Interacting with Computers 23 (2011) 363–371

Contents lists available at ScienceDirect

Interacting with Computers

journal homepage: www.elsevier.com/locate/intcom

Too good to be bad: Favorable product expectations boost subjective usability ratings

Eeva Raita, Antti Oulasvirta *

Helsinki Institute for Information Technology HIIT, Aalto University and University of Helsinki, PO Box 19800, 00076 Aalto, Finland

ARTICLE INFO

Article history:
Received 22 March 2010
Received in revised form 18 April 2011
Accepted 20 April 2011
Available online 27 April 2011

Keywords:
Usability testing
Product expectations
Subjective usability ratings
Usability evaluation

ABSTRACT

In an experiment conducted to study the effects of product expectations on subjective usability ratings, participants ($N = 36$) read a positive or a negative product review for a novel mobile device before a usability test, while the control group read nothing. In the test, half of the users performed easy tasks, and the other half hard ones, with the device. A standard usability test procedure was utilized in which objective task performance measurements as well as subjective post-task and post-experiment usability questionnaires were deployed. The study revealed a surprisingly strong effect of positive expectations on subjective post-experiment ratings: the participants who had read the positive review gave the device significantly better post-experiment ratings than did the negative-prime and no-prime groups. This boosting effect of the positive prime held even in the hard task condition where the users failed in most of the tasks. This finding highlights the importance of understanding: (1) what kinds of product expectations participants bring with them to the test, (2) how well these expectations represent those of the intended user population, and (3) how the test situation itself influences and may bias these expectations.

1. Introduction

Integral to user-centered design is the measurement of *usability*, defined by a commonly used standard (ISO 9241-11, p. 2) as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." One of the most common methods for evaluating usability is to deploy a laboratory-based usability test either during the development of a prototype (*formative usability test*), or after it (*summative usability test*). While this study concentrates on the latter, all usability tests involve a study wherein users perform tasks in a controlled setting and the processes as well as the outcomes are evaluated. Measurements in summative testing can be divided into at least two categories: (a) objective measures that indicate the effectiveness and efficiency of use and are derived from performance and overt behavior and (b) subjective measures that reflect opinions and experiences and are measured via verbal accounts and ratings. Both are used by usability practitioners (Bark et al., 2005; Gulliksen et al., 2004; Mao et al., 2005).

To better understand factors contributing to usability, studies have explored correlations between different types of measures. These studies have found mixed evidence for the relatedness of objectively and subjectively measured usability, but they have also been criticized for methodological weaknesses (Hornbæk and Lai-Chong Law, 2007). Addressing many of these weaknesses, Hornbæk and Lai-Chong Law (2007) recently analyzed the raw data from 73 usability studies and found that subjective usability measurements rarely correlate with objective ones. This evidence challenges the view that users' subjective assessment reflects objectively measured usability (Nielsen and Levy, 1994), as well as the attempts to reduce usability to one simple usability score formed by summarizing subjective and objective measures (Kindlund and Sauro, 2005). Moreover, the finding means that there really are (at least) two "usabilities," and the challenge is to understand the underlying factors of each (see Hertzum, (2010), for a recent discussion).

One explanation for the difference between subjective and objective usability comes from studies of aesthetics and usability. In a well-cited paper "What Is Beautiful Is Usable" Tranctinsky et al. (2000) show that visual appeal of an interface can be dominant over objectively measured usability in post-experiment perceptions of usability. While visual appeal might partly explain why subjective and objective usability do not always go hand-in-hand, this paper studies the effects of *product expectations*, by which we refer to beliefs and/or emotions related to a product that are formed before its actual use. The importance of product expectations is related to the fact that even when a technology is novel to its users, certain predispositions still shape their actions and experiences. While aesthetic appeal might be one indicator of the quality of a product (Tractinsky, 2004), broader product expectations are influenced by many sources—advertisements,

* Corresponding author. Tel.: +358 50 3841561; fax: +358 9 694 9768.
E-mail address: antti.oulasvirta@hiit.fi (A. Oulasvirta).

0953-5438/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.intcom.2011.04.002

brands, word of mouth, product reviews, discussion forums, and exposure to related products, for example. Previous work in HCI has shown the effect of *framing* product presentation. For example, telling users that 90% of others like product is good has a different effect than telling that 10% dislike the product (Hartmann et al., 2008). Contrary to the framing effect, which is based on keeping information the same but manipulating the way it is presented, we are interested in what happens when the product description is changed from positive to negative. Another difference is that the framing effect was exposed in a study where users did not have an opportunity to interact with the product, meaning they had very little reason to deviate from the prior information they were given (Hartmann et al., 2008).

We are not the first to acknowledge the issue of expectations: the effect of expectations on perceptions has been studied for years in various mother disciplines of HCI. We discuss these studies in greater depth in the following section (“Previous Studies of Expectations”), but, simply put, theories make two kinds of predictions concerning the relationship between expectations and perceptions: perceptions become oriented either in line with the valence of expectations or, on the contrary, against them. The notion that perceptions align with expectations stems from multiple sociological and psychological theories, such as the theory of *self-fulfilling prophecy* (Merton, 1968), which refers to a prediction (e.g., the user expecting a product to be usable) that causes itself to become true; expectations make one behave in such a way that the prediction is verified. On the other hand, in consumer psychology, post-purchase satisfaction is often seen to result from a *comparison* between expectations and actual performance. According to the *expectation–confirmation theory* (Bhattacherjee, 2001; Thong et al., 2006), expectations exist as a norm against which “actual experience” is compared. High expectations in combination with poor performance should lead to a very negative evaluation.

While expectations are often cited as an integral part of human conduct and the formation of experiences, to our knowledge, there is only one prior empirical study looking at the effects of expectations on subjective usability ratings. In a pioneering study that has gone unnoticed, Bentley (2000) let half of the subjects know that a to-be-used web site was previously tested for usability while the other half was told that no prior testing was conducted. Telling about previous testing slightly increased usability ratings (SUMI), particularly for perceived helpfulness and controllability, but did not affect users’ performance in the tasks. The results suggest that there may be a difference in the antecedents of objective and subjective usability, which would call into question attempts to derive a single summative usability score for a product as well as the use of objective usability measures alone as indicators of overall usability. Therefore, it is necessary to first replicate Bentley’s (2000) finding in other contexts. Extensive effort has been invested into developing valid usability questionnaires (for a recent study, see Brinkman et al., 2009), but there is far little understanding about potential biases that take place before and during the usability tests. Studies of expectations should eventually help practitioners to eliminate potential sources of biases by means of experimental design. Furthermore, practitioners are interested in a variety of interaction effects, for example if expectations influence usability perceptions in interaction with task difficulty. For instance, do *positive* expectations boost usability perceptions only when tasks are easy and performed well or also in a situation where tasks are difficult and performed poorly?

