Peak-End Effects on Player Experience in Casual Games

Carl Gutwin[1], Christianne Rooke[1], Andy Cockburn[2], Regan L. Mandryk[1], Benjamin Lafreniere[3]
[1]University of Saskatchewan, Saskatoon, Canada
[2]University of Canterbury, Christchurch, New Zealand
[3]Autodesk Research, Toronto, Canada
gutwin@cs.usask.ca, christianne.r@usask.ca, andy@cosc.canterbury.ac.nz, regan@cs.usask.ca, ben.lafreniere@gmail.com

ABSTRACT
The peak-end rule is a psychological heuristic observing that people's retrospective assessment of an experience is strongly influenced by the intensity of the peak and final moments of that experience. We examine how aspects of game player experience are influenced by peak-end manipulations to the sequence of events in games that are otherwise objectively identical. A first experiment examines players' retrospective assessments of two games (a pattern matching game based on Bejeweled and a point-and-click reaction game) when the sequence of difficulty is manipulated to induce positive, negative and neutral peak-end effects. A second experiment examines assessments of a shootout game in which the balance between challenge and skill is similarly manipulated. Results across the games show that recollection of *challenge* was strongly influenced by peak-end effects; however, results for *fun, enjoyment,* and *preference to repeat* were varied – sometimes significantly in favour of the hypothesized effects, sometimes insignificant, but never against the hypothesis.

Author Keywords
Peak-end theory; player experience; game design.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI).

INTRODUCTION
Peak-end theory suggests that people's retrospective evaluation of an experience is strongly influenced by the *peak* and *end* moments of that experience. That is, rather than objectively considering the experience as a whole, people tend to base their assessments on the most intense and final moments of the experience. For example, Kahneman and colleagues have shown that when people undergo a painful experience (such as colonoscopy), their memory of the overall experience is substantially coloured by the final moments of the procedure [21, 28, 29]. This overweighting of the final moments of experience can lead to surprising outcomes, such as reliable preferences for longer experiences that contain more total pain (but finish with lower pain levels) over shorter experiences with less total pain that finish abruptly.

HCI researchers have recently shown that peak-end theory also affects people's evaluations of their experiences with interactive systems. Cockburn and colleagues [6] manipulated *effort* in an interactive task by changing the number of sliders that had to be set on each of a sequence of pages. Their study showed that people significantly preferred sequences that had a less effortful peak (only a few sliders on a middle page) and a less effortful end (few sliders on the last page) compared to other sequences with an identical number of sliders but in a different arrangement. In other work, Hassenzahl and Sandweg [18] showed that the terminating level of mental effort during an interaction negatively correlated with perceived usability, and Harrison et al. [16, 17] showed that variations in the pace for visual feedback influenced people's preferences for progress bars.

Peak-end theory suggests that there are other kinds of experiences in human-computer systems that could be affected by the ordering and sequencing of events. In particular, computer games are an area where different qualities of the experience – such as difficulty or balance, and their variation over time – could have strong effects on the player's subjective assessments of the game's challenge, their own performance, or their preferences. Games provide an interesting platform for examining peak-end effects because the activities involved are designed to be highly engaging and immersive (e.g., fighting a series of enemies); games may also generate stronger momentary experiences than other forms of daily interactive tasks, allowing more sensitive experimental examination of peak-end effects. In addition, game designers are already interested in understanding how game event sequencing affects experiences – they consider order when positioning cut-scenes and pacing when designing levels [27], and they explicitly evaluate play experience [3, 30, 34]. Finally, replay and retention rate is a major concern for game designers, and a player's willingness to replay a game is strongly connected to their memory of the experience.

There is little information, however, about whether or how peak-end theory affects game experience, or what variables in games lead to effects. Early peak-end studies manipulated pain [2, 21, 28, 29]; others have examined pleasurable experiences [9, 11]; and HCI studies have shown effects

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858419

when manipulating effort that is predominantly mental [18] and physical [6]. In games, there are several variables that could be manipulated. First, *difficulty* is a fundamental part of the game experience [27], and games are often set up to have increasing difficulty (e.g., the "boss fight" at the end of a sequence). Second, the *balance* between difficulty and skill has been discussed as an important element of game design [1]; optimal balance occurs when difficulty and ability are well matched (which can lead to the user experiencing "flow" [8, 34]).

To provide game designers with information about the effects of peak-end theory in games, we carried out two studies. Because peak-end experiences could have effects on more than just overall preference, we tested several aspects of play experience including perception of performance, challenge, tension, effort, and enjoyment. The first study manipulated difficulty level in two different games ("Match-3", based on Bejeweled, and "Whac-A-Germ", which is a point-and-click reaction game). Players rated their experiences following paired game sequences that altered the timing of periods of difficult and easy play (to create hypothetical peak-end effects), while equalizing the overall difficulty of each sequence.

The second study manipulated the balance between difficulty and skill using a shootout game and a computer-controlled opponent. Participants played a series of calibration rounds to determine their skill level, and then compared pairs of games that presented a sequence of enemies. The sequences were configured to induce positive and negative peak-end effects by distributing periods of difficulty/skill balance and imbalance through the sequence.

In both studies, we asked participants to rate various aspects of their experience after each sequence, and we also asked them to make forced-choice comparisons between them (e.g., which was more fun, which was more difficult, and which would you choose to repeat).

Results of the studies show that peak-end effects can strongly influence several aspects of player experience. First, in two of the three games (Match-3 and Shootout), we found significant effects on people's perceptions of the game and of their own performance – finishing with more difficult play made participants think that the sequence overall was more difficult, and led to changes in their perceptions of performance and enjoyment. Second, in the Match-3 game, we also found that the sequences had a significant effect on which of the two versions people saw as more fun and which they would prefer to repeat. The Whac-A-Germ game, in contrast, showed only an effect of perceived challenge.

These are important results for game designers, demonstrating that the sequence of game difficulty and challenge can have strong effects on several aspects of people's retrospective assessment of play experience, as predicted by the peak-end rule. While game designers already have some intuitions about sequencing difficulty (e.g., matching skill and challenge, or providing increasing difficulty in successive levels), our work provides a theoretical framework for explaining current practice, and opens opportunities for well-founded further research.

RELATED WORK

Peak-End Effects

Kahneman et al.'s [21] 'cold-pressor' experiment had subjects immerse their arm in painfully cold water in two conditions: 'short' and 'long'. Both conditions started identically with the arm submerged in 14° C water for one minute. In the 'short' condition the participants then removed their arm from the water, but in the 'long' condition their arm remained submerged for an additional 30 seconds during which the water gradually warmed to a still painful 15° C. When asked which condition the subjects would prefer to repeat, a significant majority (69%) chose to repeat the 'long' condition – preferring more total pain with a less painful ending, to less pain overall.

