The Effects of Interaction Sequencing on User Experience and Preference

Andy Cockburn , Philip Quinn , Carl Gutwin

PII: S1071-5819(17)30105-2
DOI: 10.1016/j.ijhcs.2017.07.005
Reference: YIJHC 2139

To appear in: *International Journal of Human-Computer Studies*

Received date: 7 December 2016
Revised date: 5 June 2017
Accepted date: 18 July 2017

Highlights

- Presents results two studies into the influence of primacy and recency effects in interaction.
- Shows that recency effects significantly influence preferences.
- Primacy effects were not shown to influence preferences.
- Discusses factors influencing recency effects in interaction.

The Effects of Interaction Sequencing on User Experience and Preference

Andy Cockburn[1†], Philip Quinn[1], Carl Gutwin[2]
[1]Department of Computer Science and Software Engineering
University of Canterbury
Christchurch, New Zealand
andy@cosc.canterbury.ac.nz, philip.quinn@canterbury.ac.nz

[2]Department of Computer Science
University of Saskatchewan
Saskatoon, Canada
gutwin@cs.usask.ca

ABSTRACT

Interactive computing tasks are composed of sequences of sub-interactions (or "moments"), each of which yields a slightly different user experience. Prior work, predominantly from the psychology literature, suggests that the order of these moments can affect people's retrospective evaluation of experiences. Several kinds of sequencing effects have been examined, including *primacy*, *recency,* and *peak-end* effects. We review previous research on sequencing effects and their potential application in Human-Computer Interaction, of which prior work has found mixed results regarding the influence of interaction sequence on preference – possibly because the magnitude of experiential changes caused by interactive tasks are weaker than those studied in psychological experiments. However, sequencing effects are still of great importance to interface design, because when they occur, they have the potential to substantially change user preferences for common interactions. To explore the subtlety of sequencing effects in HCI, we describe two experiments that examined user preferences for series of interactions with different orderings that created positive and negative recency and primacy effects. Positive and negative experiences were created with simulated system assistance that either worked well (aiding the user in drag-and-drop tasks) or worked poorly (hindering the user). In both experiments, the series differed only in the order of positive and negative momentary experiences. Results of Experiment 1 were mixed: the study provided some support for recency effects, but without strong evidence. Experiment 2 modified the experimental method to better accentuate the positive and negative experiences, and produced results showing strong effects of recency, but not of primacy. We discuss reasons for these results, consider overall explanations for the subtle nature of sequencing effects on HCI tasks, and provide an agenda for further research and design lessons regarding recency effects. Overall, we contribute new understanding of a phenomenon that can have a substantial impact on user experience, but that is currently underexplored in HCI.

Keywords: peak-end rule, primacy, recency, subjective experience, interface preferences.

[†] Contact author

1 Introduction

Subjective experience is a major factor in the success of an interactive system – a person's willingness to use (or re-use) an application is strongly influenced by the perceived enjoyment of doing so (Carroll, 1987; Davis *et al.*, 1992; Hassenzahl *et al.*, 2000; Malone, 1980; Venkatesh *et al.*, 2003). Understanding the factors that influence user assessment of subjective experience is therefore a key objective for research in Human-Computer Interaction (HCI).

Studies of subjective user experience in HCI have focused on several different aspects of interaction, including: methods for measuring user satisfaction (Chin *et al.*, 1988; Hornbaek, 2006), the role of interaction aesthetics (Hassenzahl, 2004; Porat and Tractinsky, 2012; Tractinsky *et al.*, 2000; Udsen and Jørgensen, 2005), improving engagement through gamification (Deterding *et al.*, 2011), and interaction features that delight users (Levesque *et al.*, 2011; Shneiderman, 1987; Zhang and von Dran, 2001). Others have developed models of users' acceptance of technology, such as the Technology Acceptance Model (Davis, 1989; Venkatesh *et al.*, 2003).

An important feature in studies of subjective experience is the difference between the user's *instantaneous* experience (as the interaction is happening), and their *recollected* experience (after the interaction is complete). Although both are important, recollected experience is more of a concern for designers, because the user's memory of their experience will influence their willingness to repeat an interaction. For example, user retention at an e-commerce website is likely to be strongly influenced by users' recollection of their shopping and checkout experiences at the site.

One factor in user experience studies that has not been examined closely is the way that changes to the *sequencing* of sub-interactions (called "moments") can influence people's memory of an interactive experience. Psychological research has found that a person's recollection of an experience is not always faithful to what actually happened. In particular, three factors – *primacy, recency,* and *peak-end* effects – are known to influence people's memory of experiences. Primacy effects refer to an enhanced memory and an over-weighted influence of the initial moments of an experience (Murdock Jr, 1962; Shteingart *et al.*, 2013; Zauberman *et al.*, 2006), while recency and peak-end effects refer to an over-weighted influence of the terminating (*end*) and most intense (*peak*) moments (Kahneman *et al.*, 1993). These effects have been demonstrated to influence hedonic assessment (i.e. assessment of pain or pleasure) in several settings, including memory of pain during medical procedures (Redelmeier and Kahneman, 1996), the retrospective assessment of payment sequences (Langer *et al.*, 2005; Nasiry and Popescu, 2011), the remembered enjoyment of holidays (Kemp *et al.*, 2008), and assessments of quality of life (Diener *et al.*, 2001) or material goods (Do *et al.*, 2008). Further consequences of these effects are that the *duration* of an experience has relatively little effect on the remembered experience (*duration neglect*), and that adding positive experiences (or, less-negative experiences) to the end of a generally negative event can improve its overall subjective evaluation (Fredrickson and Kahneman, 1993). That is, people's memories are generally insensitive to the duration of an event, but are instead concentrated on the sensations felt near its end.

Despite the wide range of psychological studies that show a robust influence of sequencing effects, HCI has seen few experiments examining their influence on user experience. In addition, attempts to directly test the existence of peak-end effects in interactive settings have produced equivocal results, with hypothesised preferences observed in only one of three conditions involving form-filling interfaces (Cockburn *et al.*, 2015) and only one of eight tested conditions with computer games (Gutwin *et al.*, 2016).

The previous work in HCI provides only limited evidence that sequencing effects influence user preferences in interaction. However, if validated and understood in interactive contexts, sequencing effects could have important implications for interface design: for example, they might reveal opportunities for designers to influence a user's memory of an interactive experience simply by altering the sequence of elements within an interaction, potentially leading to a more favourable view of the system. As stated above, it is a user's *retrospective* assessment of experience that is crucial for design, because their willingness to repeat an interaction will be influenced by their memory of it.

In this paper we examine whether sequencing effects influence user preferences for interactions with computer systems. We first review prior literature examining related effects across various disciplines, including HCI, and then describe two empirical studies in which participants were asked to choose which of two sequences they would prefer to repeat, where the sequences manipulated only the ordering of otherwise identical tasks. The individual tasks involved drag-and-drop manipulations of objects – with positive, neutral, and negative momentary sensations induced through *positive*, *control*, and *negative* snap-to-grid assistance while aligning the dragged object with the target. During *positive* tasks, the dragged object and target were aligned with a coarse grid that made task completion easier and faster; during *control* tasks, the dragged object moved under direct-manipulation (i.e. without snapping); and in *negative* tasks, the dragged object snapped to a grid, but with the target purposely misaligned with the grid – requiring participants to drop the dragged object and reacquire it with the 'control' key pressed, which reverted to direct-manipulation (*control*) dragging.

The first experiment (n=46) compared sequences with a positive start to sequences with a positive end, and sequences with a negative start to sequences with a negative end (in both cases, comparing recency vs. primacy effects). Although participant preferences overall were in line with the recency hypothesis (i.e. that people are influenced more by moments at the end of a sequence), only one significant difference was found on any measure (frustration/satisfaction ratings were significantly worse for a "negative end" compared to a "negative start").

The second experiment (n=73) modified the experimental method in order to accentuate the magnitude of momentary sensations, and added new conditions in order to isolate primacy and recency effects. The second study found significant effects of recency on preferences for both positive and negative endings – that is, participant preferences were influenced more strongly by the final moments of the sequence. No effects of primacy were observed – that is, positive or negative moments at the start of the sequence did not reliably influence preferences.

In the discussion, we examine the results from these and related studies to identify interaction conditions and stimuli likely to influence user experience through sequencing effects, and we identify implications for UX designers and HCI researchers. The existence of sequencing effects, even for simple repetitive tasks in user interfaces, shows that designers must think carefully about the ordering of experience in addition to more traditional aspects of design such as usability, utility, and aesthetics.

Overall, this research provides new evidence that sequencing effects can have a significant impact on user experience and user preferences, even in common interactive tasks. We also contribute to a new understanding of how, where, and why these effects can influence preference, including a discussion of the types of interactive stimuli that are likely to induce reliable effects, as well as establishing an agenda for further research. Our studies suggest that consideration of sequencing effects can provide new opportunities for improving user experience in interactive systems, as well as warnings to designers and researchers who study UX and usability.

2 Related Work

When people evaluate or assess experiences (such as surgical procedures, periods of their lives, or interaction tasks), they rely on their memory of what happened and how they felt during the experience. Although people often believe that their memories faithfully represent experience, psychological research has identified systematic differences between what is objectively experienced during an episode, and the later recollection of those experiences. For example, memories of pain are strongly influenced by the severity and the recency of the discomfort, rather than the total amount or duration (Kahneman, 2000a; Kahneman, 2000b). These systematic biases between objective and remembered experience can influence people's *judgement* about the experience – for example, whether they would voluntarily repeat it, or would prefer it to some alternative experience.

In this section we review prior work from psychology that models how we experience and remember our experience of events and episodes, and focus in particular on theories of *recency, peak-end*, *duration neglect,* and *primacy*. We also review HCI research that seeks to improve our understanding of hedonic and affective experience in interactive contexts, with an emphasis on prior studies of sequencing effects in interaction.

2.1 Experience by Moments

Kahneman, Wakker and Sarin (1997) presented a model of experience that discretises an event into a series of *moments* (further developed in Kahneman, 2000a; Kahneman, 2000b). They argued that people experience events (such psychological experiments, surgical procedures, and personal relationships, as further described below) as a series of discrete *moments*, rather than as a continuous stream. The construction of these moments, and biases in one's memories of them, introduces discrepancies between the recollection of an experience and what was actually experienced.

The idea of momentary memory implies that we do not store our experiences in perfect experiential and temporal fidelity – rather, memories are formed from snapshots of the representative moments in an experience, with a bias towards the most significant or interesting episodes. Researchers who analyse these moments consider three types of *utility* (i.e. subjective worth as considered by economists; Kahneman *et al.*, 1997):

- *Instant utility*. The subjective experience of a moment at the instant that it is experienced: a "measure of hedonic and affective experience, which can be derived from immediate reports of current subjective experience" (Kahneman *et al.,* 1997, p. 376).
- *Total utility*. A temporal integral of the instant utilities for the moments that constitute an event: an objective measure of the total amount of pleasure or pain that was actually experienced.
- *Remembered utility*. A retrospective account of the total utility measure: what a person remembers and is able to recall about their experience of an event.

Total utility is considered an objective measure: it is an account of the total pleasure and pain experienced by a person during an event from real-time measures as it is experienced. Kahneman et al. (1997) describe two rules that govern total utility:

1. *Separability*. The order in which moments are experienced does not affect total utility.
2. *Time neutrality*. All moments are weighted equally, and total utility does not diminish over time. That is, any time gap between the experience and an assessment of that experience does not change total utility.

However, experimental research has found that these two rules do not hold for remembered utility (i.e. the retrospective account of total utility; Varey and Kahneman, 1992). Instead, remembered utility is vulnerable to psychological biases and frailties of memory, three of which we focus on here:

1. *Peak-end*. The most intense (i.e. most pleasurable or most painful) and terminating moments of an event have a disproportionally high influence on its remembered utility.
2. *Duration neglect*. The total duration of an event has little influence on its remembered utility.

3. *Violations of temporal dominance*. The remembered utility of a negative event can be increased by extending it with moments that reduce the average pain, even if they increase the overall (summed) pain.