2. Previous work on expectations and perceptions

Work on expectations has been carried out in various fields of study. Here we concentrate on those studies that make clear predictions about the relationships among expectations, performance, and perceptions. In short, there are two contrasting predictions: perceptions may be oriented in line with expectations or against them. We discuss these theories in relation to the predictions made, starting with theories that predict perceptions’ alignment with expectations.

In social psychology, it has been acknowledged that prior beliefs influence social interaction (Snyder and Stukas, 1999). The phenomenon called the *self-fulfilling prophecy* (Merton, 1968) refers to a prediction that causes itself to become true; a certain definition of a situation evokes behaviors that make this conception come ‘true’. In a usability study, a user hearing that “the device is very usable” could affect the way in which the device is interacted with—called behavioral confirmation in the social behavior context—or the way in which events are perceived, called perceptual confirmation (Snyder and Stukas, 1999). Users with positive expectations should therefore evaluate an interface more favorably than negatively manipulated or non-manipulated users, but if behavioral confirmation takes place, it could also affect the way the interface is interacted with. The same outcome is predicted by theories of *conformity and obedience* (Cialdini and Goldstein, 2004). Reading a positive review could, instead of inducing a prediction, make the user want to conform to the opinions of the trusted author who wrote the review. Compliance can result from two processes: information effect (stemming from belief in the accuracy of the information) and comformative effect (response to a felt need to conform) (Cialdini and Goldstein, 2004). These theories predict that *the valence of the expectation* (positive or negative) affects interaction, perceptions of interaction and the device, or both. Therefore, we should observe differences between the positive and the negative prime in performance and/or ratings. The framing effect (Hartmann et al., 2008) makes the specific prediction that positive framing of product information would increase post-experiment evaluations, but it does not state what happens if users actually fail to accomplish tasks with the interface.

Expectations have been studied in consumer psychology, with an emphasis on customer satisfaction and product judgment. According to the expectation–confirmation theory (ECT), post-purchase satisfaction is a result of *comparison* between expectations and actual performance. In other words, expectations exist as a norm against which “actual experience” is compared. This theory makes a prediction differing from those of the previously mentioned theories specifically in the case where actual performance differs from the norm: high expectations in combination with poor performance should lead to a (very) negative evaluation. By the same token, if users have low expectations of a prototype that performs well, they should be *more* satisfied than those who have had high or moderate expectations (Bhattacherjee, 2001; Thong et al., 2006). This “pendulum” idea can be tested by including in the experiment a comparison between easy and hard tasks.

Experience of use can be affected also by *a transient psychological state induced by previous experience*. For example, reading a positive review might make one happy and therefore induce an individual to provide more favorable ratings (Carver and Scheier, 1998), on the assumption that the state persists to the moment when ratings are given. Reading a positive review might also, without conscious awareness, activate a self-regulatory mechanism. Bargh (1994) has argued that environmental cues in this manner guide individuals to contextually appropriate behaviors. In a typical experiment, an individual reads text containing adjectives denoting positive elements (“good,” “fast,” etc.) or negative ones, after which behavior is measured. This might be the speed of walking or judgments of other people. A product review, similarly, could induce either an emotional state or an unconscious regulatory mechanism, with effects on actual use, judgments, or both.

3. Approach and research questions

The design of our experiment is analogous to that of a typical laboratory-based usability test. Forming the core of the study is the manipulation of two factors: (1) product expectations (positive, negative, and control) and (2) task difficulty (easy vs. hard tasks). To minimize effects from aesthetics, brand, etc. the same system, the HTC Touch Diamond (see Fig. 1), was deployed throughout the tests. Expectations were manipulated prior to use with either a negative or a positive "product review from the Internet." In the control group, no pre-information was given. In the test, users performed either easy or difficult tasks with the system and their perceptions of usability were measured after each task and at the end of the test.

This setting was utilized to answer the three research questions stated below. First, we manipulated product expectations and measured subjective usability ratings to answer our first research question: do product expectations influence subjective usability ratings? (**RQ1**). Second, we manipulated expectations as well as task difficulty in order to test the differing hypothesis following from the theories introduced. According to the social psychological theories of the self-fulfilling prophecy and compliance, expectations can influence behavior as well as judgments, but both should be aligned with expectations. On the other hand, the expectation–confirmation theory predicts expectations and performance to influence perceptions in an interaction, and, for example, positive expectations in combination with poor performance should lead to negative perceptions.

Our second research question is "Does the effect of product expectations on subjective usability ratings change in relation to task performance (success or failure in a task)?" (**RQ2**). Third, we utilized both post-task and post-experiment measurements, because we wanted to gain a better understanding of the relationship between users' experience of tasks and their overall perceptions of the device assessed at the end of the experiment. Thus, our third research question is "Do product expectations influence both task-specific ratings such as workload measurements and system-specific ratings such as System Usability Scale SUS, or only one of the two?" (**RQ3**).

Fig. 1. HTC Touch Diamond. This picture was used in the product reviews given to users before the trial (primes).

Table 1
Experimental design.

Prime	Task difficulty	
No prime	Hard ($N = 6$)	Easy ($N = 6$)
High expectations	Hard ($N = 6$)	Easy ($N = 6$)
Low expectations	Hard ($N = 6$)	Easy ($N = 6$)

4. Method

4.1. Participants

The participants ($N = 36$) volunteered for a cellular phone study announced in university students' mailing lists. Participants were selected to represent one smartphone user group, young educated adults. There were 21 females and 15 males, aged 20–30 years, with a mean age of 25 years. All participants had an upper-secondary-school education, and 16 had a bachelor's degree as well. Thirty-one participants were studying full-time toward an academic degree (a bachelor's or master's degree), one was taking a 1-year break from his university studies to complete his civil service obligation, and four listed working as their main duty, since they were studying only part-time.

