possible to categorize what patterns and components were easy or hard for the test participants to understand, and also, why. Through test participants' think-aloud comments we were able to perform quantitative analysis focusing the accordance between pre-stated expectancy and actual interaction outcome. We were also able to perform qualitative analysis on test participants' think-aloud comments stated during the tasks. These results could be considered in re-designing components in the design system at hand.

The results show that the presented usability test method provides tangible data that enable HCI practitioners to analyze components and interaction patterns in order to assess their fitness in the design system. The result of the first impression testing provides interesting insights into the perceived usability of the tested solution. In the results from the first impressions testing, we found more negative expressions than positive of what the test-participants observed, but after having completed the whole test session test-participants in this study had a mean SUS score of 75,3, which is higher than the cut-off values of the SUS score of 68 [40]. The SUS result from this test indicates that the design system is perceived as usable even though the majority of the open word choice inputs of the first impressions were negative.

A pattern we were able to identify in the analysis of the results was that many of the interactions in which the test participants failed to give the correct expectancy; the interaction outcome was still described as positive. Even though the test participants expected a specific outcome, they were glad that it did not happen that way as their expectancy was based on previous experiences in these types of systems. It was explained that some of the new components and patterns were unaccustomed but superior. As this study is the one using the *Atmosphere methodology,* we did not hypothesize the results of the identification rate, rate of understanding or rate of interaction outcome. The results in this study should be interpreted as a baseline for how users identified, understood and experienced components and their interactions in the first phase of testing.

We observed that a few components were the reason behind the identification rate and expectancy rate not being higher. For some components, none of the test participants expected the correct outcome. There were also a few components that most of the test participants were unable to identify, which was the prominent reason behind a drop in identification rate. The conclusion of these observations is that we, through *Atmosphere methodology,* were able to identify which components were intuitive and which were not. It can be stated that a handful of usability issues is responsible for diminishing the score rather than fundamental flaws in the design. Through the *Atmosphere methodology* it was proven that a few faulty components can cause users to miss a lot of functionality and information in a system, and if the component is being continuously re-used throughout the system there will be user mistakes and frustrations.

V. Conclusions and Future Directions

The present study is set out to investigate whether a usability test method could be used to assess the fitness of components and interaction patterns of design systems. It can be concluded that in applying *Atmosphere methodology*, key features of design systems can be evaluated and give crucial input to design enhancement. *Atmosphere methodology* has given us a deep understanding for each component and pattern in the design system dedicated to the EHR system concerning perceived usability in first impressions, visual affordance and expectancy.

The Atmosphere methodology is a usability test method created specifically for evaluating design systems. It provides a valuable approach into investigating the users' expectancy and impressions on a highly detailed level for every component used in a design system. *Atmosphere methodology* has a wider area of use as it can replace usability test methods that are typically biased, like heuristic evaluation [9]-[11]. As user data is a collective information of users' attributes and actions with your product it is also the voice of the customer, and we should look for ways to obtain more user data and move away from possibly biased expert evaluations. The primary consideration for developing user-centered products should be assuring that the users can complete intended tasks successfully. In using *the Atmosphere methodology* during the development phase of systems and products, HCI practitioners can ensure that the foundational components and patterns are aesthetic, intuitive and usable. In this study the method was used for evaluating a desktop solution. Many design systems, for example Google's Material design, are designed to support mobile applications. Further research is needed to explore the usefulness of the *Atmosphere methodology* in evaluating mobile solutions.

References

[1] Alexander, C., Ishikawa, S., & Silverstein, M. (1977). A Pattern Language: Towns, Buildings, Construction. Oxford University Press. ISBN 978-0195019193.

[2] Nudelman, G. (2013). Android Design Patterns. Edited by Wiley & Sons Inc. ISBN 978-1-118-39415-1.

[3] Tidwell, J. (1999). Common Ground: A Pattern Language for Human-Computer Interface Design. MolPharmacol, 61, 710–719. Retrieved fromhttp://www.mit.edu/~jtidwell/common_ground.html

[4] Wesson, J. L., Cowley, N. L. O., & Brooks, C. E. (2017). Extending a Mobile Prototyping Tool to Support User Interface Design Patterns and Reusability. Proceedings of the South African Institute of Computer Scientists and Information Technologists, SAICSIT 2017, Thaba Nchu, South Africa, September 26-28, 2017

[5] Sussmann, J. M., & Goodman, R. V. (1968). Implementing ICES module management unfer OS/360. Modular Programming: Proceedings of a National Symposium, 1968, 69-84.

[6] Google. (2014). Material Design is a unified system that combines theory, resources, and tools for crafting digital experiences. Material Design Website, (2014). Retrieved from https://material.io/

[7] Frost, B., (2016). Atomic Design. Pittsburgh: Brad Frost.

[8] Nielsen, (1994). Summary of Usability Inspection Methods. Retrieved from http://www.useit.com/papers/heuristic/inspection_summary.html

[9] Hertzum, M., Jacobsen, N. E., & Molich, R. (2002). Usability inspections by groups of specialists: perceived agreement in spite of disparate observations. In CHI'02 extended abstracts on Human factors in computing systems, 662-663, ACM.