2.2 Recency and Peak-End Effects

The *peak-end* rule states that people's memory of an experience is influenced by its peak moment (either positive or negative) and its final moments. The rule therefore encompasses the influence of two different types of momentary experience (peak intensity and terminating experience), although experiments into the effect often address only one of the two.

In a now-famous experiment investigating the influence of the terminating moments of experience, Kahneman et al. (1993) conducted a 'cold-pressor' experiment, which had participants submerge their hands in water under two conditions: *short* and *long*. Both began with the participant's hand in unpleasantly cold water (14°C) for one minute. In the *short* condition, participants then removed their hand from the water; but in the *long* condition, they kept their hand submerged for an additional 30 seconds while the water was surreptitiously warmed to a slightly-less-unpleasant 15°C. When participants were asked which trial they would prefer to repeat, most chose the *long* condition: that is, they preferred a longer unpleasant experience that had a slightly less unpleasant ending, effectively choosing an experience with more pain over one with less.

Participants were also asked to report a real-time measure of their discomfort (their instant utility) using a potentiometer. The discomfort of the first minute was comparable between conditions, but the final 30 seconds of the *long* trial had a significant drop in discomfort for most participants (the main finding was robust when replicated without this measure). That is, participants were able to recognise the drop in discomfort at the end of the *long* condition, and despite also correctly identifying that the trial lasted longer, they felt it had less overall discomfort (their remembered utility). This apparent conflict was attributed to the peak-end effect: reports of subjective experience were dominated by the most intense moment and the terminating moment.

The peak-end rule was also observed in retrospective reports from patients of a colonoscopy procedure (Redelmeier and Kahneman, 1996; Redelmeier *et al.*, 2003), whose experience was dominated by their peak discomfort and by their discomfort at the end of the procedure – and was unrelated to the procedure's duration (which varied between 4 and 69 minutes). Ariely (1998) also observed the rule in two experiments that applied pain to participants using a heating element or vice: they found a peak-end effect in global retrospective evaluations of the experience, and mixed results for the influence of the experience's duration.

Judgements about pleasurable experiences are also subject to the peak-end rule. For example, Do et al. (2008) examined the perceived pleasure in receiving gifts. In their first experiment, two lists of DVDs were produced ('A' and 'B') and given to participants as part of a raffle. The 'A' list was populated with high-rated movies; the 'B' list with low-rated movies. Participants reported their pleasure on receiving either: (a) one 'A' movie,

(b) an 'A' movie and a 'B' movie later, (c) a 'B' movie and an 'A' movie later, or (d) one 'B' movie. Receiving a 'B' movie alone was rated positively, but receiving an 'A' movie followed by a 'B' movie was rated worse than receiving an 'A' movie alone. These findings were further supported in a second experiment that gave sweets to children on Halloween. They concluded their paper with advice for gift giving: "you might consider giving only the best one – or at least making sure that you give the best one last" (p. 98).

2.2.1 Duration Neglect and Violations of Temporal Dominance

Recollections of experience are also subject to *temporal* biases – such as duration neglect and violations of temporal dominance – in which people's memory of time does not accurately reflect the actual duration of events. These effects are closely related to the peak-end rule – for example, results of both the cold-pressor and colonoscopy studies found that participants ignored the duration of the events in their attention to the peak and end moments. However, duration neglect and violations of temporal dominance involve different effects on memory.

The assumption of *temporal dominance* suggests that adding negative moments to an experience should make the overall experience more negative. However, studies have consistently found that this assumption is incorrect. For example, Fredrickson and Kahneman (1993) exposed participants to films containing either pleasant or aversive imagery, and recorded both their real-time and retrospective global affect ratings. They found that the duration of the films had only a small effect on the retrospective evaluation, and that evaluations appeared to be based on a weighted average of the moments in the experience – that is, adding less-negative imagery to a substantially negative experience improved the reported evaluation of the overall experience.

Schreiber and Kahneman (2000) conducted a similar series of experiments using unpleasant sounds of varied loudness and duration. Although they found peak-end responses to the stimuli, they did not find duration neglect: the duration of the experience had an additive effect on remembered utility. However, they did observe violations of temporal dominance: extending an aversive sound with a less aversive one improved the remembered utility of the total experience.

Related effects have been observed in many other fields, including in article pricing (Nasiry and Popescu, 2011) and effortful study (Finn, 2010). However, the effects are sometimes subtle: for example, experiments involving simulated payment sequences (Langer *et al.*, 2005) found that peak-end effects were not observed when participants were focused on the experimental manipulation, but that they became significant when participants were concurrently engaged in a distractor task. Experiments on gastronomic experiences validated duration neglect and demonstrated preference for increasing pleasantness across the courses of a meal, but failed to validate reliable effects of peak or end experience (Rode *et al.*, 2007).

2.2.2 Primacy Effects

In contrast with *peak-end* and *recency* effects, *primacy effects* cause stronger memories for experiences that are encountered first. Many experiments have confirmed that in free recall tasks, the probability of recalling an item follows a roughly U-shaped curve across the serial position of cued items (Craik and Lockhart, 1972; Murdock Jr, 1962; Page and Norris, 1998): initial items are recalled more reliably than those in the middle of the set, and terminating items are recalled most reliably. Primacy effects are also reflected in experiments on anchoring effects, which have shown that initial data items or experiences have a strong influence on the outcome of decisions, judgments and computations. For example, Tversky and Kahneman (1974) found that participants' rapid estimation of the product of the integers 1 through 8 was significantly lower (a median answer of 512) than that for the reversed order 8 to 1 (a median answer of 2250). Similarly, a study by Zauberman et al. (2006) examined participants' ratings of job applicants after watching a dynamic graphical pattern representing the applicant's correct and incorrect responses to 30 test questions. The patterns were manipulated to show increasing or decreasing rates of correct responses. Their findings showed strong primacy effects for informational assessments (i.e. reporting the count of questions answered correctly) rather than affective ones (i.e. overall satisfaction with the candidate); in contrast, recency effects showed the reverse (strongest for affective, less strong for informational).

Whereas the underlying causes of recency effects are generally attributed to the volatile limitations of short-term memory – as new data or experiences arise, they overwrite previously stored information (Craik and Lockhart, 1972) – the memory systems underlying primacy are not well understood. Opinions differ on whether initial items are recalled better because they offer increased opportunities for rehearsal (Rundus, 1971), because they lack interference with other items (Crowder, 1982), or because of the contributions of other memory systems (Davelaar *et al.*, 2005; Henson, 1998).

Regardless of the mechanisms underlying primacy and recency effects, they are robustly observable in various forms of memory tests (Montgomery and Unnava, 2009; Weiss *et al.*, 2014).

2.3 Expectation Theory

The effects described above (regarding primacy, recency, and peak-end) involve human memory systems, and this is the theoretical framework that we use in this paper. An alternative theory, however, has also been proposed to explain sequencing effects – using the idea that manipulating user expectations can influence people's assessment of experiences.

For example, a recently study by Michalco et al. (2015) examined assessments of video games. Expectations about video game quality were set through positive or negative reviews; participants then played the games and rated them. The study found that participants who felt a game failed to meet expectations rated the game negatively (on several affective measures), while those who felt it exceeded expectations gave positive ratings. Michalco et al. framed their studies around the idea of *expectation* *disconfirmation* (Oliver, 1977) – a framework for contemplating the relationship between expectations and outcomes that was prevalent in marketing research in the 1970s and 80s. Expectation disconfirmation does not predict any empirical outcomes; rather, various adaptation theories (including contrast theory (Sherif, 1961), assimilation theory (Sherif, 1961), and cognitive dissonance theory (Festinger, 1957)) explain how disconfirmation is resolved into subjective preferences. For example, under contrast theory the surprise of a disconfirmation causes perceptions to be exaggerated (an outcome that is positively disconfirmed is perceived as more positive than is objectively the case, and vice versa), while under assimilation theory disconfirmation is psychologically uncomfortable and is resolved by distorting the perception of performance to bring it closer to what was expected.

Expectation theory can appear to provide a useful framework for explaining sequencing effects as well. However, the proposed theories are incompatible with one another, and their application has been criticised (Yi, 1990). Even if studies of sequencing effects can be interpreted within the framework of expectation disconfirmation, the framework remains non-predictive (or rather, it predicts every outcome). For example, in Kahneman et al.'s cold-pressor experiment, participants might have set their expectation for the terminating moments of experience during the first minute. The lessening of pain during the extended condition might have exceeded expectations leading to positive ratings (under contrast theory), or the experience may have been adapted to match expectations leading to neutral or negative ratings (under assimilation theory).

In Michalco et al.'s study, the difficulty with expectation theory becomes clear, because in interpreting their findings, Michalco et al. accept two *opposing* adaptation theories to explain their pattern of results: "when expectations were disconfirmed, contrast theory appears to apply ... and when expectations were confirmed, assimilation theory seems to apply" (p. 615). As a result of these theoretical incompatibilities and the lack of predictive power, we do not consider expectation theory further, and interpret our empirical results using the memory-based theories of sequence effect described above.

2.4 Sequencing Effects in Interaction

Although much of the research into sequencing has taken place outside of HCI, there are a few studies that consider these effects with a variety of interactive sequences. Hassenzahl and Sandweg (2004) examined correlations between perceived usability of an interactive system and various measures characterising mental effort during an interactive task. Measures of mental effort included intensity of the whole series, variations in intensity, end intensity, peak intensity divided by end intensity, and trend (reducing or increasing mental intensity). Results found that of these measures, end intensity was best correlated with summary measures of perceived usability ($r = -.49$). They described this as a *recency effect*, stating: "Individuals construct their summary assessment on the basis of what comes to their mind about the episode they just experienced. The more recent a detail, the more easily it comes into mind" (p. 1285). Our studies expand on this finding in several ways, including analysis of *preferences* between interface series and use of tasks that are not immediately tied to mental effort.

In a study of progress bar behaviours, Harrison et al. (2007) demonstrated that experimental participants' rated progress bars of objectively-equal duration as being faster when their rate of progress accelerated across time. Harrison et al. attributed this finding to peak-end effects. A later study (Harrison *et al.*, 2010) examined animation effects within progress bars, finding generally similar results (i.e. accelerating animations were generally perceived as faster than others) – although the animation that was perceived as fastest used a decelerating backwards pulse.

Cockburn et al. (2015) directly examined user preferences derived from peak-end effects. Their conditions varied the distribution of otherwise identical components of mundane physical workload in interactive tasks, with participants making a preference choice between two interactive series. Both series presented five screens containing a number of sliders that had to be set to exact numerical values; there were always a total of 25 sliders, but the number of sliders on each page was manipulated to create peak, end, and peak-and-end experiences. Pages with few sliders were hypothesised to induce momentarily more positive experiences (as they involved less effort) than pages with many sliders – for example, the 'positive peak-and-end' condition used only 2 and 3 sliders on the second and last pages, while the 'negative peak-and-end' condition used 8 and 6 slides on the last two pages. Participants' preference choices conformed to the peak-end rule in a condition that combined peak and end effects, but the conditions that separately manipulated end and peak effects did not show significant effects on preference.

Gutwin et al. (2016) examined the application of the peak-end rule in user experience with interactive computer games. A first experiment manipulated the difficulty in two different games to induce positive, neutral and negative peak-end effects. Of the six tested preference effects, only one (positive vs. negative in a Bejeweled-style game) supported the hypothesised peak-end effect. A second experiment then manipulated the sequence of skill-challenge balance by first assessing a participant's ability and then generating game sequences that altered the challenge to produce sequences that varied between *appropriate challenge, too easy,* and *too hard.* Significant effects on the participants' perception of challenge were observed, but there were no effects on preference.

To our knowledge, primacy effects have received little attention in HCI. In a study of user preferences for computer-synthesised voices, Lee et al. (2011) explicitly refer to primacy as an explanation for their participants' general preference for the first voice heard. Similarly, Harrison's (2007) progress bar study (described above) observed a significant preference for the first observed behaviour, despite a counter-balanced design – however, they do not refer to primacy, or any other explanation, for the result.

