CrowdUX: A Case for Using Widespread and Lightweight Tools in the Quest for UX

Hanna Schneider[a], **Katharina Frison**[b], **Julie Wagner**[a], **Andras Butz**[a]

LMU Munich, Germany
[a]firstname.lastname@ifi.lmu.de
[b]katharina.frison@googlemail.com

ABSTRACT

User studies and expert reviews are established methods for evaluating usability and user experience (UX) in user-centered design. However, practitioners often struggle to integrate these often time-consuming and costly methods in their design processes. As technological products and services are becoming increasingly mobile, their contexts of use are increasingly diverse and less predictable. While this changing context is hard to capture in lab studies, the same mobile technologies also provide possibilities for new study methods. In this paper we advocate lightweight mobile tools for crowdsourcing UX feedback. In cooperation with a design agency, we built two apps that allow users to express feedback with text, ratings and pictures whenever using a product. The second app assigns feedback to categories, while the first does not. In a case study we compared the quantity, relevance, and nature of the feedback collected with both UX evaluation apps to traditional evaluation methods. The feedback collected with the apps was considered highly useful by designers and provided more user stories and context than traditional evaluations.

ACM Classification Keywords

H.5.2. User Interfaces: Evaluation/methodology

Author Keywords

User Experience; Evaluation; Experience Sampling; User-centered Design; Crowdsourcing

THE DILEMMA OF MOBILE UX

The field of HCI has produced a large body of research on user-centered design strategies and methods and their use in developing interaction designs and concepts [38]. Many of these methods were originally developed for the design processes of desktop applications or web pages, but they are less suitable for interaction design or interface design beyond the desktop: For example, testing a headphone gesture interface in the lab might produce qualitative feedback or even quantitative measures about the usage itself and thereby document usability, but it will not provide any user stories or the type of feedback that will arise when users test the same headphone gestures in their everyday life, e.g., in a subway, and embarrass themselves in front of other passengers. What works well in the lab might be totally unacceptable in the wild.

In order to address this dilemma, we worked closely with designers at an international design agency[1]. Usability testing in the lab and expert reviews are part of their standard repertoire. However, running user studies on a regular basis during the short-cycled design process is made difficult by the necessary time for preparing the study, recruiting people, running the study, and then drawing proper conclusions from the results. In addition to laboratory studies, diary studies help them to understand how people use a product or a service in its real context of use. In diary studies participants regularly report events and experiences in their daily lives [6, 8]. However, diary studies are also time- and labour-intensive to prepare and analyse. Thus practitioners – especially smaller design agencies – generally often struggle with applying this method in their daily work [2, 17].

Together with the designers of our partner agency, we therefore developed two prototypes of a UX evaluation tool in the form of a mobile app, by which we were hoping to simultaneously address several of the problems above. If used consistently, we expected this approach to have the potential to decentralize the entire process of running UX evaluations and shift feedback generation to a much larger audience. Hence, we picked the term *CrowdUX* for our app as well as for the title of this paper.

Mobile Technology: A Double-Edged Sword

Pure usability is often regarded a commodity today [30]. As technologies are becoming increasingly mobile, predicting and understanding the context in which they will be used is increasingly difficult. Nevertheless, understanding this context and the overall user experience (UX) is essential for predicting the success of such a product.

[1]http://www.designaffairs.com

DIS'16, June 04 - 08, 2016, Brisbane, QLD, Australia

ACM 978-1-4503-4031-1/16/06 $15.00
DOI: http://dx.doi.org/10.1145/2901790.2901814

On the other hand, the same technological progress also created new opportunities: Today most people carry smartphones wherever they go. These devices are powerful tools for capturing the varying context of the user and for recording feedback with little overhead.

Collecting Feedback in the Wild?

In this situation, we were invited by a design agency to explore together with them the question "What will come after the Usability lab?". The agency develops designs and interaction concepts for a wide range of products and services from industrial designs (e.g., headphones and fridges) to purely virtual products (e.g., mobile and desktop software).

After analysing the work context, goals, and challenges the design agency faces, we formulated the following vision: What if we trusted a wider base of product users to give feedback autonomously (loosely speaking, by *crowdsourcing* the UX feedback)? Could this save money and at the same time capture both the context and the overall UX better? Can such a *Crowd-UX* become a new paradigm in UX engineering, just like *Crowdsourcing* and *Crowdfunding*?

To explore this vision we built a tool that is (1) technically lightweight for scalable roll-out and (2) facilitates a lightweight interaction, so that participants can easily integrate its use in their daily lives. The basic tool can be extended by gamification and motivational features to further encourage participation. With this prototype we then explored three research questions:

RQ1 Is the feedback collected with a UX evaluation app useful and relevant for designers?

RQ2 How does the interface design of the UX evaluation app (free-form vs. structured) influence feedback?

RQ3 What is the difference between expert and non-expert feedback when given through a UX evaluation app?

RELATED WORK

Before discussing *UX feedback* in the wild, we will first establish a common understanding of *UX* as opposed to pure *usability*. Starting from the literature, we will review shortcomings of existing UX evaluation methods and potential advantages of a lightweight UX evaluation app.

Usability vs. User Experience

The goal of the proposed UX evaluation application is to collect feedback on true *user experience* in addition to collecting feedback purely focused on *usability* (as usually done in lab studies) [43]. However, measuring and evaluating UX is a major challenge [25]. To test whether our proposed method is more or less suitable than traditional methods for collecting holistic UX feedback, we need to define how *UX* is different from *usability*. ISO 9241-210 offers a clear distinction between Usability and User Experience Evaluation. *UX* is defined as a "person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service" [1]. However, a global understanding of *UX* has yet to emerge. One source of discrepancies is the gap between academia and industry [15].

In industry the term *UX* is often used as a collective term for efficiency, effectiveness, usability, user-centered design, or overall quality-of-use [3, 5, 15, 25]. In academia, various researchers have proposed frameworks to define and frame *UX*, e.g., Norman [32], McCarthy and Wright [27], Jordan [21], Hassenzahl [16], and Desmet and Hekkert [10].

In our work, we use a framework developed by Kort et al. [24, 42] (Figure 1) that merges ideas of the work from McCarthy and Wright [27], and Desmet and Hekkert [10]. This framework presents three aspects that are relevant for *UX*, namely *compositional aspects*, *aspects of meaning*, and *aesthetic aspects* and incorporates the temporal phases of an experience presented by McCarthy and Wright [27].

