

TOO GOOD TO BE BAD: THE EFFECT OF FAVORABLE EXPECTATIONS ON USABILITY PERCEPTIONS

Eeva Raita and Antti Oulasvirta
Helsinki Institute for Information Technology HIIT
Aalto University and University of Helsinki
Finland

The most common measurements in the task-based usability evaluation paradigm include behavioral (e.g., completion times or errors) and subjective measures (e.g., ratings). Previous work has shown that success and performance in the test tasks do not dictate subjective usability perceptions, which instead are affected by appraisals of the system such as those of its aesthetic appeal. While these appraisals are an outcome of the exposure to and the interaction with the system, less is known about the effect of predispositions (expectations) formed before any exposure. To understand how expectations influence usability perceptions, we devised an experiment wherein 36 subjects read a positive or a negative product review for a novel mobile device (while a control group read nothing) before a usability test. The results demonstrate a surprisingly strong amplifying effect of the positive expectation on the post-experiment ratings, which, interestingly, held even in a condition where the users failed in all of the tasks. We briefly discuss implications of this finding.

INTRODUCTION

The motivation for the present study came from the usability practitioners of a device manufacturer who have repeatedly come across a problematic phenomenon: in a usability test, a prototype with a good “objective usability” may nevertheless receive poorer ratings than a less usable product on the market (used as a baseline). This effect, which the practitioners dubbed the “blondes have more fun effect,” may be related to the effect that the visual appeal of an interface can be dominant over objectively measured “actual usability”. In the widely cited experiment by Tractinsky et al. (2000) visual appeal strongly affected usability perceptions even when usability was hampered via experimenter-induced problems, and the satisfaction with the system resulted from the perceived aesthetics. There have been numerous follow-ups to this study. Many of them explain the finding in terms of a “halo effect”; that is, the perception of a good quality “leaks” to the evaluation of other qualities (e.g., De Angeli et al., 2006).

This phenomenon is also associated with a broader question concerning the interdependence of objective and subjective usability measures. According to a recent meta-analysis, subjective perceptions rarely correlate with behavioral (“objective”) measurements of usability, such as task success and time (Hornbaek & Lai Chong-Law, 2007). Researchers are investigating which factors affect subjective and which affect objective measures of usability. In combination with findings such as that of Tractinsky et al. (2000), the evidence suggests that subjective measures of usability are at least partly affected by different factors than objective measures.

In this paper, we explore one such factor not empirically studied previously. While the study of aesthetic appeal requires some sort of exposure to the system, little is known about the effects of *expectations* that emerge *before any exposure to the system*. Expectations are of importance, because users are never naïve in their relationship with technology. Even when the technology is novel to them, there are predis-

positions that shape their action and experience, and there are many sources that might have affected expectations prior to actual use—advertisements, brands, word of mouth, product reviews, discussion forums, and exposure to related products, to name a few. It would be useful if practitioners knew how expectations affect subjective ratings to such an extent that they could avoid unwanted bias in their results. The problem is particularly relevant in comparative usability studies (say, where a prototype is compared to a product on the market).

Researchers, too, have long suggested that expectations affect usability perceptions, but empirical evidence pertinent in particular to usability testing is scarce. Most of the relevant work has been done in the study of information system (IS) usability. It has been empirically demonstrated that employees’ expectations of usability are predictive of the probability of adoption (Venkatesh et al., 2003). Additionally, IS research has shown that user satisfaction is influenced partly by the confirmation of expectations from prior IS use (Bhattacharjee, 2001) and that positive expectations increase satisfaction measured after use while unrealistic expectations are “worn out” with more experience (Szajna & Scamell, 1993).

In this paper, we attempt to understand this issue in the particular context of usability testing. The experiment follows the approach used previously to study the effect of aesthetic appeal on usability perception, with the exception that in prior studies the user interface was manipulated, whereas our experiment kept the user interface constant and manipulated the information given before the actual test. Our subjects read a positive, a negative, or no product review before the hands-on test. In the test, we deployed commonly used subjective measures: NASA-TLX, PANAS, SUS, and AttrakDiff.