3 Evaluating Sequencing Effects in Interaction

The prior results on recency, primacy, and peak-end effects in interaction provide evidence that sequencing can influence user preferences – however, the interaction studies described above are much more equivocal than earlier experiments in psychology. One factor contributing to the difference in effect between these two domains could be the very different stimuli evaluated in each domain. All of the stimuli examined in studies of interactive sequences are arguably less visceral than the stimuli typically examined in psychology experiments: for example, HCI studies have looked at mental workload (Hassenzahl and Sandweg, 2004), perception of progress bars (Harrison et al., 2007; 2010), physical workload (Cockburn et al., 2015), and game difficulty (Gutwin et al., 2016); in contrast, psychological studies have used physical stimuli such as the pain of cold water immersion or surgical procedures, or the pleasure of receiving gifts.

Interactive experiences are not usually 'pleasurable' or 'painful' in the same respect as the stimuli used in these psychological studies. Rather, momentary subjective experiences are derived from qualities of the interaction, such as the frustration or satisfaction associated with task completion through the interface. These experiences may lead to greater variance in the effects of sequencing – but it is important to note that this variability does not reduce the importance of studying sequencing as an HCI phenomenon. Prior studies have shown that when a sequencing effect occurs, it can have a substantial influence on preference – the variability, therefore, arises primarily in terms of *when* and *where* sequencing has effects, rather than the impact of those effects. This situation makes it particularly difficult for designers to know what to do – sequencing effects can have important effects on user retention and user experience, but it is not clear exactly when and where the effects will occur, and not clear how other factors in the interactive setting may interact with sequencing.

Given the small number of studies in interactive contexts, it is important to further explore the applicability and strength of sequencing effects. In the studies described below we report on one such further exploration that examines these issues for a simple interactive task. To begin, we consider two issues that are critical to the design of experiments that test sequence effects in interaction: (1) determination of an interactive stimulus that can reliably induce momentary positive and negative sensations in users; and (2) determination of a method for observing and measuring sequence effects.

3.1 Stimulus for Momentary Experience

To provide a reliable interactive stimulus for positive and negative momentary sensations, we selected *assistive user interfaces,* which can be either successful or unsuccessful in helping the user. Assistive interfaces involve 'smart' interactive features that attempt to infer a user's intention and reformulate their input to match that inference. Examples include text entry auto-correct on mobile phones, auto-formatting in word processors (e.g. inferring that a bullet or enumerated list is intended), and various forms of 'snap-to' object manipulation (e.g. snap-to-grid). When the interface correctly infers the user's intention (e.g. replacing a mistyped word), the user is likely to feel a positive momentary sensation of satisfaction due to the interface's task assistance. However, when the interface incorrectly infers the user's intention (e.g. replacing a correctly typed word with an incorrect one), the user is likely to experience a negative sensation due to the need to correct the system's mistake.

The two experiments described below explore snap-to-grid assistance during drag-and-drop interactions. In our simulated assistance system, successful snap-to-grid (where the target aligns with the grid) assisted task completion, and unsuccessful snap-to-grid (where the target is misaligned with the grid) hindered task completion – requiring the user to drop the dragged object and reacquire it with a modifier key that overrides the assistance. We chose this snap-to-grid mechanism for four main reasons:

1. Snap-to-grid is a widely used and easily understood interaction; this familiarity provides participants with a clear impression of the quality of assistance from the interface.

2. It is easy to precisely control the magnitude of the assistance or hindrance by altering the grid resolution.

3. Snap-to-grid eliminates many uncontrolled variables that occur in other assistive domains. For example, if text-entry autocorrect were used, it would be difficult to control the timing and presence of user errors, potentially requiring artificial constraints on error-free performance that may influence the results (Quinn and Zhai, 2016).

4. A recent analysis of snap-to-grid behaviours showed that this manipulation can provide a reliable and sensitive method for analysing user preferences (Quinn and Cockburn, 2016).

3.2 Measurement of Sequencing Effects

In addition to determining a stimulus for creating momentary positive and negative subjective experiences, a method is also required for measuring the influence of sequencing effects. Prior psychological studies have asked simple and broad questions in order to measure participants' subjective experience – e.g. "how much discomfort did you experience?" on a five-point Likert scale (Fredrickson and Kahneman, 1993). In contrast, HCI studies of subjective experience have often targeted more specific qualities of subjective experience: for example, pragmatic vs. hedonic quality (Hassenzahl *et al.*, 2015); or perceived competence, autonomy, relatedness, and enjoyment from self-determination theory (Ryan *et al.*, 2006). This added specificity could be valuable in studying user experience with sequence effects – however, an issue with measuring the magnitude of a subjective experience is ensuring reliability of the measure (that is, ensuring all participants interpret the question in the same way, and that their response to the measurement scale is comparable). This issue is particularly problematic when the manipulations are subtle and may not influence all participants equally – for example, whereas Kahneman et al.'s original cold-pressor experiment adjusted water temperature by an objective 1˚C, the gains provided by an assistive pointing technique will vary with each user's pointing ability.

These difficulties with scalar measurements lead us to forced preference choices between two alternatives. This is also a common method used in psychology for measuring subjective experience – e.g. "for today's third trial, you can pick which of the previous cold-water trials you will repeat" (Kahneman *et al.*, 1993). A forced-choice preference measure is valuable because it is not open to interpretation about any particular qualities of the experience – it is a truly holistic measure. While it does not give an indication about the magnitude of the preference, it can be tested repeatedly with different pairs of interfaces to explore different aspects of the experience. In addition, choosing which sequence the participant would prefer to repeat matches the overall design problem of creating user experiences that maximize user retention and likelihood of re-use. These advantages of forced preference choice lead us to use it as the primary dependent measure in both of the studies described in the next two sections.

4 Experiment 1

The goal of the first experiment was to determine whether user preferences for interactive experiences are more influenced by recency or primacy effects. The experiment used the simple and common interactive task of drag-and-drop: selecting an object with a mouse cursor, dragging it to a target location, aligning it precisely with the target, and dropping it. Participants completed sequences made up of several individual drag-and-drop tasks, and for each pair of sequences, they were then asked which series they would choose to repeat (i.e. a forced-preference choice). The collection of individual tasks in each series within a pair was identical, but the order of presentation was manipulated to create a different ordering of positive and negative momentary experiences.

There were three types of individual tasks that could be combined to form a sequence, each designed to create a different type of momentary experience:

1. *Control* dragging (neutral momentary experience). When dragged with the mouse cursor, the object moved on a fine-grained grid (4×4 px), which approximated the continuous dragging used in many traditional drag-and-drop settings. The 4×4 px grid was used to simplify the final drop action, so that participants did not need to position the object perfectly on the target before dropping. The fine-grained grid did not detract from the smoothness of object movement, however, and was not obvious to participants.

2. *Positive* snapping (positive momentary experience). The object moved on an 84×84 px grid while it was being dragged, and the target was aligned with the grid (see Figure 1). *Positive* snapping assisted participants in completing their drag-and-drop task (compared to *Control*) by reducing the pointing accuracy required to bring the object into alignment with the target;

3. *Negative* snapping (negative momentary experience). The object moved on a 224×224 px grid during dragging, but the target was misaligned with the grid.

After dropping the object as near as they could move it towards the target, participants had to reselect it while holding down the 'control' key, which temporarily switched the interface back to *Control* dragging (see Figure 2). *Negative* snapping hindered the user in completing the drag-and-drop task by introducing this extra step to each task.

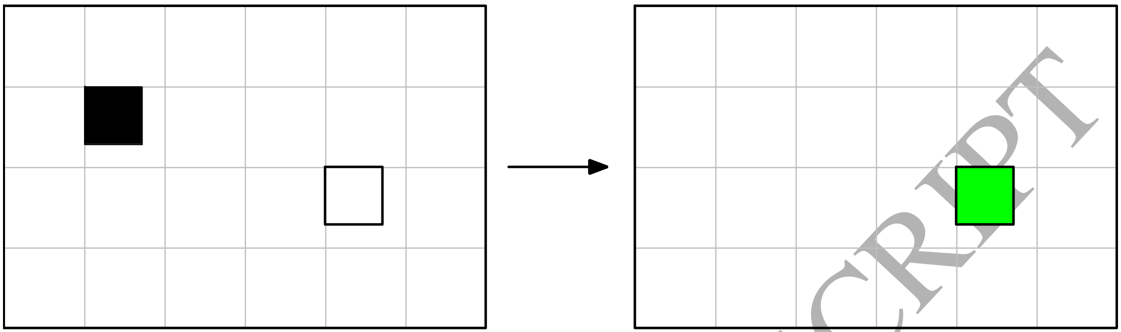


Figure 1. Positive snapping: the dragged object snaps to the grid, and the target is aligned with the grid.

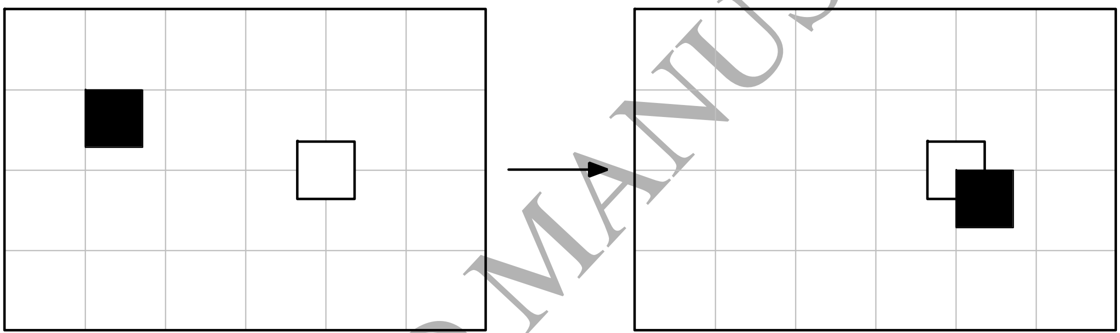


Figure 2. Negative snapping: the target box is misaligned with the snapping grid. To complete the drag-and-drop task (not shown here), the user must drop the object and reacquire it with the 'control' key pressed, at which point the object can be moved with a fine granularity (on a 4×4 px grid).

Control dragging served as a baseline behaviour, with the intention that *Positive* and *Negative* snapping trials would induce momentary positive and negative sensations relative to it.

The experiment asked participants to complete four pairs of drag-and-drop sequences, and make preference choices for one sequence in each pair. First, participants completed the two Manipulation Check (MC) conditions mentioned above:

1. **MC1.1** *Positive snapping is preferred to Control dragging.* This condition is included to validate the assumption that participants prefer *Positive* snapping tasks to *Control* dragging tasks.

2. **MC1.2** *Control dragging is preferred to Negative snapping.* This condition is included to validate the assumption that participants prefer *Control* dragging tasks to *Negative* snapping tasks.

These were followed by two conditions that tested whether recency is stronger than primacy:

3. **H1.3** *A positive end is preferred to a positive start.* The two series are identical, except for the location of four *Positive* snapping tasks. In the *+Start* series, the *Positive* tasks are the first four, and in the *+End* series they are the final four.

4. **H1.4** *A negative start is preferred to a negative end.* This condition is the negative equivalent of the positive manipulation in H1.3. The two series are identical, except for the location of four *Negative* snapping tasks. In the *-Start* series, the *Negative* tasks are the first four, and in the *-End* series they are the final four.