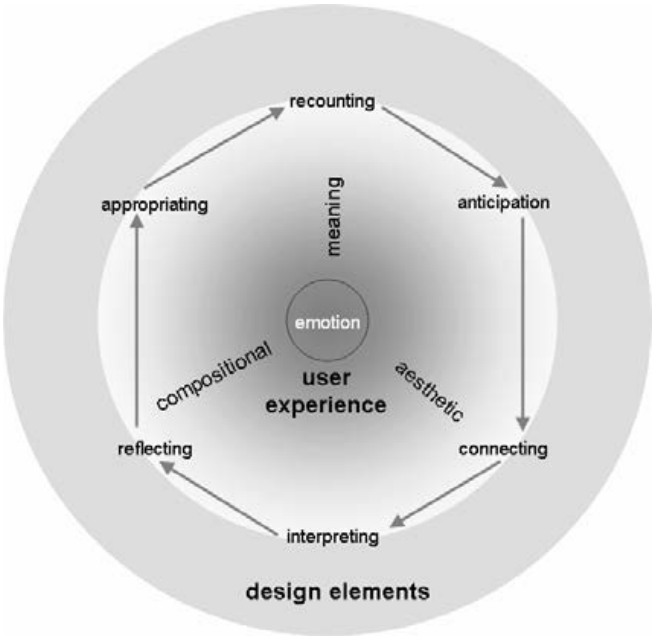

Figure 1. UX framework by Kort et al.[24] stressing *compositional*, *aesthetic*, and *meaning-related* aspects. Figure based on Kort et al.[24]

Compositional aspects address pragmatic properties of a product, e.g., usability and effectiveness. *Aesthetic aspects* address people's sensory perception, e.g., look and feel, sound or coloring of a product. This includes the user's understanding of how a product is working. *Aspects of meaning* address the hedonic properties of the product, which are cognitive processes concerning the users' higher goals, e.g., needs and desires. All these aspects are represented and can be influenced by properties of design elements of a product or service. To consider the user experience holistically, all three aspects need to be observed.

In their study, Vermeeren et al. [42] used these three aspects to categorize feedback. Their goal was to evaluate which method is most suitable to generate different kinds of feedback on a peer-to-peer television system. They found that lab studies and expert reviews were suitable to collect feedback on compositional aspects of a product while logging and sensing data was not. Both aesthetic aspects and aspects of meaning were found mainly in the laboratory study through

spontaneous *think-aloud* utterances, as well as in retrospective interviews. We adopt a similar approach as Vermeeren et al. [42]. However, while they [42] compared traditional laboratory studies and expert reviews to logging and sensing data, we compare them to a UX evaluation app that trusts users to give feedback autonomously.

Why Crowdsourcing UX?

Formative feedback during the design process is important for understanding the cognitive and emotional impact of design decisions as well as for inspiring designers [4, 7, 37, 39]. A plethora of design research methods exists, as evidenced by the views of both Roto et al. [35] and Vermeeren et al. [43]. However, while many of these methods include receiving feedback from potential end-users, they do not meet practitioners' needs for *practicability*, understanding *context of use* and *scientific quality* [43]:

(1) *Practicability:* Many methods were developed in an academic context [22, 34, 41, 43]. However, in practice, designers require time-efficient, easy-to-use evaluation methods [41]. In their evaluation work, practitioners are often constrained by budgets and limited access to potential subjects (leading to low statistical reliability of their evaluation) [19, 23]. Hence, evaluation methods have to be adaptable to industry-specific settings and skills.

(2) *Lack of Context of Use:* Methods that collect momentary data (e.g., self-reports) are highly valuable in the design process [3, 43] . However, such methods are difficult to carry out in a valid and rigorous way and are therefore rarely used in practice [43]. Hence, designers often lack an understanding of the actual context of use of the products they are designing.

(3) *Lack of Long-Term Studies:* Evaluating the whole user experience from anticipation to recounting (as proposed by McCarthy and Wright [27]) is usually not practicable and too expensive [3, 18], even in bigger agencies.

In summary, existing research calls for *practicable* methods which allow to collect *long-term* data about the *context of use*. We investigate crowdsourcing UX feedback with a lightweight UX evaluation app as a method that aims to better meet these three challenges. We actually do not use the term *crowdsourcing* in its strict sense here, referring to micro task platforms such as Mechanical Turk. Instead, we use it as a somewhat catchy term for the opportunity to leverage the opinions and contribution of many potential users and to combine and evaluate them systematically. As participants have to be equipped with the product under evaluation, the appropriate number of participants for such a study is limited by the availability of the product (A software product may be installed remotely while a physical product has to be mailed).

How to Crowdsource UX

The opportunities of crowdsourcing have previously been discussed in the fields of design [44, 45] and product development [36]. Our goal was to find out whether this method is useful for a design agency concerned with the design of digital and physical products (RQ1). Moreover, practitioners who build and use a UX evaluation app have to make several decisions, e.g., whether and how feedback is structured or categorised (RQ2) and what expertise level their participants should bring (RQ3). To inform these decisions, we built on previous work on mobile diary studies [8, 9, 14], mobile ethnographies [28], and crowd-testing [26, 44, 45].

RQ1: Usefulness of a UX Evaluation App

The basic functionalities of our UX evaluation app are taking pictures, typing text (as recommended by [8, 20]) and evaluating the overall experience on a likert-scale (as successfully used by [28]). Even though others reported positive results with similar methods [8, 20, 28] a lack of participant engagement and a lack of guidance on how to give feedback present potential problems of the proposed method.

Fading participant commitment is a well-acknowledged limitation and challenge of diary studies [6, 14, 33]. Researchers typically address this by keeping diary entries short to complete [6], by using context-sensing to minimise the effort for the user [14], or by automatically reminding participants to stop, reflect on, and report the current experience [9, 14, 20, 33] (known as the experience sampling method, ESM).

Moreover, in the proposed method, the investigator is not present and can't give guidance when participants are unsure how to give feedback. For example, Muskat et al. [28] found that some participants would have needed guidance on *how* to give feedback and *what* to give feedback *on*. Even though their basic application (with the options to rate the overall experience, add a description, and media content) helped to capture critical events and opportunities of improvement of the museum experience under evaluation, participants were often uncertain what to capture and mentioned many different aspects in one feedback entry [28]. These problems might be reduced by providing guidance through structure in the UX evaluation - a design decision that we aim to evaluate in RQ2.