4.2. Experimental design

The study was a 3×2 between-subjects experiment with *expectations* (no prime, high vs. low expectations) and *task difficulty* (easy vs. hard tasks). Each cell of the design has six participants (see Table 1).

4.3. Tasks and materials

The tasks were performed with the HTC Touch Diamond phone, a Windows-Mobile-6.1-based Pocket PC with a TouchFLO 3D interface (for specifications, see http://www.htc.com/www/product/touchdiamond/overview.html). HTC Touch Diamond was chosen because the device and its producer were relatively unknown in Finland at the time of the study. This was necessary, as we wanted to ensure that we could influence participants' product expectations with priming reliable. If a more familiar device would have been chosen, it was more likely that the participants would already possess strong product expectations and also differ in these expectations. Moreover, participants' familiarity with the device could not be used as a criterion in the selection of participants, because this would have exposed them to the device and might have led some of them to search for more information about it before the test. Instead another approach was used: participants' familiarity with the device was checked in the post-experiment interview, where it was found that, as expected, the device was unknown to the participants—only a few of them had even heard of it, and only one[1] suspected that he might have tried it, once quickly in a store.

The product review we used for priming was designed to resemble an authentic product review from the Internet. What

[1] We ran a statistical test to ensure that this participant did not differ from other users. We removed the participant from the data and no difference was found in results.

Table 2
Contents of the "product reviews from the Internet".

Positive prime	Negative prime
Easy-to-use touchscreen	Hard-to-use touchscreen
Stylish, top-notch design	A magnet for fingerprints
Useful basic buttons	Quite useless basic buttons
A technically well-equipped unit	Technical problems with 3G
Good-looking graphics, with PC-like resolution of pictures	Too much brilliance, for which the phone does not have enough power
Light unit that is pleasant to hold	Small battery that must be charged very often
Intuitive user interface	Interface that is slow to use

Fig. 2. The test setup. Tape marks the area within which the users were asked to operate the device. The task is described on the piece of paper. Image from actual data.

was changed between the two conditions was how the features of the phone were described (see Table 2). In the study, users were given task goals on pieces of paper (see Fig. 2). Half of the participants received directions for tasks we knew would be easy and the other for tasks that were hard (see Table 3). Our manipulation of task difficulty was based on "variants" of tasks, each judged easy or hard. These were found through exploration—i.e., testing and use of the device—and are specific to the model used. For example, typing umlauts (e.g., an in ä or ö) is hard with the touchpad-based keyboard, which does not present umlauts in the default interface. Keeping the task *goal* constant (or at least very similar) while manipulating task difficulty was hard to realize in practice but important, because goals are most likely important factors in expectations and experience (Hassenzahl and Ullrich, 2007).

The physical setup of the lab is illustrated in Fig. 2. All tasks were completed in a room that was light-controlled, and the door was kept closed.

Table 3
Directions for easy and hard tasks.

#	Easy version	Difficult version
1	Write a text message that says "hi sister"	Write a text message that says "hi mom" (includes umlaut in Finnish)
2	Watch a video from YouTube	Watch a video from a journal's site
3	Listen to saved music from the phone	Listen to the radio
4	Put the phone in silent mode	Change the ringtone

4.4. Procedure

The study consisted of four phases, through which each participant proceeded separately. All tests were videotaped and afterwards used to analyze and calculate both the completion rates and the times for each task. (1) In the first phase, participants completed a pre-use questionnaire about their background (sex, age, etc.). (2) In the second phase, two thirds of the participants were given a review, which aimed to prime them to have either low or high expectations for the phone tested, and one third were not given any prime, since they served as a control group for the prime. Special care was taken to introduce the prime in the same way every time. Participants were told that they could read the review at their own pace while the researcher made the last technical arrangement for the study. The researcher continued purposefully with the arrangements until the participant indicated that he or she had finished reading. (3) In the third phase, participants performed four easy or hard tasks with the phone. Participants were not told about the tasks or their difficulty beforehand. They were told that they would have seven minutes to perform each task but that they should not worry if unable to complete a task in the maximum time allowed. After every task, the participant filled in a questionnaire composed of the NASA task load index (NASA-TLX) and PANAS questionnaires. (4) In the last phase, participants filled in a closing questionnaire with questions about previous experience with mobile phones and user experience scales.

4.5. Measurements

We utilized two objective usability measures: task success and task completion time. For subjective usability, we deployed a task-specific and a post-experiment questionnaire. The task-specific questionnaire included the (1) NASA-TLX and (2) PANAS items, and the post-experiment questionnaire involved (1) the System Usability Scale (SUS) and (2) AttrakDiff. The questionnaires were translated into Finnish by the first author in several iterations where colleagues at HIIT were presented with the English original and asked to evaluate a translation.

4.5.1. Task performance

Task success and task completion time were measured based on video recordings. The task ended when the user indicated having solved the problem. If the task was not completed within the 7 min, it was marked as a failure. There were no situations in the data where the user would have thought that he/she has completed a task but the indicated answer was incorrect.

4.5.2. Post-task questionnaire

NASA-TLX is a subjective workload assessment tool (Hart and Staveland, 1988). It is a multi-dimensional rating procedure that gives an overall workload score based on a weighted average of six sub-scales. For the past 20 years, NASA-TLX has been widely used in studies of interface design and evaluation (Hart, 2006). The questionnaire consists of six questions that are answered with a rating on a seven-point scale. Alongside NASA-TLX, task-specific emotions were measured with the PANAS questionnaire, developed for the measurement of positive and negative affect (Watson et al., 1988). Consisting of 20 items, of which half concern positive and the other negative emotions, it is one of the most widely used contemporary measurements of emotions, with more than 2000 citations (Schimmak and Crites, 2005). The PANAS questionnaire addresses emotions that range from excited to strong, and participants are asked to rate how much they are feeling them, on a five-point scale ("very much" to "very little").

