ABSTRACT
A method for primed product comparisons was developed, based on the methodological considerations of emotional appraisal process and affective mental contents. The method was implemented as a computer tool, which was utilised in two experiments (N = 18 for both). Ten adjectives served as primes, and five drinking glass pictures as stimuli. Participants’ task was to choose a preference between two glasses, given the priming adjective. The results validate the method by providing test-retest reliability measures and showing convergence with questionnaires. Further, different evaluation times between the primes and the stimuli reveal the existence of different mental processes associated with various aspects of product experience, as predicted by appraisal theory. The results have various implications for experience research and development in HCI, as they demonstrate how the method can be used for product evaluation and the analysis of the mental processes, which users use to evaluate the products.

Author Keywords
product experience; affective mental contents; primed product comparisons

ACM Classification Keywords
H.5.2. User Interfaces: Evaluation/methodology, User-centered design.
of HCI research, the self-report itself allows only a limited access to the largely unconscious experience process. It is, of course, possible to analyse latent experience structures via statistical analysis of questionnaire data [34, 35], but clever experimentation should also be used to explore the unconscious cognitive processes behind experience in HCI. While the subjective self-report is necessary for understanding the meaning of the experience, more objective measures of this experience process are required to postulate and test general mechanisms of experience.

Experience has many dimensions, but one of the most important is emotion [2, 34]. There are different psychological theories of emotion, such as basic emotions [19], core affect [3], and appraisal theory [39]; here the interest is in the appraisal theory, because of its ability to elaborate between the emotion process and emotional experience. Appraisal is a cognitive analysis of an event, which establishes the personal significance of the event [23, 25, 38, 39]. This analysis proceeds as a process, which involves multiple component levels [39], different sources [41], distinct dimensions [40], and complex interactions between these elements [39, 41].

While all five components of appraisal (cognitive, neurophysiological, motivational, motor expression, and subjective feeling) are relevant in HCI research [44], here the focus is on the cognitive and the subjective feeling components. The cognitive component integrates information about the appraised event from different sources, and proceeds mostly non-consciously. Subjective feeling is the mentally represented, consciously experienced part of the emotion process, analysable as the affective contents of mental representation [33, 39]. Affective mental content refers therefore to the information, which users have about their feelings in their conscious mental representations. In the method developed below, the connection between conscious product experience and the non-conscious cognitive processing is explored.

The information sources for the appraisal process are perceptual stimuli, associative processing, and reasoning [41]. Perceptual stimuli are directly detected and quickly processed events, such as pain sensations, and do not involve mentally complex processes. Associative processing is fast and automatic, but involves memory to associate meanings with events. Reasoning is slow and consciously controlled, and constructs linguistically encoded meanings. This threefold ranking of appraisal sources is similar to Norman’s [28] three-level visceral-behavioural-reflective framework of product experience, where the physical dimensions of the products are appraised on the visceral level, while the appraisals using more culturally interpreted criteria happen on reflective level.

Because the three sources of appraisal have different computational demands [41], it is possible to study experimentally how these levels are involved in product experience. The processing time from a stimulus event to conscious experience should be dependent on how linguistically and culturally complex elements the evaluation involves. Further, as the memory-based associative processing depends on spreading activation and priming [41], experiments utilising priming as an intervention should be able to posit and test causal hypotheses concerning the formation of product experience.

In previous experimentation in HCI, users have been shown to be capable of reliably judging the appeal of stimulus even after exposures as short as 50 milliseconds (ms) [26]. The subjective ratings of visual appeal are highly consistent between exposures of 500 ms and 10 seconds [45]. It seems that this phenomenon extends to various forms of evaluation, not only liking or disliking the object under evaluation [29], and suggests that a mental representation with various possible affective contents results quickly after the stimulus onset, hinting at fast non-conscious processes. However, these studies have not discussed the computational demands between different product evaluations, as implied by the three sources of appraisal.

Appraisal theory has been used to articulate theories of user or product experience [e.g., 16, 44], but the possibilities of this psychologically well-established and richly modelled theory have not yet been fully realised in HCI research. For example, the implications of the relationship between the cognitive information processing and the subjective feeling components of appraisal have not been investigated. Grounding the research of experience in HCI on appraisal theory would help to clarify the various conceptualisations and models in user and product experience research. At best, this could lead to testable causal models concerning the formation of experience in HCI.

Primed product comparisons
Perception influences subjective experience via conscious and unconscious processes [27]. This distinction can be demonstrated with a priming effect. Primes can be presented above the threshold of conscious awareness (supraliminal priming), or below it (subliminal priming; but note that this ‘threshold’ is not static). Both levels produce observable changes in the mental process [18]. In addition to the theoretical plausibility (explicated above), priming has empirically been shown to influence assessments of various dimensions of HCI, such as perceived aesthetics, quality, and usability [8, 31]. It is therefore a promising experimental technique for investigating the conscious and unconscious parts of product experience. However, the framework for using priming in user and product experience research is still in development (for progress on the framework, see e.g., [8, 50]).