RQ2: Structured vs. Free-Form Feedback

While Muskat et al. [28] designed their mobile ethnography application to be open to all kinds of feedback, others developed and evaluated more structured feedback tools, e.g., allowing researchers to define per-capture questions that participants answer about each captured photo before uploading it [8]. Some feedback tools are structured according to a set of guidelines. Luther et al. [26], for example provided a feedback structure consisting of seven key design principles: Readability, Layout, Emphasis, Balance, Simplicity, Consistency, and Appropriateness. Similarly, Xu et al. [44, 45] structured feedback in five categories: elements, first notices, impressions, goals, and guidelines. They found that impressions feedback was perceived as most helpful by designers, followed by feedback on goals, with guidelines rated as the least helpful feedback [44]. Both the free-form feedback in Muskat et al.'s study [28] and the structured feedback in Luther et al.'s [26] and Xu et al.'s [44, 45] studies contained valuable information. In our study we aim to better understand the consequences of providing structure by comparing both versions in one case study.

RQ3: Experts vs. Non-experts

The proposed UX evaluation app can be used for tests with both UX and usability experts and non-experts. Expert feedback is often regarded as the gold standard [26, 44, 45]. However, the superiority of expert feedback is at least debatable. Existing research suggests that crowd critique (by a non-expert crowd) matches expert critique better than random critique [26] and in studies where large amounts of feedback were collected (e.g., large-scale prototype testing), non-expert crowd feedback was even more useful than feedback from a limited number of experts [23]. Furthermore, Luther et al. found that many assumed "false positives", issues identified by crowd workers but not by experts, pointed out legitimate issues [26]. Similarly, Nielsen and Molich [31]) revised their list of usability issues to include points only identified by novices but not by experts.

Taking a closer look at the nature of expert and non-expert feedback, it seems that expert feedback has essentially different qualities than non-expert feedback. For example, in Luther et al.'s [26] study a participant commented that novices provided "good emotional feedback" while experts offered "a higher-level technical critique". In our study we aim to gain a deeper understanding of the different nature of expert and non-expert feedback (RQ3).

In summary, similar approaches (e.g., mobile diaries, mobile ethnography, and crowd testing) have been successfully applied to gather user feedback in various contexts e.g., to evaluate the use of ubiquitous and mobile technology [8, 9, 14], the museum experience [28], or graphic design work [44, 45, 26]. In light of this promising previous work, we set out to get a better understanding of the applicability of this approach to the context of a design agency. Furthermore, we aim to investigate how the design of the UX evaluation app (free-form vs. structured) and the expertise (UX and usability expert vs. non-expert) influence the nature, the quantity and the quality of collected feedback.

DESIGN OF A LIGHT-WEIGHT UX EVALUATION TOOL

In a workshop with three designers of our partner agency, we defined requirements for our future UX evaluation tools:

(I) UX evaluation tools should collect 'lots of' and 'good' feedback with a high level of diversity.

(II) The time spent in study preparation and post-hoc analysis of the results should be minimized.

(III) Users should be contacted directly without the detour through expensive agencies; and users should spend less time in user studies leading to cheaper reimbursement.

In this paper, we primarily focus on the first requirement of the designers. We investigate how feedback quality, quantity and diversity compares between traditional user testing and our evaluation tools.

According to the designers we spoke to, UX evaluation tools of the future shall collect feedback of comparable (1) *quantity*, (2) *quality*, and (3) *diversity* to traditional user testing methods. Designers value (1) *feedback quantity*, i.e., the number of critical statements users provide: they are a source for qualitative quotes and provide a quantitative basis to argue for design decisions with managers and costumers. Designers value (2) *feedback quality* as an important inspiration for new design impulses and ideas which push the design process further. Designers value (3) *diverse* feedback, i.e., when the sum of feedback statements provides insight into various design aspects of the product such as usability, aesthetics or user experience stories [42] and thereby provides a holistic perspective on a product or service.

UI for Unmoderated Feedback Collection

Our partner design agency works on a heterogeneous set of products and services, ranging from traditional web pages, to services for renting bikes, to gestural interaction concepts for headphones. Traditional remote user studies applied in desktop environments are one way of unguided feedback collection. They are, however, not applicable for products where interaction takes place on other devices or in different places, e.g., outdoors. We decided to use mobile phones as a means for providing feedback as we can reasonably expect people to have them in reach in most situations. Users can provide text, video, images or audio to quickly describe the context, their feelings and thoughts.

Figure 2 shows the user interface of our prototype *CrowdUX*: users find a list of feedback items which can be edited or removed (see Fig. 2a) and a button to add new feedback (see Fig. 2b). New feedback is provided using a dialog wizard (see Fig. 2c) proposing – not enforcing – that users provide a title, a capture (photo, audio, or video), a description and a feedback rating. The rating is performed on a 7-point Likert scale from 1-very negative to 7-very positive.

Figure 3 shows a variation of *CrowdUX*, which we called *SortedCrowdUX*. It introduces two additional features: (A) a user-assessed categorization of feedback items (see Fig. 3a) and (B) an indication (on a color scale from red via white to green) of the amount of already provided feedback.

With the user-assessed feedback categorization, we expected users to provide more diverse feedback because the interface suggested a variety of potential feedback aspects. Users first decide on a category, e.g., branding or packaging (see Fig. 3 left) and then see a list of items provided for that category. Note that all actions from Figure 3 right are just as in *CrowdUX* (see Fig. 2). The 18 categories are derived from the HUX (Holistic User Experience) Framework [40] but we excluded some product-irrelevant categories (e.g., smell). As a promising side-effect, the categorized feedback reduces the post-processing time for designers because categories don't need to be assigned post-hoc. With the red-to-green visual feedback about users' *'performance'*, we expected to motivate users to provide more feedback in order to push the number up and the color from red to green.