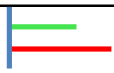
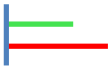
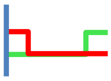
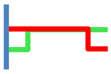
	Hypotheses		**Experience**	**Preferred (%)**			**Frust.-Satis. (median/mean)**			
	H. Preferred (A)	**H. Reject (B)**		**A**	**B**	***p***	**A**	**B**	***Z***	***p***
1.1	Positive 10P	Control 10C		87	13	< **.001**	4 3.97	2 1.13	5.5	< **.001**
1.2	Control 10C	Negative 10N		72	28	= **.004**	1.5 1.57	-1 -1.13	4.9	< **.001**
1.3	+End 10(CN)+4P	+Start 4P+10(CN)		63	37	= .1	0 0.0	0 -0.28	1.2	= 0.23
1.4	-Start 4N+10(PC)	-End 10(PC)+4N		63	37	= .1	1 0.85	0 0.39	2.0	< **.04**

Table 1. Experiment 1 hypotheses: the hypothesised preferred (A, green) and rejected (B, red) series, and their task construction – for example, 10(CN)+4P indicates 10 *Control* and 10 *Negative* trials (in a random order), followed by 10 *Positive* trials. The 'experience' plot in the middle of the table shows the hypothesised subjective experience across time for the predicted preferred (green) and rejected (red) series. Preference choice data and responses on the Frustration-Satisfaction scale are shown on the right.

The specific sequences and their associated hypotheses are summarised in Table 1. In the table, C, P, and N respectively indicate *Control, Positive,* and *Negative* tasks, with numerals indicating the number of each task to be completed. For example, the notation "10(CN)+4P" indicates 10 *Control* and 10 *Negative* tasks (in a random order), followed by 4 *Positive* tasks. The baseline experiences within hypotheses 1.3 and 1.4 consisted of a mixture of control and negative (10(CN)) or control and positive (10(PC)) trials because in our pilot testing we felt that this mixture increased the salience of the key manipulation, which was the timing within the sequence of the four positive or four negative trials. For example, a positive trial is more noticeably positive when contrasted with a mixture or control and negative trials, as compared to its contrast with a series comprising only control trials. The 'Experience' column in Table 1 shows characteristic plots of the subjective experience (positive/negative) for each of the hypothetically preferred (green) and not preferred (red) task series for each of the conditions (described below).

4.1 Experimental Task

All drag-and-drop trials involved selecting a 50×50 px black-filled square and dragging it onto a target object demarked by a 50×50 px unfilled square with a 1 px black outline. The target was 252 px away from the starting location. When selected, the black square turned blue, and when exactly aligned with the target it turned green.

Control dragging tasks used a fine-grained 4×4 px grid (approximately 1.3×1.3 mm on our experimental apparatus) in order to ease the difficulty of placing the object precisely onto the target. Dragging an object on this fine-grained grid was not noticeable to participants as 'snapping' – movement appeared to be smooth and continuous. The tasks in this condition had a theoretical Fitts' Law index of difficulty (ID) of 6 bits (Fitts, 1954; MacKenzie, 1992). The top plot of Figure 3 illustrates typical task completion with *Control* dragging: the user rapidly reduces the distance to the target, may overshoot the target, and slowly aligns the object with the target.

Positive snapping tasks used an 84×84 px grid (28×28 mm), with the target object always aligned to the grid. As the user dragged the object, it snapped between locations on the grid (the grid was never visible – i.e. the gridlines in Figures 1 and 2 were not shown to participants). The index of difficulty of *Positive* dragging tasks was 2 bits, theoretically permitting faster task completion than *Control* tasks (because the grid makes the target substantially larger in motor space). The middle plot of Figure 3 characterises task completion with *Positive* snapping: the distance to target reduces in discrete steps (point ❶ in the figure), with the final step rapidly completing the task.

Negative snapping tasks used a 224×224 px grid, with the target object always misaligned from the grid by 28 px – this meant that the closest that the user could drop the object was 28 px from the target. As the user moved the object, it snapped to the grid (point ❶ in the bottom plot of Figure 3), but the misalignment required participants to drop the object and reacquire it with the control key pressed (point ❷), at which point the object could be dragged using the fine-grained grid of the *Control* condition (point ❸). Final object placement behaved identically to *Control* tasks. The total index of difficulty for the two pointing components of negative snapping (initial movement of the object to 28 pixels from the target ❶ and final placement using the fine-grained grid ❸) was 4 bits. However, this difficulty value does not reflect the total time needed for the *Negative* tasks, since time was also required to drop and reacquire the object (region ❷).

4.2 Method and Procedure

Participants were instructed that they would make several choices between two series of drag-and-drop tasks, choosing which they would prefer to repeat if asked to do so. They then completed a series of 18 practice tasks to gain familiarity with the *Control*, *Positive* and *Negative* snapping behaviours. These practice trials made clear to participants that in *Negative* conditions, the object could not be placed directly on the target, and all participants correctly adopted the strategy of moving the object near to the target, then reacquiring it with the control key and completing the task.

Participants then moved on to the four experimental conditions (Table 1). Conditions MC1.1 and MC1.2 were completed first, with the recency conditions H1.3-1.4 completed in a random order. For each of the conditions, the order of exposure to hypothetically preferred and non-preferred series was counterbalanced.

Upon completing the first series in each condition, participants used a slider to rate how frustrating (extremely, -5) to satisfying (extremely, +5) they found that series. After completing the second series, they used the same slider to rate that series, and then responded to our primary question by choosing which of the two series (first or second) they would repeat if asked to do so. On selecting their preferred series, they continued to the next pair of series after a rest period of 10 seconds.

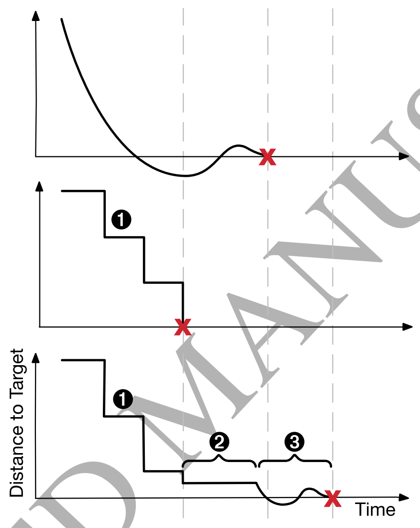


Figure 3. Characteristic plots of distance to target by time for *Control* dragging (top), *Positive* snapping (middle) and *Negative* snapping (bottom).

4.2.1 Apparatus

The study ran on Intel Core i7 computers running Linux Mint 14, with a wired Logitech optical mouse and a 22″ monitor at 1680×1050 px. Experimental software was written in Python, and recorded all user actions and responses.

4.2.2 Participants

Forty-six undergraduate computer science students volunteered to participate (9 female; aged 17–46, mean 20.2), each receiving a $5 café voucher.

4.2.3 Design

The primary dependent variable was participants' preference choice, separately analysed using two-tailed binomial sign tests for each of the four experimental hypotheses (manipulation checks **MC1.1** and **MC1.2,** and recency-versus-primacy hypotheses **H1.3** and **H1.4**). Timing data for the conditions were also analysed using paired *t*-tests to confirm there was no significant difference in time between series in the recency conditions that may explain preference choices (details given below).

4.3 Results

Results are analysed separately for each of the four conditions. Overall proportions choosing each series in each condition are shown in Table 1.

Manipulation Check 1.1: Positive vs. Control. As intended, a significant majority of participants (40 of 46 people, 87%, 95% CI [.74, .95]) chose the *Positive* snapping series rather than the *Control* series (binomial sign test, $p < .001$). Part of the preference for positive snapping may be attributable to its much faster performance (*Positive* mean time 7.62 s, 95% CI [7.06, 8.18]) compared to *Control* (mean 17.15 s, 95% CI [16.27, 18.02]): $t_{45} = 20.78$, $p < .001$, $d = 3.8$. Similarly, participants' ratings of frustration-satisfaction for the *Positive* series (median 4.0) were significantly higher (indicating less frustration) than they were for *Control* (median 2.0): Wilcoxon Signed-Ranks Test $Z = 5.5$, $p < .001$.

We therefore accept **MC1.1**, and conclude that the *Positive* manipulation was effective in creating a positive experience (since *Positive* snapping trials were preferred to *Control* trials).

Manipulation Check 1. 2: Control vs. Negative. Also as intended, a significant majority of participants (33 of 46 participants, 72%, 95% CI [.57, .84]) chose the *Control* series over the *Negative* series (binomial sign test, $p = .004$). Like MC1.1, part of this preference may be attributed to the *Control* series being much faster (mean time 17.08 s, 95% CI [16.40, 17.75]) than the *Negative* series (mean 27.48 s, 95% CI [24.77, 30.19]): $t_{45} = 7.80$, $p < .001$, $d = 1.6$. Similarly, participants' ratings of frustration-satisfaction for the *Negative* series (median -1.0) were significantly lower (indicating greater frustration) than they were for *Control* (median 2.0): Wilcoxon $Z = 4.9$, $p < .001$.

We therefore accept **MC1.2**, and conclude that the *Negative* manipulation was successful (since *Control* trials were preferred to *Negative* trials).

Hypothesis 1.3: +End vs. +Start. Twenty-nine of the 46 participants (63%) chose the *+End* condition in preference to the *+Start* condition, but the difference was not significant (binomial sign test, $p = .1$). The two series, which were the inverse of one another in terms of task order, required similar completion times, as intended (a mean difference of 0.23, 95% CI [-1.47, 1.94], $t_{45} = 0.28$, $p = .78$, $d = 0.03$). Participants' ratings of frustration-satisfaction for *+End* (median 0) were similar to those for *+Start* (median 0), giving no significant difference (Wilcoxon $Z = 1.2$, $p = 0.23$).

We therefore are unable to accept **H1.3**, as the hypothesised preference for the *+End* condition was not significant. A majority of participants conformed to the hypothesis (63%), and this proportion is not far from that reported in Kahneman's cold-pressor study

(69%). Nevertheless, there is no clear evidence that participants are more strongly influenced by the final moments of a sequence than they are by the initial moments.

Hypothesis 1.4: -Start vs. -End. Twenty-nine of the 46 participants (63%) chose the *-Start* condition in preference to the *-End* condition (binomial sign test, $p = .1$). As with hypothesis H1.3, the two series are order inverses of one another, so the times required for each condition were similar, as intended (mean difference 0.66 s, 95% CI [-0.70, 2.02], $t_{45} = 0.98$, $p = .33$, $d = 0.14$). Participants' ratings of frustration-satisfaction for *-End* (median 0.0) were significantly lower (i.e., more frustrating) than those for *-Start* (median 1.0): Wilcoxon Z = 2.0, $p = .04$.

Therefore, while we cannot accept **H1.4**, there are indications that a different experiment might expose the effect. On the primary preference measure, a larger proportion of participants preferred the *–Start* sequence, but the proportion is again insufficient to show a statistically-significant difference. There was a significant difference in the frustration-satisfaction scale, but overall, we do not have clear evidence that participants are influenced more by final moments than by initial moments.

4.4 Discussion

The main objective of Experiment 1 was to test whether recency effects strongly influence preferences in interaction (i.e. they can outweigh primacy effects). Although the manipulation checks confirmed that we were correctly creating positive and negative experiences, the experimental conditions provided only weak evidence for the influence of recency effects. We next examine possible reasons for the overall lack of evidence, before addressing these limitations in the design of Experiment 2.

4.4.1 Weak Momentary Stimuli

The *Positive* and *Negative* tasks were designed to create momentary positive and negative experiences for the user. Although the manipulation check conditions showed the anticipated preferences for the *Positive* and *Negative* tasks in isolation, they were not as strongly preferred as anticipated. While 87% preferred the *Positive* condition to *Control* (MC1.1), only 72% preferred the *Control* condition to *Negative* (MC1.2).

When considering that snapping allowed the *Positive* MC1.1 series to be completed in less than half the time of the *Control* series (7.6 s vs. 17.2 s), and with much less manipulation effort, it is surprising that 13% of participants preferred the *Control* condition.

It is also surprising that 28% of participants selected the *Negative* MC1.2 series as preferred, especially when considering that it took 61% longer than the *Control* series to complete (27.5 s vs. 17.1 s). It also added a mechanical burden in having to drop and reacquire the object to override the misaligned grid. However, *Negative* trials were not without some element of assistance. In particular, the initial snapping behaviour allowed participants to attain a location close to the target (24 px distant) more quickly than they could without it. It is possible that this rapid attainment of an 'almost there' state was valued.