We seek to answer three research questions:

1. Do expectations influence usability ratings?
2. Do task performance and subjective ratings operate as independent factors or in interaction?
3. Is there a difference between task-specific and system-specific ratings?



Figure 1: HTC Touch Diamond. The same picture was used in the product reviews (primes, see Table 1).

METHOD

Participants

The participants (N=36) were volunteers who enrolled for a mobile phone usability study announced via eight mailing lists for university students. They were 21 females and 16 males, aged between 20 and 30 years, with the mean age being 25 years. All participants had an upper secondary-school education, and 16 had academic degrees also; 31 participants listed studying as their main duty, one was completing his national service, and four were working.

Experimental Design

The study was a 3 x 2 between-subjects experiment with expectations (positive, negative, or neutral) and task difficulty (easy or hard) as the factors. The cell size was six participants.

Tasks and Materials

Tasks were performed with the HTC Touch Diamond phone, a Windows-Mobile-6.1-based Pocket PC with a TouchFLO 3D interface (see Figure 1). The device was launched in Europe in 2008 and at the time of the study was unknown to participants. The product review used for priming was designed to resemble an authentic product review from the Internet. What was changed between the conditions was how the features of the phone were described (see Table 1).

Table 1: Contents of the primes, “product reviews from the Internet”

Positive prime	Negative prime
Easy-to-use touchscreen	Hard-to-use touchscreen
Stylish, top design	A magnet for fingerprints
Useful basic buttons	Quite useless basic buttons
Good technical equipment	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
Lightness and pleasantness to hold	Small battery that has to be charged really often
Intuitive user interface	Slow interface to use

Table 2: Instructions for easy and hard tasks

#	Easy version	Hard version
1	Write a text message that says “hi sister”	Write a text message that says “hi mom” (in Finnish, includes an umlaut that was not preset for the keyboard)
2	Watch a video from YouTube	Watch a video from a journal’s site
3	Listen to saved music	Listen to the radio
4	Put the phone in silent mode	Change the ringtone

In the study, users were given task goals on pieces of papers. Half of the participants received directions for easy and the others for hard tasks (see Table 2). Tasks were found by exploration and chosen such that they varied only in difficulty in this particular device.

Procedure

The study consisted of four phases. All participants were tested individually, and all tests were videotaped with participants’ informed consent for later analysis of task performance. 1) In the first phase, participants completed a pre-test questionnaire about their demographics. 2) In the second phase, two thirds of the participants were given a positive or a negative review, and one third were not given any prime, since they served as a control group. Special care was taken that the prime be introduced 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 said that he or she had finished reading. 3) In the third phase, participants performed four easy or hard tasks with the phone. The participants were not told about the tasks or their difficulty in advance. They were told that they would have seven minutes to perform each task and that they should inform the researcher when they had completed each. Participants filled in a short questionnaire after every task. 4) In the last phase, they completed a closing questionnaire with questions about previous experience with mobile phones and usability perceptions.

MEASUREMENTS

Task-specific Questionnaires

NASA-TLX is a subjective workload assessment tool that gives an overall workload score based on a weighted average from six sub-scales. It employs six questions touching upon the various aspects of task load, which are answered with a rating on a seven-point scale (Hart & Staveland, 1988). Along with NASA-TLX, task-specific emotions were measured with the PANAS questionnaire, developed for the measurement of positive and negative affect and consisting of 20 items. The PANAS questionnaire addresses emotions ranging from excitement to distress, and participants are asked to indicate how much they are feeling these, on a five-point scale (Watson et al., 1988).

Final Questionnaire

In SUS, there are 10 statements concerning the ease of learning and using the system that are rated on a five-point Likert scale. The overall score is calculated by summing statement-specific scores and then multiplying the sum by 2.5. The final SUS score has a range of 0–100 (Brooke, 1996).