4.5.3. Post-experiment questionnaire

SUS was developed in 1996 to meet the demands of measurement in industrial contexts (Brooke, 1996). From 1996 to 2006, it was used in at least 206 studies, of Web applications, networking equipment, phones, etc. (Bangor et al., 2008). SUS consists of 10 statements rated on a five-point Likert scale. The overall score is calculated by first summing the score contributions for all of the individual statements, which range from 0 to 4. The sum of the scores is then multiplied by 2.5, and the final SUS score has a range of 0–100. A score under about 60 is claimed to reflect relatively poor usability (Tullis and Albert, 2008).

AttrakDiff (Hassenzahl et al., 2003) is a questionnaire designed to measure pragmatic and hedonic qualities of user experience, as well as the attractiveness of a product. Pragmatic Quality is related to users' ability to achieve goals with the product, while Hedonic Quality is about how motivating, interesting, and identifiable the product is. The questionnaire consists of 28 pairs of words, sets of opposites—for example, "easy–hard"—which users evaluate on a seven-step scale ranging from −3 to +3.

4.6. Prime verification

To ensure that the primes functioned as expected, we carried out a brief independent study. We sent invitations to a student mailing list, asking participants to complete a short survey. Participants ($N = 85$) first answered questions about their background and then were randomly chosen to read either a positive or negative review of the HTC Touch Diamond (the same as the ones used in the final test). After reading the review, participants rated the device on the SUS and AttrakDiff scales. Verifying our primes, there was a statistically significant difference between prime groups: those who read the positive review rated the device higher in their SUS scores than did those who read the negative review, $F(1, 83) = 25.23$, $p < .001$, $\eta^2 = .233$. In addition, the positive-prime group rated the device higher on Pragmatic Quality than negative prime group, $F(1, 83) = 34.24$, $p < .001$, $\eta^2 = .292$, more attractive than the negative-prime group, $F(1, 83) = 11.82$, $p < .001$, $\eta^2 = .125$, and higher on Hedonic Identification than the negative-prime group, $F(1, 83) = 19.97$, $p < .001$, $\eta^2 = .194$. The prime groups did not differ in the mean scores for Hedonic Stimulation, $F(1, 83) = 0.30$. Mean scores for SUS and AttrakDiff scores are displayed in Table 4.

5. Results

For statistical testing, we utilized a 3×2 analysis of variance (ANOVA) with *prime* and *difficulty* as the two main factors. Analysis of variance assumes the homogeneity of variances, which was tested with Levene's test of equality of variances. Levene's test of equality of variances proved significant for three of the analysis: task success, $F(5, 30) = 2.56$, $p < .05$, task completion time $F(5, 30) = 2.85$, $p < .05$, and positive affect, $F(5, 30) = 2.7$, $p < .05$. Therefore, following the practice proposed by Keppel and Wickens (1991), we set a more stringent alpha (.01) for these particular tests. Levene's test of equality of variances was not significant for any other analysis and, thus, we utilized an alpha of 0.05 for all other tests. When comparing groups of more than two, we utilized Tukey's HSD for post hoc test. The F, p, and η^2 values for all analysis are presented, with the means in Table 5. The eta-squared (η^2) statistic is an estimate of effect size that describes the proportion of total variability attributable to a factor.

5.1. Task performance

Task performance comprised of task success and task completion time. As an overview the following significant effects were found: task difficulty had significant effects on both task success and task completion time. In addition, the interaction of prime and task difficulty had a significant effect on task success.

5.1.1. Task success

The effect of prime on task success was not significant. Task difficulty influenced completion rate significantly in such a way that easy tasks were completed more often than hard tasks, $F(1, 30) = 120.10$, $p < .001$, $\eta^2 = .800$. The interaction of prime and task difficulty with respect to task success was significant, $F(2, 30) = 10.98$, $p < .001$, $\eta^2 = .423$.[2] Fig. 3 displays the results.

5.1.2. Task completion time

The prime did not have a significant effect on task completion time. Task difficulty did have a significant effect on average completion time: easy tasks were completed more quickly than hard tasks, $F(1, 30) = 135.44$, $p < .001$, $\eta^2 = .819$. There was no interaction effect of prime and task difficulty on task completion time.

5.2. Post-task ratings

Post-task ratings comprised of two scales: NASA-TLX (task load) and PANAS (negative and positive affect). One significant effect was found: task difficulty had a significant effect on task load.

5.2.1. Task load

The effect of prime on task load was not significant. Task difficulty had a significant effect on task load: users in the hard task condition reported a greater overall task load than those in the easy task condition, did $F(1, 30) = 35.12$, $p < .001$, $\eta^2 = .539$. The interaction between prime and task difficulty was not significant. Fig. 4 displays the results.

5.2.2. Emotions

Priming did not have a significant effect on positive or negative affect. Task difficulty had a borderline-significant effect on positive affect, $F(1, 30) = 3.48$, $p = .072$, $\eta^2 = .104$, but not on negative affect. There were no significant interaction effects of prime and task difficulty on positive or negative affects.

5.3. Post-experiment ratings

Post-experiment ratings included SUS and AttrakDiff scores. Both post-experiment measurements reflected the same pattern: priming and task difficulty had independently significant effects on SUS and AttrakDiff scores (except for Hedonic Stimulation),

Table 4
Mean scores (standard deviation in parenthesis) for SUS and AttrakDiff ratings.

Prime	SUS	Pragmatic quality	Attractiveness	Hedonic identification
Positive	60.59 (12.97)	.82 (1.04)	1.16 (1.06)	0.91 (0.94)
Negative	46.08 (13.10)	−0.49 (0.99)	0.41 (0.93)	−0.01 (0.92)

[2] For a closer look at the interaction, we split the data by task difficulty and ran the statistical tests separately for easy and hard tasks. For the easy tasks, the assumptions underlying analysis of variance were not fulfilled, Levene's test: $F(2, 15) = 15.25$, $p < .001$. The prime did not have a significant effect on task success for the easy tasks. For the hard task condition Levene's test did not prove significant, $F(2, 15) = 0.45$, and the effect of prime on task success was significant. The negative-prime group completed *more* hard tasks ($M = 2.67$, $SD = 0.52$) than did the positive-prime ($M = 1.83$, $SD = 0.75$) and no-prime groups ($M = 1.00$; $SD = 0.63$), $F(2, 15) = 10.14$, $p < .005$. In the post hoc test the negative-prime group differed significantly from the no-prime group ($p < .005$), but not from the positive-prime group ($p > .05$).