In Experiment 2, the magnitude of the Negative momentary experiences is accentuated by altering the grid resolution to reduce any positive sensation associated with rapid target approach. This is achieved by using an initial starting state that was misaligned with a large grid size, causing the object to actually snap further from the target than when it is first selected.

4.4.2 Unstable Baseline Experience

The main experimental conditions H1.3 and H1.4 altered the ordering of four *Positive* or *Negative* tasks (4P and 4N in Table 1) with respect to a baseline of mixed trials. In H1.3, the baseline consisted of 10 *Control* and 10 *Negative* tasks; and in H1.4 it consisted of 10 *Positive* and 10 *Negative* tasks. Although the intention of using a mixture of trial types within the baseline was to increase the salience of the key manipulation (the ordering of the four positive or negative trials), it is possible that the mixture caused the opposite effect. The three different types of snapping behaviour within one series of tasks may have created a pattern of experience that was more complex than intended, thereby reducing the salience of the intended manipulation.

In Experiment 2, we modify the baseline experience in each series so that it consists only of *Control* trials, providing a more consistent reference point for assessment.

4.4.3 Primacy Effects are Stronger than Anticipated

In creating task series that were objectively identical in total content, we relocated the terminating *end* experience (4P or 4N) to the *start* of the other series. Based on previous work in psychology, our hypothesis was that recency effects would be stronger than primacy effects – but the conflict between these two effects may have reduced the differences in Study 1 to the point where they were no longer statistically significant.

Experiment 2 adds conditions to isolate recency and primacy effects by comparing sequences that have the positive or negative stimulus located at the start, middle, and end of the series.

5 Experiment 2

The aim of Experiment 2, like that of Experiment 1, was to provide evidence that sequencing effects influence user preferences of interactive series. Experiment 2 used a modified version of the method used in Experiment 1. It maintained the forced-choice preference selection between pairs of sequences (again using drag-and-drop tasks), as summarised in Table 2. Following Manipulation Check conditions MC2.1 and MC2.2, participants completed three paired task sequences that examined both positive and negative versions of the following comparisons: End vs. Start, Middle vs. End, and Start vs. Middle. For example, condition H2.3 examined preferences for a series with a positive end (*+End*) against a series with a positive start *(+Start*), and H2.4 compared a positive end (+*End*) with a positive middle (+*Mid*). Experiment 2 also increased the magnitude of the negative momentary experience and stabilised the baseline experience, as described above.

Specific paired series conditions are summarised in Table 2. All participants first carried out the two Manipulation Check (MC) conditions that were similar to the MC conditions in Experiment 1:

1. **MC2.1** *Positive snapping preferred to control dragging.* This condition was included to validate the assumption that participants preferred *Positive* tasks to *Control* tasks. The two series contained 8 *Positive* tasks (8P) and 8 *Control* tasks (8C).

2. **MC2.2** *Control dragging preferred to negative snapping.* This condition was included to validate the assumption that participants prefer control dragging to negative snapping.

The experiment tested three different *Positive* stimulus hypotheses:

3. **H2.3** *A positive end is preferred to a positive start.* The two series were identical in total content, but four *Positive* snapping tasks occurred at either the start or the end of the series. This hypothesis is similar to H1.3 in Experiment 1, and is intended to compare the effects of primacy versus recency (but with the altered methodology of Experiment 2).

4. **H2.4** A *positive end is preferred to a positive middle* (positive recency without opposing primacy). Four *Positive* tasks occurred either at the middle or the end of the series. In contrast to the series tested in H2.3, any primacy effect (accentuating the positive start) should be eliminated. This hypothesis therefore tests recency effects in the absence of potentially opposing primacy effects.

5. **H2.5** *A positive start is preferred to a positive middle* (positive primacy without opposing recency). Four *Positive* tasks occurred either at the start or the middle of the series. This hypothesis tests a primacy effect – whether a positive start is remembered more strongly than the positive middle.

The experiment also tested three different *Negative* stimulus hypotheses:

6. **H2.6** *A negative start is preferred to a negative end.* Four *Negative* tasks occurred at either the start or end of the series. This hypothesis is similar to H1.4 in Experiment 1, and is intended to again compare the effects of primacy versus recency.

7. **H2.7** *A Negative middle is preferred to a negative end* (negative recency without opposing primacy). This hypothesis is the equivalent of H2.4, but using *Negative* tasks instead of *Positive* ones. In contrast to the series tested in

H2.6, any primacy effect (accentuating the *Negative* start) should be eliminated. The hypothesis tests recency effects in the absence of potentially opposing primacy effects.

8. **H2.8** *A negative middle is preferred to a negative start* (negative primacy without opposing recency). This hypothesis is the *Negative* equivalent of H2.5, testing possible primacy effects of a *Negative* start. If the *Negative* start is strongly remembered, then participants should prefer the *Negative* middle.

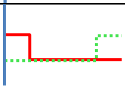
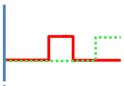
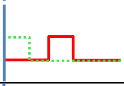
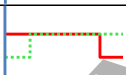
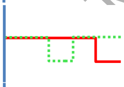
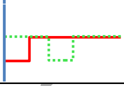
	9. Hypotheses		Experience	Preferred (%)			Frust.-Satis. (median/mean)			
	H. Prefer (A)	**H. Reject (B)**		**A**	**B**	***p***	**A**	**B**	***V***	***p***
2.1	Positive 8P	Control 8C		81	19	< **.001**	5 3.85	2 2.22	1773	< **.001**
2.2	Control 8C	Negative 8N		95	5	< **.001**	2 2.23	-2 -1.67	2217	< **.001**
2.3	+End 12C+4P	+Start 4P+12C		68	32	= **.047**	1 1.51	1 0.95	297	= **.008**
2.4	+End 12C+4P	+Mid 6C+4P+6C		81	19	< **.001**	2 1.97	1 1.24	330	= **.01**
2.5	+Start 4P+12C	+Mid 6C+4P+6C		51	49	= 1	2 1.32	1 1.13	222.5	0.42
2.6	-Start 4N+12C	-End 12C+4N		74	26	= **.006**	1 0.23	-1 -1.06	222.5	= **.008**
2.7	-Mid 6C+4N+6C	-End 12C+4N		60	40	= .31	0 -0.03	0 -0.63	270	= **.05**
2.8	-Mid 6C+4N+6C	-Start 4N+12C		37	63	= .18	0 0.34	0 0.23	86.5	= 0.98

Table 2. Experiment 2 hypotheses: the hypothesised preferred (A) and rejected (B) series, and their exact task makeup – for example, 12C+4P indicates 12 *Control* trials followed by 4 *Positive* trials. The 'experience' plot in the middle of the table shows the hypothesised subjective experience across time for the predicted preferred (green) and rejected (red) series (time axis not to scale). Preference choice data and responses on the Frustration-Satisfaction scale are shown on the right side of the table.

We also analyse the influence that the order of exposure to the series had on participants' preferences. Although series order was counterbalanced across participants, it is possible that preferences may vary dependent on presentation order – for example, it is possible that the influence of a negative ending to a sequence may dissipate over time, resulting in a stronger aversion to negative ending when it is more recently encountered (i.e., when experienced as the second series, rather than the first).

5.1 Experimental Task

As in Experiment 1, all trials involved selecting and dragging a 50×50 px black-filled square precisely onto a target object demarked by a 50×50 px unfilled square with a one-pixel black outline. The target was placed 252 px away from the initial location of the object. *Positive* and *Control* tasks in Experiment 2 were identical to those in Experiment 1. Negative trials were modified such that the grid had a resolution of 1022 px, with the target placed in the centre of the grid. Consequently, when the user first moved the object in *Negative* tasks, it snapped to align with the grid at a location that was *further* from the target (see Figure 4), requiring users to immediately drop and reacquire the object. This modification was used to remove the positive element of more rapid target approach from *Negative* tasks, and thereby increase their subjective negativity.

Baseline tasks within each series were also modified from Experiment 1. All series contained 12 *Control* tasks (rather than the mixture of 20 tasks used in Experiment 1), plus 4 *Positive* or 4 *Negative* trials. The experimental modification was made to create a simpler and more uniform baseline experience.

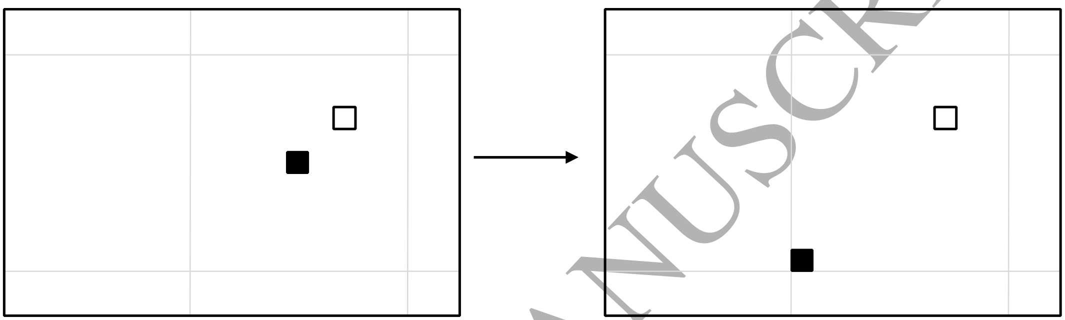


Figure 4. In *Negative* trials in Experiment 2, the black square was initially placed at the centre of the snapping grid of resolution 1022 px. The target was initially 252 px away from the black square (left). When first selected, the black square snapped to align with the grid, moving it further away from the target (right).

5.2 Method and Procedure

As in Experiment 1, participants were instructed that they would make several choices between two series of drag-and-drop tasks, choosing which they would prefer to repeat. They then completed a series of 18 practice drag-and-drop tasks to gain familiarity with the *Control*, *Positive* and *Negative* snapping behaviours.

Unlike Experiment 1, the positive (H2.1–H2.3) and negative (H2.3–H2.5) experimental conditions were divided as a between-subjects factor (i.e. each participant completed either the positive or negative conditions, but not both). This amendment to the within-subjects design of Experiment 1 was made to reduce the number of series encountered by each participant, and to avoid them becoming overwhelmed by the number of comparisons (16 series and 8 comparisons if exposed to all conditions) and therefore insensitive to the manipulation – a known confound in forced-choice experiments (Keren and Raaijmakers, 1988).

Therefore, after the practice trials, participants moved on to five experimental conditions, beginning with manipulation check conditions MC2.1 and MC2.2. Each participant then completed three pair comparisons based on their assignment to the positive or negative group. The three pairs (represented by H2.3–H2.5 and H2.6–H2.8) were presented in random order; in addition, the ordering of the sequences within each pair was also counterbalanced.

Participants' subjective experience was assessed after each series using the 11-point scale of Experiment 1 (extremely frustrating -5 to extremely satisfying +5). Following each pair of series, participants made a forced-choice response selecting the series they would prefer to repeat.

The apparatus and participant pool was the same as Experiment 1. Seventy-three undergraduate computer science students volunteered to participate (12 female; aged 18-44, mean 20.2). Preference choice, task completion time, and subjective frustration-satisfaction measures were analysed as in Experiment 1.

5.3 Results

Results for the eight experimental conditions are separately analysed in the following sections. In summary, three of the four hypothesised recency effects were supported, but none of the four hypotheses about primacy effects were supported. Table 2 summarises preference choices and mean/median Likert scale results, together with statistical outcomes. Figure 5 graphically summarises overall preferences choices for each of the eight conditions. Order effects for preference choices are summarised in Figure 6.