Implementation of the Prototypes

For the implementation we used angularJS ESCMA Script 6 and HTML5. The feedback is stored locally with PouchDB[1]

[1] http://pouchdb.com/

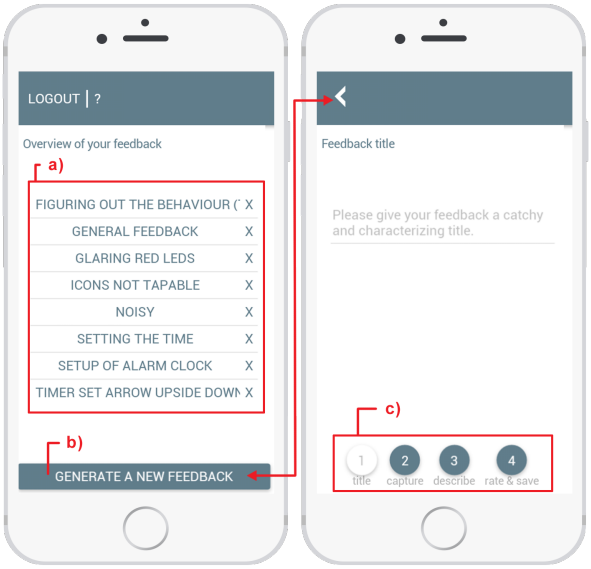

Figure 2. With *CrowdUX* users can provide feedback to products and services: left, (a) list of submitted feedback items, (b) button to create new feedback; right, dialog wizard guiding through feedback (c) title, capture (e.g., photo), description and rating.

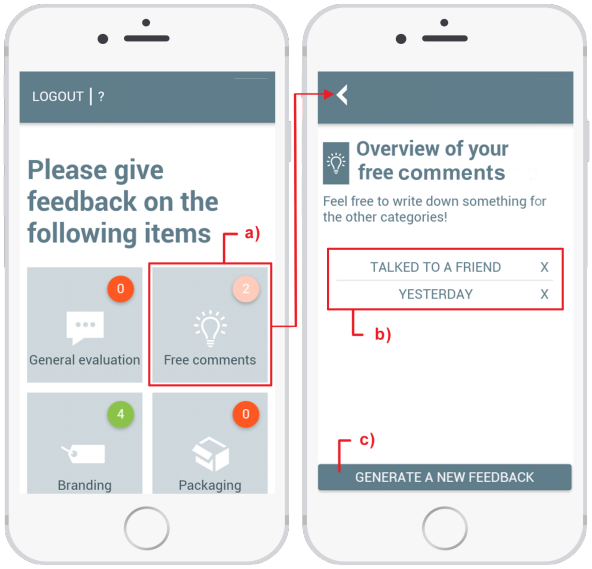

Figure 3. In addition to *CrowdUX*, with *SortedCrowdUX* users can categorize their feedback: left, (a) example tile representing the category "free comments"; right, (b) list of submitted feedback items in category "free comments", (c) button to create new feedback in category "free comments".

and remotely with CouchDB[2]. PouchDB is optimized to run in browsers and stores data if the user is offline. In case there is a connection, the data will be synchronized with CouchDB, a document-based database in which feedback is stored as single documents. Each user has her own personal database, to which she has to log in when starting the web-app. In this database feedback can be added, edited and deleted. Each entry has a time stamp. The administrator has access to the results of all participants on the CouchDB side. For the UI we used and adapted components of Angular Material[3].

CASE STUDY: APP-BASED VS. TRADITIONAL

The Designers in our partner agency traditionally involve two user groups, novices and experts, in their user testing: With novices they perform a laboratory study, while they ask experts to write an expert review. In our study, we compared collecting feedback through these traditional methods with collecting feedback through our prototypes. We were particularly interested in how feedback quantity, quality and diversity would differ between both types of evaluation and between both user groups.

Participants

We recruited 30 paid (5€) participants (14 female), and assigned 5 per condition, based on Nielsen's recommendation to test with 5 users for the best benefit-cost ratio [29]. We chose a lower compensation than most agencies would award their participants, (e.g., 50-100€), as a higher compensation is likely to reveal equally good or better results. 15 participants were UX-experts, defined as someone who has experience with both conducting and participating in user experience studies. Experts were employees of our partner agency, e.g., as UX designers or usability engineers and researchers in the area of UX and usability, assigned randomly to the three conditions. The other 15 participants were non-experts, defined as someone who has no experience with conducting and participating in user experience studies. Non-experts were, for example, students, teachers, or employees in Communication Management, PR, or Business Development. The average age of all participants was 30.

Method

We ran a between-subjects study with a 3 *StudyType* (*Traditional* vs. *CrowdUX* vs. *SortedCrowdUX*) × 2 *Expertise* (*non-experts* vs. *experts*) design. Each expert and non-expert in each *StudyType* received the travel alarm clock *KLOCKIS* from IKEA[4] for evaluation (see Fig. 4). In all conditions, we gave users 5 explicit tasks with the clock.

StudyType Traditional

The traditional procedure of our partner design agency includes user testing with non-experts and expert reviews by experts. In our study, non-experts were asked to participate in a 30 minute experiment. The experimenter asked them to perform 5 tasks (e.g., set the alarm) in a think-aloud fashion and collected feedback in subsequent interviews. We provided each expert with a clock and asked them to write a 1-page review of the product based on the same 5 tasks.

[2]http://couchdb.apache.org/
[3]https://material.angularjs.org/
[4]http://www.ikea.com

Table 1. Tasks and motivational messages sent to study participants via text messages though out the course of the 8-day study

SMS 1	SMS 2	SMS 3	SMS 4	SMS 5
Hi tester, welcome to the test week! Have you already set up the time of your new clock? How do you like your new clock? Cheers, testaffairs	Hey tester, have you already set the alarm for tomorrow? How do you like it? Use the app for your feedback! Good night and sleep well! Cheers, testaffairs	Hi tester, how are you? Are you still using testaffairs? Don't forget to give feedback. We need your opinion! Cheers, testaffairs	Hi tester, today is Sunday! I hope you don't need a clock! But how about eggs for breakfast? What do you think about the timer? Please use the app to give feedback! Thanks and cheers, testaffairs	Hi tester, the week is almost over, Thank you for your feedback! So, how do you generally evaluate the clock? Last chance to give feedback :D Everything matters! Cheers, testaffairs

Figure 4. Each participant received a travel alarm clock *KLOCKIS* from Ikea for evaluation in a box with instructions.

StudyType CrowdUX and StudyType SortedCrowdUX
Both non-experts and experts were introduced to the feedback collection tools; they received an alarm clock and were asked to use the product at home during an 8-day period. All participants received another 5 tasks via text message on their phone and were reminded to keep posting feedback (see Tab. 1).

