ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/281229002

An Exploration of the Relation Between Expectations and User Experience

Article *in* International Journal of Human-Computer Interaction · July 2015

DOI: 10.1080/10447318.2015.1065696

CITATIONS
20

READS
580

3 authors, including:

Jaroslav Michalco
University of Copenhagen
2 PUBLICATIONS **26** CITATIONS
SEE PROFILE

Kasper Hornbæk
University of Copenhagen
188 PUBLICATIONS **6,649** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

WallViz View project

Human Bayesian Inference View project

All content following this page was uploaded by Jaroslav Michalco on 20 September 2015.

The user has requested enhancement of the downloaded file.

Intl. Journal of Human–Computer Interaction, 31: 603–617, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 1044-7318 print / 1532-7590 online
DOI: 10.1080/10447318.2015.1065696

Taylor & Francis
Taylor & Francis Group

An Exploration of the Relation Between Expectations and User Experience

Jaroslav Michalco, Jakob Grue Simonsen, and Kasper Hornbæk
Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

Downloaded by [89.173.14.71] at 12:19 01 September 2015

Before using an interactive product, people form expectations about what the experience of use will be like. These expectations may affect both the use of the product and users' attitudes toward it. This article briefly reviews existing theories of expectations to design and perform two crowdsourced experiments that investigate how expectations affect user experience measures. In the experiments, participants saw a primed or neutral review of a simple online game, played it, and rated it on various user experience measures. Results suggest that when expectations are confirmed, users tend to assimilate their ratings with their expectations; conversely, if the product quality is inconsistent with expectations, users tend to contrast their ratings with expectations and give ratings correlated with the level of disconfirmation. Results also suggest that expectation disconfirmation can be used more widely in analyses of user experience, even when the analyses are not specifically concerned with expectation disconfirmation.

1. INTRODUCTION

A largely neglected aspect of user experience is that of temporality. As users engage with interactive products, they become increasingly familiar with them, and this over time may change their experience. Furthermore, prior to using a product, users are in anticipation—they form expectations about what would the experience be like (Karapanos, Zimmerman, Forlizzi, & Martens, 2009). Such expectations may play an important role in the subsequent experience (Anderson & Hair, 1972) and may eventually affect the usability ratings of a product. However, only a few studies have investigated the effects of expectations on subjective usability ratings or user experience (e.g., Al Sokkar & Law 2013; Raita & Oulasvirta, 2011). Studies concerning the effect of expectations on satisfaction and behavior are mainly part of other research disciplines, such as psychology (Levin, Schneider, & Gaeth, 1998) or marketing (Anderson, 1973; Cardozo, 1965; Churchill & Surprenant, 1982; Cohen & Goldberg, 1970; Levin & Gaeth, 1988; Oliver & Linda, 1981; Olshavsky & Miller, 1972; Olson & Dover, 1976), and often use nontechnological products as the object of study.

One type of an interactive product about which users form expectations beforehand is computer games. Before playing a game, some users read reviews or ratings to find out what the experience would be like. Such reviews often contain information on both usability and user experience but use terms distinct from reviews of other kinds of software. For example, computer games are often evaluated on playability, a term encompassing both usability and other factors such as game story or game mechanics (Desurvire, Caplan, & Toth, 2004). The facts that (a) computer games are a popular facet of pop culture, (b) are interactive, and (c) users are in general familiar with reviews as a source of information prior to interacting with the product make them an apt class of interactive products for studying the effects of expectation on usability and user experience by means of a sizable user group.

The goal of this article is to study how expectations affect user experience ratings of online games. We report two experiments where participants were presented with an online game, played it, and afterward rated it with respect to user experience and to whether their expectations were met. These experiments were conducted online to obtain a large pool of participants. In contrast to earlier work on user experience, we use a variety of products and several measures of both expectations and user experience. This allowed for an exploration of the effect of expectations on user experience. The results showed that expectations affect ratings of user experience substantially and that the notion of expectation disconfirmation helps separate what appear to be different influences of failing to meet expectations and surpassing them. We believe that these results have implications for user experience research, have open questions for future work, and influence how practical user experience evaluations should be conducted.

2. RELATED WORK

2.1. Theories of Expectation

Expectations of a product may impact the subsequent experience. Expectations can be either confirmed (when the

Address correspondence to Kasper Hornbæk, Department of Computer Science, University of Copenhagen, Njalsgade 128-132, DK-2300 Copenhagen, Denmark. E-mail: kash@di.ku.dk

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hihc.

Downloaded by [89.173.14.71] at 12:19 01 September 2015

experience with the product meets the expectations) or disconfirmed (when the experience is different than expected). There are two types of disconfirmation—*positive* and *negative* disconfirmation. The expectations are *positively disconfirmed* when the experience exceeds them and *negatively disconfirmed* when the experience does not meet the expectations (Oliver, 1977). If some product features exceed the expectations, these features become increasingly important in the overall satisfaction; if some features do not meet the expectations, they seem to be less relevant. Thus, when the expectations conflict with the experience users appear to behave in one of two ways: They either (a) adapt their perception of their prior expectations to match the experience or (b) adapt the experience to match the expectations.

The first of the aforementioned behaviors—adapting perception of prior expectations to meet the actual experience—is explained by *contrast theory,* which states that the user distorts their perception of their prior expectations to match the performance of the product. This implies that a negative disconfirmation should result in a negative rating, whereas a positive disconfirmation should lead to a very positive rating (Oliver, 1977). This prediction is also in line with the *expectation-confirmation theory*, which states that postpurchase satisfaction is the result of comparison between expectations and the actual performance (Raita & Oulasvirta, 2011).

On the other hand, *assimilation theory* and *dissonance theory* state that users distort the disconfirming experience to match the expectations. If expectations were formed from a product review, users who read an overstatement of the product will rate it better than those with unbiased expectations, and users who read an understatement of the product will rate it lower. This theory is also in agreement with the theory of *self-fulfilling prophecy*, that is, when a prediction causes actors to, consciously or not, make the prediction true due to the prediction being explicitly stated (Oliver, 1977; Olshavsky & Miller, 1972; Raita & Oulasvirta, 2011). In short, *contrast theory* states that the *ratings are a function of disconfirmation level,* whereas *assimilation theory* states that the *ratings are a function of expectations* (Oliver, 1977).

These theories have mainly been studied in fields outside of human–computer interaction (e.g., psychology and marketing). These studies use various products as objects of interest, such as pens (Anderson, 1973; Cardozo, 1965), audio recorders (Olshavsky & Miller, 1972), coffee (Cohen & Goldberg, 1970; Olson & Dover, 1976), beef (Levin & Gaeth, 1988), video players (Churchill & Surprenant, 1982), or pajamas (Oliver & Linda, 1981). We compared nine studies based on their approach to expectation manipulation. Of these, six manipulated expectations to some degree (Anderson, 1973; Cardozo, 1965; Churchill & Surprenant, 1982; Levin & Gaeth, 1988; Olshavsky & Miller, 1972; Olson & Dover, 1976), whereas only two kept the expectations constant (Cohen & Goldberg, 1970; Oliver & Linda, 1981). However, only five studies altered the quality of the product (Churchill & Surprenant, 1982; Cohen & Goldberg, 1970; Oliver & Linda, 1981; Olshavsky & Miller, 1972; Olson & Dover, 1976), whereas three studies kept the quality constant (Anderson, 1973; Cardozo, 1965; Levin & Gaeth, 1988). A special case (the ninth article) is the study by Oliver (1977), which looked at the level of disconfirmation of expectations, and both expectations and level of disconfirmation were inferred from participants' responses. It is worth mentioning that in eight of nine studies the authors used only one product, for which the researchers altered its performance, instead of using multiple products of different quality, for instance, using one recorder and only altering the quality of the recording as in Olshavsky and Miller (1972).

2.2. Expectations and User Experience

Expectations of interactive products have been studied mainly in research on information systems. For instance, Bhattacherjee (2001) studied the difference between the initial decision to use a system and the continued use of the system using expectation-confirmation theory. A model was developed that posited an influence of expectation confirmation (measured as the post-use response to questions on whether performance was better than expected) on both satisfaction and perceived usefulness. Bhattacherjee found that expectation confirmation was a strong predictor of satisfaction but that it also predicted perceived usefulness well. Later studies in information systems have shown that disconfirmation in general has a negative influence on attitudes (e.g., Venkatesh & Goyal, 2010) and that expectations may be analyzed separately for usefulness and ease of use. Brown, Venkatesh, and Goyal (2012) developed a model and coded the responses of 1113 employees in an organization adopting an Intranet-based knowledge-sharing system in two phases: prior to the program being introduced and 6 months after introduction. Using a nonlinear model of expectation disconfirmation, they found that disconfirmation is a predictor of level of use, but that for low levels of disconfirmation, the level of expectation is a stronger predictor. Similarly, positive disconfirmation was a strong predictor of perceived usefulness. Many other studies of expectations in information systems exist (e.g., Brown, Venkatesh, & Goyal, 2014; Brown, Venkatesh, Kuruzovich, & Massey, 2008; Lee, 2010).