Fredrickson and Kahneman [13] similarly observed that the duration of negative experiences had little impact on retrospective evaluation. Subjects viewed unpleasant and pleasant film clips that varied in intensity and duration, giving continuous ratings of their momentary experiences. They were then asked to rank the clips in order of overall pleasantness. Findings showed *duration neglect* – rankings were unaffected by the duration of experience, as indicated by a correlation between rankings and a weighted average of momentary experience (which is independent of duration), and an absence of correlation between rankings and the total pleasantness/unpleasantness (which depends on duration).

In subsequent work, Kahneman et al. [19, 20, 22] presented three different types of 'utility' that might be used to characterize an experience: *instant utility*, which is the experience of a particular moment; *total utility*, which is the sum total of momentary experiences, and *remembered utility*, which is the user's memory of the experience. Importantly, Kahneman's studies demonstrate that *remembered utility* and *total utility* are very different measures of experience. As we substantially live in our memories, and make judgments and choices based upon them, *remembered utility* is particularly important.

The overweighted influence of the peak and ending moments of experience has been confirmed in many studies beyond painful and unpleasant stimuli. This includes the pleasure of receiving gifts [11], assessment of life quality (James Dean and Alexander Solzhenitzyn effects [9]), article pricing [23, 26], effortful study [12], enjoyment of gambling [7], and customer service experiences [10].

Relatively few studies in HCI have examined peak-end effects. Hassenzahl and Sandweg [18] examined correlations between perceived usability and various manipulations of mental effort during an interactive task, including conditions that manipulated peak and terminating mental effort. Results showed that the terminating level of mental effort correlated negatively with usability assessment, which they attributed to a recency effect. In their studies of progress bars, Harrison et al. [16, 17] referred to peak-end effects when explaining user preferences for progress bars that showed accelerating rates of progress. Finally, Cockburn et al. [6] recently directly examined peak-end effects as a method for manipulating users' retrospective assessment of interactive experiences. Their experiment examined user preferences for interfaces that altered the distribution of a constant total workload across a series of interactive pages. Results showed a significant preference for a series that combined a positive peak and end over a series that combined a negative peak and end. However, results were not significant for either series where only peak or end was manipulated.

Observing and experimenting with peak-end effects requires two components: first, experimenters must manipulate the sequence of subjects' momentary experiences, while keeping the objective total content of those experiences equitable; second, experimenters require a way to measure subjects' retrospective assessment of experience. When experiments involve direct pain, straightforward manipulations are possible (although with ethical challenges), including cold water immersion [21] and pressure induced by a vice-clamp [2]. Furthermore, the effectiveness of these measures can be easily validated through real-time subject reporting on the pain received [21]. Retrospective assessments are typically assessed through rating scales and through the use of forced preference choices, which ask participants which of two experiences they would choose to repeat.

Controlling and validating the momentary affective experience during human-computer interaction is challenging because users can respond differently to controlled elements. For example, when manipulating page workload in a study of peak-end effects in HCI [6], some participants reported a positive sense of accomplishment when a page contained high workload, while others reported a negative sense of burden. Neurological measures (e.g., EEG and fMRI) and autonomic activity measures (e.g., heart rate, pupil dilation, and skin conductivity) are promising technologies for understanding subjective experiences, but they remain largely under development for these applications (see [5] for a review).

Our studies, described below, use various measures of subjective experience (including self-report measures on various dimensions of experience and overall preference) to examine retrospective assessment of play experience in games where we manipulate objective game difficulty and difficulty-skill balance.

Sequencing, Challenge, and Player Experience in Games

Games are intended to provide a very different experience compared to other types of interactive systems – they are designed to entertain rather than to support stand-alone productivity or communication tasks [27]. As a result, there is considerable research into ways of evaluating player experience in games, much of which has focused on aspects of experience related to enjoyment, immersion, self-determination, fun, motivation, and suspense [3, 27, 30].

Although sequencing is an important part of game design (e.g., in following a narrative path, or in designing the ultimate goal of winning the game), the effects of different sequencing approaches on measures of experience has not been studied as widely. Work in other domains, however (notably in research into gambling) has looked at the idea of sequencing. Researchers have found that gamblers' memories are biased towards extreme events in a sequence, such as wins or jackpots [14] and have found strong peak-end effects on people's retrospective assessment of a sequence of events at a slot machine [35]. The reward structures of these games (e.g., the payout interval) has strong effects on people's willingness to keep playing, and the idea of reward structures has also been explored in the context of video games [3, 27]. However, "Creating a game that establishes immediate and continued motivation to continue playing over long periods of time is a very complex issue" ([27], p.6). Finally some researchers have looked at the ways that points of repeated failure in games (such as repeatedly dying at a particular spot) can negatively affect perception of an otherwise fun game [4].

Several games researchers have examined how a game's *challenge* should be designed in order to provide a good experience [1, 34]. Some games have progressively-harder levels, often culminating in a "boss fight" against a very difficult computer opponent (e.g., the Bowser level in Super Mario 64, see Figure 3). However, simply increasing difficulty level is not the only way to design challenge – researchers have noted that challenge should be matched to the player's level of skill (i.e., game balancing). As Pagulayan and colleagues state, "Too easy is boring and too hard is unfair. Either perception can make a person cease playing" ([27] p.6). When challenge and skill are appropriately matched, players can enter a "flow" state of intense focus and control, and a loss of awareness of self and time [1]. Similarly, researchers have shown that game suspense is greatest when players are slightly behind in a competitive match, and that enjoyment is highest when the outcome is uncertain [1]. The concept of balance does not state how different difficulty levels will affect retrospective experience, other than to suggest that balanced challenge and skill leads to positive experiences.

In contrast to the idea of skill/difficulty balance, there are several examples of games that are successful even though they are extremely easy or difficult. First, some games provide almost no challenge to the user, and instead focus on regular rewards, allowing the user to enter a state of "blissful productivity" with little effort [25]. For example, "Zen Mode" in *Bejeweled* (bejeweled.popcap.com, see Figure 1) allows players to continue indefinitely, with no time limit and with the game always providing at least one move on the board. Similarly, a genre of "clicker" games (such as *Cookie Clicker*) has very low difficulty – all the player needs to do

Figure 1. Bejeweled

Figure 2. Clicker Heroes

Figure 3. Super Mario 64

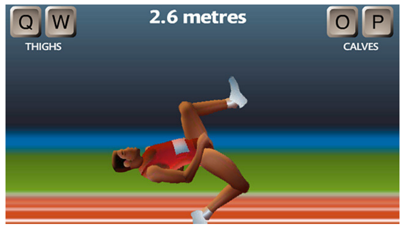

Figure 4. QWOP

to progress is to click the mouse button. These games are surprisingly popular – there are dozens of titles tagged as "Clicker" on the Steam platform, and one game, *Clicker Heroes*, has tens of thousands of players and a "very positive" rating (Figure 2).