The method for primed product comparisons stems from the notion that people are able to make conscious aesthetic preferences between products. The user is presented a prime, which is used to make a judgment between two
similar products. The use of a prime as the criterion for comparison allows investigating the relationship between conscious and unconscious product experience, as the prime can be presented either supra- or subliminally. In this study, the focus is on creating and validating the method, and primes over the conscious threshold are used; further studies will focus on subliminal priming. If the method produces valid evaluations of products, the results should correlate with other methods of product evaluation; this serves as the basis for the first hypothesis of the study.

**H1.** There is a strong positive correlation in the evaluation of stimuli between primed product comparison and questionnaire responses.

H1 serves to validate the method, but does not argue why questionnaires should be replaced with primed product comparisons. In primed product comparisons, the participants are asked to make their judgment as fast as possible: this allows the study of processing times associated with different aesthetic judgments. This kind of information is useful in HCI, as it allows researchers and designers to better understand how different product properties are mentally processed. The notion that the three sources of appraisal involve different processing times should be visible in longer judgment times with such appraisals, which have more linguistically and cultural complex associations. For example, visceral judgments such as ‘This object is heavy’, should be processed faster than evaluations of beauty or modernity of an object. We therefore propose that the method of primed product comparisons should produce differences between the primes, depending on how computationally demanding the subsequent evaluation task is. Further, while the evaluation time is affected by the prime, it should also be affected by the product pair.

**H2.** In primed product comparisons, mean reaction times are different between the primes.

**H3.** In primed product comparisons, mean reaction times are different between the products.

In addition, the effects of the second and the third hypotheses are also expected to interact, because for certain pairs, certain primed evaluations should be easier to process than for others. For example, the weight of two objects of clearly different size should be easier to appraise than of two very similar objects.

**H4.** In primed product comparisons, mean reaction times of a same prime are different depending on the product pair.

These hypotheses are tested in a laboratory environment with software developed for primed product comparisons. The use of computer software for assessing product or user experience is not a new idea, and many experience metrics have been computerised. Tools for assessing user experience or usability in interactive environments, such as web sites, are examples of this (e.g., UserTesting [47], Usabilityhub [21]). However, the methods for acquiring and analysing data used by these tools are not always based on rigorous scientific theories and operationalisation. Further, often such tools provide data via user feedback (e.g., communicative feedback, success rate of tasks or questionnaires), and the results focus on improvements of specific websites. Therefore, their use in studying the cognitive experience process is limited and new tools for this purpose are needed.

There are also computerised evaluation tools, which are developed based on scientific theory and operationalisation. Generally, these tools fall into two categories: tools that utilise self-report experience metrics for product and service evaluation, and tools that collect objective, such as psychophysiological data. Some tools serve as web-based research environments, and include data collection, data analysis, and result reporting (e.g., LEMtool [13], PrEmo [30], optimalSort [42], or AttrakDiff [1, 10]). Other tools have implemented the traditional ‘pen and paper’ -method in software environment (e.g., UEQ [46]).

One prospect of software-aided user research is that it makes collecting data and producing results faster and easier than the more traditional methods. Such tools are hence often emphasised when the focus is in fast results, and not necessarily in scientific analysis (e.g., in e-commerce environments with tools such as Google Analytics [24]). However, when considering software tools in the context of basic scientific research, it is vital to be aware of their methodological assumptions and evaluate the circumstances in which these tools are suitable for providing answers to the research problems. Here, the methodological assumptions have been made clear.

**METHOD**

**Equipment & Procedure**

The equipment for primed product comparisons consisted of a computer, a computer screen, and a reaction time (RT) switch with two buttons. A computer application (tool henceforth), was programmed for presenting the primes and the stimuli. Given a number of product pictures (stimuli) and words (primes), the tool iterated through all possible combinations of pairs of stimuli and primes. Primes and the stimulus pairs were presented on a computer screen.

In the two experiments reported here, one task consisted of one prime (one of ten adjectives) and one stimulus pair (two of five product pictures). A prime was first presented on the screen for two seconds, which is well enough for conscious recognition, making the prime supraliminal [27]. Then the prime was replaced with a stimulus pair, side to side in a randomised arrangement. The task of the participant was to press the RT button on the side of the stimulus, which more corresponded to the prime. For example, if the prime was festive, the task was to choose which one of the pair of stimulus pictures is more festive. This choice is called here preferential match or just preference, which is used instead
of plain ‘match’ to emphasise the subjective nature of appraisal process.