Whereas the aforementioned studies have allowed detailed models and analysis approaches to be developed for expectations, they share a number of limitations with respect to their relevance and implications for user experience research. First, the aforementioned studies are mainly correlational, making drawing conclusions about causality difficult. Second, the studies are mainly based on the modeling approach associated with the technology acceptance model (Davis, 1989). Although this model has had significant influence in information systems, it is rarely related to or used in user experience research. Third, the studies rarely use key measures from user experience research, in particular those relating to the perception of software quality (e.g., AttrackDiff) or software specific experiences (e.g., play

Downloaded by [89.173.14.71] at 12:19 01 September 2015

experiences). For these reasons, we turn to review studies of user experience that study expectations empirically.

Only a few earlier publications in user experience research have studied expectations. These studies manipulated the expectations of participants by using one of two techniques: framing, which keeps the information the same but alters the way this information is presented (e.g., saying that 90% of users liked an app and saying that 10% did not like the app contains the same information but presents it differently), and by priming, that is, increased sensitivity to certain stimuli due to prior experience (e.g., an increased sensitivity to certain features of a product after reading positive reviews).

Framing was, for instance, used in the work of Hartmann, De Angeli, and Sutcliffe (2008), in which website quality was evaluated on three criteria: usability, look & feel, and content & service quality. The authors used two online experiments. In the first, participants ($N = 67$) read a description of a website (without actually seeing it) and were asked to rate it based on the written description. In the second experiment, participants ($N = 360$) read a description of a website, were shown the website, and then were asked to rate it. In both experiments the participants who were exposed to positively framed information provided better ratings of the website than those in the control group. Also, the participants who were in the negatively framed group rated the website lower than the control group, which shows that framing does affect the subsequent quality ratings (perceived usability: $\eta^2 = .25$ in the first experiment; $\eta^2 = .06$ in the second experiment[1]).

Bentley (2000) investigated how expectations affect the ratings of a website. Twenty-four participants were presented with a description of a website describing it either in a positive or in a negative way. After the participants were primed, the experiment consisted of these stages. At first, the participants filled in a questionnaire about their background; then they were given 5 min to explore the interface. Then they were given several tasks to complete. Upon completion of all tasks the participants filled in Software Usability Measurement Inventory (Kirakowski & Corbett, 1993). The results showed no significant difference between participant groups in terms of performance, suggesting that performance was independent of both type of presentation and participant experience. However, participants who had the website presented in positive way rated the design better than those in the negative group on controllability ($\eta^2 = .12$), helpfulness ($\eta^2 = .39$), and general usability ($\eta^2 = .33$). As there was no difference in performance, the difference observed was only the result of priming, suggesting that different expectations do affect usability ratings.

Raita and Oulasvirta (2011) studied how product expectations influenced usability ratings. Thirty-six participants went through five stages of an experiment: (a) They filled out a questionnaire about their background; (b) the participants not in the control group were given either a positive or a negative review of an HTC phone; (c) all participants were given the phone and asked to perform a series of tasks with different levels of difficulty; (d) after every task, they were asked to fill a questionnaire based on NASA task load index (NASA-TLX) and Positive and Negative Affect Schedule; and (e) participants were given a postexperiment questionnaire, which involved the System Usability Scale and AttrakDiff. The results showed that product expectations influenced usability ratings such as pragmatic quality ($\eta^2 = .376$) and hedonic identification ($\eta^2 = .277$). These results are in line with the phenomenon of self-fulfilling prophecy (participants with high expectations rated the product better), though Raita and Oulasvirta deemed this theory too broad to properly explain the results obtained. In contrast, expectation-confirmation theory (the negatively biased users should give very high ratings when their expectations are exceeded, and vice versa) was disconfirmed in this study.

Al Sokkar and Law (2013) modeled expectations and user experience in shopping on a furniture site (ikea.com). The models developed comprised measures of aesthetics, instrumental qualities, and experiential qualities of the site, as well as issues of trust, purchase intention, and overall satisfaction. Their modeling combined user experience work, the technology acceptance model, and expectation-confirmation theory. These approaches support hypothesizing about the different influences among the measures collected before interaction, during the interaction, and after interaction. The key results of structural modeling suggest that all measures collected before interaction impact the measures obtained during interaction. The impact of measures obtained before and during interaction on the phase after interaction was negligible; only intention to buy impacted overall satisfaction.

Earlier work that bridges research on expectations and user experience research has several limitations. First, it has studied only a small number of products in single experiments. This might cause issues with generalizability of the results, as the results of only one product cannot be generalized to all products in the same category. The drawback of doing user experience research with just one product instance has been carefully discussed with mitigation attempted through studying several products (Hassenzahl & Monk, 2010).

Second, experimental studies of expectations and their influence on user experience are rare; we are aware only of the studies of Hartmann et al. (2008) and Raita and Oulasvirta (2011) that employs framing and priming to affect expectations.

Third, across work in information systems and user experience, many studies manipulate only either expectations (e.g., Cardozo, 1965; Churchill & Surprenant, 1982; Levin & Gaeth, 1988; Olshavsky & Miller, 1972; Raita & Oulasvirta, 2011) or quality (Cohen & Goldberg, 1970; Oliver & Linda, 1981). This causes concern because only when controlling both variables can we study the levels of disconfirmation. This in turn is required to investigate contrast theory, which states that ratings

[1]When reporting effect sizes we use the following thresholds for η^2: $\eta^2 = .01$ – small effect, $\eta^2 = .06$ – medium effect, $\eta^2 = .14$ – large effect (Cohen, 1988, Chapter 8).

are a function of disconfirmation level. Alternatively, one could use the design suggested by Oliver (1977), where one variable is expectations and another is disconfirmation level.

The experiments presented next purport to explore addressing these three limitations.

3. EXPERIMENTAL SETUP

The intention of the two experiments was to investigate how expectations affect ratings of online games. However, expectations are formed in many ways, such as by advertisements or reviews, which then have an influence on the ratings of an interactive product. When researching expectations, this sometimes becomes a problem, as the participants often bring these expectations to the lab room as well (Raita & Oulasvirta, 2011). One solution is to

> try to influence expectations at the beginning of the study by giving users information about product that will "override" the influence of prior knowledge . . . [for example] by advertisements or other product information they would be naturally exposed to before trying and buying the product. (Raita & Oulasvirta, 2011, p. 370)

We set up two experiments where participants were presented with an online game, played it, and afterward rated it on different measures of user experience. In the first experiment we controlled expectation levels by letting participants receive information about the game prior to playing it. The goal of this experiment was to investigate how expectations affect various user experience measures in a setup where it is unlikely that participants would have prior knowledge about the game. In the second experiment the expectations were not controlled but inferred from the difference between preexposure and postexposure ratings and by asking participants. The intention of the second experiment was to investigate the role of expectations without using priming but relying on two different measures of participants' expectation disconfirmation.

To reduce the effect of prior knowledge, the participants played free online games that they were unlikely to be familiar with; hence participants were unlikely to have preexisting expectations or attitudes towards these games. As part of the experiment, participants were primed by product information (using game reviews, ratings, and rankings) that they would often be exposed to, or seek out, when finding a new game to play. This works as a way to control expectations (Bentley, 2000; Olshavsky & Miller, 1972; Raita & Oulasvirta, 2011). After being primed, the participants were asked to play certain well-rated or poorly rated free online games for at least 5 min and afterward to rate it on nine different measures and write a short game review. The control group was not given any information about the game before playing.

We used several techniques to ensure that participants were playing the game for at least 5 min, including measuring the time the participant spent on the website, calculating keystrokes (or mouse clicks) the participant made during the interaction, and asking the participant questions about the gameplay after playing the game.

The participants rated the games on a variety of user experience measures; Table 1 contains a summary. The ratings include a simple 10-point scale about the game; items to measure goodness and beauty (Hassenzahl, 2004); the reduced version of AttrakDiff2 (Hassenzahl & Monk, 2010), consisting of eight pairs of antonyms describing the experience with subscales of pragmatic and hedonic quality; and a scale measuring playability (the Game Play Questionnaire; Ryan, Rigby, & Przybylski, 2006). The Game Play Questionnaire contains several subscales constructed from various different questionnaires such as Player Experience of Need Satisfaction or Intrinsic Motivation Inventory. From this questionnaire, only in-game autonomy, in-game competence, and game enjoyment were used.