At the other end of the challenge spectrum, some games are designed to be extremely difficult yet are also popular. For example, there are numerous lists of "hardest video games of all time" that celebrate extreme difficulty, and some games such as QWOP (foddy.net, Figure 4) appear to have been designed to prevent players from succeeding, while drawing on physical comedy [33]. These games can still provide a satisfying player experience, however, because games create a high degree of failure tolerance in players, and in some cases failure can be enjoyable (e.g., an accomplishment may be more rewarding after repeated failure). For example, researchers found that a particularly difficult game level, in which players repeatedly died, was rated as the most enjoyable [4].

STUDY 1: MANIPULATING DIFFICULTY

Our first study examines peak-end effects caused by different sequences of game difficulty. We chose game difficulty as our manipulation of momentary experience for two reasons. First, difficulty is a parameter that game designers can typically easily control by manipulating factors such as the number, speed, and firing rate of enemies, the rate of play (e.g., scroll speed or time per move), or the number of solutions available at any point. Second, previous interaction studies have demonstrated that different forms of 'difficulty' (mental effort [18] and physical workload [6]) have successfully induced peak-end effects. Together, these suggested that game difficulty could be a widely applicable and sensitive means for examining peak-end effects.

Methods

Games: Match-3 and Whac-A-Germ

We developed two custom games for this study, called Match-3 and Whac-A-Germ. Match-3 (Figure 5) is based on Bejeweled – the player clicks on sets of at least three contiguous tiles that are the same color. Unlike Bejeweled, tiles do not slide down once cleared – instead, each correctly selected set is grayed out. The objective of the game is to click as many sets as possible within five seconds. A falling white ball shows the remaining time for each game screen (Figure 5, right).

We created sequences of varying difficulty by algorithmically altering the number and size of the sets of contiguous tiles available within each game screen. The availability of many large sets makes the game easier, whereas fewer, smaller sets make the game harder. Figure 5 shows a medium difficulty game screen containing 9 sets and 34 tiles within them; Figure 6 (left) shows an easy screen containing 12 sets and 60 tiles; and Figure 6 (right) shows a hard screen containing 4 sets and 13 tiles.

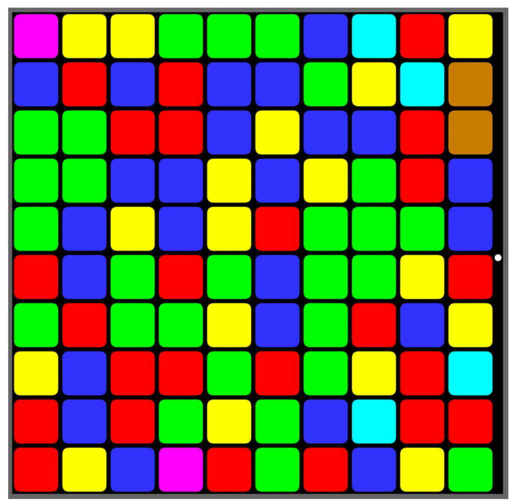

Figure 5. Match-3, medium difficulty. Goal of the game is to click on as many contiguous regions of three same-color tiles as possible before the timer (falling ball at right) finishes.

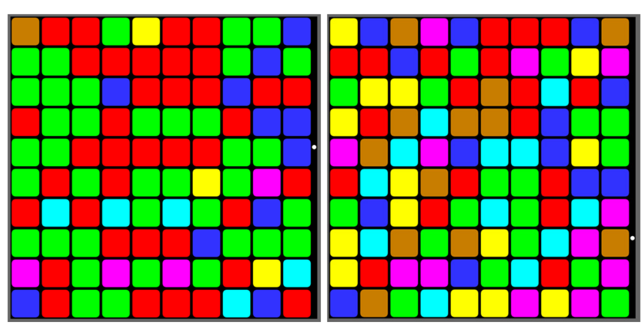

Figure 6. Match-3, easy (left) and hard (right) difficulty.

Match-3 games contained 15 screens, with the difficulty sequences described below. In all games, the total number of tiles in the sets was identical.

The second game was a targeting task motivated by the fairground game Whac-A-Mole. In our version, called Whac-A-Germ (Figure 7), players clicked as quickly as possible on a series of "germs" appearing on the screen. There were 15 germs in each round of the game. Difficulty was manipulating using the Fitts' Law index of difficulty (ID) – in easy tasks the germ was large and close to the player's cursor; and in hard tasks the germ was small and far away. The sequences of difficulties followed the patterns described below. The sum of the difficulties for each complete sequence was identical.

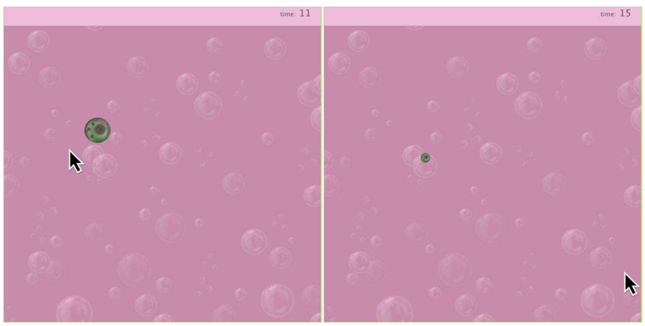

Figure 7. Whac-A-Germ, easy (left) and hard (right). Other screens were between these difficulties.

Difficulty Sequences (Experimental Conditions)

We designed three difficulty sequences to test hypothetically positive, negative, and neutral peak-end effects. The Match-3 sequences are shown in Figure 8 (the y-axis shows the total number of tiles available to be cleared in each screen). The *Positive* sequence contains easy screens (many tiles available) at its middle and end. The *Negative* sequence has screens with high difficulty at the middle and end. The *Neutral* sequence has screens with identical difficulty throughout. Note that the total objective difficulty of all sequences is identical. Whac-A-Germ sequences were the same shape as those shown in Figure 8 (achieved by manipulating Fitts' Law index of difficulty).

Participants and Apparatus

Twelve participants were recruited from a local university (7 female; mean age 27.7 years). Seven participants stated that they did not play computer games, and five stated that they played less than three hours per week. All were familiar with mouse input (>8 hrs/week).

The two games were written in Processing and were run on a Windows 7 computer, with a 21-inch 1080p monitor. Questionnaire responses were recorded on the same computer using an online form.

Procedure

Participants completed a demographics questionnaire and were introduced to the games. Participants then played twelve games in each system (Match-3 and Whac-A-Germ, order counterbalanced), divided into six pairs. The six pairs comprised three comparisons in two possible orders for the conditions shown in Figure 8: Negative vs. Positive, Neutral vs. Positive, Neutral vs. Negative.