Data Collection
We stored all feedback in digital form: In *Traditional* we transcribed the interviews of non-experts and received digital reports from experts whereas in all other conditions we received the feedback they had posted digitally.

Data Cleaning and Postprocessing
We processed all textual feedback and separated it into *statements*. Semantically similar statements were consolidated into one *insight*.

Statements: we define statements as feedback that (1) contains one specific design aspect (2) of the alarm clock, and (3) is comprehensible. We split up feedback text, when it contained multiple statements: For example, a user might post one feedback: *"the alarm button is not visible, but the clock frame looks beautiful"*; in this case, we would have separated the feedback into two statements. We excluded 80 text statements of *Traditional*, 4 text statements of *CrowdUX* and 9 text statements of *SortedCrowdUX* because they were sarcastic, incomplete, incomprehensible, or did not contain feedback regarding *KLOCKIS*.

Insights: we define insights as semantically distinct information extracted from statements with relevance for the redesign of the product. Multiple statements often referred to the same issue or insight. For example, one user might have said *"the alarm button is not visible"* in *Traditional* and another user might have written *"I can not find the alarm button"* in *CrowdUX*. In this case both statements refer to the same insight.

Feedback Quality
Three designers from our partner agency, who did not participate in the study, assigned a priority ranging from 1 (low priority) to 5 (high priority) to each distinct insight. When designers assigned different priorities they discussed the insight together and agreed upon a priority together.

RESULTS

Highest Feedback Quantity with Traditional User Studies
Those conditions in which feedback was given orally yielded significantly more feedback than all other conditions. This is evidenced by an interaction effect between *Expertise* and *StudyType* $F_{2,24}$=10.29, $p < 0.0001$ on *feedback quantity* (see Fig. 5): Non-experts provided significantly more feedback in *Traditional* – on avg. 48.6 statements (CI=[41.22, 55.98]) – than in *CrowdUX* (*mean*=14.2, CI=[6.82, 21.58]) and *SortedCrowdUX* (*mean*=16.2, CI=[8.82, 23.58]). A Tukey post-hoc test reveals that experts show no significant difference between *StudyType*s: on average, experts provided 13.2 statements in *Traditional* (CI=[5.8, 20.6]), 7.2 statements in *CrowdUX* (CI=[−0.2, 14.6]), and 8.6 statements in *SortedCrowdUX* (CI=[1.2, 16.0]).

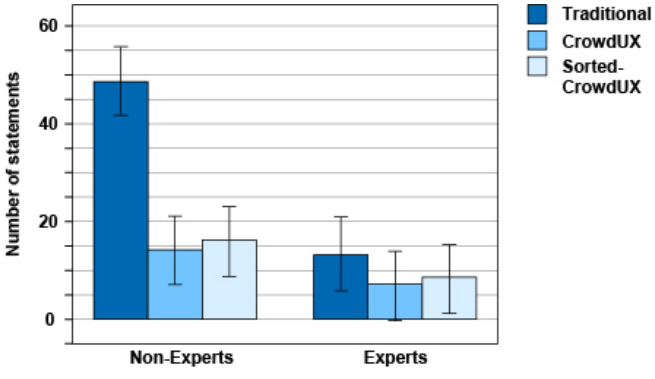

Figure 5. Number of feedback statements in each condition

This interaction effect can be explained by the fact that traditional user studies were the only evaluation method in

our study in which feedback was given orally. In all other conditions participants had to express their feedback in writing, which takes more time and effort. Hence, participants wrote down only the most relevant feedback. This explanation is also consistent with the finding that written feedback on average was more relevant to designers than oral feedback.

Higher Relevance of App Feedback

Participants gave more feedback of priority 5 with *CrowdUX* and *SortedCrowdUX* than in the *Traditional* conditions. However, this difference is not statistically significant. On average 55.4% of the feedback in *SortedCrowdUX* (CI=[42.2, 68.6]), 48.3% in *CrowdUX* (CI=[35.1, 61.5]) and 39.6% in *Traditional* (CI=[26.4, 52.8]) had priority 5 (Fig. 6). Furthermore, expert feedback in *SortedCrowdUX* had a significantly higher percentage of priority 5 feedback (*mean*=70.3, CI=[57.1, 83.5]) than non-expert feedback (*mean*=42.6, CI=[28.9, 55.8]).

Receiving more dense feedback (less feedback with low priority and more feedback with high priority) saves design agencies the time and money needed to process and analyze feedback of low priority.

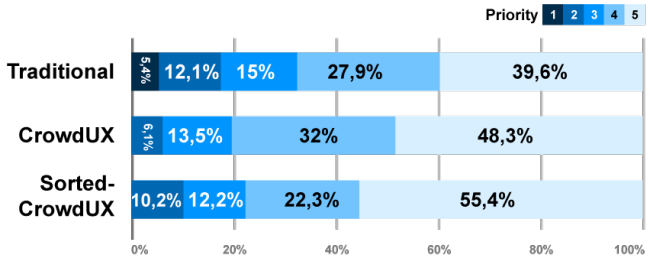

Figure 6. The average percentage of priority 5 feedback (out of all statements of a participant) was highest in *SortedCrowdUX*.

When looking closely, the higher density of feedback in *CrowdUX* and *SortedCrowdUX* was not only caused by the absence of low priority feedback, but also by additional high priority feedback that was not discovered in *Traditional*.

No Method Found All Important Insights

Out of 31 priority 5 insights in total, 23 were discovered in *Traditional* and 20 in both *CrowdUX* and *SortedCrowdUX*. Conversely, 8 priority 5 insights were missed in *Traditional*, and 11 each in *CrowdUX* and *SortedCrowdUX*. Hence, with 10 participants no single method found all important insights, indicating that combining methods might yield best results. Remarkably, even though *Traditional* yielded *3x* more feedback than *CrowdUX* and *SortedCrowdUX*, those statements contained only *1.15x* more priority 5 insights.