In addition, expectation disconfirmation was measured in a fashion similar to Oliver (1977). Participants were asked to indicate on a scale from −3 to 3 whether the game was worse than expected (negative values), as good as expected (zero), or better than expected (positive values). Participants who selected negative values were assigned to the negative disconfirmation group, participants who selected zero were assigned to the zero disconfirmation group, and participants who selected positive values were assigned to the positive disconfirmation group.

4. FIRST EXPERIMENT

The first experiment studied how expectations affect ratings of both well-rated and poor-rated games. We primed the participants with a short game description, then had them play an online game from the website armorgames.com and subsequently had them rate the game on the nine measures described in Table 1. The participants were recruited by the crowdsourcing website crowdflower.com and upon completion they were given a small monetary reward.

The participants were randomly assigned to a group based on prime (positive, negative, or control group) and game quality (well-rated or poorly rated games). For each game quality group, three games with comparable ratings were chosen to have a larger variety of products. This means that, in total, there were six games used in the study, for which two game descriptions were written (one positive and one negative description). The purpose of using several games was to avoid trying to reach general conclusions with just one instance of game; this problem has haunted other areas of user experience research (Hassenzahl & Monk, 2010).

4.1. Game Selection

To select games for the experiment we performed a prestudy with six well-rated games and six poorly rated games based on Armorgames rankings (see Table 2). Candidates for the well-rated games were in the 90th percentile, and the poorly rated games were in the 10th percentile of the ranking of all games on the website; the final games were chosen to be the highest (or lowest, respectively) ranked games that (a) did not require registration, (b) had a first "level" or subdivision that could be

Downloaded by [89.173.14.71] at 12:19 01 September 2015

TABLE 1
Overview of Measures

Measure	Questionnaire, Reference	Items	Anchors	Range
Game rating		1	worst/best	1 – 10
Goodness	(Hassenzahl, 2004)	1	bad/good	1 – 7
Beauty	(Hassenzahl, 2004)	1	ugly/beautiful	1 – 7
Pragmatic quality	Reduced AttrakDiff2 (Hassenzahl & Monk, 2010)	4	multiple	1 – 7
Hedonic quality	Reduced AttrakDiff2 (Hassenzahl & Monk, 2010)	4	multiple	1 – 7
In-game autonomy	PENS (Ryan et al., 2006)	3	not at all/very much	1 – 7
In-game competence	PENS (Ryan et al., 2006)	3	not at all/very much	1 – 7
Game enjoyment	IMI (Ryan. Mims, & Koestner, 1983)	4	not at all true/very true	1 – 7
Expectation disconfirmation	(Oliver, 1977)	1	worse/better than I thought	−3 – +3

Note. PENS = Player Experience of Need Satisfaction; IMI = Intrinsic Motivation Inventory.

TABLE 2
Overview of Games Used in the Prestudy, Their Genre, and Rating

Game Name	Genre	Armorgames Rating (Game Quality)
Crystal Crisis	platform	5.1 (bad)
I Am Godzilla	action	5.1 (bad)
Block World	maze	5.1 (bad)
Flagman	platform	5.1 (bad)
Super Pig	platform	5.1 (bad)
Machine Man	shooter	5.1 (bad)
Balloon Invasion	shooter	8.6 (good)
Tactical Assassin Substratum	shooter	8.6 (good)
Super Adventure Pals	platform	8.6 (good)
Strand	puzzle	8.6 (good)
Cyber Chaser	platform	8.6 (good)
Sieger	puzzle	8.6 (good)

Note. The games in bold were used in the main experiment.

Downloaded by [89.173.14.71] at 12:19 01 September 2015

completed in 5 min or less, and (c) were not sequels of other candidate games satisfying (a) and (b). Afterward, a prestudy was set up on Crowdflower where participants ($N = 50$) played the games for 5 min and then rated them on a 10-degree scale. This prestudy used a between-group design with two conditions (well-rated and poorly rated games) for one independent variable (game quality). From this prestudy a subset of three well-rated and three poorly rated games was chosen by sorting the six well-rated games and six poorly rated games by average rating from Crowdflower and choosing the three best-rated games and three worst-rated games.

An independent samples t-test was used to check the effect of game quality on game ratings for the six games. There was a significant difference in ratings between the two groups, $t(23) = 3.175$, $p < .005$ (two-tailed), with well-regarded games ($M = 6.85$, $SD = 1.864$) scoring higher than poorly regarded games ($M = 4.25$, $SD = 2.221$) and a large observed effect (Cohen's $d = 1.271$, $\eta^2 = .305$).

4.2. Priming of Expectations

After choosing the games, two primed descriptions (positive and negative) were written for each game. The game descriptions used for priming comprised a fake product description, a rating, and a ranking. As participants played only one game each, we wrote a generic game description that could be used for all games for each prime group. Reviews for both primes contained similar wording but with opposite polarity for adjectives. For instance, if the positive prime contained the word *awesome*, the negative prime contained the word *terrible*. To identify the polarity of English words, we used the list by Hu and Liu (2004) containing around 6,800 English words and their connotations. The sentences explaining the purpose of the game were taken from the Armorgames website.

An example of a game description can be seen in Figure 1, which shows both negatively and positively primed game descriptions for a game called Sieger that was also shown to participants. This example was inspired by the description of

Downloaded by [89.173.14.71] at 12:19 01 September 2015

Please, read an excerpt from a well- known gaming website about the game called Sieger.

Sieger - 1.2/10

The purpose of the game is to kill all the castle defenders and save the hostages by carefully choosing what supporting blocks of the castle to smash. The game contains simple graphics, is not very fun to play and has difficulties keeping players entertained.

Player reviews:
john001 (1/10):
TERRIBLE GAME!!!

tom973 (1/10):
Sieger is so boring:(

jane_66 (1/10):
The worst game I have ever played.

Position in ranking (Total: 255 games):

#	Game	Rating
252	Flagman	1.3
253	Machine Man	1.2
254	Strand	1.2
255	Sieger	1.2

Please, read an excerpt from a well- known gaming website about the game called Sieger.

Sieger - 9.8/10

The purpose of the game is to kill all the castle defenders and save the hostages by carefully choosing what supporting blocks of the castle to smash. The game contains advanced graphics, is fun to play and can keep players entertained for hours.

Player reviews:
john001 (10/10):
AWESOME GAME!!!

tom973 (10/10):
Sieger is so much fun:)

jane_66 (10/10):
The best game I have ever played.

Position in ranking (Total: 255 games):

#	Game	Rating
1	Sieger	9.8
2	Flagman	9.8
3	Machine Man	9.8
4	Strand	9.7

FIG. 1. Examples of primed reviews: a negative prime (top) and a positive prime (bottom).

the game on App Store and contains a short introduction, rating of the game, a short description of the gameplay, customer's ratings, and the ranking of the game. The game description for each game in our study contained appropriate description of its gameplay.

After writing these descriptions another prestudy was set up on Crowdflower where 122 participants read the game description and then rated the game on beauty and goodness using two single-item, 7-point Likert scales. The purpose of this prestudy was to verify that the primes work as intended.

Downloaded by [89.173.14.71] at 12:19 01 September 2015

Indeed, two independent samples *t*-tests showed a large effect of game descriptions for both beauty ($\eta^2 = .498$) and goodness ($\eta^2 = .750$).

4.3. Experiment Design and Procedure

The first experiment used a 3 × 2 between-group design where participants were randomly assigned to a group based on prime (positive, negative, control group) and game quality (bad, good) with three variants of good and bad games.

After being assigned to an experimental group the participants went through the following stages of the experiment:

1. Participants in the positive and negative conditions read a description of the game they were going to play. Participants in the control group did not read any game description.
2. Participants played the game for at least 5 min.
3. Participants filled out a questionnaire containing the measures shown in Table 1.
4. Participants filled out a questionnaire on demographics.

The participants were 183 Crowdflower users residing in the United States at the time of the experiment. Seven participants had prior experience with the game they were assigned to and were excluded from statistical analysis, reducing the number of participants to 176. Table 3 shows the distribution of participants over conditions.

For all metrics a 3 × 2 factorial analysis of variance was conducted to assess the impact of prime and game quality on each given metric. Afterward, a post hoc Tukey Honestly Significant Difference Test was conducted to compare the scores among prime groups.