AttrakDiff applies a user experience scale that has been widely used in practical evaluation work. In AttrakDiff, pragmatic (goal achievement) and hedonistic quality (stimulation and identification) as well as attractiveness are measured with 28 word pairs evaluated on a seven-step scale. The middle value (4) creates scale values (Hassenzahl et al., 2003).

Prime Verification

We conducted a brief independent survey to make sure that our primes worked as expected. We sent invitations to a student mailing list, asking participants to take part in a brief survey. In the questionnaire, participants were first asked for their background information, after which they were asked to read a review of the HTC Touch Diamond (the same as the ones used in the main test) and then rate it with the SUS and AttrakDiff evaluations. In total, 87 participants enrolled for the study, and they were randomly assigned to read either the negative or the positive review. Verifying our primes, there was a statistically significant difference between the groups: Those who read the positive review rated the device higher in SUS scores ($M=63.86$, $SD=13.82$) than those who read the negative review ($M=48.03$, $SD=13.48$), with $F(1,85)=28.25$, $p < 0.001$. There were analogous results for AttrakDiff scales.

RESULTS

For statistical testing, we utilized a 3 x 2 Analysis of Variance (ANOVA) with *prime* and *difficulty* as the two main factors. An alpha of 0.05 is utilized throughout, unless otherwise mentioned.

Task Performance

Task success: Levene's test of equality of variances proved significant, $F(5,30)=2.56$, $p=0.048$. Therefore, we set a more stringent alpha (0.01) for this particular test (see Keppel and Wickens, 1991). The effect of prime on task success was not significant at this α level, $F(2,30)=5.09$. Task difficulty influenced completion rate significantly in such a way that easy tasks were completed more often ($M=3.78$, $SD=0.43$) than hard tasks ($M=1.83$, $SD=0.92$), $F(1,30)=120.10$, $p < 0.001$. The interaction effect of prime and task difficulty on task success was significant, $F(2,30)=10.98$, $p < 0.001$.

Task completion time: Levene's test of equality of variances was significant: $F(5,30)=2.85$, $p=0.032$, so we set the alpha to a more stringent level (0.01). Prime did not have a significant effect on task completion time, $F(2,30)=1.24$. Task difficulty had a significant effect on average completion time: easy tasks were completed more quickly ($M=477.9$, $SD=220.9$) than hard tasks ($M=1297.1$, $SD=211.2$),

$F(1,30)=135.44$, $p < 0.001$. There was no interaction effect of prime and task difficulty on task completion time, $F(2,30)=1.56$.

Post-task Ratings

Task load: Levene's test of equality of variances was not significant for the following analyses. Prime had a close-to-significant effect on task load, $F(2,30)=3.07$, $p=0.061$, in such a way that the no-prime group felt the highest task load. Task difficulty had a significant effect on task load. Users in the hard task condition reported a greater task load ($M=15.81$, $SD=3.58$) than those in the easy task condition ($M=9.98$, $SD=2.91$), $F(1,30)=35.12$, $p < 0.001$. The interaction effect of prime and task difficulty was close to significant, $F(2,30)=2.72$, $p=0.082$. Figure 2 depicts the situation.

Emotions: Levene's test of equality of variances was significant for positive affect reported in the third task, $F(5,30)=3.43$, $p=0.014$, and for the overall positive affect, $F(5,30)=2.7$, $p=0.037$, so we set the alpha to a more stringent level (0.01) for these particular tests. Levene's test was not significant for other analyses. Prime did not have significant effects on task-specific or overall affect (all F s < 2.15). Task difficulty had a significant effect on the negative affect reported in the third task, $F(1,30)=4.93$, $p < 0.05$, and positive affect reported in the second task $F(1,30)=5.27$, $p < 0.05$, since easy tasks led to reports of more positive and hard tasks to reports of more negative affect. Task difficulty had a close-to-significant effect on the overall positive affect, $F(1,30)=3.48$, $p=0.071$. Task difficulty did not have significant effects on other task-specific or overall affect. There were no significant interaction effects of prime and task difficulty on task-specific or overall affect (all F s < 1.53).