More Context with CrowdUX

Looking closer at the high priority insights missed in traditional user studies, it becomes evident that 62.5% of these insights are related to the context of use. One subject for example reported "I thought this clock has a radio clock and I do not have to set the time manually. Unfortunately I recognized - after waiting for five minutes - that this function does not exist". Other participants complained that the

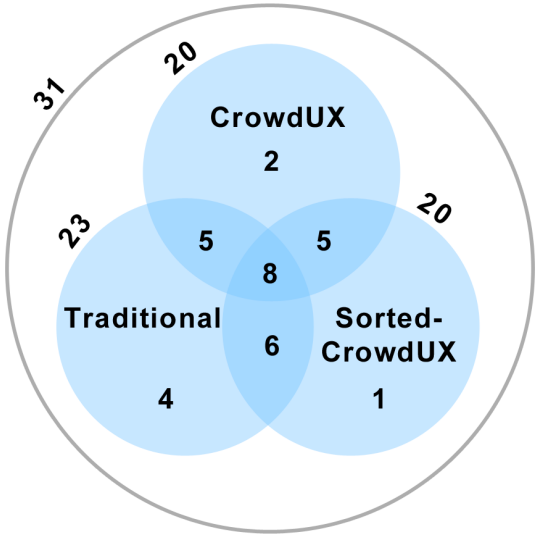

Figure 7. Of 31 insights with the highest priority, 23 were discovered in *Traditional* and 20 in each *CrowdUX* and *SortedCrowdUX* - no method discovered all important insights with 10 participants.

beeping of *KLOCKIS* (when being turned 90 degrees) is unpractical, when carrying it in a suitcase; they commented on the lack of time-zone adaption and automatic time setting; and they mentioned how much they enjoyed using the temperature function throughout the day. In contrast to usability issues, which are easily discovered in lab studies, this type of feedback entails scenarios of use, context of use and users' personal values and feelings when using the product.

To quantify what kind of feedback each method reveals, we distinguished between aesthetic, compositional and meaning feedback (according to the framework by Kort et al. [24]). Figure 8 shows the average amounts of each feedback aspect per condition. *CrowdUX* yielded the biggest proportion of *meaning* feedback (on average 37.3% of feedback given by a participant (CI=[26.4, 48.2]), which is significantly more than in *Traditional* ($F_{2,27}$=4.03, p = 0.026). In *Traditional* the average percentage of *meaning* feedback was 15.8% (CI=[4.9, 26.8]) and in *SortedCrowdUX* 25.6% (CI=[14.6, 36.5]).

This effect confirms that light-weight UX evaluation tools are in fact suitable for gathering context-rich user stories.

Traditional yielded the biggest proportion of *compositional* feedback (on average 63.3% of feedback given by a participant (CI=[49.4, 77.1]). In *CrowdUX* the average percentage of *compositional* feedback per participant was 38.7% (CI=[24.8, 52.5]) and in *SortedCrowdUX* 24.1% (CI=[10.2, 37.9]). This result is not surprising as studies in a usability lab are by nature focused on usability problems. *SortedCrowdUX* yielded the biggest proportion of *aesthetic* feedback per participant (on average 50.9% of feedback (CI=[36.7, 64.1]). In *Traditional* the average percentage of *aesthetic* feedback was 20.9% (CI=[7.2, 34.5]) and in *CrowdUX* 24.0% (CI=[10.4, 37.7]).

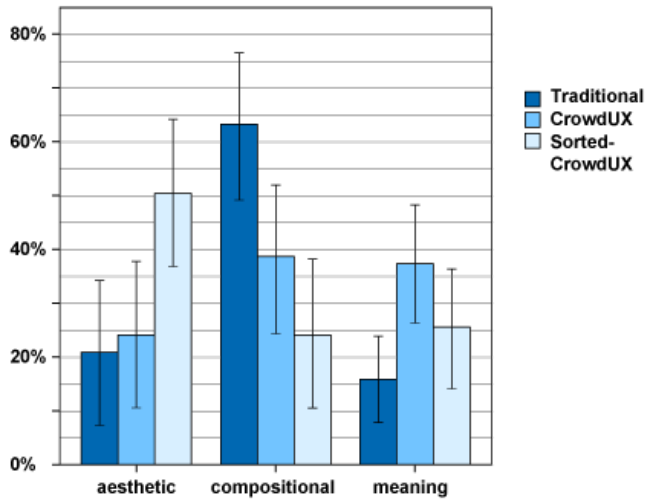

Figure 8. *CrowdUX* revealed more user stories and context (concerning the *meaning* aspect) while *Traditional* revealed more feedback about usability issues (concerning *compositional* aspects) and *SortedCrowdUX* yielded feedback about *aesthetic* aspects.

We suspect that this effect was partly caused by the feedback categories suggested by *SortedCrowdUX*. Hence, providing categories indeed seems to be a powerful tool to guide feedback. In particular, the feedback provided by experts was also found to be more useful when categories were provided.

Questions and Tasks Do Not Lead to More Feedback

The relatively smaller amount of feedback per participant with *CrowdUX* and *SortedCrowdUX* indicates that engaging users to give more feedback still is a major challenge of this type of mobile diary studies. In *CrowdUX* and *SortedCrowdUX* we sent participants text massages with questions, reminders, and tasks (in total five messages over the course of the 8-day study). Figure 9 shows the percentage of feedback items and observations triggered by each of these prompts: The percentage of self-initiated feedback per participant was significantly higher than the percentage of feedback triggered by other sources over all conditions $F_{1,29}$=67.33, $p < 0.0001$. In *Traditional* 13.0% of the feedback was triggered by a scripted question (CI=[7.7, 18.3]), 5.1% of the feedback was triggered by a task (CI=[−1.0, 11.2]), 11.5% of the feedback was triggered by a spontaneous question the instructor asked (CI=[9.0, 14.1]), and 64.1% of the feedback was self-initiated (CI=[56.1, 72.0]). In addition, the instructor made notes when she observed that the participant had difficulties. These instructor observations accounted for 6.3% of the discovered insights (CI=[4.7, 7.9]). In *CrowdUX* 15.6% of the feedback was triggered by a scripted question (CI=[10.3, 21.0]), 12.0% was triggered by a task (CI=[5.9, 18.1]), and 72.4% was self-initiated (CI=[64.4, 80.4]). In *SortedCrowdUX* 1.4% of the feedback was triggered by a scripted question (CI=[−3.9, 6.8]), 0.8% by a task (CI=[−5.3, 7.0]), and 97.3% was self-initiated (CI=[89.3, 105.3]). As there was no investigator present in *CrowdUX* and *SortedCrowdUX*, there were no observations or spontaneous questions in these conditions.