Two experiments were conducted with small differences, and are reported here together. In the first experiment, after the priming adjective had been displayed for two seconds, the pair of stimuli was presented for three seconds, after which the screen was cleared and the prime presented again. The participant had to indicate preference by pressing the RT switch as quickly as possible (the participants were asked to keep their index fingers at the buttons at all times). In the second experiment, the pair of stimuli was not cleared, and the participant was asked to indicate preference as soon as possible after having been presented the stimulus pair. In both experiments, after the preference, a message ‘OK’ was shown for two seconds, after which a new task was presented starting with the prime. The order of the tasks was randomised at the start of each trial. The participants were given two rest periods, the first after completing one third, and the second after completing two thirds of the tasks. The rest duration was up to the participant, but the rests were short, usually less than 15 seconds.

After conducting the tasks, the participants were shown each stimulus picture, one at a time in a randomised order, and asked to appraise the stimuli with semantic differential (SD) questionnaires. The questionnaire consisted of ten adjective pairs so that one adjective of each pair corresponded to an adjective used as a prime in the first part of the experiment. The scale of the questionnaire was from one to nine, one indicating close resemblance of the left-hand adjective, and nine indicating close resemblance to the right-hand adjective. Therefore, the participants rated each stimulus first using the primed product comparisons method, and then with a traditional SD questionnaire. This procedure allowed for testing H1.

Participants & Stimuli

For both experiments, \( N = 18 \) participants were recruited using a mailing list for those interested in participating in user psychological experiments. For the first experiment, the mean age of the participants was 21.8 years (SD = 2.6, age range 19–28). Fourteen participants were men, and four women. For the second, the mean age was 24.3 years (SD = 3.0, age range 20–31). Seven participants were men, and eleven women. There were no common subjects between the experiments.

The priming adjectives for the experiments were chosen from a study of drinking glass user experience [35]. From the 31 candidates, ten adjective pairs were chosen to represent various important drinking glass characteristics and to involve different sources of appraisal (H2), while keeping the amount of primes limited. The adjectives were also named important by professional glass designers. From each pair, one adjective was used as a prime, resulting in ten priming adjectives. The adjective pairs were following (boldface indicates the chosen prime): festive – mundane, light – heavy, modern – traditional, durable – fragile, practical – impractical, fleeting – timeless, angular – curvy, general-purpose – specific-purpose, decorated – undecorated, and grappable – ungrappable. Both adjectives were present in the SD questionnaire.

In both experiments, five pictures of drinking glasses were used as stimuli. The glasses were ‘Essence Plus’, ‘Essence’, ‘Tapio’, ‘Ultima Thule’, and ‘Senta’ (displayed below in the results section). Drinking glasses were chosen, because they are familiar, everyday products. To influence processing times associated with similarity (H3), some of the glasses were similar to each other, and some very different from each other, for instance regarding the shape of the product. The pictures were scaled so that on the screen, their size was close to their real-life size. The number of all possible pairwise comparisons of the five glasses was ten.

Data analysis

The primed product comparison data of one participant consisted of 100 preferential matches: ten stimulus pairs multiplied by ten priming adjectives. To prepare the data for analysis, comparative preference percentages for each glass were calculated. Ranging from zero to one, the preference percentage compares the stimulus against any other stimulus on a given prime. For example, in the first experiment, 15 participants out of 18 preferred ‘Essence’ to ‘Essence Plus’ on timeless. Hence, the preference percentage of ‘Essence’ to ‘Essence Plus’ on timeless was \( \frac{15}{18} \approx 0.83 \) (83% preferred ‘Essence’), and the preference percentage of ‘Essence Plus’ to ‘Essence’ on timeless was \( \frac{3}{18} = 1 - \frac{15}{18} \approx 0.17 \) (17% preferred ‘Essence Plus’).

The preference percentage provides a standardised quantitative way for describing the overall comparative stimulus preferences, and its scale (0–1) does not depend on the number of participants. By averaging all preference percentages of a stimulus, a preference score (PS) can be calculated. For ‘Essence Plus’ the PS of timeless in the first experiment was 0.61, meaning that, on average, ‘Essence Plus’ was preferred about 61% of time when compared to other glasses on timeless. The closer the PS is to 1.00, the more there was preference for the glass on a given adjective. Conversely, a PS of 0.00 would equate to the glass being never chosen on a given adjective. Table 1 presents example PSs.