Post-experiment Ratings

Both post-experiment measures of usability perception reflected the same pattern. Levene's test was not significant for any of the following analyses.

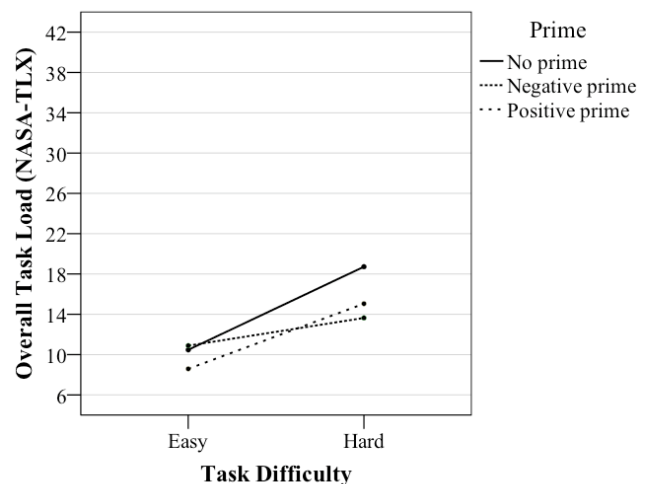


Figure 2: The effect of prime and task difficulty on task load (min.6, max. 42).

DISCUSSION

SUS scores: Prime had a significant effect on SUS ratings: The positive-prime group rated the device more positively ($M=55.83$, $SD=18.23$) than the negative-prime ($M=33.13$, $SD=12.89$) and no-prime groups ($M=31.04$, $SD=17.01$), $F(2,30)=16.17$; $p < 0.001$. Task difficulty had a significant effect on SUS ratings—users in the easy task condition ($M=50.56$, $SD=16.88$) rated the device more positively than did users in the hard condition ($M=29.44$, $SD=15.98$), $F(1,30)=28.58$, $p < 0.001$. The interaction effect of prime and task difficulty on SUS ratings was not significant, $F(2,30)=1.57$. The situation is illustrated in Figure 3, below.

AttrakDiff: Prime had a significant effect on the means of pragmatic quality, hedonistic identification, and attractiveness. The positive-prime group rated the device more pragmatically suitable ($M=0.41$, $SD=0.63$) than the negative-prime ($M=-0.63$, $SD=0.68$) and no-prime groups ($M=-0.79$, $SD=0.89$), $F(2,30)=10.47$; $p < 0.001$. The positive-prime group rated the device higher for hedonistic identification ($M=0.33$, $SD=0.91$) than the negative-prime ($M=-0.42$, $SD=1.1$) and no-prime groups ($M=-0.92$, $SD=1.1$), $F(2,30)=5.74$; $p < 0.01$, and more attractive ($M=0.79$, $SD=0.92$) than did the negative-prime ($M=0$, $SD=1.0$) and no-prime groups ($M=-0.46$, $SD=0.78$), $F(2,30)=6.78$, $p < 0.01$.

Task difficulty had a significant effect on the means of pragmatic quality, hedonistic identification, and attractiveness. Users in the easy task condition rated the device more pragmatic ($M=-0.03$, $SD=0.83$) than did users in the hard task condition ($M=-0.64$, $SD=0.89$), $F(1,30)=6.84$, $p < 0.05$. Users in the easy task condition rated the device higher for hedonistic identification ($M=0.17$, $SD=1.00$) than users in the hard condition ($M=-0.83$, $SD=1.07$) $F(1,30)=10.87$, $p < 0.01$, and more attractive ($M=0.47$, $SD=0.99$) than did users in the hard task condition ($M=-0.25$, $SD=0.94$), $F(1,30)=6.63$, $p < 0.05$.

The interaction of prime and task difficulty was not significant for any AttrakDiff scores (all $F_s < 1.31$), and neither of them had a significant effect on the mean score for hedonistic simulation (all $F_s < 0.15$).