4.4. Results

Table 4 shows an overview of the influence of priming and game quality on dependent measures. The effect of prime was significant for five measures—beauty, $F(2, 170) = 4.245$, $p < .05$, $\eta^2 = .040$; goodness, $F(2, 170) = 4.314$, $p < .05$, $\eta^2 = .021$; game rating, $F(2, 169) = 5.362$, $p < .01$, $\eta^2 = .052$; pragmatic quality, $F(2, 170) = 3.521$, $p < .05$, $\eta^2 = .038$; and hedonic quality, $F(2, 170) = 3.045$, $p < .05$, $\eta^2 = .032$. Priming did not affect the game-experience qualities: in-game competence, $F(2, 170) = 1.686$, *ns*; in-game autonomy, $F(2, 170) = 1.755$, *ns*; and game enjoyment, $F(2, 170) = .685$, *ns*. Thus, priming affects how users rate the game across a variety of measures.

TABLE 3
Number of Participants in Each Experimental Group for Experiment 1

	Positive Prime[a]	Negative Prime[b]	Control Group[c]
Well-rated game[d]	34	28	31
Poorly rated game[e]	27	30	26

[a] $n = 61$. [b] $n = 58$. [c] $n = 57$. [d] $n = 93$. [e] $n = 83$.

Table 4 also shows that game quality influences user experience measures. The effect of game quality was significant for all measures—beauty, $F(1, 170) = 32.878$, $p < .0005$, $\eta^2 = .154$; goodness, $F(1, 170) = 21.637$, $p < .0005$, $\eta^2 = .107$; game rating, $F(1, 169) = 21.653$, $p < .0005$, $\eta^2 = .106$; pragmatic quality, $F(1, 170) = 7.389$, $p < .01$, $\eta^2 = .040$; hedonic quality, $F(1, 170) = 10.731$, $p < .001$, $\eta^2 = .057$; in-game competence, $F(1, 170) = 8.356$, $p < .0005$, $\eta^2 = .095$; in-game autonomy, $F(1, 170) = 19.582$, $p < .0005$, $\eta^2 = .101$; and game enjoyment, $F(1, 170) = 12.655$, $p < .0005$, $\eta^2 = .069$. These results are not surprising but serve as a manipulation check to show that there were significant differences between the two game-quality conditions.

The interaction effect of prime and game quality was not significant for any measure, which is consistent with the findings of both Olshavsky and Miller (1972) and Raita and Oulasvirta (2011). This result suggests that the interaction between game quality and prime is not important in a user's perception of the product and that there are other factors that affect ratings.

Table 5 shows the influence of expectation disconfirmation on user experience measures. One-way analyses of variance were conducted with expectation disconfirmation as independent variable and the eight user experience measures as dependent variables. We find a significant difference between expectation disconfirmation groups for all measures (with a large effect, $\eta^2 > .14$) except for pragmatic quality. We find larger effect sizes for expectation disconfirmation than for game quality or prime. The scores for the positive disconfirmation group were the highest, whereas the scores for negative disconfirmation group were the lowest.

For the effect of expectation disconfirmation and prime on ratings, participants in the high-expectation, positive-disconfirmation group scored highest on all measures. When the disconfirmation level decreased (i.e., the game was worse than expected), most of the measures decreased significantly. Participants in the low-expectation, negative-disconfirmation group scored lowest on most measures, but the ratings dramatically improved when the disconfirmation was positive (i.e., when the game was better than expected). To give an example of how disconfirmation affects scores, consider in-game competence in Table 5, $F(2,173) = 29.390$, $p < .0005$, $\eta^2 = .254$. For positively disconfirmed participants, the average score was 5.07 ($SD = 1.285$, $N = 87$), for zero-disconfirmed participants it was 3.47 ($SD = 1.629$, $N = 38$), and for negatively disconfirmed participants it was 3.24 ($SD = 1.743$, $N = 51$).

Note that expectation disconfirmation is treated as an independent variable in the previous discussion; Table 6 shows the Pearson correlation coefficients for the relationship between expectation disconfirmation and each measure to show that there is indeed a strong relation among expectation

TABLE 4
Results of Experiment 1

Measure	Prime Group				Game Quality Group		
	Negative[a]	Control[b]	Positive[c]	η^2	Bad[d]	Good[e]	η^2
Game rating	4.6 (2.64)	5.6 (2.69)	6.2 (2.71)	.052	4.5 (2.69)	6.4 (2.55)	.106
Beauty	3.4 (1.77)	3.6 (1.49)	4.3 (1.70)	.040	3.1 (1.61)	4.4 (1.50)	**.154**
Goodness	3.7 (1.88)	4.5 (1.92)	4.7 (1.89)	.021	3.6 (1.50)	4.9 (1.76)	.107
Hedonic quality	3.4 (1.61)	3.7 (1.59)	4.2 (1.57)	.032	3.4 (1.68)	4.2 (1.45)	.057
Pragmatic quality	4.3 (1.21)	4.6 (1.16)	4.6 (1.21)	.038	4.4 (1.25)	4.9 (1.13)	.040
In-game competence	3.9 (1.71)	4.2 (1.70)	4.5 (1.76)	*ns*	3.6 (1.78)	4.7 (1.52)	.095
In-game autonomy	3.2 (1.62)	3.5 (1.78)	3.9 (1.83)	*ns*	3.0 (1.65)	4.1 (1.68)	.101
Game enjoyment	3.8 (1.67)	4.1 (1.89)	4.0 (1.81)	*ns*	3.5 (1.73)	4.5 (1.77)	.069

Note. Each row in the table represents a measure. For each prime group and game quality group a mean rating with a standard deviation (in parentheses) is shown. Effect sizes are measured with η^2 for primes and game quality. Boldface text represents large ($\eta^2 > .14$) observed effect. Note that game rating uses a scale from 1 to 10; all other measures use a scale from 1 to 7.

[a]$N = 58$. [b]$N = 57$. [c]$N = 61$. [d]$N = 83$. [e]$N = 93$.

TABLE 5
Results From Experiment 1 for Expectation Confirmation

Measure	Expectation Disconfirmation			
	Negative[a]	Zero[b]	Positive[c]	η^2
Game rating	2.9 (2.21)	4.9 (2.44)	7.2 (1.78)	**.454**
Beauty	2.6 (1.59)	3.4 (1.53)	4.7 (1.29)	**.289**
Goodness	2.5 (1.69)	3.9 (1.61)	5.5 (1.17)	**.460**
Hedonic quality	2.3 (1.36)	3.4 (1.27)	4.8 (1.02)	**.467**
Pragmatic quality	4.1 (1.42)	4.3 (1.21)	5.0 (.91)	.117
In-game competence	3.2 (1.74)	3.5 (1.63)	5.1 (1.29)	**.254**
In-game autonomy	2.1 (1.39)	3.0 (1.37)	4.6 (1.34)	**.406**
Game enjoyment	2.3 (1.36)	3.6 (1.47)	5.2 (1.21)	**.482**

Note. Each row in the table represents a measure. For each prime group a mean rating with a standard deviation (in parentheses) is shown. Effect sizes are measured with η^2. Boldface text represents large ($\eta^2 > .14$) observed effect. Note that game rating uses a scale from 1 to 10, all other measures uses a scale from 1 to 7.

[a]$N = 51$. [b]$N = 38$. [c]$N = 87$.

TABLE 6
Correlations Between Expectation Disconfirmation and Measures for Experiment 1

	Beauty	Goodness	Game Rating	Pragmatic Quality	Hedonic Quality	Autonomy	Competence	Game Enjoyment
r	.534*	.678*	.673*	.683*	.332*	.632*	.477*	.693*

*$p = .01$.

disconfirmation and user experience measures even when analyzed in this way.

For a different view on disconfirmation, Figure 2 shows normalized average ratings per disconfirmation group for all eight measures—beauty, goodness, overall rating, pragmatic quality, hedonic quality, in-game autonomy, in-game competence, and game enjoyment. The positive disconfirmation group is colored in blue, the zero-disconfirmation group is in green, and the negative disconfirmation group is in red. The averages of the positively disconfirmed groups appeared farther from zero-disconfirmation than the negatively disconfirmed groups.