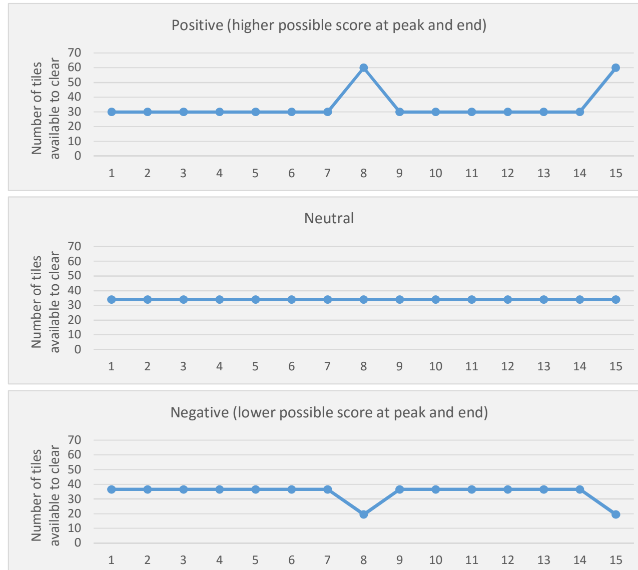

Figure 8. Difficulty sequences for Match-3 game showing number of tiles available per screen (more tiles mean lower difficulty). All three sequences have the same total difficulty.

After each game, participants completed a questionnaire asking about their experience in that game. For Match-3, we asked about perceived performance, ability to see matches on the board, player speed, and game difficulty; for Whac-A-Germ, we asked about perceived performance, perceived accuracy, player speed, and game difficulty. All questions used a 1-9 scale (low to high).

After each pair, participants answered four forced-choice questions, selecting between the two games of the pair – which game was more fun, which was more interesting, which was more challenging, and which the player would prefer to repeat. Once all pairs and questionnaires for one system (Match-3 or Whac-A-Germ) were finished, the player repeated the process for the other game.

Design and Research Questions

Questionnaire responses on the 1-9 scale were analysed using Friedman tests. Forced-choice responses were analysed using a binomial test. Our hypotheses for the study were predominantly exploratory, but we anticipated that participants' rankings and preferences would be positively influenced by our Positive conditions, and negatively influenced by the Negative conditions.

Results

Match-3: Post-Game Responses

We asked four questions (1-9 scale) at the end of each Match-3 game, regarding players' perception of their performance, their speed, their ability to see matches on the board, and the difficulty of the game. Results are shown in Figure 9.

Friedman tests showed significant differences for Player speed (χ^2=4.83, p<.05) and Difficulty (χ^2=10.29, p<.01), but not for Performance (χ^2=5.29, p=.071) or Ability to see matches (χ^2=2.38, p=.30).

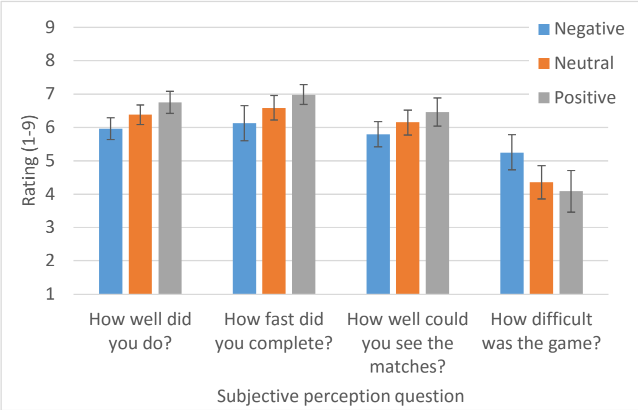

Figure 9. Match-3: mean post-game participant ratings ±s.e.

Match-3: Comparison of Sequences

After each pair of sequences, we asked participants four forced-choice questions about the pair (see Table 1). Binomial tests on the counts found several significant differences (p<.05). For the pair with the largest expected difference (Positive to Negative), there were significantly more responses choosing the Positive sequences as more fun, more interesting, and easier. In addition, there were significantly more responses choosing the Positive sequence to repeat. For the other pairs with smaller expected differences (Positive-Neutral and Neutral-Negative), players stated that Positive was more interesting than Neutral, that Neutral was more challenging than Positive, and Negative more challenging than Neutral.

	Of the two games you just played, which: was more fun?	was more interesting?	was more challenging?	would you repeat?
Positive peak and end	**19**	**18**	**5**	**18**
Negative peak and end	**5**	**6**	**19**	**6**
Binomial p-value	**.0033**	**.011**	**.011**	**.011**
Positive peak and end	16	**18**	**6**	14
Neutral (no peak or end)	8	**6**	**18**	10
Binomial p-value	.076	**.011**	**.011**	.27
Neutral (no peak or end)	15	9	**5**	13
Negative peak and end	9	15	**19**	11
Binomial p-value	.15	.15	**.0033**	.42

Table 1. Participant counts for comparison questions (significant differences in bold).

Whac-A-Germ: Post-Game Responses

We asked four questions (1-9 scale) at the end of each game (perceived performance, accuracy in clicking, player speed, and the difficulty of the game). Results are shown in Figure 10. Friedman tests did not show any significant effects of Sequence on any of these questions (Performance: χ^2=0.5, p=.78; Accuracy: χ^2=3.29, p=.19; Speed: χ^2=4.67, p=.097; Difficulty: χ^2=0.54, p=.76).

Whac-A-Germ: Comparison of Sequences

After each pair of sequences, we asked participants four choice questions about the pair (see Table 2). Binomial tests on the counts found only one difference – significantly more participants chose the Negative sequence as more challenging than the Positive sequence (p<.05).

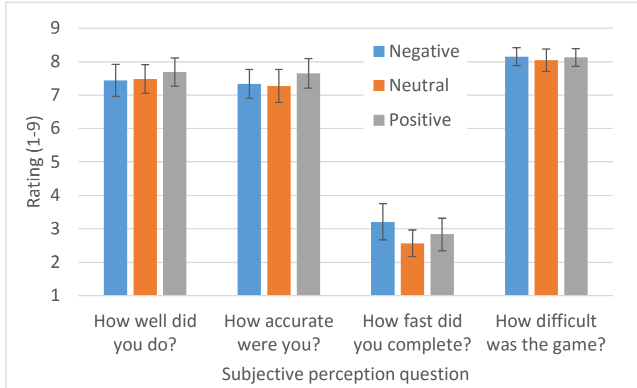

Figure 10. Whac-A-Germ: mean post-game participant ratings

	Of the two games you just played, which: was more fun?	was more interesting?	was more challenging?	would you repeat?
Positive peak and end	12	12	**6**	15
Negative peak and end	12	12	**18**	9
Binomial p-value			**.011**	.14
Positive peak and end	13	12	8	13
Neutral (no peak or end)	11	12	16	11
Binomial p-value	.42		.076	.42
Neutral (no peak or end)	11	11	8	10
Negative peak and end	13	13	16	14
Binomial p-value	.42	.42	.076	.27

Table 2. Participant counts for comparison questions (significant differences in bold).