[10] Hertzum, M., Molich, R., & Jacobsen, N. E. (2014). What you get is what you see: revisiting the evaluator effect in usability tests. Behaviour & Information Technology, 33(2), 144-162.

[11] Molich, R., Bevan, N., Butler, S., Curson, I., Kindlund, E., Kirakowski, J., & Miller, D. (1998). Comparative evaluation of usability tests.

Usability Professionals Association 1998 Conference, 22-26 June 1998 (Washington DC: Usability Professionals Association), pp. 189-200.
[12] Olmsted-Hawala, E. L., Murphy, E. D., Hawala, S., & Ashenfelter, K. T. (2010). Think-aloud protocols: A comparison of three think-aloud protocols for the use in testing data-dissemination web sites for usability. Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, USA, 2381-2390.
[13] Ericsson, K. A., & Simon, H. A. (1996). Protocol Analysis: Verbal Reports As Data. (Revised ed.) MIT Press, Cambridge, MA, USA.
[14] Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. Transactions on Professional Communication, 43(3), 261-278.
[15] Dumas, J., & Redish, J. A. (1999). Practical Guide to Usability Testing. Intellect Press, Portland, OR, USA.
[16] Gronier, G. (2016). Measuring the first impression: Testing the Validity of the 5 Second Test. Journal of Usability Studies, 12 (1), 8-25.
[17] Lee, S., & Koubek, R. J. (2010). Understanding user preferences based on usability and aesthetics before and after actual use. Interacting with Computers, 22 (6), 530-543.
[18] Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. Interacting with computers, 13 (2), 127-145.
[19] Norman, D. A. (2002). Emotion and design: Attractive things work better. Interactions Magazine, ix (4), 36-42.
[20] Ilmberger, W., Held, T., & Schrepp, M. (2008). Cognitive processes causing the relationship between Aesthetics and Usability. In: HCI and usability for education and work. Heidelberg: Springer, 43-54.
[21] Liu, C., White, R. W., & Dumais, S. (2010). Understanding web browsing behaviors through Weibull analysis of dwell time. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '10 (pp. 379-386). New York, NY: ACM.
[22] Guo, F., Wang, X-S., Shao, H., Wang, X-R., & Liu, W-L. (2019). How User's First Impression Forms on Mobile User Interfaces?: An ERPs Study, International Journal of Human-Computer Interaction, DOI: 10.1080/10447318.2019.1699745
[23] Grishin, J., & Gillan, D. J., (2019). Exploring the Boundary Conditions of the Effect of Aesthetics on Perceived Usability. Journal of Usability Studies, 14 (2), 76-104.
[24] Norman, D. A. (1988). The Psychology of Everyday Things. New York: Basic Books.
[25] Della Sala, S., Marchetti, C., & Spinnler, H. (1991). Right-sided anarchic (alien) hand: A longitudinal study. Neuropsychologia, 29(11), 1113–1127.
[26] Phillips, J. C., & Ward, R. (2002). S-R correspondence effects of irrelevant visual affordance: Time course and specificity of response activation. Visual Cognition, 9 (4/5), 540-558.
[27] Albert, W., & Dixon, E. (2003). Is this what you expected? The use of expectation measures in usability testing. Proceedings of Usability Professionals Association 2003 Conference, Scottsdale, AZ.
[28] Rich, A. & McGee, M. (2004). Expected Usability Magnitude Estimation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 48(5), 912–916.
[29] Albert, B., & Tullis, T. (2013). Measuring the User Experience: Collecting, Analyzing and Presenting Usability Metrics. Second edition. Morgan Kaufmann, 2013.
[30] Van Ryzin, G. (2013). An experimental test of the expectancy-disconfirmation theory of citizen satisfaction. Journal of Policy Analysis and Management, 32 (3), 597-614.
[31] Oliver, R. L. (1996). Satisfaction: A behavioral perspective on the consumer. New York: McGraw Hill.
[32] Voss, G. B., Parasuraman, A., & Grewal, D. (1998). The roles of price performance, and expectations in determining satisfaction in service exchanges. Journal of Marketing, 62(4), 46-61.
[33] World Medical Association, World medical association declared of Helsinki: Ethical principles for medical research involving human subjects, Jama, 310 (2013), 2191-2194.
[34] Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. Behavior Research Methods, Instruments & Computers, 35, 3, 379- 383.
[35] Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. Proceedings of INTERCHI 1993, ACM Press, 206-213.
[36] Spool, J. & Schroeder, W. (2001). Testing web sites: five users is nowhere near enough. CHI '01 Extended Abstracts on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 285–286. DOI: https://doi.org/10.1145/634067.634236.
[37] Brooke, J. (1996). SUS: A "quick and dirty" Usability Scale. Usability Evaluation in Industry, 189, 4-7.
[38] Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. International Journal of Human-Computer Interaction, 24(6), 574–594.
[39] Orfanou, K., Tselios, N., & Katsanos, C. (2015). Perceived usability evaluation of learning management systems: Empirical evaluation of the System Usability Scale. The International Review of Research in Open and Distributed Learning, 16(2). https://doi.org/10.19173/irrodl.v16i2.1955
[40] Sauro, J. (2011). A practical guide to the System Usability Scale (SUS): Background Benchmarks & Best Practices. Measuring Usability LLC, Denver, 2011.