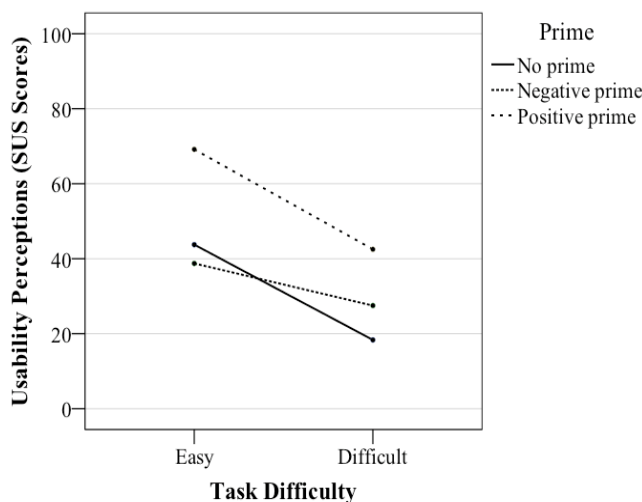


Figure 3: The effects of prime and task difficulty on usability perceptions (min. 0, max. 100).

Before an attempt at explanation, let us sum up the findings. For post-task ratings, we found that, confirming the validity of the usability manipulation, tasks were completed more often, more quickly, and with a lighter task load in the easy task condition. For post-experiment ratings, we found that the positive prime amplified SUS ratings by 74% in comparison to the negative-prime and no-prime groups; the amplification was uniform across both task difficulty groups. This pattern was analogous for all AttrakDiff components except hedonistic stimulation.

The challenge for explaining the findings is this: Why did the valence of expectations (positive vs. negative) appear as an inert factor for post-task load while it had so strong an effect on the post-experiment rating? Explanations for our findings can be sought from the numerous theories that touch upon expectations, and we will shortly mention a few of them. In consumer psychology, the widely applied *expectation confirmation theory* (ECT) states that post-purchase satisfaction is a result of *comparison* between expectations and performance. That is, expectations exist as a norm against which experiences are compared. According to this theory, high expectations in combination with poor performance should lead to a negative evaluation and vice versa (e.g., Bhattacharjee, 2001). This theory did not get verification from our study, where, on the contrary, high expectations boosted ratings regardless of task success.

Another possible explanation can be drawn from social psychology and the theory of *self-fulfilling prophecy*, which refers to a prediction that causes itself to become true; positive expectations of an interaction partner make one see the other favorably (Wilkins, 1976). If the user is told, that “the prototype is very usable”, it could affect the way in which the device is interacted with—called behavioral confirmation—and/or the way in which events are perceived, called perceptual confirmation. This was partly in line with our results, except that in our study priming influenced only post-experiment ratings and not users’ performance or post-task ratings.

One tentative explanation builds on the idea derived from cognitive psychology that expectations are a belief system recruited differently in actual task performance from in post-experiment ratings. In the study, expectations were manipulated by reviews with an evaluative content. The reason the final questionnaire was so different is probably that it required users to form an evaluative opinion of the system *as a whole*. Providing a stable opinion of a briefly used system is difficult, and it is natural to refer to prior knowledge from authoritative sources. By comparison, the post-task ratings required only the evaluation of one’s immediate experience.

Interestingly, the no-prime group was at the same level in overall SUS ratings as the negatively primed group. The explanation for this is still a question mark and would require a study of the participants’ prior knowledge of HTC as a brand and of smartphones in general. It may have been the case, for example, that these non-tech-savvy users’ prior expectation was that smartphones are hard to use so the no-prime group had negative expectations also.