STUDY 2: MANIPULATING BALANCE

The results of Study 1 consistently showed that participants' perception of *challenge* was influenced by manipulation of peak-end experience. The relationship between the challenge of an experience and a person's skill levels are well known to strongly influence engagement in that experience [8, 34] – high levels of challenge with insufficient skills will lead to frustration; insufficient challenge with high skills leads to boredom; and a *balance* between challenge and skill can optimize user engagement.

As the balance between challenge and skill is a key consideration for game designers, Study 2 specifically uses *skill-challenge balance* as the experiential factor that is manipulated through peak-end event sequencing. An initial calibration session is used to determine the participants' skill levels, and peak-end experiences are constructed by creating game sequences that vary the timing of moments that have *appropriate balance*, are *too easy*, and are *too hard*, relative to the users' skill.

Methods

Game: Space Shootout

We built a simple one-on-one shooting game with spaceships at the top and bottom of the screen (bottom controlled by the player, top controlled by the computer). The player moved their ship left and right with the 'h' and 'l' keys, fired the gun with the 'k' key, and invoked a temporary shield (one second duration) with the 'z' key. The object of the game was to shoot the opponent's ship, and when this occurred, the round finished; the game was designed to have several rounds (so that we could create different difficulty sequences).

The computer enemy was designed to have a variable level of difficulty (from 0 to 20) through programmatic control of four elements: the enemy's speed, the enemy's fire rate, the likelihood that the enemy would use its shield when needed, and the likelihood that the enemy would change direction in the face of incoming fire. The algorithm was hand-tuned in pilot testing until it provided a very low starting difficulty at level 0 (trivial for any player to beat), a gradual increase from level 0 to 20, and an extremely high difficulty at level 20 (essentially impossible to beat).

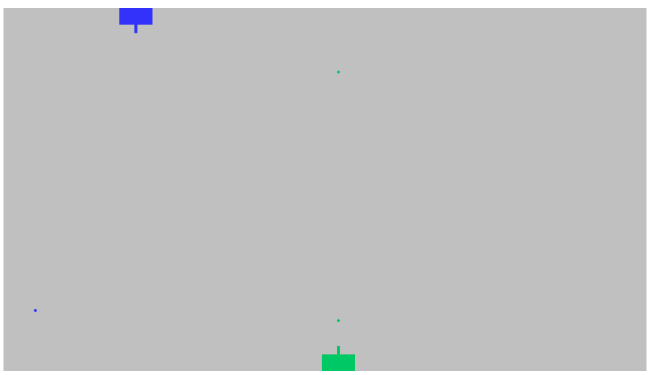

Figure 11. Shootout game: player at bottom, enemy at top.

Balance Sequences (Experimental Conditions)

To test for peak-end effects when manipulating skill-challenge balance (as opposed to simple difficulty in Study 1), we created sequences that used an appropriate balance as the positive momentary experience (corresponding to Study 1's easy peak-and-end sequence), and an inappropriate balance as the negative momentary experience (corresponding to Study 1's difficult peak-and-end condition). Importantly, there are two forms of inappropriate balance: too easy and too hard.

The sequences are illustrated in Figures 12 and 13. The overall hypothesis for these sequences is that people may remember sequences that end with an appropriate balance differently than sequences that are unbalanced at the end. Furthermore, there may be differences between different forms of imbalance (too easy or too hard).

Participants and Apparatus

Twelve participants were recruited from a local university (5 female; mean age 27.5 years). Gaming experience varied: four people stated that they did not usually play computer games, four stated that they played less than three hours per week, and four played between 3 and 10 hours/week.

As in the first study, the game system was written in Processing and ran on a mid-range Windows 7 computer, with a 21-inch 1080p monitor. Questionnaire responses were recorded on the same computer using an online form.

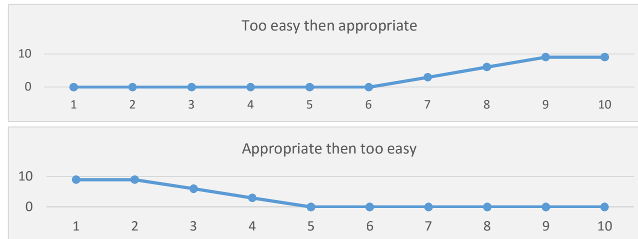

Figure 12. Example difficulty levels for sequences in shootout game (easy version): sequences are exactly inverted, with appropriate difficulty at either start or end. Note: overall line was adjusted up or down based on player calibration.

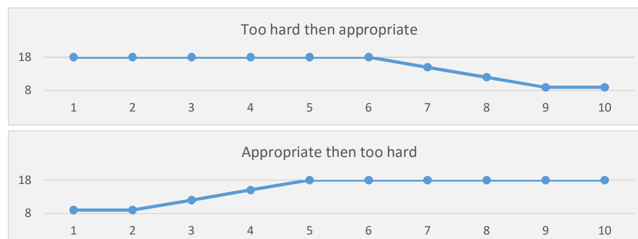

Figure 13. Example difficulty for sequences in shootout game (hard version).

Procedure

Participants completed a demographics questionnaire and were introduced to the shootout game. Participants then completed a calibration session in which they played 20 rounds against enemies of increasing difficulty – the object of the calibration was to determine three difficulty levels: one that was too easy, one that was too hard, and one where the player was evenly matched to the computer opponent. The experimenter determined the balanced level by choosing a round that was one of the longest (neither player was able to win easily) and where the participant either won the round or had won in the previous round. This level was used to set the difficulty levels of the sequences as described above.

In the second phase of the study, participants played pairs of games with matched but inverted difficulty sequences. Instead of simply changing difficulty as in the first study, we manipulated whether the challenge was too easy, too hard, or evenly matched. The *Too-Easy* pair (see Figure 12) had one sequence that had enemies that were too easy then finished with enemies that were evenly matched to the player, and one sequence that began evenly matched and then became too easy. The *Too-Hard* pair (Figure 13) had one sequence that was too hard then evenly matched, and one that was even and then too hard.