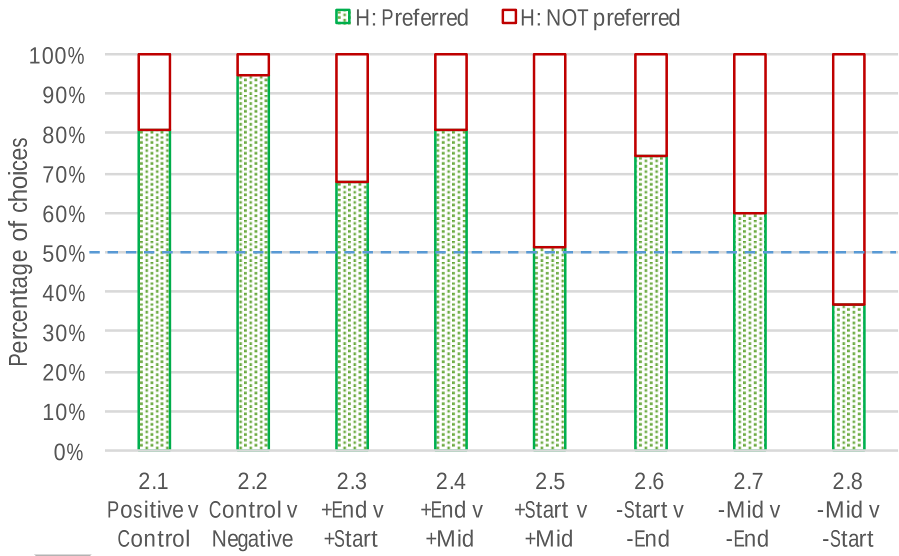


Figure 5. Experiment 2 preference choice results. Percentage of choices for the hypothesised preferred (green, bottom; first named series) and not-preferred alternatives for each of the eight hypotheses.

5.3.1 Manipulation Check 2.1: Positive vs. Control

As intended, a significant majority of participants chose the *Positive* series over the *Control* series: 59 of 73 people (81%), 95% CI [.70, .89], binomial test $p < .001$. They also rated the *Positive* series as more satisfying than the *Control* series (means 3.85 and 2.22 s, Wilcoxon $p < .001$). The mean time taken to complete the Positive series was 6.89

s, 95% CI [6.4, 7.4], which was significantly shorter than the Control series at 14.2 s, 95% CI [13.5, 14.9]: $t_{72} = 24.1$, $p < .001$, $d = 2.9$.

Analysis of the effect of order (Figure 6) shows that a significantly larger proportion of participants selected the *Positive* series when it was completed second (90%) rather than first (68%) – two-sample test for equality of proportions, $\chi^2 = 4.57$, $p = .03$. This suggests that a recency effect may apply, even in the manipulation check condition – participants were more likely to select the *Positive* series if it was experienced more recently.

5.3.2 Manipulation Check 2.2: Control vs. Negative

A significant majority of participants chose the *Control* series over the *Negative* series: 69 of 73 people (94.5%), 95% CI [.87, .98], $p < .001$. Mean frustration/satisfaction ratings for the two series were substantially different at 2.23 for *Control* and -1.67 for *Negative* (Wilcoxon $p < .001$). Mean times with the *Control* and *Negative* series were 14.3 s, 95% CI [13.6, 14.9], and 34.9 s, 95% CI [30.1, 39.7], and the difference was significant: $t_{72} = 8.98$, $p < .001$, $d = 1.4$. There was no effect of order ($\chi^2 \approx 0$, $p = 1$).

In comparison with Experiment 1, in which only 72% preferred the Control series to Negative, the experimental modification to increase the magnitude of sensation associated with Negative trials was successful.

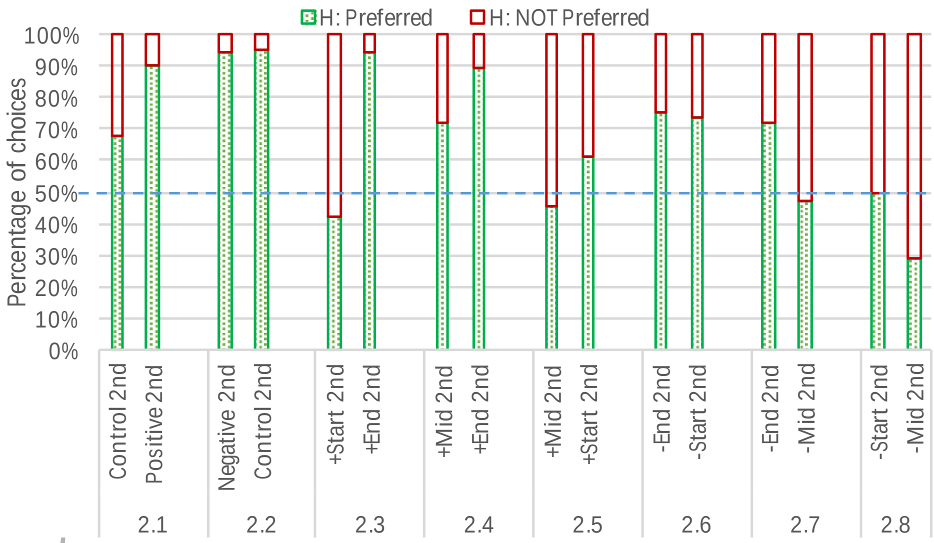


Figure 6. Experiment 2 preference choice results by series order. Percentage of choices for the hypothesised preferred (green, bottom; right of the two series in each condition) and not-preferred series, dependent on which series was completed second, for each of the eight hypotheses.

5.3.3 Hypothesis 2.3: Positive Recency vs. Positive Primacy

A significant majority of participants chose the *+End* series (25 of 37, 68%, 95% CI [.50, .82]) in preference to the *+Start* series: binomial test, $p = .047$. We therefore accept **H2.3.** Responses on the frustration/satisfaction scale were similar to the forced-choice responses, with a higher mean for the *+End* series (1.51) than the *+Start* series (0.95): Wilcoxon $p = .008$). As intended, the mean time to complete the series was similar at 24.67 s, 95% CI [23.2, 26.1], and 24.1 s, 95% CI [22.2, 25.9], for *+End* and *+Start:* $t_{36} = 0.84$, $p = 0.41$, $d = 0.11$.

As shown in Figure 6, participants were much more likely to select the *+End* series when completed second (94%) rather than first (42%), giving a significant effect of order ($\chi^2 = 9.29$, $p = .002$). This again conforms to a recency effect – a *Positive* ending has a stronger effect on subjective experience when more recently experienced.

5.3.4 Hypothesis 2.4: Positive Recency without Opposing Primacy

A significant majority of participants chose the *+End* series (30 of 37, 81%, 95% CI [.65, .92]) in preference to the *+Mid* series – binomial test, $p < .001$. We therefore accept **H2.4.** Frustration/satisfaction responses are again similar to preference selection, with means of 1.97 for *+End* and 1.24 for *+Mid* (Wilcoxon $p = .01$).

As intended, the mean time to complete the series was similar at 23.9 s, 95% CI [22.7, 25.1], and 24.2 s, 95% CI [23.1, 25.2], for *+End* and *+Mid:* $t_{36} = 0.47$, $p = 0.64$, $d = 0.07$.

There was no significant effect of order ($\chi^2 = 0.85$, $p = 0.36$), although the trend reflects that of conditions H2.1 and H2.3, with a larger proportion of participants preferring the positive ending when completed more recently (89.5%) rather than first (72%).

5.3.5 Hypothesis 2.5: Positive Primacy without Opposing Recency

The test of primacy showed no significant effects. Approximately half of the participants (51%) selected the *+Start* series. Mean responses on the frustration/satisfaction scale were similar at 1.32 for *+Start* and 1.13 for *+Mid* (Wilcoxon, $p = .41$). There was also no significant effect of order ($\chi^2 = 0.32$, $p = 0.57$). We therefore fail to accept **H2.5.**

5.3.6 Hypothesis 2.6: Negative Recency vs. Negative Primacy

Conforming to our hypothesis, a significant majority of participants chose the *-Start* series (26 of 35, 75%, 95% CI [.57, .88]) in preference to the *-End* series: binomial test, $p = .006$. We therefore accept **H2.6.**

Responses on the frustration/satisfaction scale were similar to preference choices, with the mean response for *–Start* (0.23) substantially higher than that for *–End* (-1.06): Wilcoxon $p = .008$. The mean times to complete the series were very similar at 35.7 s, 95% CI [33.3, 38.1], and 35.7 s, 95% CI [33.4, 38.0], for *–Start* and *–End:* $t_{34} = 0.05$, $p = 0.96$, $d = 0.007$.

There was no significant order effect ($\chi^2 \approx 0$, $p = 1$), with a similar proportion of participants preferring the *–Start* series, regardless of whether it was completed first (75%) or second (74%).

5.3.7 Hypothesis 2.7: Negative Recency without Opposing Primacy

There was no significant overall preference for the *–Mid* series (21 of 35, 60%, 95% CI [0.42, 0.76]) over the *–End* series: binomial test, $p = .31$. We therefore fail to accept **H2.7.** Frustration/satisfaction responses again show significantly more frustration with the *–End* series (mean -0.63) than *–Mid* (-0.03): Wilcoxon $p = .05$.

There was no significant order effect (Figure 6), with a similar proportion of participants preferring the *–Start* series, regardless of whether it was completed first (75%) or second (74%): $\chi^2 = 1.34$, $p = 0.24$.

Although there was no significant effect of order, the proportion of choices reflect the earlier observed significant effects, with 52% preferring *–End* to *–Mid* when *–End* was completed first, but only 28% preferring *–End* (and 72% selecting *–Mid*) when *–End* was completed second. As discussed later, these values conform to an explanation that recent negative experiences are more likely to be rejected.

5.3.8 Hypothesis 2.8: Negative Primacy without Opposing Recency

As with the *Positive* primacy hypothesis (H2.5), there were no significant effects of *Negative* primacy: although 63% of participants stated a preference for *–Mid*, the difference was not significant (binomial $p = .18$). Responses on the frustration/satisfaction scale also show no significant effect (Wilcoxon $p = 0.98$). We therefore fail to accept **H2.8.** There was no significant effect of order ($\chi^2 = 0.86$, $p = 0.35$).

6 General Discussion

Together, the main findings of Experiments 1 and 2 are:

1. Positive and negative recency effects can significantly influence user preferences for interactive series (Experiment 2).

2. The experiments provided no evidence that primacy effects influence preferences for interactive series (Experiment 2).

3. Recency effects were strongest when the positive or negative stimuli was recently encountered (experimentally), possibly suggesting that the preference effects may be transitory (Experiment 2).

4. Experimentally observing recency effects on preference can be difficult due to weak experiential stimuli (Experiment 1 and related work), an unstable baseline

experience (Experiment 1), and possible transitory nature of the effects (Experiment 2).

In this section, we consider issues that are relevant to the further study and use of sequencing effects in HCI, including a comparison of the results from the two experiments, the types of experiences that trigger the effects, possibilities for further investigation of primacy effects, and the potential for making use of sequencing effects in design.

6.1 Why Were Recency Effects Stronger in Experiment 2 than 1?

Experiments 1 and 2 both included hypotheses concerning preference for a positive end over a positive start (hypotheses H1.3 and H2.3) as well as preference for a negative start over a negative end (hypotheses H1.4 and H2.6). Neither hypothesis was supported in Experiment 1, but both were supported in Experiment 2.

We believe two factors contributed to the difference in outcome between Experiments 1 and 2. First, Experiment 2 was more powerful than Experiment 1 because it involved many more participants (73 vs. 46). For example, H1.3 was not accepted (binomial test $p = .1$) with 63% of participants selecting a preference for *+End* over *+Start.* If exactly the same proportion of the 73 participants in Experiment 2 had selected *+End*, a binomial test would have accepted H2.3, with $p = 0.03$ – in fact, the response was stronger in Experiment 2, at 68%, yielding $p = .002$.

These preference rates of 63 and 68% in Experiments 1 and 2 are not dissimilar to the 68.5% preference rate observed in Kahneman's et al.'s (1993) cold-pressor experiment. This suggests that future HCI researchers should carefully consider the number of participants required to reveal significant effects on preference.