To test H1 (a strong positive correlation in the tool and questionnaire responses), the PS and the mean SD questionnaire score were correlated for each glass on all adjectives. First, Pearson correlation coefficients were calculated to indicate the amount of agreement between the product evaluations made using the tool and using the questionnaire. Next, the stimuli were ranked within the primes on their PS and mean SD values, and these ranks
correlated. Coefficients close to 1.00 indicate strong agreement, and close to 0.00 little agreement between the tool and the SD questionnaire. Coefficients over .50 were expected for the correlation to be strong [7]. Further, test-retest reliability of the method was assessed by correlating the PSs between the experiments. As in testing H1, both Pearson correlation and correlation of ranks within a prime were calculated. High correlations (over .50) indicate that the stimuli were evaluated similarly in the two experiments, which means that the tool provides repeatable results.

In the second experiment, in addition to preference data, RTs from the stimulus onset to the preference were recorded. In order to test H2 (different mean RTs between the primes), H3 (different mean RTs between the pairs), and H4 (different mean RTs for a given prime between the stimulus pairings), a multilevel model predicting RT was constructed [12]. Main effects in the model were the prime (H1) and the stimulus pair (H2), and a two-way interaction effect between these two was added as a third term (H4). The procedure corresponds to testing the main effects of the prime and the stimulus pair using one-way analyses of variance (ANOVA), and testing the interaction with a two-way ANOVA. However, a multilevel model was used instead of repeated measures ANOVA, because it better handles violations of sphericity [12], often associated with RT data, and works better with testing interaction effects within nested data. In all tests reported here, the level of statistical significance was \( \alpha = .05 \) (Sidak adjusted in multiple comparisons). All tests were two-tailed.

While the focus of the study is in the affective process of product experience, the method of primed product comparisons provides also means for product evaluation. In order to demonstrate this, qualitative conclusions about the stimuli are presented at the end of the results. High (> 0.7) and low (< 0.3) PSs (averaged between the two experiments) were highlighted to characterise the stimuli. These product descriptions are not associated with the hypotheses of the study, but they demonstrate how the method can be used in product experience evaluation.

# RESULTS

The participants were able to conduct the tasks without problems. On average, completing all 100 tasks in the first experiment (excluding the rest periods) took 13.3 minutes. The participants of the second experiment completed the tasks, on average, in 11.3 minutes (excluding the rest periods). The second experiment was faster on average, because the participants were asked to indicate their preference as quickly as possible, while the participants in the first task waited for three seconds before this.

Example PSs and rankings of the five drinking glasses on timeless and durable are presented in Table 1. Space limitation permits printing all PSs, but these values were only of instrumental interest, not the end result of the study. ‘Senta’ is clearly a timeless glass, while ‘Ultima Thule’ was considered not timeless. The PS correlation between the two experiments was \( r = .92 \) (\( p < .001 \), \( df = 48 \)), and the correlation of PS ranks between the two experiments was \( \rho = .83 \) (\( p < .001 \), \( df = 48 \)). The results show high agreement between the participants in the two experiments, indicating test-retest reliability for the method of primed product comparisons.

Concerning H1, the Pearson correlation between the PSs and mean SD responses was, for the first experiment, \( r = .82 \) (\( p < .001 \), \( df = 48 \)), and for the second experiment, \( r = .80 \) (\( p < .001 \), \( df = 48 \)). A scatter-plot, illustrating the strong correlation, is shown in Figure 1. The correlations of the ranks were \( \rho = .90 \) (\( p < .001 \), \( df = 48 \)) and \( \rho = .74 \) (\( p < .001 \), \( df = 48 \)). All correlations were high, indicating that the participants made coherent evaluations between the tool and the SD questionnaire, supporting H1.

Regarding H2, the RT differences between the primes are illustrated in Figure 2 (left), which shows the mean RT

|                | Timeless | Durable |
|----------------|----------|---------|
|                | Exp1     | Exp2    | Exp1    | Exp2    |
| Essence Plus   | .61 (2)  | .61 (2) | .10 (5) | .14 (5) |
| Essence        | .40 (3)  | .58 (3) | .44 (3) | .43 (3) |
| Senta          | .82 (1)  | .72 (1) | .28 (4) | .32 (4) |
| Tapio          | .40 (3)  | .39 (4) | .83 (2) | .86 (1) |
| Ultima Thule   | .26 (5)  | .19 (5) | .85 (1) | .75 (2) |

Table 1. Example PSs of the drinking glasses on timeless and durable. The glass rank is in parenthesis.
deviation from the grand mean RT between the adjectives of the study. The grand RT mean of the study was 1.76 seconds. The difference in RTs between the adjective primes was statistically significant, $F(9, 1700) = 5.0$, $p < .001$. Preferences on timeless, for example, took longer on average than with other adjectives, and judgments on durable and light were the fastest to make. The same effect was observed for the stimulus pairs (H3; Figure 2, right): certain pairs took, on average, longer than others to compare, $F(9, 1700) = 8.5$, $p < .001$. Whereas comparisons of ‘Essence P’ and ‘Senta’ took the longest, comparisons between ‘Essence’ and ‘Thule’ were fastest.