Downloaded by [89.173.14.71] at 12:19 01 September 2015

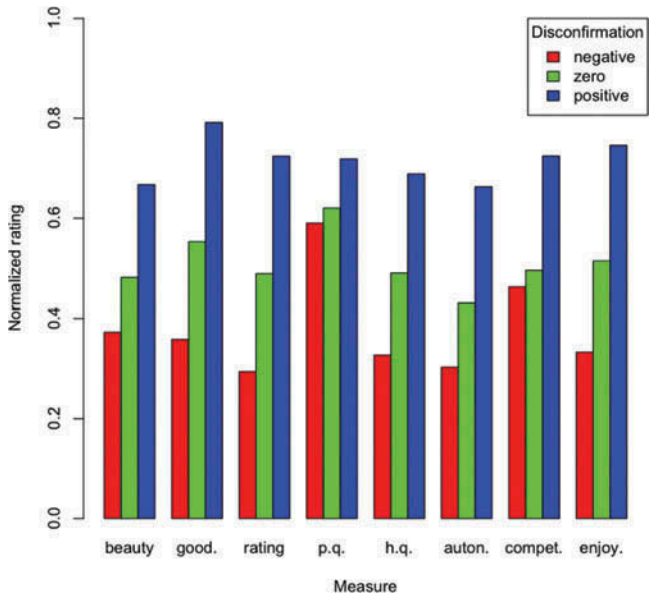

FIG. 2. Average ratings per disconfirmation group in Experiment 1.

To further investigate this pattern, we compared the average ratings for both expectation disconfirmation groups relative to the zero-disconfirmation group. Figure 3 shows the average change in ratings (in %) for both expectation disconfirmation groups in comparison to the zero-disconfirmation group. The pattern of disconfirmation is complex, but positive disconfirmation seems to be stronger than negative disconfirmation, with positive disconfirmation increasing the ratings by 42.47% across all measures (with respect to the zero disconfirmation group) and negative disconfirmation on average decreasing the ratings by only 25.24%. In addition, positive disconfirmation with a negative prime resulted in the largest change in ratings for most measures (average increase was 62% across all measures with respect to zero disconfirmation). This pattern can be illustrated with the ratings of in-game autonomy, where positive disconfirmation increased the ratings by on average 54%, whereas negative disconfirmation decreased the ratings by on average 28.67%. Also, positive disconfirmation with a negative prime was associated with a 67% increase in rating of in-game autonomy with respect to zero disconfirmation.

The results of this experiment suggest that both assimilation theory and contrast theory apply to a certain extent when studying the effect of expectations on user experience. It seems, however, that the most important factor that affects ratings of online games is not expectations per se, but rather whether, and to what degree, these expectations are confirmed.

Because expectation disconfirmation seems to be the most important factor that affects ratings of online games, we set up a second experiment that focused on expectation disconfirmation measured in two different ways—by comparing preexposure and postexposure ratings and by asking to what extent the expectations were confirmed (as in the first experiment).

5. SECOND EXPERIMENT

The second experiment studied expectation disconfirmation without the use of priming. Not only did we ask about disconfirmation at the end of the experiment (as in the first experiment), but we assessed expectations as differences between preexperiment and postexperiment questionnaires (similar to Oliver, 1977). We call the difference between preexperiment and postexperiment game rating *rating change* to avoid confusion with expectation disconfirmation.

5.1. Design

This experiment implemented a between-group design with game quality as an independent variable and user experience ratings as dependent variables. In addition, when analyzing results, preexposure ratings were used as an independent variable representing the level of expectations.

We collected two measures of expectation disconfirmation: (a) the change of overall rating between preexposure rating and postexposure rating, and (b) the answer to the question about expectation disconfirmation used in the first experiment. In an ideal scenario these ratings should correlate strongly as both of them measure the same phenomenon—expectation disconfirmation.

In addition, we split participants ($N = 77$) into three *expectation groups*—high expectation, medium expectation, and low expectation—based on a comparison of the participants' individual preexposure ratings with the median preexposure score. The participants who gave the game a higher score than median were assigned to a high-expectation group, those who gave a lower score than median were assigned to a low-expectation group, and those who gave exactly the median score were assigned to a medium-expectation group. The average preexposure rating was $M = 6.21$, $Mdn = 6.00$, $SD = 1.780$. After the split, the low-expectation group consisted of 29 participants, the medium-expectation group of 12 participants, and the high-expectation group of 36 participants.

For simplicity, we used only one poorly rated and one well-rated game, in contrast to the first experiment, in which we used three games of similar quality. The choice of the games was made from the set of games that was used in the main experiment: From the well-rated games we used the game that scored highest on most measures (*Cyber Chaser*), and from the poorly rated games we chose the game that scored the lowest on most measures (*I Am Godzilla*).

Because we did not use primes in this experiment, the participants were shown neutral game descriptions from Armorgames, which were three sentences long and categorized the game into three genres.

Thus, in this experiment the participants (a) at first read a short, unbiased game description; (b) rated how they expect the

Downloaded by [89.173.14.71] at 12:19 01 September 2015

Downloaded by [89.173.14.71] at 12:19 01 September 2015

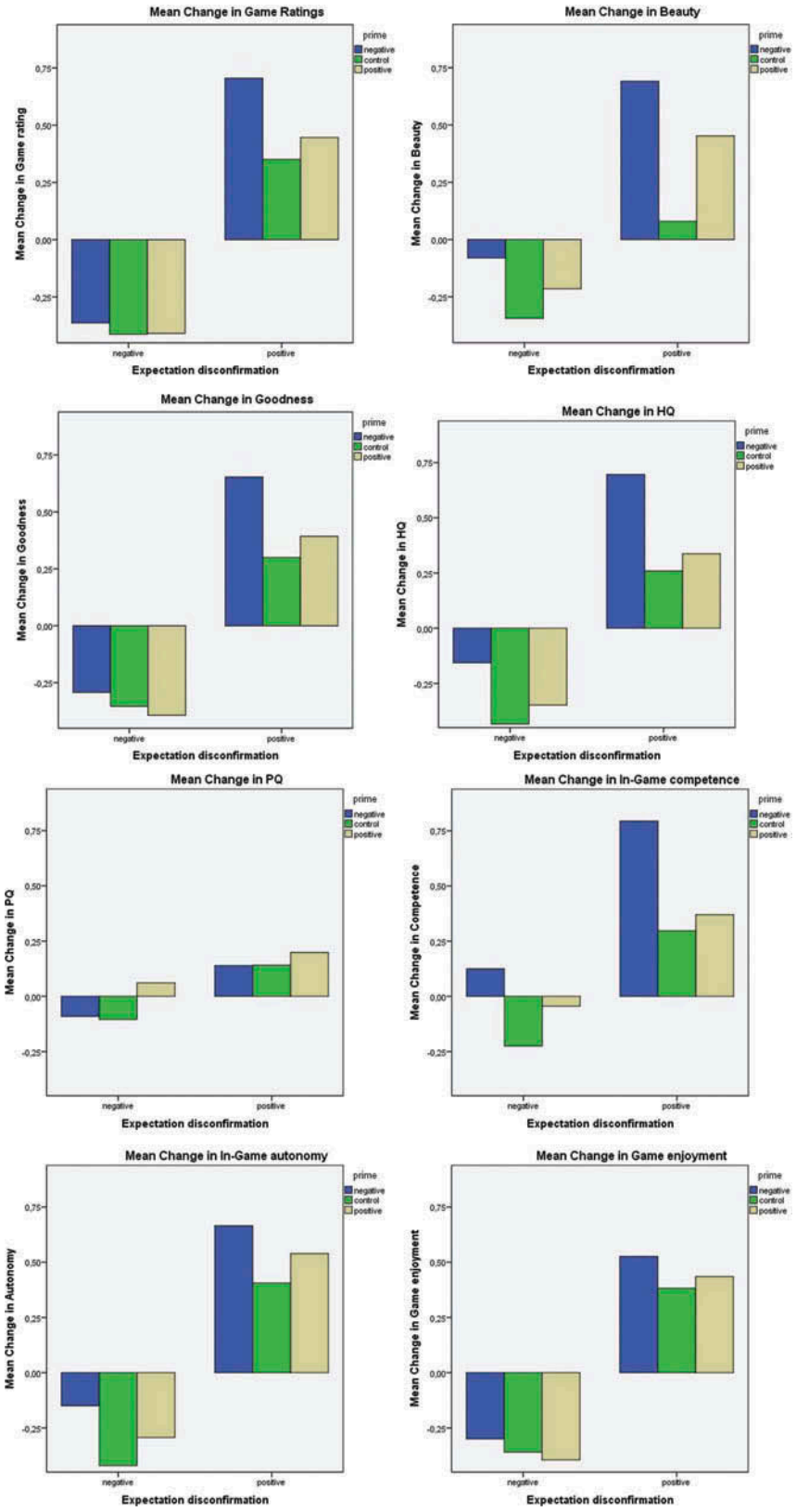

FIG. 3. Effect of expectation disconfirmation on dependent measures on measures of beauty, goodness, game rating, pragmatic quality, hedonic quality, in-game autonomy, in-game competence, and game enjoyment.