Table 5
Means (standard deviation in parenthesis) for task success, task completion time, task load, positive affect and negative affect, SUS and AttrakDiff scores in relation to priming and task difficulty.

	Task success	Task completion time	Task load	Positive affect	Negative affect	SUS	Pragmatic quality	Hedonic identification	Attractiveness	Hedonic stimulation
Positive-prime	2.92 (1.24)	829.2 s (474.9)	11.82 (4.36)	96.33 (9.89)	45.75 (22.00)	55.83 (18.23)	0.37 (0.62)	0.33(0.91)	0.80 (.98)	0.33 (.73)
Negative-prime	3.08 (0.67)	886.3 s (413.8)	12.26 (3.30)	88.25 (18.55)	54.42 (25.71)	33.13 (12.89)	−0.50 (0.64)	−0.42(1.06)	−0.02 (.1.04)	0.13 (1.20)
No-prime	2.42 (1.56)	956.0 s (535.4)	14.60 (5.08)	92.58 (22.13)	61.17 (25.64)	31.04 (17.01)	−0.750 (0.86)	−0.92(1.14)	−0.48 (.78)	0.23 (1.33)
Easy tasks	3.78 (0.43)	477.9 s (220.9)	9.98 (2.91)	97.78 (14.68)	47.67 (18.69)	50.56 (16.88)	0.01 (0.79)	0.17(1.00)	0.46 (1.03)	0.30 (1.0)
Hard tasks	1.83 (0.92)	1297.1 s (211.2)	15.81 (3.58)	87.00 (18.68)	59.89 (28.68)	29.44 (15.99)	−0.60 (0.81)	−0.83(1.07)	−0.26 (.98)	0.17 (1.17)

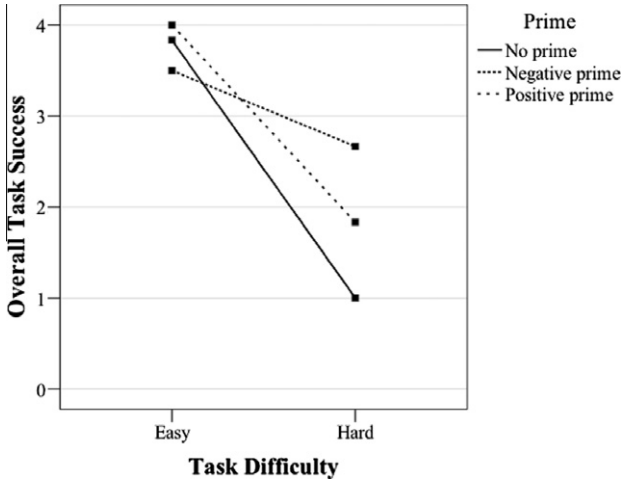

Fig. 3. The effect of prime and task difficulty on task success (min. 0, max. 4).

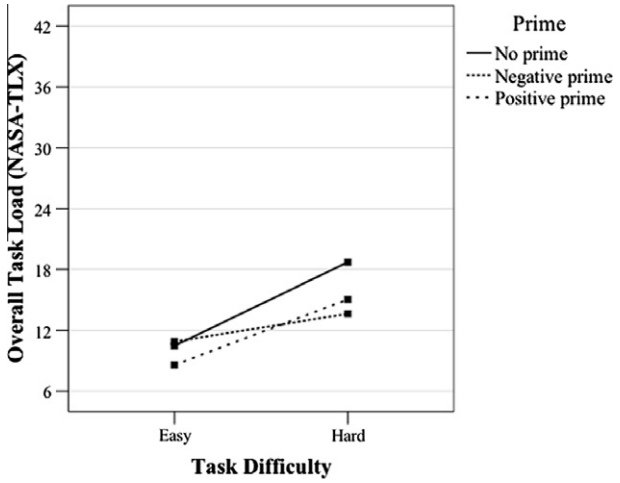

Fig. 4. The effect of prime and task difficulty on task load (min. 6, max. 42).

but there were no significant effects of the interaction of priming and task difficulty.

5.3.1. SUS scores

Prime had a significant effect on SUS ratings. The positive-prime group rated the device more positively than the negative-prime and no-prime groups did, $F(2, 30) = 16.17$; $p < .001$, $\eta^2 = .519$. In the post hoc test the positive-prime group differed significantly from the negative-prime ($p < .01$) and no-prime groups ($p < .001$). Task difficulty had a significant effect on SUS ratings—users in the easy task condition rated the device more positively than did users in the hard condition, $F(1, 30) = 28.58$, $p < .001$, $\eta^2 = .488$. The interaction effect of prime and task difficulty on SUS ratings was not significant. The situation is reflected in Fig. 5.

5.3.2. AttrakDiff

Priming had a significant effect on the mean scores for Pragmatic Quality, Hedonic Identification, and Attractiveness. The positive-prime group rated the device higher on Pragmatic Quality than the negative-prime and no-prime groups did, $F(2, 30) = 9.06$; $p < .001$, $\eta^2 = .376$. The post hoc tests verified that the positive prime group differed significantly from the negative-prime ($p < .05$) and no-prime groups ($p < .005$). The positive-prime group rated the device higher for Hedonic Identification, as compared to the negative-prime and no-prime groups, $F(2, 30) = 5.74$, $p < .01$, $\eta^2 = .277$. In the post hoc test the positive prime group differed significantly from the no-prime group ($p < .005$), but not from the negative-prime group ($p > .05$). The positive-prime group found

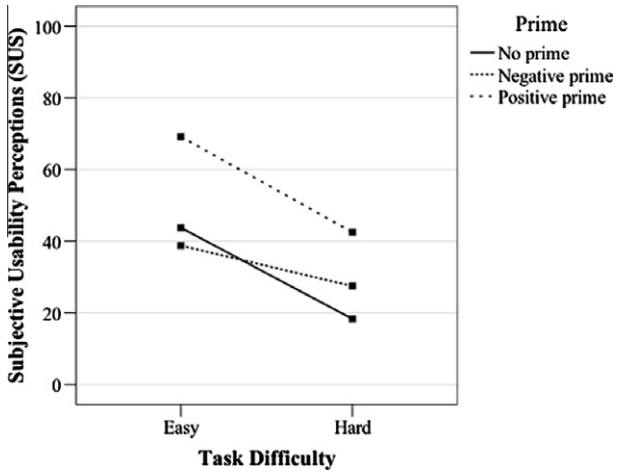

Fig. 5. The effects of prime and task difficulty on SUS ratings (min. 0, max. 100).

the device more attractive than the negative-prime and no-prime groups did $F(2, 30) = 6.44$, $p < .01$, $\eta^2 = .301$. In the post hoc test the positive-prime group differed from the negative-prime group ($p < .005$), but not from the no-prime group ($p > .05$).