The interaction effect between the prime and the stimulus pair (H4) was also statistically significant, $F(81, 1700) = 1.8$, $p < .001$, indicating difference in mean RTs of the same prime between stimulus pairs. For example, judging durability between ‘Essence’ and ‘Essence Plus’ took longer than between ‘Essence’ and ‘Thule’, but judgments on timeless took as long with both pairs. The overall results of the multilevel model support all three hypotheses H2–H4, and provide evidence that there are differences in the appraisal process times depending on the nature of the resulting affective mental content.

Finally, the PSs were used to highlight the qualitative characteristics of the stimuli. These highlights are presented in Figure 3, in which each glass is given a qualitative description based on high and low PSs (mean of the two experiments). Adjectives with PSs less than .30 are presented as their opposite adjective from the SD questionnaire, and adjectives with PSs more than .70 as the adjective itself. The results show, for example, that from the five glasses used as stimuli, ‘Senta’ is considered a timeless, traditional type of glass, while ‘Tapio’ is a practical product for everyday purposes.

**DISCUSSION**

The large correlations between the preference scores (PSs) and the semantic differential (SD) questionnaire scores support H1. The correlations were high in both experiments ($r_1 = .82$, $r_2 = .80$). Although the participants of the first experiment had a mandatory three second contemplation period before preferential match, the participants of the second experiment, making the preference as quickly as possible (1.76 s., on average), were able to validly use the tool for evaluating the stimuli. Further, high correlations between the PSs of the two experiments ($r = .92$) give test-retest reliability for the method, and indicate that with the primes and the stimuli used, 18 participants was enough to attain consistency and generality in the results.

The observed convergence between the two sources of data (the tool and the questionnaire) indicates that the method of primed product comparisons was able to elicit similar affective mental contents as the traditional pen and paper SD questionnaire. This is expectable, as the appraisal mechanism for the both types of product evaluations are similar. The participants were asked to consider given stimuli with given criteria. As the results of both sources of data were similar, it seems reasonable to assume that the affective contents of the participants’ mental representations during the tasks were coherent, and that the participants were able to report them reliably.

One of the main benefits of primed product comparisons compared to SD questionnaire is in the control that the method gives to the experimenter, and the subsequent possibilities at data collection. The differences in RTs between the primes (H2) and between the stimulus pairs (H3) support this conclusion. Certain primes allowed for faster judgments, indicating smaller information processing...
time and therefore different appraisal sources. The same is true for the stimulus pairs: pairs similar to each other took longer, on average, to compare, than dissimilar pairs.

The interaction effect between prime and stimulus pair (H4) further supports the analysis of product experience as an appraisal process. Some stimulus pairs were easier to compare on a given adjective, which means that the time it takes to process the stimuli and arrive at a preferential match, while being primed with certain adjective, is dependent on the interplay of the stimuli and the evaluation criteria. The important notion here is that it is not possible to explain these results without reference to both the information processing requirements and the contents associated with the primed product comparisons. The appraisal process is mostly non-conscious, but the subjective meaning of the prime and the stimulus influence this non-conscious process [33, 39].

While studies on quick exposure times suggest that very short (50 ms) exposures to stimuli are enough to arrive at reliable judgments [26, 45], the model of human visual system [43] and the notion of the three sources of appraisal with different computational demands [41] encourage longer exposure times at least with more detailed or complex evaluations. The observed differences in RTs in our experiment suggest that very short exposure times do not sufficiently take into account the different processing times associated with different affective contents [cf. 26, 45]. Taking into account the confidence intervals of the results, it seems that primed product comparisons require approximately from 1.3 to 2.2 seconds. Very short exposures are perhaps enough for processing low spatial frequencies, in which the overall outline of the stimulus is encoded [43]. This low-spatial frequency information is then used in guiding the analysis of the stimulus in detail in high spatial frequencies. High spatial frequencies and complex affective associations require more processing time, which translates into more detailed evaluation.

The largest RT difference within the adjective primes was between timeless (largest average RT) and light (smallest average RT). This is in line with the discussion, which lead to H2–H4. Different processing times between the primes can be considered both from the perspective of three sources of appraisal and the Norman’s visceral-behavioural-reflective framework [28, 41]. Judgments concerning the basic physical characteristics of an artefact (e.g., light or durable) require less reasoning and associative processing than the more linguistically and culturally complex judgments (e.g., timeless or traditional) [41]. Further studies are therefore encouraged to individuate the processing of primes according to their source or level, and to connect this to what the primes actually mean and how they make sense with the given stimuli.