Second, we intentionally increased the magnitude of negative stimulus in Experiment 2 to ensure that negative trials were perceived negatively. In the manipulation check condition of Experiment 1 (MC1.2), 28% of participants selected the Negative series as preferred to the Control series. By increasing the magnitude of performance impairment encountered during negative trials, in the manipulation check condition of Experiment 2 (MC2.2), only 5% of participants selected the Negative series in preference to the Control series. We suspect that the increased intensity of negative sensations during negative trials increased the sensitivity of Experiment 2, allowing it to confirm a preference for *+Start* over *+End* (H2.6, 74%, $p = .006$) when Experiment 1 failed to do so (H1.4, 63%, $p = .1$).

6.2 Stimuli for Sequencing Effects

There are many questions for further research concerning the types, magnitudes, and durations of interactive stimuli that contribute to sequencing effects. Previous studies of sequencing effects outside of HCI have examined strong hedonic stimuli such as pain (Ariely, 1998; Fredrickson and Kahneman, 1993; Kahneman *et al.*, 1993). However, within HCI the stimuli for momentary experience tend to be weaker, such as mental or physical effort, or the annoyance and frustration associated with successful and failed assistance, such as that examined in our experiments.

Understanding the subjective outcomes of successful and failed attempts at assistance is important because interfaces are increasingly seeking to explicitly offer suggestions, adapt their behaviour, or modify the user's input to match the system's inference of the user's goal. Previous studies have shown asymmetries between objective performance and subjective experience of assistive user interfaces (Quinn and Cockburn, 2016; Quinn and Zhai, 2016), but not the influence of sequencing their successes and failures, as examined here.

Beyond the positive and negative experiences derived from assistive interfaces, it is likely that other forms of momentary stimuli will also induce reliable sequencing effects. Table 3 summarises prior studies in HCI relating to sequencing effects, identifying the primary momentary stimulus examined. The table shows that recency effects have generally been confirmed across a variety of stimuli, including mental effort (Hassenzahl and Sandweg, 2004), visible progress (Harrison *et al.*, 2007), physical effort (Cockburn *et al.*, 2015), game difficulty/challenge (Gutwin *et al.*, 2016), and the satisfaction/frustration of user interface assistance and assistance failures (this paper). Fully separating the experiential influence of the stimuli in these experiments is difficult – for example, a manipulation of mental effort may also induce frustration and annoyance – however, the indications across studies are that recency effects occur across a wide range of different types of interactions.

The stimuli examined across the studies shown in Table 3 can be viewed as conforming to different points on the NASA TLX categorisation of workload measures (Hart and Staveland, 1988), which includes dimensions for mental, physical and temporal demand, as well as measures for performance, frustration and overall effort. Future studies might examine sequencing effects with other forms of affective stimuli during interaction, such as different points on the valence/arousal classification schema (Russell, 1980). Understanding the influence of broad types of momentary stimuli could be relevant to the increasingly wide range of user interface applications, such as interfaces for behaviour change and health/mood monitoring (Gurman *et al.*, 2012; Maher *et al.*, 2014). For example, if recency effects are validated in these domains, then a short positive experience at the end of a session with a behaviour change application may have a disproportionately strong influence on the likelihood that users later return to the application.

Another area for further research related momentary experiences concerns the development, validation and deployment of new methods for determining the user's affective state. In particular, biometric data captured through sensors such as galvanic skin response and functional magnetic resonance imaging can provide indications of the user's momentary state (Cowley *et al.*, 2016; Fairclough, 2009; Kim *et al.*, 2016), and these methods might be adapted to study and potentially influence sequencing effects during interaction – for example, interfaces might adapt to the user's affective state, potentially curtailing or extending an interactive session until a particular threshold affective state is attained.

Study	Examining	Stimulus	Outcome measure	Sequence manipulation	Results
Hassenzahl & Sandweg (2004)	Correlations between mental effort and perceived usability	Mental effort (intensity and timing)	Perceived usability	Recency Peak Trend Peak/End	Strongest correlation between recency and perceived usability
Harrison et al. (2007)	Perception of progress bars as being 'faster'	Visible progress (of a progress bar)	Choice of 'faster' progress (between two)	Various rates of progress	Significant effect of recency: accelerating progress preferred
Cockburn et al. (2015)	Preference for page sequences	Physical workload (number of sliders to be set per page)	Preference choice	Recency Peak Peak-and-End	Marginal effects of recency. None for peak; significant effect of peak-and-end
Gutwin et al. (2016)	Experience of game sequences	Game difficulty Game challenge	Preference and various experiential measures	Peak-and-End Recency	Mixed results, depending on game
This paper	Preference for sequences with assistive/failed-assistance	Satisfaction/frustration of assistance (failures) and associated workload.	Preference choice	Recency Primacy	Significant effects of recency; no effects of primacy

Table 3. Summary of prior studies examining sequencing effects, including the primary stimuli used to induce momentary experiences, the measure of experiential outcome, and a summary of the results.

6.3 Absence of Primacy

Primacy effects have been previously observed in interactive contexts – for example, Lee et al. (2011) attributed a choice preference for the first of two synthesised voices to primacy, and Harrison (2007) reported a general preference for the first of two progress bars (in their experimental order), but without reference to primacy.

In our second experiment, neither of the conditions testing primacy (H2.5 and H2.8) showed evidence in support of their associated hypotheses. Indeed, H2.8 indicated the opposite effect: a majority of participants preferred the *–Start* series to the *–Mid* series. We suspect this may be due to the short duration of the series used in the experiment, causing experiences in the middle and the end of the series to become indistinct. In particular, the relatively strong preference (71%) for *–Start* condition when the *–Mid* series was completed second (Figure 6) is consistent with a recency effect: participants may have been averse to the relatively recent experience of the negative trials in the middle of the recently completed *–Mid* series.

One promising avenue for further work on primacy effects in interaction concerns user recollection for different *types* of information in an interactive experience. Zauberman et al. (2006) showed that recency effects were particularly strong when participants made *hedonic* assessments of an experience (i.e. overall affective judgments), but that primacy effects were strong when participants were asked to make *informational* assessments of the same experience. For example, after watching a dynamic display representing job applicants' correct and incorrect test responses, participants' overall satisfaction with the job applicant (a hedonic measure) was subject to recency effects, while their assessment of the number of correct responses was most strongly influenced by primacy effects.

Our experiments were predominantly concerned with hedonic assessment (comparative satisfaction with the two series), which according to Zauberman's finding should be most strongly influenced by recency effects. There are therefore opportunities for exploring the influence that different question phrasing has on participants' recollection of interactive experience. Outcomes from such studies have potentially important applications in interaction – for example, users' decisions might be different if prompted with the text "if you are not satisfied with auto-correct, you can turn it off" (hedonic, emphasising recency) versus "if auto-correct makes frequent errors, you can turn it off" (informational, emphasising primacy).

6.4 Transience and Endurance of Recency Effects

Experiment 2 (summarised in Figure 6) showed that recency effects were strongest when the positive or negative ending was experienced most recently. For example, in conditions H2.3 and H2.4, the *+End* condition was preferred in 94% and 89% of choices when it was completed second, compared to only 42% and 72% when completed first. Similarly, in condition 2.7, the *–End* condition was rejected in 72% of choices when it was completed second, but only 47% when completed first.

These findings are consistent with those observed in previous work, with Montgomery et al. (2009) reporting that global evaluations immediately following an experience give "conditions under which recency effect is most likely to be obtained". Montgomery et al. postulated that primacy effects are more enduring than recency effects, and they experimentally investigated the prediction by testing recollection of experiences immediately afterwards and after a delay. Results showed an interaction effect of primacy/recency recall with delay – recollection of items at the beginning of a series degraded less over time compared to items later in the series.

There are opportunities for further work in examining various temporal aspects related to interactive series and preferences. These include examining series of much longer duration than those tested in our experiment, as well examining recollection of those experiences after a delay. For example, we would like to compare users' recollection of simulated online shopping experiences that are manipulated to contain positive and negative effects of primacy and recency, and examine how immediately-reported preferences change (if at all) over days and weeks.

6.5 Sequencing Effects as a Design Risk and a Design Opportunity

A user's subjective experience of an interactive task is likely to play an important role in their willingness to repeat the experience, and so understanding mechanisms that have a reliable impact on preferences is valuable for designers. Sign-up and checkout processes on websites are notorious for causing user frustration (Cho, 2004), and making the memory of these sequences more pleasant has a clear benefit for site designers. Our experiments demonstrate that simple interface manipulations can contribute to reliable preference outcomes, suggesting that designers should be cautious to not inadvertently cause a negative effect. That is, even if designers are not interested in attempting to create a more positive experience, they must be careful to not accidentally cause a more negative one (e.g. by putting an additional demand or extra step at the end of a checkout sequence). The potential damage to user experience is accentuated because it is often the final stages of a sequence where a failure occurs (e.g. several pages of information input are only checked at the end of the sequence).

The results reported in this paper, and those of related work on sequencing effects, are relevant for many forms of interaction design. When the interaction involves a series of actions, and there is likely to be momentary differences in the user's subjective experience of those actions, designers should carefully consider the possibility for recency effects to influence the user's overall remembered experience. Designers are often relatively unconstrained regarding the distribution of elements across an overall interaction – certain data must be collected, but the order of collection may be malleable. Our results suggest that designers should think carefully about recency effects when considering the distribution of interactive elements. Doing so may improve users' subjective experience and increase their willingness to use the system again in the future; conversely, failing to consider peak-end effects may cause adverse user reactions.

7 Conclusion

Sequencing effects such as primacy, recency, and the peak-end rule are known to influence people's recollection of experiences. Given their prevalence in the basic teachings of psychology, there has been relatively little research into sequencing effects in HCI.

We reviewed prior research on sequencing effects and presented new experimental results exploring these phenomena and demonstrating that recency effects can significantly influence user's preference for interfaces that are otherwise identical in their objective interaction requirements. Together, the extensive findings from psychology, limited prior findings in HCI, and the results reported in this paper show that recency effects offer designers an opportunity to improve users' memories of their subjective experience. Conversely, failing to attend to the timing of negative interaction elements may cause much stronger adverse memories than anticipated. We advocate a better understanding of sequence effects as both an opportunity and a risk for interaction design and user experience.

REFERENCES

Ariely, D., 1998. Combining Experiences Over Time: The Effects of Duration, Intensity Changes and On-Line Measurements on Retrospective Pain Evaluations. Journal of Behavioral Decision Making **11**: 19-45.

Carroll, J., 1987. The adventure of getting to know a computer. Readings in Human-Computer Interaction: A Multidisciplinary Approach. Baecker, R. and Buxton, W. Morgan Kaufmann**:** 639-648.

Chin, J. P., Diehl, V. A. and Norman, K. L., 1988. Development of an instrument measuring user satisfaction of the human-computer interface. Proceedings of CHI'88: SIGCHI Conference on Human Factors in Computing Systems. ACM, Washington, DC**:** 213-218.

Cho, J., 2004. Likelihood to abort an online transaction: influences from cognitive evaluations, attitudes, and behavioral variables. Information and Management **41**(7): 827-838.

Cockburn, A., Quinn, P. and Gutwin, C., 2015. Examining the peak-end effects of subjective experience. Proceedings of CHI'15: ACM Conference on Human Factors in Computing Systems, Seoul, Korea, 357-366. ACM Press.

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., Barral, O., Kosunen, I., Valtonen, T., Huotilainen, M., Ravaja, N. and Jacucci, G., 2016. The Psychophysiology Primer: A Guide to Methods and a Broad Review with a Focus on Human–Computer Interaction. Foundations and Trends® Human–Computer Interaction **9**(3-4): 151-308.

Craik, F. and Lockhart, R., 1972. Levels of processing: A framework for memory research. Journal of Verbal Learning and Verbal Behavior **11**: 671-684.

Crowder, R. G., 1982. The demise of short-term memory. Acta Psychologica **50**(3): 291-323.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J. and Usher, M., 2005. The Demise of Short-Term Memory Revisited: Empirical and Computational Investigations of Recency Effects. Psychological Review **112**(1): 3-42.

Davis, F., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly **13**: 319-339.