Task difficulty had a significant effect on mean scores for Pragmatic Quality, Hedonic Identification, and Attractiveness. Users in the easy task condition rated the device higher on Pragmatic Quality

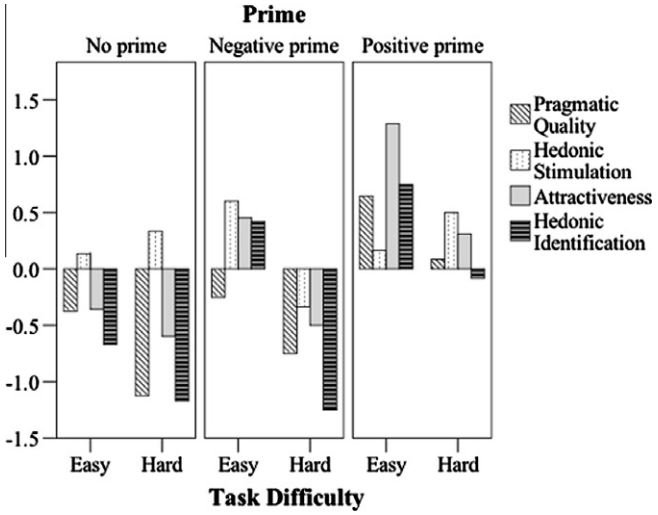

Fig. 6. The effects of prime and task difficulty on AttrakDiff ratings (min. −3, max. +3).

than did users in the hard task condition, $F(1, 30) = 7.25$, $p < .05$, $\eta^2 = .195$. Users in the easy task condition also rated the device higher for Hedonic Identification than did users in the hard condition, $F(1, 30) = 10.87$, $p < .01$, $\eta^2 = 0.266$, and more attractive than did users in the hard task condition, $F(1, 30) = 6.05$, $p < .05$, $\eta^2 = .168$.

The interaction of prime and task difficulty was not significant for any of the dimensions of AttrakDiff, and neither had a significant effect on the mean score for Hedonic Stimulation. The results are displayed in Fig. 6.

5.4. A non-parametric test

A problem associated with the small cell size of our study is that ANOVA relies on variables being normally distributed. To test if the results hold without this assumption, we run non-parametric tests for the two independent variables. The tests confirm the results obtained by ANOVA.

In the Mann–Whitney U test, *task difficulty* was found to have significant effect on task performance as measured with task success ($p < .001$) and task completion time ($p < .001$). Task difficulty had significant effect on task load ($p < .001$). Task difficulty had significant effect on the following post-experiment measures: SUS ratings ($p < .001$), Pragmatic Quality ($p < .05$), and Hedonic identification ($p < .01$). Task difficulty did not have a significant effect on overall negative or positive affect, Hedonic Stimulation, Attractiveness. In the Kruskal–Wallis test, *prime* was found to have a significant effect on the following post-experiment measures: SUS ratings ($p < .01$), Pragmatic Quality ($p < .01$), Attractiveness ($p < .05$) and Hedonic Identification ($p < .05$). The prime did not have significant effects on task success, task completion time, overall task load, overall negative or positive affect, or Hedonic Stimulation.

6. Discussion

In the present study, our goal was to understand the effects of expectations in a setting that resembles a typical usability test. In the study, users read either a very positive or a very negative product review, and a control group received no prior information. After reading the review, half of the participants in each group performed tasks known to be hard and the other completed analogous tasks known to be easy. Usability was measured both with objective usability metrics (task success and task completion time) and subjective post-task (NASA-TLX and PANAS) and post-experiment (SUS and AttrakDiff) questionnaires.

Considering our first research question, we found that product expectations have an influence on usability ratings (RQ1). The positive prime amplified SUS ratings by 74% in comparison to the baseline (negative-prime and no-prime groups), with the amplification being similar in the two task difficulty groups (easy and hard). The obtained effect sizes are much larger than in the prior study (Bentley, 2000). The pattern was analogous for three components of AttrakDiff: Pragmatic Quality, Hedonic Identification, and Attractiveness. The only AttrakDiff component not influenced was Hedonic Stimulation. Second, no support was found for the hypothesis that task difficulty would alter the effect of product expectations on subjective usability ratings (RQ2). In the study, task difficulty had a statistically significant effect on task performance and task load: easy tasks were completed more successfully, more swiftly, and with less task load than hard tasks were. The interaction effect of task difficulty and priming on subjective post-task ratings was not found significant: both were found to have only independent effects on post-task ratings.

Third, priming had a significant effect on the system-specific post-experiment ratings but no significant effect on the task-specific post-task ratings was found (RQ3). In other words, positive priming boosted the subjective usability ratings given at the end of the experiment but no significant influence on post-task measures of task load or emotions was found.

6.1. Appraisal of theoretical alternatives

Our results are in line with the theory of the self-fulfilling prophecy to the extent that participants with high expectations did perceive the device more positively than did other groups (negative and no-prime). However, priming independently influenced only the post-task ratings, not the interaction with the device (as measured in terms of task success and completion time) or task-specific ratings. Moreover, the self-fulfilling-prophecy theory seems too broad to explain the specific pattern observed in the study or the possible processes behind it. By contrast, ECT (Bhattacherjee, 2001; Thong et al., 2006) was not verified in our study. According to ECT, the negatively primed users should have given *higher* ratings when the tasks were easy, thus exceeding their expectations, and, vice versa, the positively primed users' post-experiment ratings should have "crashed" with the hard tasks.

In addition, the pattern observed is not compatible with theories that predict a transient psychological state to influence experiences (see: Bargh, 1994; Carver and Scheier, 1998). First, no significant effects of priming or task difficulty on emotions were found. Second, the task performance did not reflect the valence of prime, but instead the two valences produced comparable effects. However, the utilized setting does limit these conclusions to some extent: the study tested the effects of expectations evoked with primes including versatile product information (to mirror real-life situation) and not the influence of adjectives accumulating into one key factor as in Barghs (1994) original tests.