These results show that questions, tasks and reminders triggered only a small amount of feedback. Hence, these measures do not solve the problem of low participant engagement of mobile diary studies. Strategies that could help to tackle this problem include gamification [11], creative reimbursement strategies [33] or persuasive system design [13]. An alternative strategy for collecting sufficient feedback, of course, is to increase the number of participants.

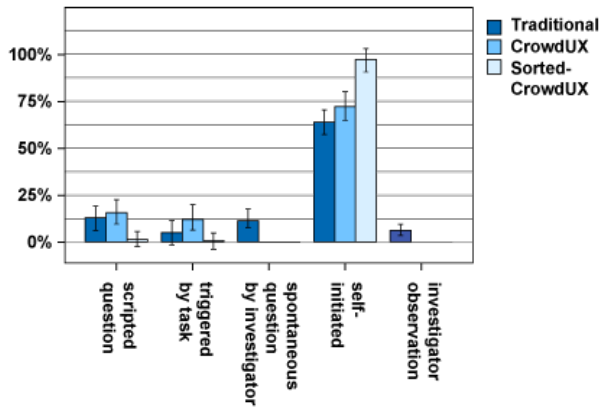

Figure 9. Most of the feedback in all three conditions was self-initiated by participants and not triggered by questions or tasks.

DISCUSSION AND CONCLUSION

In this paper, we have discussed methods for collecting user feedback through the use of light-weight tools in the form of mobile apps. In collaboration with a design agency, we conducted a case study comparing the feedback generated with a UX evaluation app by both experts and non-experts to feedback generated in traditional user tests and expert reviews. The goal of our case study was to better understand how the choice of method and participants influences the (quality and quantity of) feedback obtained from the perspective of a design agency.

RQ1: Lab Studies vs. UX Evaluation App

Our first research question was to evaluate whether feedback collected with a UX evaluation app is valuable for designers. Overall our results showed that this was the case: Feedback was understandable, rich, and relevant for designers. Moreover, evaluations with both *CrowdUX* and *SortedCrowdUX* yielded feedback that did not emerge in *Traditional* (Fig. 7). In particular, this feedback contained rich user stories and context of use while the *Traditional* method revealed more usability feedback.

RQ2: Structured vs. Free-Form Feedback

Our second goal was to understand the influence of design properties of the UX evaluation app. For this, we compared two versions of a UX evaluation app, one in which participants were asked to specify the category of feedback (*SortedCrowdUX*) and one in which they were not

(*CrowdUX*). We found that *SortedCrowdUX* yielded less user stories and context of use and more feedback regarding the aesthetics of the product. These results may reflect the fact that some of the proposed categories were related to aesthetics e.g., colour and haptics. While we used categories to guide feedback in certain directions, we were not aware that they could also prevent users from entering feedback that does not fit into any category, despite the existence of a category called *Other Comments*. Another drawback of providing categories was that users might have misunderstood some categories and therefore assigned feedback erroneously.

RQ3: Experts vs. Non-Experts

Our third research question was about the differences between expert and non-expert feedback. In *CrowdUX* we did not find significant differences in feedback quality between expert and non-expert feedback. However, *SortedCrowdUX* yielded especially useful feedback from experts.

Compared to non-experts, experts had a better understanding for the feedback categories defined by designers. Compared to expert reviews, evaluations with *SortedCrowdUX* allowed experts to spend time *living* with the product and they were more likely to discover issues related to the context of use, in comparison with evaluating the product in an artificial setting within 30 minutes. We hence argue that expert evaluations with *SortedCrowdUX* can leverage some of the same advantages as autoethnographies [12]. Simultaneously they can also provide the same benefits as heuristics in expert reviews [31]: Designers can choose the categories relevant for the design of a specific product before the evaluation.

Outlook: A Wide Spectrum of Usage Scenarios

We believe that lightweight UX evaluations as the ones in our case study can be a useful tool in many usage scenarios, including a variety of products and services in different phases of the development cycle, and including both local and global, as well as short- and long-term studies. The presented prototypes are extensible in various directions. First, in future work we plan to extend the prototype itself with the option to provide audio feedback and with gamification and creative reimbursement strategies and to investigate the effects. Second, in future work parameters such as the expertise of participants, length of study and categories could be integrated in the setup process of a study. In this way, designers could simultaneously leverage the benefits of three established methods: expert reviews, traditional user tests, and diary studies. As a third way to further explore the potential of the presented method, the prototypes could be used with remotely recruited crowdsourced participants, e.g. asking for feedback on a software product.

The Case for Widespread Lightweight Tools

In summary, our case study shows that lightweight UX tools have become technically feasible and present a valuable addition to traditional evaluation methods. Practicability and the possibility to collect information about the context of use are substantial advantages of the proposed method, making it especially suitable for projects with a tight budget and the need for rigorous evaluation.

ACKNOWLEDGMENTS

We thank Alexander Peters and Johannes Huber, designers at our partner agency, who contributed to our research project with insightful stories, experiences and opinions about the work context and the design processes in their design agency. Furthermore, we thank all participants for their time, feedback, and engagement throughout the study.

REFERENCES

1. ISO 9241:210:2010: Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems. Standard, International Organization for Standardization, Geneva, CH, March 2010.
2. Rui Alves, Pedro Valente, and Nuno J. Nunes. The state of user experience evaluation practice. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, pages 93–102. ACM, 2014.
3. Javier A. Bargas-Avila and Kasper Hornbæk. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2689–2698. ACM, 2011.
4. Nigel Bevan. Classifying and selecting UX and usability measures. In *International Workshop on Meaningful Measures: Valid Useful User Experience Measurement*, pages 13–18, 2008.
5. Nigel Bevan. UX, usability and ISO standards. In *26th Annual Conference on Computer Human Interaction (CHI), ACM SIGCHI*, 2008.
6. Niall Bolger, Angelina Davis, and Eshkol Rafaeli. Diary methods: Capturing life as it is lived. *Annual review of psychology*, 54(1):579–616, 2003.
7. Michael Burmester, Kilian Jäger, Laura Festl, and Marcus Mast. Studien zur formativen evaluation der user experience mit der valenzmethode. *Reflexionen und Visionen der Mensch–Maschine-Interaktion–Aus der Vergangenheit lernen, Zukunft gestalten*, 9:567–572, 2011.
8. Scott Carter and Jennifer Mankoff. When participants do the capturing: the role of media in diary studies. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 899–908. ACM, 2005.
9. Sunny Consolvo and Miriam Walker. Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Computing*, (2):24–31, 2003.
10. Pieter Desmet and Paul Hekkert. Framework of product experience. *International journal of design, 1 (1) 2007*, 2007.
11. Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: defining gamification. In *Proceedings of*

the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, pages 9–15. ACM, 2011.