Further, both the notion of appraisal as subjective, relational evaluation [23], and Norman’s [28] suggestion that subjective interpretations have computational impact on the reflective, but not on the visceral level, encourage research of individual differences in product experience using the method of primed product comparisons. Assuming this hypothesis, past experiences with the products, for example, should have an impact on the results of primed product comparisons (preferences and RTs), but only with primes, which require information at the associative and reasoning levels. Brand experience, for example, has been shown to influence customer satisfaction and loyalty [6], and therefore individual differences in RTs and preferences, based on previous brand experience, should be observable.

The method of primed product comparisons rests on a large number of repeated measures of short exposures. Studies...
utilising repeated exposures of the same stimuli have been criticised for not taking into account the mere exposure effect, which refers to the increased likelihood of positive evaluation as the same stimulus is presented repeatedly [5, 26, 49]. However, in the experiments reported here, the mere exposure effect was countered by having the participants always make their preference based on a pair of two stimuli. Therefore, although the participants were increasingly familiar with the stimuli as the experiment proceeded, no single stimulus was given special treatment and the effects of familiarity countered each other. In addition, no systematic decrease – or increase – in the RTs was observed during the experiment.

The stimuli for the experiments reported here consisted of drinking glasses, and the visual variable in the stimulus material was the shape of the products. The role of shape in appraisal and subsequent conscious affective mental contents was visible in the qualitative results given in Figure 3. These results, while not related to the hypotheses of the study, give support to the method of primed product comparisons in product evaluation context, as product shapes and forms have been considered essential in determining product success [e.g., 4, 35].

Drinking glasses are not ordinary stimuli in HCI studies, but there is no reason not to generalise the method of primed product comparisons for other product domains, such as computer interfaces, web sites, and mobile phones, all of which can be represented as pictures and hence as stimuli. This will of course require choosing appropriate primes, and the method can also be used to analyse how different primes work with different product domains. As suggested in the introduction, current interest in experience in HCI necessitates methodologically solid instruments for studying the affective process of people, who interact with technology.

Information on affective mental contents of products provide valuable support for design [33], and it is suggested here that due to its ease and speed of use, the method of primed product comparisons will serve as an important design tool especially in the early prototyping and iterative evaluation in HCI. Compared to traditional pen and paper methods, the tool allows the analysis of the thinking process that occurs when these products are evaluated, in addition to the evaluation of the products themselves.

Of course, affective and pleasurable product design also includes other product properties than shape. Therefore, future steps in developing the tool presented here should include more visual variables, such as colour. In addition, the stability and the content of the affective responses after short exposures also depend heavily on the context [37]. Experimenting with the nature of primes, such as using concept pictures or textual narratives, instead of adjectives, can be used to explore how different affective mental contents are elicited. This would serve purposes of both basic research and experience design. Further experimenting with subliminal primes and masking is also warranted, especially for the purposes of studying the cognitive process of appraisal behind experience in HCI.

As compared with other software used for user and product experience evaluation [e.g., 1, 13, 30], the tool for primed product comparisons presented here provides more versatile uses for basic and applied contexts. The combination of priming and stimulus pairing offers ways to investigate the experience process in detail, and the tool provides a relatively quick method for analysing products. In product evaluation, both early and final stage evaluations are possible, conducted according to desired evaluation criteria. Further, the theoretical background of appraisal process and mental contents provides coherence and validity for the purposes of evaluation.

Among the popular and scientifically grounded tools of user and product experience evaluation, many are based on problematic methodological assumptions. The methodological background for AttrakDiff [1], for example, is in Osgood’s psycholinguistics, which uses factor analysis to reveal dimensions of affective experience [10]. However, this methodology is not supported by modern psychological interpretation of affective processes [34]. On the other hand, such tools as Lemtool [13] and Premo [14] utilise the methodology from the theory basic emotions, which assumes that there is a core set of universal emotions with corresponding physiological patterns. However, this methodological assumption is also questionable, as shown in meta-analyses of studies of basic emotions [3]. Of course, basic emotion words are still useful in research of emotional experience in HCI [34].

Although appraisal theory is likewise contested in current psychological discourses [3, cf. 39], it currently provides the most theoretically coherent account of emotion as a cognitive process, while being detailed at the same time. The ability of the theory to predict RT differences in primed product comparisons gives support to its usefulness in user and product experience research. Hence, it is maintained that the study of cognitive processes and affective mental contents with methods such as primed product comparisons is critical for scientific analysis of experience in HCI [35]. Although the experience evaluation tools, which are based on questionable methodologies, can produce relevant information for product evaluation, the explication of the connection between psychological theories and tools for experience evaluation should be more enforced in the user and product experience community.