Downloaded by [89.173.14.71] at 12:19 01 September 2015

game to be (similar to the game rating question in Table 1 but worded using "expect"); (c) played the game for at least five minutes; and (d) rated the game on the same measures as in the first experiment (see Table 1).

5.2. Results

Table 7 shows the results for this experiment. The expectation group represents whether the participant was in high-expectation, medium-expectation, or low-expectation group. The results show that the naturally occurring expectations affected all measures of user experience, except pragmatic quality, with large effects (η^2 values over .148). Of interest, these effects were larger than the corresponding effects of priming used in the first experiment.

Table 7 also summarizes data on rating change, that is, the change in ratings between preexposure and postexposure ratings (i.e., the alternative way of calculating expectation disconfirmation). Rating change had a strong effect on all measures— on overall ratings, $F(2, 74) = 53.873$, $p < .0005$, $\eta^2 = .593$; beauty, $F(2, 74) = 6.134$, $p < .005$, $\eta^2 = .142$; goodness, $F(2, 74) = 27.149$, $p < .0005$, $\eta^2 = .423$; hedonic quality, $F(2, 74) = 23.581$, $p < .0005$, $\eta^2 = .389$; pragmatic quality, $F(2, 74) = 10.642$, $p < .0005$, $\eta^2 = .223$; in-game autonomy, $F(2, 74) = 29.504$, $p < .0005$, $\eta^2 = .444$; in-game competence, $F(2, 74) = 22.401$, $p < .0005$, $\eta^2 = .377$; and game enjoyment, $F(2, 74) = 34.042$, $p < .0005$, $\eta^2 = .479$. Thus, rating change (similar to expectation disconfirmation in the first experiment) is associated with a large effect on user experience measures.

The relationship between the expectation disconfirmation (as in the first experiment) and rating change was investigated using the Pearson correlation coefficient. There was a strong positive correlation between the two variables, $r = .771$, $N = 76$[2], $p < .0005$, $r^2 = .594$. This suggests that one can ask the participants either to what degree their expectations were confirmed or to rate the product both before and after interaction and infer the disconfirmation levels from the difference between these ratings.

The scatter plot in Figure 4 unpacks rating change by showing the preexposure ratings on the *x*-axis and postexposure ratings on the *y*-axis. The color represents game quality, and the shade of color represents number of participants with same ratings (darker shade = more participants). The dashed line represents expectation confirmation, the upper half-plane represents positive disconfirmation, and the lower half-plane represents negative disconfirmation. Figure 4 shows that most of the participants who played a poorly rated game had negatively disconfirmed expectations and that most of the participants who played a well-rated game had positively disconfirmed expectations, the same result as in the first experiment.

Table 8 shows the results from Experiment 2 for game rating. As anticipated and similar to the first experiment, game quality had large effects on the user experience measures (except pragmatic quality). This served as a manipulation check of the manipulation of game quality.

6. OVERALL DISCUSSION

Our results showed that expectations significantly influence ratings of user experience; this held across different games and across experiments. In particular, this effect depended on whether expectations were positively or negatively disconfirmed. The experiments also allowed us to compare using priming to study expectations against using the more naturally occurring expectations as in Experiment 2. Finally, we claim that our study has implications for future research on expectations and, more generally, on user experience. Next we discuss these findings in turn.

[2]One participant did not provide the rating change, but did provide all the other data.

TABLE 7
Results of Experiment 2

	Expectation Group				Rating-Change Group			
Measure	Low[a]	Medium[b]	High[c]	η^2	Negative[d]	Zero[e]	Positive[f]	η^2
Game rating	4.1 (2.3)	6.0 (2.09)	6.7 (2.48)	**.203**	3.4 (1.84)	6.8 (1.40)	7.6 (1.62)	**.593**
Beauty	3.2 (1.51)	4.0 (1.28)	5.0 (1.32)	**.248**	3.5 (1.67)	4.3 (1.49)	4.8 (1.27)	**.142**
Goodness	3.3 (1.77)	4.8 (1.29)	5.3 (1.77)	**.226**	3.1 (1.76)	5.3 (1.27)	5.7 (1.21)	**.423**
Hedonic quality	3.2 (1.40)	4.0 (1.66)	4.7 (1.42)	**.187**	2.9 (1.37)	4.9 (1.02)	4.9 (1.22)	**.223**
Pragmatic quality	4.2 (.92)	4.9 (.46)	4.8 (1.29)	*n.s.*	4.0 (1.08)	5.4 (.89)	4.8 (.89)	**.389**
In-game competence	2.9 (1.54)	4.2 (1.37)	4.4 (1.88)	**.148**	2.6 (1.49)	5.2 (1.20)	4.6 (1.48)	**.444**
In-game autonomy	2.6 (1.51)	3.9 (1.66)	4.3 (1.62)	**.185**	2.3 (1.31)	4.8 (1.18)	4.6 (1.47)	**.377**
Game enjoyment	3.3 (1.48)	4.3 (2.05)	4.8 (1.44)	**.156**	2.8 (1.17)	5.2 (1.07)	5.2 (1.34)	**.479**

Note. Each row in the table represents a measure. For each group a mean rating with a standard deviation (in parentheses) is shown. Effect sizes are measured with η^2 for expectations and rating change. Boldface text represents large ($\eta^2 > .14$) observed effects.

[a]$N = 19$. [b]$N = 12$. [c]$N = 36$. [d]$N = 34$. [e]$N = 11$. [f]$N = 32$.

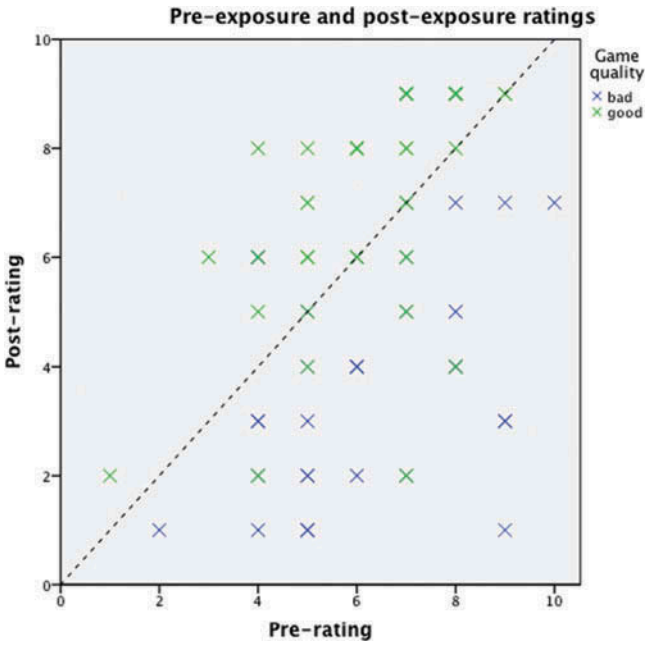

FIG. 4. Preexposure and postexposure ratings.

TABLE 8
Results From Experiment 2 for Game Quality

	Game Quality Group		
Measure	Bad[a]	Good[b]	η^2
Game rating	4.4 (2.62)	6.8 (2.04)	**.203**
Beauty	3.5 (1.78)	4.7 (1.10)	**.151**
Goodness	3.7 (2.09)	5.1 (1.42)	**.145**
Hedonic quality	3.4 (1.69)	4.6 (1.24)	**.184**
Pragmatic quality	4.4 (1.29)	4.7 (.86)	*ns*
In-game competence	3.3 (1.94)	4.4 (1.52)	.077
In-game autonomy	3.0 (1.75)	4.2 (1.51)	**.160**
Game enjoyment	3.4 (1.67)	4.9 (1.33)	**.232**

Note. Each row in the table represents a measure. For each prime group a mean rating with a standard deviation (in parentheses) is shown. Effect sizes are measured with η^2. Boldface text represents large ($\eta^2 > .14$) observed effect. Please note that game rating was rated on a scale from 1 to 10, all other measures were rated on a scale from 1 to 7.

[a]$N = 38$. [b]$N = 39$.

6.1. Priming Expectations and User Experience

The experiments showed that expectation affected ratings of user experience. This was shown mainly through the use of priming. In the first experiment, the effect of prime was significant ($\eta 2 \in [.021, .052]$) for five measures related to user experience—game rating, beauty, goodness, pragmatic quality, and hedonic quality. In the second experiment, the effect of expectations was significant and large for all measures except for pragmatic quality ($\eta 2 \in [.148, .248]$).