Davis, F. D., Bagozzi, R. P. and Warshaw, P. R., 1992. Extrinsic and Intrinsic Motivation to Use Computers in the Workplace1. Journal of Applied Social Psychology **22**(14): 1111-1132.

Deterding, S., Dixon, D., Khaled, R. and Nacke, L., 2011. From game design elements to gamefulness: defining "gamification". Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments. ACM, Tampere, Finland**:** 9-15.

Diener, E., Wirtz, D. and Oishi, S., 2001. End effects of rated life quality: the James Dean Effect. Psychological Science **12**(2): 124-8.

Do, A., Rupert, A. and Wolford, G., 2008. Evaluations of pleasurable experiences: The peak-end rule. Psychonomic Bulletin & Review **15**(1): 96-98.

Fairclough, S. H., 2009. Fundamentals of physiological computing. Interacting with Computers **21**(1-2): 133-145.

Festinger, L., 1957. A Theory of Cognitive Dissonance. Harper & Row, New York.

Finn, B., 2010. Ending on a High Note: Adding a Better End to Effortful Study. Journal of Experimental Psychology: Learning, Memory, and Cognition **36**(6): 1548-1553.

Fitts, P. M., 1954. The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement. Journal of Experimental Psychology **47**: 381-391.

Fredrickson, B. and Kahneman, D., 1993. Duration neglect in retrospective evaluations of affective episodes. Journal of Personality and Social Psychology **65**(1): 45-55.

Gurman, T. A., Rubin, S. E. and Roess, A. A., 2012. Effectiveness of mHealth Behavior Change Communication Interventions in Developing Countries: A Systematic Review of the Literature. Journal of Health Communication **17**(sup1): 82-104.

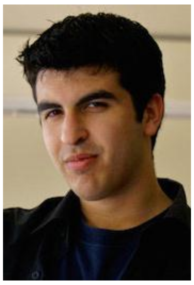
Gutwin, C., Rooke, C., Cockburn, A., Mandryk, R. L. and Lafreniere, B., 2016. Peak-End Effects on Player Experience in Casual Games. Proceedings of CHI'16: ACM Conference on Human Factors in Computing Systems, San Jose, California, 5608-5619. ACM.

Harrison, C., Amento, B., Kuznetsov, S. and Bell, R., 2007. Rethinking the progress bar. Proceedings of the 20th annual ACM symposium on User interface software and technology, Newport, Rhode Island, USA, 115-118. ACM.

Harrison, C., Yeo, Z. and Hudson, S. E., 2010. Faster progress bars: manipulating perceived duration with visual augmentations. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA, 1545-1548. ACM.
Hart, S. and Staveland, L., 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. Human Mental Workload. Hancock, P. and Meshkati, N.**:** 139-183.
Hassenzahl, M., 2004. The Interplay of Beauty, Goodness, and Usability in Interactive Products. Human-Computer Interaction **19**(4): 319--349.
Hassenzahl, M., Platz, A., Burmester, M. and Lehner, K., 2000. Hedonic and ergonomic quality aspects determine a software's appeal. Proceedings of CHI'00: ACM SIGCHI conference on Human Factors in Computing Systems. ACM, The Hague, The Netherlands**:** 201-208.
Hassenzahl, M. and Sandweg, N., 2004. From mental effort to perceived usability: transforming experiences into summary assessments. CHI '04 Extended Abstracts on Human Factors in Computing Systems. ACM, Vienna, Austria**:** 1283-1286
Hassenzahl, M., Wiklund-Engblom, A., Bengs, A., Hagglund, S. and Diefenbach, S., 2015. Experience-Oriented and Product-Oriented Evaluation: Psychological Need Fulfillment, Positive Affect, and Product Perception. International Journal of Human-Computer Interaction **31**: 530-544.
Henson, R. N. A., 1998. Short-Term Memory for Serial Order: The Start-End Model. Cognitive Psychology **36**(2): 73-137.
Hornbaek, K., 2006. Current practice in measuring usability: Challenges to usability studies and research. International Journal of Human-Computer Studies **64**(2): 79-102.
Kahneman, D., 2000a. Evaluation by Moments: Past and Future. Choices, Values and Frames. Kahneman, D. and Tversky, A. Cambridge University Press New York.
Kahneman, D., 2000b. Experience Utility and Objective Happiness: A Moment-Based Approach. Choices, Values and Frames. Kahneman, D. and Tversky, A. Cambridge University Press New York.
Kahneman, D., Fredrickson, B., Schreiber, C. and Redelmeier, D., 1993. When More Pain Is Preferred to Less: Adding a Better End. Journal of Personality and Social Psychology **65**(1): 401-405.
Kahneman, D., Wakker, P. and Sarin, R., 1997. Back to Bentham? Explorations of Experienced Utility. The Quarterly Journal of Economics **112**(2): 375-405.
Kemp, S., Burt, C. D. B. and Furneaux, L., 2008. A test of the peak-end rule with extended autobiographical events. Memory & Cognition **36**(1): 132-138.
Keren, G. B. and Raaijmakers, J. G. W., 1988. On between-subjects versus within-subjects comparisons in testing utility theory. Organizational Behavior and Human Decision Processes **41**(2): 233-247.
Kim, J., Wang, J., Wedell, D. H. and Shinkareva, S. V., 2016. Identifying Core Affect in Individuals from fMRI Responses to Dynamic Naturalistic Audiovisual Stimuli. PLoS ONE **11**(9): e0161589.
Langer, T., Sarin, R. and Weber, M., 2005. The retrospective evaluation of payment sequences: duration neglect and peak-and-end effects. Journal of Economic Behavior & Organization **58**(1): 157-175.
Lee, K. M., Jung, Y. and Nass, C., 2011. Can User Choice Alter Experimental Findings in Human–Computer Interaction?: Similarity Attraction Versus Cognitive Dissonance in Social Responses to Synthetic Speech. International Journal of Human-Computer Interaction **27**(4): 307-322.
Levesque, V., Oram, L., MacLean, K., Cockburn, A., Marchuk, N. D., Johnson, D., Colgate, J. E. and Peshkin, M. A., 2011. Enhancing Physicality in Touch Interaction with Programmable Friction. Proceedings of CHI'11: ACM Conference on Human Factors in Computing Systems, Vancouver, Canada, 2481-2490. ACM.
MacKenzie, I. S., 1992. Fitts' law as a research and design tool in human-computer interaction. Human-Computer Interaction **7**: 91-139.
Maher, A. C., Lewis, K. L., Ferrar, K., Marshall, S., De Bourdeaudhuij, I. and Vandelanotte, C., 2014. Are Health Behavior Change Interventions That Use Online Social Networks Effective? A Systematic Review. J Med Internet Res **16**(2): e40.
Malone, T., 1980. What makes things fun to learn? A study of intrinsically motivating computer games. Department of Psychology. Stanford University, Palo Alto, CA.
Michalco, J., Simonsen, J. G. and Hornbaek, K., 2015. An exploration of the relation between expectations and user experience. International Journal of Human-Computer Interaction **31**(9): 603-617.
Montgomery, N. V. and Unnava, H. R., 2009. Temporal Sequence Effects: A Memory Framework. Journal of Consumer Research **36**(1): 83-92.

Murdock Jr, B. B., 1962. The serial position effect of free recall. Journal of Experimental Psychology **64**(5): 482-488.
Nasiry, J. and Popescu, I., 2011. Dynamic Pricing with Loss-Averse Consumers and Peak-End Anchoring. Operations Research **59**(6): 1361--1368.
Oliver, R., 1977. Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. Journal of Applied Psychology **62**: 480-486.
Page, M. P. A. and Norris, D., 1998. The primacy model: A new model of immediate serial recall. Psychological Review **105**(4): 761-781.
Porat, T. and Tractinsky, N., 2012. It's a Pleasure Buying Here: The Effects of Web-Store Design on Consumers Emotions and Attitudes. Human Computer Interaction **27**: 235-276.
Quinn, P. and Cockburn, A., 2016. When Bad Feels Good: Failed Assistance and Interface Preferences. Proceedings of CHI'16: ACM Conference on Human Factors in Computing Systems, San Jose, California, 4005-4010. ACM.
Quinn, P. and Zhai, S., 2016. A Cost–Benefit Study of Text Entry Suggestion Interaction. Proceedings of CHI'16: ACM Conference on Human Factors in Computing Systems, San Jose, California, 83-88. ACM.
Redelmeier, D. A. and Kahneman, D., 1996. Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. Pain **66**(1): 3-8.
Redelmeier, D. A., Katz, J. and Kahneman, D., 2003. Memories of colonoscopy: a randomized trial. Pain **104**(1): 187--194.
Rode, E., Rozin, P. and Durlach, P., 2007. Experienced and remembered pleasure for meals: Duration neglect but minimal peak, end (recency) or primacy effects. Appetite **49**(1): 18-29.
Rundus, D., 1971. Analysis of rehearsal processes in free recall. Journal of Experimental Psychology **89**(1): 63-77.
Russell, J., A., 1980. A circumplex model of affect. Journal of Personality and Social Psychology **39**(6): 1161-1171.
Ryan, R. M., Rigby, C. S. and Przybylski, A., 2006. The motivational pull of video games: A self-determination theory approach. Motivation and Emotion **30**(4): 344-360.
Schreiber, C. and Kahneman, D., 2000. Determinants of the rememberd utility of aversive sounds. Journal of Experimental Psychology: General **129**: 27-42.
Sherif, M., 1961. Social judgment : assimilation and contrast effects in communication and attitude change / by Muzafer Sherif and Carl I.Hovland. Yale U.P, New Haven.
Shneiderman, B., 1987. Direct Manipulation: A Step Beyond Programming Languages (excerpt). Readings in Human-Computer Interaction: A Multidisciplinary Approach. Baecker, R. and Buxton, W. Morgan-Kauffman**:** 461-467.
Shteingart, H., Neiman, T. and Loewenstein, Y., 2013. The role of first impression in operant learning. Journal of Experimental Psychology: General **142**(2): 476-488.
Tractinsky, N., Katz, A. and Ikar, D., 2000. What is beautiful is usable. Interacting with Computers **13**(2): 127-145.
Tversky, A. and Kahneman, D., 1974. Judgment under Uncertainty: Heuristics and Biases. Science **185**(4157): 1124-1131.
Udsen, L. E. and Jørgensen, A., Helms, 2005. The aesthetic turn: unravelling recent aesthetic approaches to human-computer interaction. Digital Creativity **16**(4): 205-216.
Varey, C. and Kahneman, D., 1992. Experiences extended across time: Evaluation of moments and episodes. Journal of Behavioral Decision Making **5**(3): 169-185.
Venkatesh, V., Morris, M. G., Davis, G. B. and Davis, F. D., 2003. User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly **27**(3): 425-478.
Weiss, B., Guse, D., Moller, S., Raake, A., Borowiak, A. and Reiter, U., 2014. Temporal Development of Quality of Experience. Quality of Experience: Advanced Concepts, Applications and Methods. Moller, S. and Raake, A. Springer.
Yi, Y., 1990. A critical review of consumer satisfaction. Review of marketing 1990. Zeithaml, V. American Marketing Association, Chicago, IL**:** 68-123.
Zauberman, G., Diehl, K. and Ariely, D., 2006. Hedonic versus informational evaluations: task dependent preferences for sequences of outcomes. Journal of Behavioral Decision Making **19**(3): 191-211.
Zhang, P. and von Dran, G. M., 2001. User Expectations and Rankings of Quality Factors in Different Web Site Domains. International Journal of Electronic Commerce **6**(2): 9-33.

Dr. Andy Cockburn is a professor in the Department of Computer Science of the University of Canterbury, Christchurch, New Zealand. He is a computer scientist with interests in modelling and empirically measuring human performance with interactive systems.



Dr. Philip Quinn is a research scientist at Google Inc. The research reported in this paper was completed during his PhD studies at the University of Canterbury. His research interests include human factors measurement and modelling.

Dr. Carl Gutwin is a professor in the Department of Computer Science and the director of the HCI lab. His main research areas are computer-supported cooperative work, interaction techniques, and surface computing.