The data poses the following challenge: why did the valence of expectations (positive vs. negative) have *no* effect on post-task measures but did have one on the post-experiment usability ratings, replicating Bentley's (2000) finding? Our tentative explanation is related to the difference between the types of evaluation expected in task-specific vs. system-specific questionnaires: in post-task questionnaires the object of evaluation was task load and emotions, in post-experiment questionnaire it was the system. The post-experiment questionnaire required users to form an *evaluative opinion of the system as a whole*. Providing a stable opinion of a briefly used system *as a whole* is inherently challenging, and it is only natural to refer to prior knowledge from authoritative and

trusted sources. Such judgments are susceptible to the well-known cognitive biases like anchoring and framing (Hartmann et al., 2008) and could be tested by including a condition where the product review was given *after* actual use. By comparison, the post-task ratings asked only for evaluation of one's immediate experience of the task, not the device. The results of the study can also be explained with the theories of conformity and compliance, suggesting a felt need to trust the opinions of an authority, who in this case can be either the author of the product review or the experimenter who gives it to the user. Reliance on socially trustable sources is reinforced regarding matters of which one possesses only limited first-hand knowledge (Cialdini and Goldstein, 2004). In sum, the results demonstrate that while users can provide fairly accurate evaluations of task-specific load in relation to task success and completion time, their perception of a device as a whole is not just a sum of task-specific experiences but more likely a combination of these with product expectations and possibly other factors. Future work should expose the relationship to relevant factors, such as mental models that are known to affect users' expectations of a site's spatial layout (Roth et al., 2010).

Interestingly, in our study the no-prime group was at the same level in overall SUS ratings as the negatively primed group. The explanation for this is speculative and would require a study of the participants' prior expectations of HTC as a brand and of smartphones in general. It may have transpired that these non-tech-savvy users' prior expectation of smartphones was that they are hard to use, leaving the no-prime group too with negative expectations. Also, that the brand was unknown to the participants might have led them to expect it not to be as good as more familiar brands. Another explanation is that the perceived usability corresponded more closely to the negative review and the device really was hard to use. We acknowledge that our participants received none of the prior guidance in the use of the device that is often given in usability tests, which might have accentuated the difference between the no-prime and primed groups.

As a final observation, there were indications of a possible effect of expectations on *intermittent task performance and experience* that could come out statistically significant with a larger sample size. For the hard task condition, the effect of the type of prime on task success was significant. The effect of prime on task load was borderline-significant. That priming affects experience and performance in a task would be an important finding, although it does not relate well to the other findings. A possible explanation is that negatively primed participants expected the task to be more difficult and therefore tried harder. For the easy task, there may be a ceiling effect. Another possibility is that users who received the prime had more preknowledge of the interface and could therefore better process unexpected problems than users in the no-prime condition for whom the interface was totally new.

6.2. Implications for usability evaluation

We conclude by discussing implications for usability evaluation. The results show that users' expectations influence usability ratings so strongly that they may overshadow good performance. This finding has implications especially for summative usability evaluation aimed at discovering not just usability problems (like formative evaluation) but also revealing how future users will experience and perceive the product in their everyday life. At first glance, an obvious avenue based on the findings might be to remove the effect of expectations. However, since it is most likely that expectations influence usability ratings not just in the test situation but also in real use, removing the effect would not lead to more reliable indicators of users' usability ratings "in the wild." Instead, the finding should spark usability practitioners' interest in better understanding (1) what kinds of product expectations participants bring with them to the test, (2) how well these expectations represent those of the intended user population, and (3) how the test situation itself influences and may bias these expectations.

The difficulty in detecting participants' expectations lies in the measurement. While prior studies have sometimes simply measured expectations as the first task after meeting the test user (Szajna and Scamell, 1993), we do not see it as a viable practice in usability testing. The problem is that studies in consumer psychology have shown that stating expectations aloud makes perceivers focus on the negative issues and shifts the evaluation to the negative side (Ofir and Simonson, 2005). There are at least two ways to measure expectations in such a manner that the problem of asking can be avoided: measure expectations either long before the trials, assuming that (a) expectations are not going to change before testing and (b) that users are not going to be aware of them any longer at the time of testing, or afterwards, assuming that they could report them in a reliable manner, no matter what happened during the actual test. Both alternatives are worth exploring empirically but have their limitations.

In order to find out what kinds of expectations the intended user population has of the product and how well participants represent these, one could try to statistically control them. This could be done by estimating product expectations, without explicitly asking about them in the test, on the basis of correlation data for background demographics. Given the age, gender, socioeconomic status, previous computer usage, etc. of a user, one could estimate expectations. This solution would require systematic collection of expectations in large populations in order to determine the kinds of expectations that exist for a product. This knowledge could then be used as a criterion for selection of participants who are representative of the desired user population in their expectations. Nevertheless, this would not eliminate random variance in expectations, especially in small-*N* studies that practioners will do anyway.

Alternatively, one could try to influence expectations at the beginning of the study by giving users information about the product that one hopes will "override" the influence of prior knowledge and match the knowledge of the intended user population. This approach has some ecological validity, because in real life people seldom try new products that they know nothing about; they read and consume prior information that functions as the basis for this choice. The challenge would be to identify information that matches the sources used in the real world. Perhaps the users could be given advertisements or other product information that they would naturally be exposed to before trying and buying the product.

The clearest implication of the study's main result is that practitioners should, as much as possible, try to avoid biasing the users themselves. Experimenter expectations can and will bias users' expectations and therefore should be controlled. In clinical medicine, where the placebo effect is taken seriously, double-blind experiments are the gold standard. Alas, usability tests have no equivalent of the "white pill" manipulation; the tester and the user will most likely know which is the prototype, and this is unavoidable because they actually interact with it. More realistically, one could follow a good practice from experimental psychology: write down everything that is said about the system in the test in order to minimize variation from trial to trial and experimenter to experimenter.