However, in the first experiment priming did not affect the gameplay-related measures, such as in-game autonomy, in-game competence, and game enjoyment. The lack of effect on gameplay-related measures is surprising because we find an overall effect of prime on game rating. One explanation could be that these measures are based on more specific questions and not subject to a general halo-effect of prime. In the second experiment, gameplay-related measures showed a large effect ($\eta 2 \in [.148, .185]$). This discrepancy might have been caused by how expectations were measured in these two experiments: In the first experiment the primed reviews were not designed to affect these qualities, but in the second experiment the participants were split into expectation groups based on their preexposure ratings, which may in some way have affected the ratings of these qualities.

It is difficult to compare the strong influence of expectations to studies in human–computer interaction, as no prior study rated products on the same measures as this study. However, one can compare the effect of prime on AttrakDiff ratings with Raita and Oulasvirta's (2011) study, where priming had a significant effect on both pragmatic and hedonic quality. This is consistent with our findings. However, the effect sizes in our study were much smaller (both effects were "small" – $.01 < \eta^2 < .06$) than those found by Raita and Oulasvirta (both effects were "large" – $\eta^2 > .14$).

Even though expectations had large effects, the effect of game quality was significant for all measures and was large on the ratings of beauty, small on AttrakDiff scores (hedonic quality and pragmatic quality), and medium on all other measures. Thus, the effect of game quality was stronger than the effect of expectations, which is consistent with the study of Olshavsky and Miller (1972), in which the authors studied the effect of expectations and product performance of tape recorders and found a large effect of performance ($\eta^2 = .30$) and a medium effect of expectations ($\eta^2 = .09$). These results indicate that game quality plays a more important role in a user's perception of the product than what other people say about the product. However, the effect of game quality on ratings was not strong enough to explain all differences between experimental groups. We investigate these differences by discussing further factors next.

Pragmatic quality (PQ) from the AttrakDiff questionnaire behaved differently from the other measures. First, game quality ($\eta^2 = .040$) and prime ($\eta^2 = .038$) had a similar effect on PQ ratings with the lowest difference in effect size among all measures. Second, the effect size of game quality on PQ was the lowest among all measures in the first experiment and not significant in the second experiment. Third, the effect of expectation disconfirmation on pragmatic quality in the first experiment was only medium ($\eta^2 = .117$), lowest among all measures. These results suggest that pragmatic qualities, such as whether a game is structured, practical, predictable, or simple, are affected neither by game quality nor by prime but rather by the game itself, as the effect of the game itself was the strongest among all

Downloaded by [89.173.14.71] at 12:19 01 September 2015

Downloaded by [89.173.14.71] at 12:19 01 September 2015

factors ($\eta^2 = .148$, "large," for both poorly rated and well-rated games). This might be due to the fact that our primes were not designed and tested for each measure individually or simply because showing primed reviews is a "weak" type of prime for pragmatic quality. However, Raita and Oulasvirta (2011) showed that using primed text did affect PQ to a large degree.

Obtaining information about expectations without affecting them is quite challenging. It is thus worth noting that there were only minor differences in the results we obtained when using constructed primes as in Experiment 1 and the way of obtaining information about expectations used in Experiment 2, which did not employ a constructed prime but rather a short neutral description from the Armorgames website and participants' own anticipations. In short, both experiments showed that participants who played poorly rated games were negatively disconfirmed and most of the participants who played a well-rated game were positively disconfirmed.

One aspect not covered by the experiment is the *strength* of expectations. For example in Experiment 2, participants not exposed to neutral reviews (and hence without expectations) were expected to give a good game good reviews and a bad game bad ones. Thus, although the results show that expectations were disconfirmed, it is not clear whether disconfirmation of expectations held with conviction (e.g., high expectations of the latest in a series of high-quality games) would affect game ratings more strongly than the more artificial ones we induced.

6.2. Disconfirmation of Expectations and User Experience

Another key finding with respect to expectations is that they affected ratings differently depending on whether they were confirmed. The effect of expectation disconfirmation was significant for all measures with large effect sizes ($\eta^2 \in [.117, .482]$), except for pragmatic quality in the second experiment. These effect sizes were much larger than those for game quality or prime. The scores for the positive disconfirmation group were the highest, whereas the scores for negative prime group were the lowest (see Figure 2). The effect of both expectation disconfirmation and prime on ratings (see Figure 3) showed that the participants in high-expectation, positive-disconfirmation group scored highest on all measures, and when the disconfirmation level decreased (the game was worse than expected), most of the measures significantly dropped. In contrast, participants in low-expectation, negative-disconfirmation group scored the lowest on most measures, but the ratings dramatically improved when the disconfirmation was high (when the game was better than expected). Thus, the experiments suggest that when expectations are confirmed, users tend to assimilate their ratings with their expectations; conversely, if the product quality is inconsistent with expectations, users tend to contrast their ratings with expectations and give ratings correlated with the disconfirmation level.

These results are consistent with Oliver (1977), who showed that negative disconfirmation resulted in a negative rating, whereas positive disconfirmation led to a very positive rating. Oliver (1977) investigated the effect of expectations and expectation disconfirmation on car ratings and found that the effect of expectation disconfirmation was significant ($p < .01$) and stronger than the effect of expectations and that scores were higher for high expectation group at all disconfirmation levels. In our study on games, the results were similar, with expectation disconfirmation having larger effect on ratings than prime or game quality. Moreover, these results were consistent with the works of Olshavsky and Miller (1972), who—even though they did not directly measure expectation disconfirmation—found that if performance was very different from expectations, users tended to contrast their evaluation with their original expectations; similarly, if the product performance was close to expected, users tended to assimilate their evaluation toward expectations. In other words, when the expectations were disconfirmed, contrast theory appears to apply—ratings were a function of disconfirmation levels—and when the expectations were confirmed, assimilation theory seems to apply—ratings were a function of expectations. These results are of great interest, as it seems that expectation disconfirmation plays an important role in how users rate products. They suggest that expectation disconfirmation is the strongest factor in a user's perception of the product and, thus, if a product underperforms, the ratings will suffer more than if users rated the product with no prior expectations. Conversely, when a product overperforms, the ratings are higher.

Our results differ from those of Raita and Oulasvirta (2011). There, the data did not corroborate the expectation-confirmation theory, which predicts that high expectations and low performance of a product (negative disconfirmation) should lead to a very negative rating, whereas low expectations and high performance (positive disconfirmation) should lead to a very positive rating. In contrast, this theory is supported by our study, as expectation disconfirmation had the highest effect on ratings. Our results also contrast those of Venkatesh and Goyal (2010), who found that disconfirmation in general was bad. In contrast, positive disconfirmation seems to boost the ratings on most dependent measures.

6.3. Limitations and Future Work

We now briefly discuss limitations of the present work and some avenues for future work. First, expectation disconfirmation can be measured in multiple ways, two of which were tested in our experiments: (a) directly asking the participants whether their experience was better or worse than they expected, and (b) inferring disconfirmation from the difference between preexposure and postexposure ratings. We have used these two methods in experiments where participants had to state their expectations immediately prior to playing a game; this introduces a possible bias as disconfirmation levels may be higher than they would be "in the wild," where users rarely state their expectations in writing. Another way to study expectation

disconfirmation would be to split the experiment into two stages that occur in different times—in the first stage the participants would be presented a game and rate how they expect the game to be; in the second stage a couple of weeks later, they would play the game and rate it again. This setting would ensure that participants are not forced to explicitly state their expectations immediately prior to playing the game, hence eliminating the potential for the aforementioned bias. This setting would be similar to the study of expectation disconfirmation for an intranet-based knowledge-sharing system by Brown et al. (2012), where the second stage occurred 6 months after the first.

Second, both measuring expectations and selecting a design for experimental work are difficult. When studying expectations and expectation disconfirmation, we recommend the study design from Oliver (1977), which was also used in our second experiment—a 2 × 3 design with two expectation levels (low, high) and three disconfirmation levels (negative, zero, positive disconfirmation). Optionally, one can also use a control group as an additional expectation group.

Third, it may be of interest to try the experimental setting in our follow-up study for other ratings than usability ratings and for other products than online games. However, as all of our measures showed large effect of expectation disconfirmation (even the general "overall rating"), we expect expectation disconfirmation to have a significant and large effect on any type of rating, as evidenced by the work of Oliver (1977), who studied expectation disconfirmation on the ratings of cars and obtained comparable results to our study.