Implications

The results show that users' expectations are a potentially confounding factor and so strong that they may overshadow good performance in tasks. Unfortunately, there is no easy solution to this problem. One obvious solution would be to measure expectations just before testing. However, studies in consumer psychology have shown that stating expectations aloud influences perceptions, shifting evaluations to the negative side (Ofir & Simonson, 2005). Another solution would be to measure expectations either long before the trials, assuming that a) expectations would not change before testing and b) that users would not be aware of them anymore at the time of testing, or afterwards, assuming that they could report them in a reliable and unbiased manner, no matter what happened during the actual test. Both alternatives are worth exploring but include limitations related to the fact that expectations and experiences mutually shape one another.

The effects of expectations can also be taken into account in the design of a usability evaluation. One way would be to avoid comparisons between prototypes and well-known products and seek other baselines for comparison, such as comparisons within a single product line or between versions of the same product. In such a set-up, expectations would be less varied and less random. One could also try to statistically control expectations, but this would require systematic collection of expectations in large populations and would not eliminate random variance especially in small N-studies.

CONCLUSION

The most important finding from the present study is the demonstration of the existence of the effect. The effect of a positive expectation was found to include a large (74%) increase in a commonly used measure of usability perception. The effect held also when users failed in all tasks, boosting the rating to the level of the non-primed group who did not fail in tasks. We call this effect the "too good to be bad effect."

In this paper, we have shown that simplistic theories that predict that expectations influence perceptions monotonically are insufficient. Our treatment of theoretical alternatives shows no evidence for one available explanation (ECT), but further research is needed to achieve demarcation with respect to the remaining explanations. A distinction that needs more attention in the future is the observed difference between in-task experiences and post-experiment ratings.

While we cannot yet offer a preeminent solution to the problem, we hope that this study will raise the awareness of researchers and practitioners alike. An important goal for future work is to develop pragmatically useful yet valid ways to deal with expectations in usability evaluations.

ACKNOWLEDGEMENTS

This work was supported by the UCIT Graduate School and Tekes project Theseus. We thank Mikael Wahlström and Antti Lindqvist for valuable comments.

REFERENCES

- Bangor, A., Kortum, P.T. and Miller, J.T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 24(6), 574-594
- Brooke, J. (1996). SUS – A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I.L. McClelland (Eds.), *Usability evaluation in industry*. Taylor and Francis, London, UK, 189-194.
- Bhattacharjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly*, 25(3), 351-370
- De Angeli, A., Sutcliffe, A. and Hartman, J. (2006). Interaction, usability and aesthetics: What influences users' preferences? In *Proceedings of the 6th conference on Designing Interactive systems*, PA, USA.
- Hart, S.G. and Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds.), *Human Mental Workload*, North Holland Press, Amsterdam, Holland, 239-250
- Hassenzahl, M., Burmester, M., and Koller, F. (2003). AttrakDiff: Ein fragebogen zur messung wahrgenommene hedonischer und pragmatischer qualität. In J.Ziegler and G. Szwillus (Eds.), *Mensch and Computer 2003. Interaktion in Bewegung*. B.G. Teubner, Stuttgart, Deutschland, 187-196.
- Hornbaek, K. and Lai Chong-Law, E. (2007). Meta-analysis of Correlations among Usability Measures. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, California, USA.
- Keppel, G. and Wickens, T. D. (1991). *Design and Analysis. A Researcher's Handbook (4th Edition)*. Prentice Hall New Jersey, USA.
- Ofir, C. and Simonson, I. (2005). The effect of stating expectations on customer satisfaction and shopping experience. *Journal of Marketing Research*, XLIV, 164–174.
- Szajna, B. and Scamell, R.W. (1993). The effects of information system user expectations on their performance and perceptions. *MIS Quarterly*, 17(4), 493-516.
- Tractinsky, N., Katz, A.S. and Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, 13(12), 127-145.
- Venkatesh, V., Morris, M.G., Davis, G.B. and Davis, F.D. (2003). User acceptance of information technology: toward a unified view. *MIS Quarterly*, 27(3), 425-278.
- Watson, D., Clark, L.A. and 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.
- Wilkins, W.E. (1976). The concept of a self-fulfilling prophecy. *Sociology of Education*, 49, 175-183.