7. Conclusion

In the long run, the effect of expectations is best addressed by researching the question of how experiences and perceptions evolve alongside technology use. This paper has shown that the

simplistic theory that predicts that expectations modulate experiences is insufficient—one needs to explain why in-task experiences and post-experiment ratings differ. An important goal for future work is to develop a pragmatic and valid way to account for expectations in usability evaluations. In addition, there is a need for further studies exploring the formation of product expectations, use experiences, and perceptions in the real-life context. For the time being, the mere existence of the "too good to be bad" effect shows that designing products with good objectively measured usability is not enough; one needs to evoke positive product expectations also, if one is to ensure positive perceptions of a product.

Acknowledgements

Parts of this work has been published in E. Raita, A. Oulasvirta, Too Good To Be Bad: The Effect Of Favorable Expectations on Usability Perceptions, Proc. HFES 2010.

References

Bark, I., Følstad, A., Gulliksen, J., 2005. Use and usefulness of HCI methods: results from an exploratory study among Nordic HCI practitioners. In: Proc. HCI 2005, 5–9 September, Edinburgh. Springer-Verlag, London, pp. 201–217.

Bangor, A., Kortum, P.T., Miller, J.T., 2008. An empirical evaluation of the system usability scale. International Journal of Human–Computer Interaction 24 (6), 574–594.

Bargh, J.A., 1994. The four horsemen of automaticity: awareness, intention, efficiency, and control in social cognition. In: Wyer, R., Srull, T. (Eds.), Handbook of Social Cognition, Basic Processes, 2nd ed., vol. 1. Lawrence Erlbaum Associates, New Jersey, pp. 1–40.

Bentley, T., 2000. Biasing web site user evaluation: a study. In: Proc. Australian Conference on Human–Computer Interaction, OZCHI 2000, IEEE Computer Society Press, pp. 130–134.

Bhattacherjee, A., 2001. Understanding information systems continuance: an expectation–confirmation model. MIS Quarterly 25 (3), 351–370.

Brinkman, W.-P., Haakma, R., Bouwhuis, D.G., 2009. Theoretical foundation and validity of a component-based usability questionnaire. Behaviour and Information Technology 2 (28), 121–137.

Brooke, J., 1996. SUS – a quick and dirty usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (Eds.), Usability Evaluation in Industry. Taylor and Francis, London, pp. 189–194.

Carver, C.S., Scheier, M.F., 1998. On the Self-Regulation of Behavior. Cambridge University Press, Cambridge, UK.

Cialdini, R.B., Goldstein, J., 2004. Social influence: conformity, and compliance. Annual Review of Psychology 55, 591–621.

Gulliksen, J., Boivie, I., Persson, J., Hektor, A., Herulf, I., 2004. Making a difference – a survey of usability profession in Sweden. In: Proceedings of NordiCHI. ACM Press, New York, NY, pp. 207–215.

Hart, S.G., 2006. NASA-task load index (NASA-TLX); 20 years later. In: Proc. HFES 2006.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (Eds.), Human Mental Workload. North Holland Press, Amsterdam, pp. 239–250.

Hartmann, J., De Angeli, A., Sutcliffe, A., 2005. Framing the user experience: information biases on website quality judgments. In: Proc. CHI 2008. ACM Press, New York, pp. 855–864.

Hassenzahl, M., Ullrich, D., 2007. To do or not to do: differences in user experience and retrospective judgments depending on the presence or absence of instrumental goals. Interacting with Computers 19 (4), 429–437.

Hassenzahl, M., Burmester, M., Koller, F., 2003. AttrakDiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. In: Ziegler, J., Szwillus, G. (Eds.), Mensch und Computer 2003. Interaktion in Bewegung. B.G. Teubner, Stuttgart, pp. 187–196.

Hertzum, M., 2010. Images of usability. International Journal of Human–Computer Studies 26 (6), 567–600.

Hornbæk, K., Lai-Chong Law, E., 2007. Meta-analysis of correlations among usability measures. In: Proc. CHI 2007. ACM Press, pp. 617–626.

ISO 9241-11.

Keppel, G., Wickens, T.D., 1991. Design and Analysis: A Researcher's Handbook, 4th ed. Prentice Hall, New Jersey.

Kindlund, E., Sauro, J.A., 2005. A method to standardize usability metrics into a single score. In: Proc. CHI 2005. ACM Press, pp. 401–409.

Mao, J.-Y., Vredenburg, K., Smith, P.W., Carey, T., 2005. The state of user-centered design practice. Communications of the ACM 48, 105–109.

Merton, R.K., 1968. Social Theory and Social Structure. The Free Press, New York.

Nielsen, J., Levy, J., 1994. Measuring usability: preference vs. performance. Communications of ACM 37 (4), 66–75.

Ofir, C., Simonson, I., 2005. The effect of stating expectations on customer satisfaction and shopping experience. Journal of Marketing Research XLIV, 164–174.

Roth, S.P., Schmutz, P., Pauwels, S.L., Bargas-Avila, J.A., Opwis, K., 2010. Mental models for web objects: where do users expect to find the most frequent objects in online shops, news portals, and company web pages? Interacting with Computers 22 (2), 140–152.

Schimmak, U., Crites, S., 2005. The structure of affect. In: Albarracin, D., Johnson, B., Zanna, M. (Eds.), The Handbook of Attitudes. Lawrence Erlbaum Associates, New Jersey, pp. 397–436.

Snyder, M., Stukas, A.A., 1999. Interpersonal processes: the interplay of cognitive, motivational, and behavioral activities in social interaction. Annual Review of Psychology 50, 273–303.

Szajna, B., Scamell, R.W., 1993. The effects of information system user expectations on their performance and perceptions. MIS Quarterly 17 (4), 493–516.

Thong, J.Y.L., Hong, S.J., Tam, K.Y., 2006. The effects of post-adoption beliefs on the expectation–confirmation model for information technology continuance. International Journal of Human–Computer Studies 64 (9), 799–810.

Tractinsky, N., 2004. Towards the study of aesthetics in information technology. In: Proc. ICIS 2004, pp. 11–20.

Tranctinsky, N., Katz, A.S., Ikar, D., 2000. What is beautiful is usable. Interacting with Computers 13 (12), 127–145.

Tullis, T., Albert, B., 2008. Measuring the User Experience. Elsevier, Burlington, Massachusetts.

Watson, D., Clark, L.A., Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of Personality and Social Psychology 54 (6), 1063–1070.