After each game, participants completed a validated scale asking about their experience in that game. We used the Intrinsic Motivation Inventory (IMI) questionnaire, which measures four dimensions of experience: enjoyment, perceived competence, effort, and tension [24] using 18 items rated on a 1-5 scale: 1=Strongly Disagree; 5=Strongly Agree. After each pair, participants also answered the forced-choice questions used in Study 1: these asked which game was more fun, more interesting, more challenging, and which the player would prefer to repeat.

Design and Hypotheses

For the post-game questions, we combined the questions from the IMI into the four constructs; the IMI is a validated instrument that allows the use of ANOVA on the combined scores for enjoyment, competence, effort, and tension [24]. For the forced-choice questions, we counted the number of participants choosing each response and compared counts using a Binomial test. Our hypotheses in Study 2 were again exploratory: we expected sequences ending with appropriate balance to be seen as preferable, but the effects on play experience are hard to predict, given the many ways that players view success and difficulty in games.

Results

Shootout: Post-Game Responses

Mean results for the IMI are shown in Figures 14 and 15 (Too-Easy and Too-Hard versions of the game respectively). RM-ANOVA showed that in the Too-Easy version, there were significant main effects of Perceived Competence ($F_{1,11}$=9.68, p=.01, η^2=.47) and Tension ($F_{1,11}$=5.47, p=.039, η^2=.33). For the Too-Hard version, there were effects of Perceived Competence ($F_{1,11}$=13.3, p=.004, η^2=.55) and Effort ($F_{1,11}$=7.8, p=.018, η^2=.41).

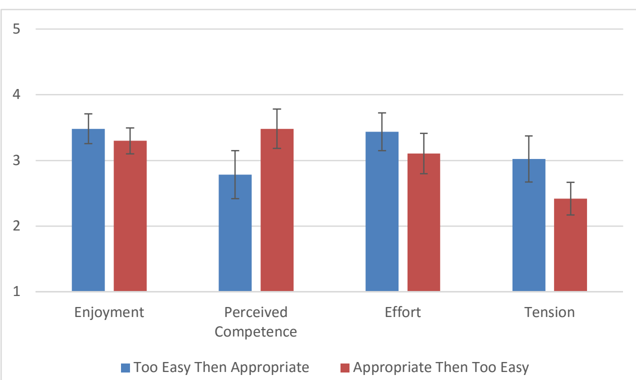

Figure 14. Shootout (too-easy version): mean post-game participant ratings (±s.e.)

These results show that players saw themselves as more competent when the sequence ended with a set of too-easy enemies, and also when the sequence ended with a set of too-hard enemies (see below for discussion of this result). In the too-easy games, players rated tension higher when the game ended with appropriate balance; and in the too-hard games, players stated that that they put in more effort when the game ended with balanced difficulty.

Shootout: Comparison of Sequences

After each pair, we asked participants four choice questions about the pair (see Table 3). Binomial tests on the counts found only one difference, in the Too-Easy version – significantly more participants felt that the sequence that ended with appropriate balance was more challenging than the appropriate-then-too-easy sequence.

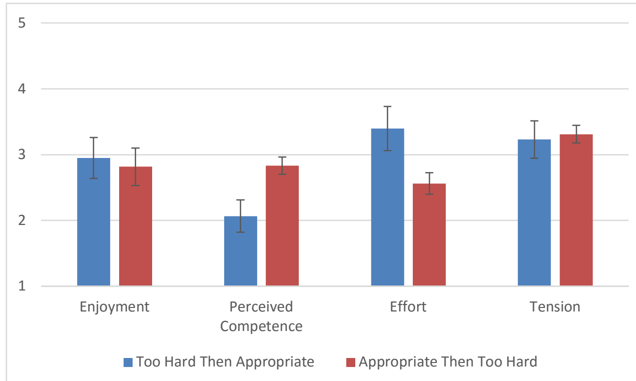

Figure 15. Shootout (too-hard version): mean post-game participant ratings

	Of the two games you just played, which: was more fun?	was more interesting?	was more challenging?	would you repeat?
Too Easy → Appropriate	5	8	**10**	6
Appropriate → Too Easy	7	4	**2**	6
Binomial p-value	.77	.19	**.019**	
Too Hard → Appropriate	6	6	5	5
Appropriate → Too Hard	6	6	7	7
Binomial p-value			.77	.77

Table 3. Participant counts for comparison questions (significant differences in bold).

DISCUSSION

The two studies provide several main findings:

- In Match-3, sequences with positive peaks and ends were judged as significantly more fun, more interesting, and less challenging than sequences with negative peaks and ends (despite identical total objective difficulty). Players also rated their own speed significantly higher with positive sequences, and were significantly more likely to choose the positive sequence to play again.
- In Whac-A-Germ, the only significant difference found was that players saw negative sequences as significantly more challenging than positive sequences.
- In the too-easy version of Shootout, players rated their own competence significantly higher when the sequence ended with too-easy enemies (compared to appropriately balanced), and also rated these sequences as less tense.
- In the too-hard version of Shootout, players rated their competence significantly higher when the sequence ended with too-hard enemies, but also rated their own effort as significantly lower in these sequences.
- The only difference in the forced choice selections between Shootout sequences was that the sequence ending with appropriate skill/difficulty balance was chosen as

being more challenging significantly more frequently than the sequence ending with easy enemies.

- There were no significant differences in enjoyment for the different sequences of Shootout, nor did players choose any one sequence as more fun or more replayable.

The following sections discuss possible reasons underlying the results, consider some of the surprising outcomes from the study, connect our findings back to larger issues of game design and peak-end theory, and identify experimental limitations and opportunities for further work.

Strong peak-end effects of difficulty in Match-3

Adjusting the difficulty of the screens in the Match-3 game led to strong differences in people's perceptions of the sequences – they thought the positive sequence was more fun, more interesting, and more worthy of replay.

Possible explanations concern the style of game, the way in which people think about difficulty in this game, and the possibility that we hit a sweet spot within the range of difficulties chosen. The gameplay mechanics are extremely simple – similar to the "clicker" games described earlier, which have very low challenge but are still seen as rewarding by players. In these games, obtaining easy rewards is a key component of gameplay, unlike other genres (such as completing a quest or achieving a goal). In Match-3, therefore, players may have felt that "easier is better," and thus may have had stronger memories of the sequence that allowed them to do particularly well (i.e., Positive peak and end). It is also possible that the difficulty of "easy" configurations was well matched to our participants' skill, giving an engaging final experience.

Few peak-end effects of difficulty in Whac-A-Germ

In contrast to Match-3, the only effect of difficulty sequence in Whac-A-Germ was that people felt that the negative sequence was more challenging than the positive sequence (despite its objectively neutral total content). It is possible that the absence of other effects arises from the way in which difficulty and reward influenced its gameplay. Even though people did feel that the negative sequences were more challenging, the manipulation of difficulty in Whac-A-Germ appears to have had a weaker effect on momentary experiences than in Match-3. In Match-3, being successful on a screen was rewarding (e.g., it was satisfying to clear a large amount of the screen), and success was directly correlated to the difficulty of the screen. In Whac-A-Germ, there was no difference in reward based on difficulty, and the game behaved the same regardless of whether the player had clicked on a small far target or a large close one.