CONCLUSION
This study presents a method of primed product comparisons, useful for basic and applied human-computer interaction (HCI) research. The method is based on the methodology of studying affective mental contents [34, 35], which are part of the appraisal process [23, 25, 38, 39]. A tool implementing the method of primed product
comparisons was constructed. Product evaluations made with the tool agreed with the results of traditional pen and paper evaluation, providing validity to the method and the tool. However, compared to traditional questionnaires, the method provides additional information relating to the mental processing associated with making affective judgments. Understanding the process of product evaluation as cognitive computations and affective contents brings us closer to a psychologically explicated theory of experience [36].

In addition to the basic research implications of the method of primed product comparisons, it has possible applied uses. User-centred design processes lean on understanding the user and the assessment of design solutions, especially in the early product design processes [32]. The proposed method and its implementation as a computer software can be utilised to meet the challenges in early design phases, as well as in different stages of iterative design processes. The priming aspect of the tool offers targeted evaluation of, for example, the experience goals of the design process. Comparing the product being designed with earlier prototypes, or with different design solutions of the same phase prototype are options, which provide information to support subsequent iterations. In addition, future versions of the tool should include tablet computer solution for agile and adaptable field evaluation. Further, online analyses of the collected data during or straight after the experiment are easy to provide in computerised environment.

ACKNOWLEDGMENTS
The research has been supported by the Finnish Funding Agency for Technology and Innovation (TEKES) and the Finnish Metals and Engineering Competence Cluster (FIMECC) programme UXUS (User experience and usability in complex systems), and Doctoral Program in Computing and Mathematical Sciences (COMAS). We thank research assistant Merja Lehtiharju.