12. Carolyn S. Ellis and Arthur Bochner. Autoethnography, personal narrative, reflexivity: Researcher as subject. 2000.
13. Brian J. Fogg. Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(Dec.), 2002.
14. Jon Froehlich, Mike Y. Chen, Sunny Consolvo, Beverly Harrison, and James A. Landay. Myexperience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the 5th international conference on Mobile systems, applications and services*, pages 57–70. ACM, 2007.
15. Marc Hassenzahl. User experience (UX): towards an experiential perspective on product quality. In *Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine*, pages 11–15. ACM, 2008.
16. Marc Hassenzahl and Noam Tractinsky. User experience – a research agenda. *Behaviour & information technology*, 25(2):91–97, 2006.
17. Mats Hellman and Kari Rönkkö. Is user experience supported effectively in existing software development processes? In *Proc. of COST294-MAUSE Workshop on Valid Useful User Experience Measurement (VUUM). Reykjavik, Island*, pages 32–37, 2008.
18. Kasper Hornbæk. Current practice in measuring usability: Challenges to usability studies and research. *International journal of human-computer studies*, 64(2):79–102, 2006.
19. Tobias Hoßfeld, Christian Keimel, Matthias Hirth, Bruno Gardlo, Julian Habigt, Klaus Diepold, and Phuoc Tran-Gia. Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. *Multimedia, IEEE Transactions on*, 16(2):541–558, 2014.
20. Stephen Intille, Charles Kukla, and Xiaoyi Ma. Eliciting user preferences using image-based experience sampling and reflection. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, pages 738–739. ACM, 2002.
21. Patrick W. Jordan. *Designing pleasurable products: An introduction to the new human factors*. CRC press, 2002.
22. Pekka Ketola and Virpi Roto. Exploring user experience measurement needs. In *Proc. of the 5th COST294-MAUSE Open Workshop on Valid Useful User Experience Measurement (VUUM). Reykjavik, Island*, pages 23–26, 2008.
23. Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.
24. Joke Kort, Arnold P.O.S. Vermeeren, and Jenneke E. Fokker. Conceptualizing and measuring user experience. In *Proc. Towards a UX Manifesto, COST294-MAUSE affiliated workshop*, pages 57–64, 2007.
25. Effie L-C. Law, Virpi Roto, Marc Hassenzahl, Arnold P.O.S. Vermeeren, and Joke Kort. Understanding, scoping and defining user experience: a survey approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 719–728. ACM, 2009.
26. Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. Structuring, aggregating, and evaluating crowdsourced design critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 473–485, New York, NY, USA, 2015. ACM.
27. John McCarthy and Peter Wright. Technology as experience. *interactions*, 11(5):42–43, September 2004.
28. Matthias Muskat, Birgit Muskat, Anita Zehrer, and Raechel Johns. Generation y: evaluating services experiences through mobile ethnography. *Tourism Review*, 68(3):55–71, 2013.
29. Jakob Nielsen. Guerrilla hci: Using discount usability engineering to penetrate the intimidation barrier. *Cost-justifying usability*, pages 245–272, 1994.
30. Jakob Nielsen. *Usability engineering*. Elsevier, 1994.
31. Jakob Nielsen and Rolf Molich. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 249–256. ACM, 1990.
32. Donald A. Norman. Introduction to this special section on beauty, goodness, and usability. *Hum.-Comput. Interact.*, 19(4):311–318, December 2008.
33. Leysia Palen and Marilyn Salzman. Voice-mail diary studies for naturalistic data capture under mobile conditions. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 87–95. ACM, 2002.
34. Virpi Roto, Marianna Obrist, and Kaisa Väänänen-Vainio-Mattila. User experience evaluation methods in academic and industrial contexts. In *Interact 2009 conference, User Experience Evaluation Methods in Product Development (UXEM'09), Uppsala, Sweden*. Citeseer, 2009.
35. Virpi Roto, Heli Väätäjä, Satu Jumisko-Pyykkö, and Kaisa Väänänen-Vainio-Mattila. Best practices for capturing context in user experience studies in the wild. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pages 91–98. ACM, 2011.

36. Jonathan Rubin. Can you crowdsource your user experience research? (online at http://www.digitalgov.gov/2014/12/09/can-you-crowdsource-your-user-experience-research/), December 2014.

37. D. Royce Sadler. Formative assessment and the design of instructional systems. *Instructional Science*, 18(2):119–144, 1989.

38. Liz Sanders. On modeling an evolving map of design practice and design research. *interactions*, 15(6):13–17, 2008.

39. Donald A. Schön. *The reflective practitioner: How professionals think in action*, volume 5126. Basic books, 1983.

40. Claude Toussaint, Marc Toussaint, and Stefan Ulrich. Hux–measuring holistic user experience. *Tagungsband UP12*, 2012.

41. Kaisa Väänänen-Vainio-Mattila, Virpi Roto, and Marc Hassenzahl. Towards practical user experience evaluation methods. *EL-C. Law, N. Bevan, G. Christou, M. Springett & M. Lárusdóttir (eds.) Meaningful Measures: Valid Useful User Experience Measurement (VUUM)*, pages 19–22, 2008.

42. Arnold P.O.S. Vermeeren, Joke Kort, Anita Cremers, and Jenneke Fokker. Comparing ux measurements, a case study. In *Proceedings of the International Workshop on Meaningful Measures: Valid Useful Experience Measurement, Reykjavik, Iceland, June*, volume 18, pages 72–78, 2008.

43. Arnold P.O.S. Vermeeren, Effie L-C. Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, pages 521–530. ACM, 2010.

44. Anbang Xu, Shih-Wen Huang, and Brian Bailey. Voyant: generating structured feedback on visual designs using a crowd of non-experts. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1433–1444. ACM, 2014.

45. Anbang Xu, Huaming Rao, Steven P. Dow, and Brian P. Bailey. A classroom study of using crowd feedback in the iterative design process. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1637–1648. ACM, 2015.