In addition, every screen in Whac-A-Germ had the same level of success – that is, clicking on the germ advanced to the next target. There was no time limit, and little to motivate or engage the player. Therefore, every screen had the same success experience regardless of difficulty. Although people's perception of overall difficulty was affected by the sequencing (small, distant targets are objectively harder than big, near ones), this may have had less effect on play experience (small, distant targets are no more 'fun' or engaging than big, near ones). Consequently, participants' preferences and ratings were largely unaffected by our manipulations.

Consistent peak-end effects on perception of challenge

People's responses in both studies showed that their perception of game challenge was consistently altered by the sequencing of difficulty in the game – by the peak and end experiences in Study 1, and by the end experience in Study 2. These results were consistent in three different kinds of games involving three different mechanics (matching, targeting, and fighting an opponent).

There are two issues to consider when interpreting this finding: what it says about the existence of peak-end effects in games, and what its implications are for game design. The consistency of the result provides strong evidence for a difficulty-based peak-end effect in games. In both studies, sequences that finished with a higher difficulty level were reliably seen as more difficult overall, even though the objective total difficulty was identical.

Although game designers may already know that the final experience of a game is important (e.g., a poor ending could "leave a bad taste in the player's mouth"), there is still little understanding of the specific effects of sequencing, and our results suggest that evaluations of play experience could be influenced in ways that designers are not considering.

For example, when running evaluation tests for a game, designers may receive feedback that a game is too hard if the evaluation session ends with a difficult set of events. This could lead designers to rebalance the overall game, which might result in a game that is too easy in sections of the sequence. Multi-player settings are another evaluation scenario that could be complicated by peak-end effects. In multi-user online play, the game has little control over the difficulty faced by the player – and our results show that differences in the sequences of enemies could lead to large differences in a player's perception of overall difficulty.

Perception of competence, effort, & tension in Shootout

There were also interesting effects on player perceptions in the Shootout game. First, it is puzzling that perceived competence went up significantly both when the sequence ended with too-easy enemies, and also with too-hard enemies. The too-easy case seems logical and conforms to peak-end theory – it is likely that participants viewed themselves as more competent during easy play, and their retrospective assessment of competence was substantially influenced by their terminating experience. However, the too-hard case seems out of place. A speculative explanation is that players may have been reacting to the near-impossibility of the game – that is, when the game is impossible to win, players may have started to attribute the reason for their performance to external sources (i.e., that the game was far too difficult) rather than to themselves. This explanation is reflected in prior psychology findings that people have a tendency to accept the credit for their successes, while attributing failure to external sources [32].

Another finding that might initially appear surprising is that participants' retrospective assessment of effort was lower following the too-hard ending. However, we suspect this is due to the players giving up when they became unable to compete due to excessive challenge – their effort was therefore very low (or nil) in the final moments. Similarly, the participants' recollections of tension were significantly lower following a too-easy ending, which reflects a lack of challenge during the final moments.

It is important to reiterate that although we cannot definitively state why people had these perceptions of their experience, their perceptions were definitely altered by the change in sequence. This implies again that game evaluators should be aware of potential peak-end effects.

Preferences for easier and harder games in Shootout

The Shootout study did not show any effects on enjoyment, fun, interest, or willingness to repeat one sequence over another. This suggests that in a game like Shootout, there is not the same connection between difficulty and enjoyment as observed for Match-3. In Shootout, there seems to be a much more complex relationship between people's preferences and the level of skill/challenge balance.

It seems surprising, for example, that in Study 2, players did not prefer sequences that ended with an appropriate difficulty level – even though it appeared that their memories of the sequences were in fact influenced by the ending balance (too-hard and too-easy). As expected, players indicated that the sequence ending with too-hard enemies was more difficult than the sequence ending with appropriate balance, however, their preference choices were split between too-hard and appropriate endings. Similarly mixed choice results were found for the too-easy sequences (except for challenge, as discussed earlier).

These conflicting results suggest that in some games, easier may be preferred to balanced; and in others, harder may be preferred. We had assumed that an appropriate balance between skill and difficulty would be seen as the preferable option (as suggested by previous literature [1,26]), but this was not the case in our study.

Despite the seemingly obvious logic behind balancing skill and difficulty, there are a few reasons why players might choose a sequence that they recall as being too-difficult or too-easy. In some settings, people actually want things to be highly challenging. As described above, some games are extraordinarily difficult, but have attracted a large following. Given that the Shootout game was very simple in terms of gameplay, it may have been that some players saw increased difficulty as the only thing that would make the game interesting over a longer term. Similarly, some of our participants may have seen the too-easy condition as a satisfying success experience, leading them to state a preference for the sequence that ended with easy enemies.

These equivocal results show that even though peak-end effects are occurring in our games, the connection between difficulty and repetition choice in our experiments is not as clear as the connection between pain and choice that Kahneman originally studied.

Experimental limitations and further work

Although several of our findings conform to the peak-end rule, others failed to show predicted effects. Reliable generalisation of the peak-end rule to game settings therefore requires further study. This includes analysis of different distributions of game content (e.g., experience curves other than those shown in Figure 8), manipulation of experiential factors such as fun or comedy [33], and studying different types of gamers (e.g., different expertise, and those with different preferences [15]). There are also opportunities to examine the effect of peak experiences in isolation, as our studies only examined combined peak and end experiences (Study 1) and end effects (Study 2).

CONCLUSIONS

The peak-end rule is a psychological heuristic stating that people's retrospective assessment of an experience is substantially influenced by the intensity of the peak and terminating moments of that experience. It has been validated in many areas involving human quality judgments, but it has received relatively little attention in HCI and less still in gaming, where the remembered quality of experience is likely to be a fundamental factor influencing potential future play.

Results from the three different games examined in this paper show that peak-end effects occurred when the sequence of difficulty and challenge-skill balance was manipulated. Users' recollection of challenge was uniformly affected by peak-end sequence manipulations, despite the tested conditions being objectively identical in total content. Peak-end manipulations also led to significant differences in recalled fun, interest and willingness to repeat in one of the three games.

This work provides initial insights into how a widely known but under used (in HCI) psychological phenomenon influences game experiences. There are abundant opportunities for further work in examining how different parameters of momentary experience induce peak-end effects – our experiments used only relatively crude manipulations of difficulty and challenge-skill balance. There are also extensive opportunities for examining and understanding the conditions under which peak-end retrospective assessments translate into desired positive outcomes, such as recalled enjoyment and fun.