Fourth, we used games as the object of study. Although this allowed for questionnaires specialized on gaming, games comprise a particular class of interactive products that are more oriented toward leisure and hedonic quality than many other products. It may be that products oriented toward other qualities (e.g., effectiveness or efficiency for work-oriented products) would show markedly different effect sizes.

Fifth, the analysis in the present article is relatively unsophisticated compared to the modeling done by Al Sokkar and Law (2013), as well as that performed in research drawing on the technology acceptance model (e.g., Venkatesh & Goyal, 2010). We have relied more on experimental manipulations and related approaches to analysis. Future work could use more sophisticated techniques for analysis, including structural equation modeling.

6.4. Some Suggestions for User Experience Work

The results of the first experiment suggest that *both* assimilation theory and contrast theory apply to some extent when studying expectations and user experience, and thus that the effect of expectations on user experience is relatively complex. It seems that the most important factor that affects user experience ratings of online games is not expectations, per se, but rather whether, and to what degree, these expectations are confirmed or disconfirmed. When expectations are confirmed, users tend to assimilate their ratings toward their expectations; if the product quality is inconsistent with expectations, users tend to contrast their ratings with expectations and rate based on disconfirmation level. This shows that when studying expectations, one should always test for expectation *dis*confirmation—either by asking participants directly or by inferring disconfirmation levels from participants' responses. Thus, researchers and practitioners could use expectation disconfirmation as an additional measure when performing statistical analysis even when the study does not specifically concern expectation disconfirmation: Disregarding this factor may show an incomplete picture of how expectations work and affect participants' user experience. In addition, as suggested by Oliver (1977), expectation disconfirmation can be used as a predictor of postexposure ratings.

7. CONCLUSION

We have investigated the effect of expectations on user experience using several extant theories of expectation and employing simple computer games to perform experiments elucidating the effect of expectation. In contrast to related studies, we have varied both expectations (using priming) and game quality. Our results show that although priming and game quality had a significant effect on most of the user experience measures we considered, this effect was smaller than the effect of expectation disconfirmation. Thus, when a game's quality is inconsistent with expectations, users tend to contrast their ratings with expectations and rate the game based on disconfirmation level, in accordance with contrast theory. However, the distribution of expectation disconfirmation among experimental groups suggests that the relationship between expectations and subsequent ratings is complex: There were some users with high expectations who played a poorly rated game and gave the game high ratings, in accordance with assimilation theory but not with contrast theory.

As expectation disconfirmation has a larger effect on subsequent user experience ratings than expectations or game quality alone, researchers studying the relationship between expectations and user experience ratings should design their experiments in such a way that they also measure expectation disconfirmation, either by directly asking the participants or by inferring disconfirmation levels from the difference in preexposure and postexposure ratings. Despite this article being focused on usability ratings in online games, we believe that expectation disconfirmation has a strong effect on ratings of other products, as our results are compatible with the results of a similar study in a different domain.

In future studies on expectations, we advocate that researchers measure expectation disconfirmation due to its strong effect on user experience ratings. Thus, expectation disconfirmation can also be used as a predictor of postexposure ratings.

Downloaded by [89.173.14.71] at 12:19 01 September 2015

ORCID

Jakob Grue Simonsen http://orcid.org/0000-0002-3488-9392

REFERENCES

Al Sokkar, A., & Law, E. (2013). Validating an episodic UX model on online shopping decision making: A survey study with B2C e-commerce. In *Proceedings of the 5th ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (EICS '13; pp. 297–306). New York, NY: ACM.

Anderson, R. E. (1973). Consumer dissatisfaction: The effect of disconfirmed expectancy on perceived product performance. *Journal of Marketing Research*, *10*, 38–44.

Anderson, R. E., & Hair, J. F. (1972). Consumerism, consumer expectations, and perceived product performance. In *SV—Proceedings of the Third Annual Conference of the Association for Consumer Research* (pp. 67–79). Duluth, MN: Association for Consumer Research.

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

Bentley, T. (2000). Biasing web site user evaluation: A study. In *Proceedings of Australian Conference on Human–Computer Interaction* (OZCHI '00) (pp. 130–143). Baulkham Hills, New South Wales, Australia: HFESA.

Brown, S. A., Venkatesh, V., & Goyal, S. (2012). Expectation confirmation in technology use. *Information Systems Research*, *23*, 474–487.

Brown, S. A., Venkatesh, V., & Goyal, S. (2014). Expectation confirmation in information systems research: A test of six competing models. *MIS Quarterly*, *38*, 729–756.

Brown, S. A., Venkatesh, V., Kuruzovich, J., & Massey, A. P. (2008). Expectation confirmation: An examination of three competing models. *Organizational Behavior and Human Decision Processes*, *105*, 52–66.

Cardozo, R. N. (1965). An experimental study of customer effort, expectation, and satisfaction. *Journal of Marketing Research*, *2*, 244–249.

Churchill, G. A., & Surprenant, C. (1982). An investigation into the determinants of customer satisfaction. *Journal of Marketing Research*, *19*, 491–504.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. B., & Goldberg, M. E. (1970). The dissonance model in post-decision product evaluation. *Journal of Marketing Research*, *7*, 315–321.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*, 319–340.

Desurvire, H., Caplan, M., & Toth, J. A. (2004). Using heuristics to evaluate the playability of games. In *Extended Abstracts on Human Factors in Computing Systems (CHI '04)* (pp. 1509–1512). New York, NY: ACM.

Hartmann, J., De Angeli, A., & Sutcliffe, A. (2008). Framing the user experience: Information biases on website quality judgement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08) (pp. 855–864). New York, NY: ACM.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human–Computer Interaction*, *19*, 319–349.

Hassenzahl, M., & Monk, A. (2010). The inference of perceived usability from beauty. *Human–Computer Interaction*, *25*, 235–260.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '04) (pp. 168–177). New York, NY: ACM.

Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J.-B. (2009). User experience over time: An initial framework. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09) (pp. 729–738). New York, NY: ACM.

Kirakowski, J., & Corbett, M. (1993). SUMI: The Software Usability Measurement Inventory. *British Journal of Educational Technology*, *24*, 210–212.

Lee, M. C. (2010). Explaining and predicting users' continuance intention toward e-learning: An extension of the expectation–confirmation model. *Computers & Education*, *54*, 506–516.

Levin, I. P., & Gaeth, G. J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. *Journal of Consumer Research*, *15*, 374–378.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, *76*, 149–188.

Oliver, R. L. (1977). Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of Applied Psychology*, *62*, 480–486.

Oliver, R. L., & Linda, G. (1981). Effect of satisfaction and its antecedents on consumer preference and intention. *Advances in Consumer Research*, *8*, 88–93.

Olshavsky, R. W., & Miller, J. A. (1972). Consumer expectations, product performance, and perceived product quality. *Journal of Marketing Research*, *9*, 19–21.

Olson, J. C., & Dover, P. (1976). Effects of expectation creation and disconfirmation on belief elements of cognitive structure. *Advances in Consumer Research*, *3*, 168–175.

Raita, E., & Oulasvirta, A. (2011). Too good to be bad: Favorable product expectations boost subjective usability ratings. *Interacting with Computers*, *23*, 363–371.

Ryan, R.M., Mims, V., & Koestner, R. (1983). Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of Personality and Social Psychology*, *45*, 736–750.

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, *30*, 344–360.

Venkatesh, V., & Goyal, S. (2010). Expectation disconfirmation and technology adoption: Polynomial modeling and response surface analysis. *MIS Quarterly*, *34*, 281–303.

ABOUT THE AUTHORS

Jaroslav Michalco is an IT Analyst at Gratex International (since 2014). He received his Bachelor degree from the Slovak University of Technology in Bratislava in 2012 and MSc degree from the University of Copenhagen in 2014. His research interests include usability, user experience, and social networks.

Jakob Grue Simonsen is associate professor at the Department of Computer Science, University of Copenhagen. He received his Ph.D. and Habilitation degrees from the University of Copenhagen in 2005 and 2012, respectively, and his MBA degree from Edinburgh Business School in 2008. His research interests include mathematical logic, computability and complexity theory, usability, and information retrieval.

Kasper Hornbæk is professor at the Department of Computer Science, University of Copenhagen. He received his Ph.D. from University of Copenhagen in 2002. His research interests concern human–computer interaction, including user experience research, interaction techniques for large-displays, and using embodied cognition to drive midair interaction.

Downloaded by [89.173.14.71] at 12:19 01 September 2015

View publication stats