Looking Back: Retrospective Study Methods for HCI

Daniel M. Russell and Ed H. Chi

The think-aloud protocol (Ericsson & Simon, 1985) has participants talk *while* doing the behavior of interest. While this approach is often used, speaking aloud during the activity can introduce social, cognitive load, and attention aberrations, creating a somewhat unnatural behavioral response (Dickson, McLennan, & Omodei, 2000; Wilson, 1994). On the other hand, as Ericsson and Simon points out (1985), Ericsson (2006), in the retrospective cued recall (RCR) approach, the amount of time that passes between mental action and recollection of that action necessarily introduces artifacts of memory and post-event processing that interfere with accurate recall. Neither is perfect.

With all this in mind, retrospective analysis is a methodology for conducting studies where the participant does their normal behavior without taking any disruptive action such as writing a diary entry, talking about their behavior, or responding to an interruption. RCR methods can be used to reconstruct participants' behaviors, rationales, affective reactions, and responses for events that have been recorded. In essence, a RCR method is whenever the participant is later asked to recall (or explain) their earlier behavior when prompted by cues such as images taken during their behavior, videos of the event, eye tracking showing what they were looking at during the task, etc. The central element of a study is that have important recollection-aiding cues have been captured during the experience and then used in post-event discussion and analysis. This method of gathering user behaviors is remarkably accurate when recollection cues and interview methods are well designed, even when there are fairly lengthy delays between action and recall.

D.M. Russell (✉) • E.H. Chi
Google, Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
e-mail: drussell@google.com; edchi@google.com

J.S. Olson and W.A. Kellogg (eds.), *Ways of Knowing in HCI*,
DOI 10.1007/978-1-4939-0378-8_15, © Springer Science+Business Media New York 2014

Retrospective Methods: Introduction

Traditional HCI study techniques tend to be very active, in the moment, and event driven, using many of the methods described in the chapters of this book. The analysis of events that have taken place during the course of an HCI study or experiment over a longer period have largely been diary studies (Czerwinski, Horvitz, & Wilhite, 2004; Rieman, 1993) experience sampling (Hektner, Schmidt, & Csikszentmihalyi, 2007; Larson & Csikszentmihalyi, 1983), or log data analysis [see Chap. "Understanding Behavior Through Log Data and Analysis"].

By contrast, while in-lab usability studies are effective at discovering some kinds of UI use patterns, it is notoriously difficult to track the more naturalistic, longer-term behavior of users in the lab or in the wild (Russell & Grimes, 2007). Specialized tracking devices, such as diaries (physical or online) or interruption studies (such as "experience sampling methods") (Brandt et al., 2007; Kuniavsky, 2003) can all materially affect user behavior by continually reminding the user that their behavior is being tracked and monitored. The very fact that the participants are in a laboratory setting (see, the Hawthorne effect (McCarney et al., 2007)), or that they are consciously updating a diary, can lead to important changes in their behavior through observer-expectancy effects (Steele-Johnson, 2000).

Similarly, log data analysis alone does not provide insights into user motivations, nor are logs capable of providing much context about what a user did before, during, or after, particularly when the behaviors of interest take place outside of the system under study.

Diary studies and interviews are difficult to sustain for a long period of time since they rely on participant motivation, which tends to decline as the study continues (Blackwell, Jones, Milic-Frayling, & Rodden, 2005). Furthermore, diaries and interviews tend to focus on specific events and tasks and therefore limit observations to the specific tasks or events, which is not ideal for studying behaviors where users are passively watching or infrequently interacting, instead of actively interacting.

For cases where observations of normal, non-lab user behavior is desired, and where researchers are interested in the context and motivations of participants in a study, retrospective methods are often useful techniques to consider. These methods are particularly useful when it comes to the need for understanding user perceptions and the ability to observe the context of user activity that is not triggered by an event or task.

Let us begin with a definition of these methods:

Definition: a *retrospective* study is one that records data about the behavior of the participant(s) over some period of time. This study-period data is reviewed by the participant afterwards, with the participant providing context and commentary on their behavior as prompted by examining the data that was collected during the course of the study.

A retrospective study is defined by several important experimental design dimensions:

1. *Data collection*: the way in which data (and what type of data) is collected for later review by the participant.
2. *Study duration* (varying from minutes to days).
3. *Review instruments*, *interview methods, and process* used by the participant to elicit recall of prior events.
4. *Sampling frequency* of data collection (samples/time-unit).
5. *Delay of review after collection*: how much time has elapsed since the data was collected and the samples reviewed.

In this chapter, we provide an analysis of retrospective methods in HCI, first discussing the nature of human memory vis-à-vis retrospective recall, then outlining the methodologies used for conducting retrospective studies, presenting a sample retrospective study analyzed along the dimensions given above, concluding with a review of the features and challenges that come with the methods.