ACKNOWLEDGMENTS

This work was supported by NSERC, the GRAND NCE, and the SurfNet research network.

REFERENCES

1. Abuhamdeh, S. and Csikszentmihalyi, M. The importance of challenge for the enjoyment of intrinsically motivating, goal-directed activities. *Personality and Social Psychology Bulletin 38*, 3 (2012), 317-330.
2. Ariely, D. Combining Experiences Over Time: The Effects of Duration, Intensity Changes and On-Line Measurements on Retrospective Pain Evaluations. *Journal of Behavioral Decision Making 11* (1998), 19-45.
3. Bernhaupt, R. (ed.), *Game User Experience Evaluation*. Springer, 2015.
4. Birk, M., Luerig, C. and Mandryk, R.L. A Metric for Automatically Flagging Problem Levels in Games from Prototype Walkthrough Data *Academic MindTrek '15*, Tampere, Finland, 2015, To appear. .
5. Brave, S. and Nass, C. Emotion in human-computer interaction. in Julie, A.J. and Andrew, S. eds. *The human-computer interaction handbook*, L. Erlbaum Associates Inc., 2003, 81-96.
6. Cockburn, A., Quinn, P. and Gutwin, C. Examining the peak-end effects of subjective experience. in *Proceedings of CHI'15: ACM Conference on Human Factors in Computing Systems*, ACM Press, (2015), 357-366.
7. Cowley, E. and Farrell, C. Retrospective evaluations: Will chance versus luck oriented individuals select different moments of an experience? *Advances in Consumer Research*, San Antonio, United States, 2006, http://www.acrwebsite.org/volumes/v33/naacr_v33_32.pdf.
8. Csikszentmihalyi, M. *Flow: The Psychology of Optimal Experience*. Harper and Row, New York, 1991.
9. Diener, E., Wirtz, D. and Oishi, S. End effects of rated life quality: the James Dean Effect. *Psychol Sci. 12*, 2 (2001), 124-128.
10. Dixon, M. and Verma, R. Sequence effects in service bundles: Implications for service design and scheduling. *Journal of Operations Management 31*, 3 (2013), 138--152.
11. Do, A., Rupert, A. and Wolford, G. Evaluations of pleasurable experiences: The peak-end rule. *Psychonomic Bulletin & Review 15*, 1 (2008), 96-98.
12. Finn, B. Ending on a High Note: Adding a Better End to Effortful Study. *Journal of Experimental Psychology: Learning, Memory, and Cognition 36*, 6 (2010), 1548-1553.
13. Fredrickson, B. and Kahneman, D. Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology 65*, 1 (1993), 45-55.
14. Griffiths, M. Fruit machine gambling: The importance of structural characteristics. *Journal of Gambling Studies 9*, 2 (1993), 101-120.
15. Hamari, J. and Tuunanen, J. Player types: A meta-synthesis. *Transactions of the Digital Games Research Association 1*, 2 (2014), 29-53.
16. Harrison, C., Amento, B., Kuznetsov, S. and Bell, R. Rethinking the progress bar. in *Proceedings of the 20th annual ACM symposium on User interface software and technology*, ACM, (2007), 115-118.
17. Harrison, C., Yeo, Z. and Hudson, S.E. Faster progress bars: manipulating perceived duration with visual augmentations. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, (2010), 1545-1548.
18. Hassenzahl, M. and Sandweg, N. From mental effort to perceived usability: transforming experiences into summary assessments *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, ACM, Vienna, Austria, 2004, 1283-1286
19. Kahneman, D. Evaluation by Moments: Past and Future. in Kahneman, D. and Tversky, A. eds. *Choices, Values and Frames*, Cambridge University Press New York, 2000.
20. Kahneman, D. Experience Utility and Objective Happiness: A Moment-Based Approach. in Kahneman, D. and Tversky, A. eds. *Choices, Values and Frames*, Cambridge University Press New York, 2000.
21. Kahneman, D., Fredrickson, B., Schreiber, C. and Redelmeier, D. When More Pain Is Preferred to Less: Adding a Better End. *Journal of Personality and Social Psychology 65*, 1 (1993), 401-405.
22. Kahneman, D., Wakker, P. and Sarin, R. Back to Bentham? Explorations of Experienced Utility. *The Quarterly Journal of Economics 112*, 2 (1997), 375-405.
23. Langer, T., Sarin, R. and Weber, M. The retrospective evaluation of payment sequences: duration neglect and peak-and-end effects. *Journal of Economic Behavior & Organization 58*, 1 (2005), 157-175.
24. McAuley, E., Duncan, T. and Tammen, V.V. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: a confirmatory factor analysis. *Res Q Exerc Sport 60*, 1 (1989), 48-58.
25. McGonigal, J. 2014. Games Can Make a Better World. from http://janemcgonigal.com/2014/01/06/transcript-games-can-make-a-better-world/.
26. Nasiry, J. and Popescu, I. Dynamic Pricing with Loss-Averse Consumers and Peak-End Anchoring. *Operations Research 59*, 6 (2011), 1361--1368.
27. Pagulayan, R.J., Keeker, K., Wixon, D., Romero, R.L. and Fuller, T. User-centered design in games. in Jacko, J., A and Sears, A. eds. *The human-computer interaction handbook*, L. Erlbaum Associates Inc., 2003, 883-906.
28. Redelmeier, D.A. and Kahneman, D. Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain 66*, 1 (1996), 3--8.

29. Redelmeier, D.A., Katz, J. and Kahneman, D. Memories of colonoscopy: a randomized trial. *Pain 104*, 1 (2003), 187--194.
30. Ryan, R., Rigby, C.S. and Przybylski, A. The Motivational Pull of Video Games: A Self-Determination Theory Approach. *Motivation and Emotion 30*, 4 (2006), 344-360.
31. Siang, A.C. and Radha Krishna, R. Theories of learning: a computer game perspective. in *Multimedia Software Engineering, 2003. Proceedings. Fifth International Symposium on*, (2003), 239-245.
32. Streufert, S. and Streufert, S.C. Effects of conceptual structure, failure, and success on attribution of causality and interpersonal attitudes. *Journal of Personality and Social Psychology 11*, 2 (1969), 138-147.
33. Švelch, J. Comedy of Contingency: Making Physical Humor in Video Game Spaces. *International Journal of Communication 8* (2014), 2530-2552.
34. Sweetser, P. and Wyeth, P. GameFlow: a model for evaluating player enjoyment in games. *Comput. Entertain. 3*, 3 (2005), 3-3.
35. Yu, E., C., Lagnado, D., A. and Chater, N. Retrospective evaluations of gambling wins: Evidence for a 'peak-end' rule *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2008, 77-82.