REFERENCES
1. AttrakDiff. http://attrakdiff.de/index-en.html.
2. Bargas-Avila, J.A., and Hornbæk, K. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In Proc. CHI 2011, ACM Press (2011), 2689-2698.
3. Barrett, L.F. Are emotions natural kinds? Perspectives on Psychological Science 1, 1 (2006), 28-58.
4. Bloch, P.H. Seeking the ideal form: Product design and consumer response. Journal of Marketing 59, 3 (1995), 16-29.
5. Bornstein, R.F. Exposure and affect: Overview and meta-analysis of research, 1968–1987. Psychological Bulletin 106, 2 (1989), 265-289.
6. Brakus, J.J., Schmitt, B.H., and Zarantonello, L. Brand experience: What is it? How is it measured? Does it affect loyalty? Journal of Marketing 73, 3 (2009), 52-68.
7. Cohen, J. A power primer. Psychological Bulletin 112, 1 (1992), 155-159.
8. Hartmann, J., De Angeli, A., and Sutcliffe, A. Framing the user experience: Information biases on website quality judgement. In Proc. CHI 2008, ACM Press (2008), 855-864.
9. Hassenzahl, M. The interplay of beauty, goodness, and usability in interactive products, Human-Computer Interaction 19, 4 (2004), 319-349.
10. Hassenzahl, M., Burmester, M., and Koller, F. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In J. Ziegler and G. Szwilus (Eds.), Mensch & Computer, B. G. Teubner, Stuttgart, 2003, 187-196.
11. Hassenzahl, M., and Tractinsky, N. User experience – a research agenda. Behaviour & Information Technology 25, 2 (2006), 91-97.
12. Hox, J.J. Multilevel analysis. 2nd ed. Routledge, Hove, 2010.
13. Huism, G., van Hout, M., van Dijk, E., van der Geest, T., and Heylen, D. LEMtool: measuring emotions in visual interfaces. In Proc. CHI 2013, ACM Press (2013), 351-360.
14. Desmet, P. Measuring emotion: Development and application of an instrument to measure emotional responses to products. In M. Blythe et al. (Eds.), Funology: From usability to enjoyment, Academic Publishers, Dordrecht, 2003, 111-123.
15. Desmet, P. Faces of product pleasure: 25 positive emotions in human-product interactions. Int. Journal of Design, 6, 2 (2012), 2-29.
16. Desmet, P. and Hekkert, P. Framework of product experience. Int. Journal of Design, 1, 1 (2007), 57-66.
17. Diefenbach, S., Lenz, E., and Hassenzahl, M. An interaction vocabulary. Describing the how of interaction. In CHI 2013 Extended Abstracts. ACM Press (2013).
18. Eimer, M., and Schlaghecken, F. Links between conscious awareness and response inhibition: Evidence from masked priming. Psychonomic Bulletin & Review 9, 3 (2002), 514-520.
19. Ekman, P., Basic emotions. In T. Dalglish and M. Power (Eds.), Handbook of cognition and emotion, John Wiley & Sons, New York, 1999, 45–60.
20. Ericsson, K.A., and Simon, H.A. Protocol analysis: Verbal reports as data. MIT Press, Cambridge, MA, 1984.
21. Five Second Test, Click Test and Nav Flow – UsabilityHub. https://usabilityhub.com/
22. Fodor, J. Fodor’s guide to mental representation. *Mind*, 94 (1985), 76-100.
23. Frijda, N. The laws of emotion. *American Psychologist* 43, 5 (1988), 349-358.
24. Google Analytics Official Website – Web Analytics & Reporting. www.google.com/analytics/
25. Lazarus, R.S. *Emotion and adaptation*. Oxford University Press, New York, 1991.
26. Lindgaard, G., Fernandes, G., Dudek, C. and Brown, J. Attention web designers: You have 50 milliseconds to make a good first impression! *Behavior & Information Technology* 25, 2 (2006), 115-126.
27. Marcel, A.J. Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive psychology* 15, 2 (1983), 197-237.
28. Norman, D.A. *Emotional design*. Basic Books, New York, 2004.
29. Papachristos, E. and Avouris, N. Are First Impressions Only Related to Visual Appeal? In *Proc INTERACT 2011, Part I, LNCS*, Springer (2011), 489-496.
30. PrEmo. http://www.premotool.com/
31. Reber, R., Schwarz, N., and Winkielman, P. Processing fluency and aesthetic pleasure: Is beauty in the perceiver’s processing experience? *Personality and Social Psychology Review* 8, 4 (2004), 364-382.
32. Roto, V., Rantavuo, H., and Väänänen-Vainio-Mattila, K. Evaluating user experience of early product concepts. In *Proc DPP109*, Compiègne University of Technology (2009), 1-10.
33. Saariluoma P. Apperception, content based psychology and design. In U. Lindemann (Ed.), *Human behaviour in design*, Springer, Berlin, 2003, 72-78.
34. Saariluoma, P., and Jokinen, J.P.P. Emotional dimensions of user experience – a user psychological analysis. *Int. Journal of Human-Computer Interaction* 30, 4 (2014), 303-320.
35. Saariluoma, P., Jokinen, J., Kuva, S., and Leikas, J. User experience as mental contents. In *Proc 10th European Academy of Design Conference*, Chalmers University of Technology (2013).
36. Saariluoma, P., and Oulasvirta, A. User psychology: Re-assessing the boundaries of a discipline. *Psychology* 1, 5 (2010), 317-328.
37. van Schaik, P. and Ling, J. The role of context in perceptions of the aesthetics of web pages over time. *Int. Journal of Human-Computer Studies* 67, 1 (2009), 79-89.
38. Scherer, K.R. What are emotions? And how can they be measured? *Social Science Information* 44, 4 (2005), 695-729.
39. Scherer, K.R. The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion* 23, 7 (2009), 1307-1351.
40. Smith, C. A. and Ellsworth, P.C. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48, 4 (1985), 813-838.
41. Smith, C.A. and Kirby, L.D. Toward delivering on the promise of appraisal theory. In K. A. Scherer, A. Schorr, and T. Johnstone, (Eds.), *Appraisal Process in Emotion*, Oxford University Press, New York, 2001, 121-138.
42. Optimal Workshop. http://www.optimalworkshop.com/optimalsort.htm
43. Thielsch, M.T., and Hirschfeld, G. Spatial frequencies in aesthetic website evaluations – explaining how ultra-rapid evaluations are formed. *Ergonomics* 55, 7 (2012), 731-742.
44. Thüring, M., and Mahlke, U. Usability, aesthetics and emotions in human–technology interaction. *Int. Journal of Psychology* 42, 4 (2007), 253-264.
45. Tractinsky, N., Cockhavi, A., Kirschenbaum, M., and Sharfi, T. Evaluating the consistency of immediate aesthetic perceptions of web pages. *Int. Journal of Human-Computer Studies* 64, 11 (2006), 1071-1083.
46. UEQ-Online. http://www.ueq-online.org/.
47. UserTesting. http://www.usertesting.com/
48. Vermeeren, A.P., Law, E.L.C., Roto, V., Obrist, M., Hoonhout, J., and Väänänen-Vainio-Mattila, K. User experience evaluation methods: Current state and development needs. In *Proc. 6th NordiCHI*, ACM Press (2010), 521-530.
49. Zajonc, R.B. Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology* 9, 2 (1968), 1-27.
50. Zhou, H., and Fu, X. Understanding, measuring, and designing user experience: The causal relationship between the aesthetic quality of products and user affect. In J. Jacko (Ed.), *Human-Computer Interaction. Interaction, Part I*, Springer, Berlin, 2007, 340-349.