Human Memories and How They Work

The retrospective analysis approach takes advantage of the well-known human ability to visually recognize images of earlier situations they had been in, and comment on them. Images, particularly images of an environment (e.g., the computer screen) that have been created as part of the normal course of work, are particularly powerful at bringing about recall of the situation at the time (Brewer, 1986, 1988). Images, especially when used during post-study interviews with the participants, can give an improved view into what was happening in the actual setting of use with nearly imperceptible intrusion. This human ability to recall situational information when retrospectively cued by some data, sound, or visual imagery obtained at the time is the key to these methods.

Yet, at the same time, human memory is notoriously fragile and subject to many kinds of recall errors; memories are often imbued with a sense of accuracy and quality they actually lack (Roediger & McDermott, 1995). A good understanding of the ways in which memory is subject to alteration can help us design retrospective methods that yield useful data for HCI studies. Important factors from these memory studies are the tendency:

1. To reconstruct a memory of a prior event according to a widely held, prototypical pattern of this event category (Van Boven et al., 2009; Schacter, 2001) rather than by accurate recall of the actual events.
2. To follow the researcher's lead in answering questions about the event (Steele-Johnson, 2000; Weisberg, Krosnick, & Bowen, 1996).
3. To make associations about events based on perceived similarities between the recalled event and other, similar experiences that influence memory of the event to be similar to those previous events (Underwood, 1965).

One of the key techniques for avoiding false memories and improving accuracy in recall is to use *cueing* techniques. Cues are used in retrospective studies (usually images or videos) from the record of the participant's behavior. It has become clear that even highly meaningful events will be inaccurately recalled if there is no cueing to orient and remind the participant about the event and that giving cues improves the accuracy of recall (Lamming et al., 1994). Shiffman et al. (1997) demonstrated the poor quality of retrospectively recalling behaviors from 12 weeks ago, even though when the events in question were actively logged by the participant on a personal handheld device, the act of simply *recording* an event seemed to have little impact on quality of recollection. By contrast, actually *seeing* contextually appropriate cues that capture salient cues of the time, place, and activity is an important piece of the method.

Given that the accuracy of uncued memory rapidly deteriorates after about 1 day, there is good reason to wonder about the accuracy of retrospective recall. Some critiques of retrospective methods question their accuracy when the recall is from more than 1 week in the past and when the recollection is performed without the use of cues to support recollection (such as with post-event interviews and surveys) (Novick, Santaella, Cervantes, & Andrade, 2012). However, the careful use of recall cues derived from the participant's own history has been shown to lead to more accurate and useful recollections from some time in the past (see Sect. "A Sample Retrospective Analysis Method" below).

Consequently, setting up the method of a retrospective study requires attention to the details of data collection and event review with the participant to avoid introducing false memories, or asking the participant to recall more than they can accurately report upon (Loftus, 1996).

Earlier Retrospective Work

There is a long tradition of using photos of key events to cue retrospective memories. Collier (1967) is mostly closely associated with the photographic technique in anthropological settings, when photos are used as both prompts and foils to elicit memories and context around some circumstance. Van Gog, Paas, and Van Merriënboer (2005) reports on the tradeoffs involved in using concurrent versus retrospective reporting of problem solving behaviors, ending with the observation that a retrospective recall is often preferred to avoid interfering with the problem-solving process as it occurs.

The idea that images can also be used for HCI recollection purposes can be seen in the work of Van House (2006) and Intille, Kukla, and Ma (2002), although these (and other similar systems) capture images of the world context, and do not provide the detailed internal tracking of events in the user's experience of the online world as a logging system could.

There have been a variety of retrospective methods developed to understand behavior by looking back at their performance. Here we discuss logging tools to track user behavior for later analysis by the participant, video recording and playback, eye tracking with post-task commentary, the Day Reconstruction Method (DRM), and the Experience Sample method.

Logging: Many systems have been built to unobtrusively log user events over time for later analysis. These range from the obvious logging analysis systems of Web behavior to client-side tooling that records user behavior in great detail, allowing the user to look at their behavior afterwards and comment accurately on what (and why) they were acting in a particular way.

LogViewer (Blackwell et al., 2005) is a tool that logs user events and screen images in Web behavior for later analysis. LogViewer also creates a tree analysis visualization of user behavior to track which clicks generate which subsequent Web page views. Their data was primarily intended to facilitate the tracing of user behavior—how many times was the back button used to return to earlier pages, and how user navigation is organized in terms of landmarks pages, etc. They also interviewed their participants with the screen captures as cues, but with a focus on gathering contextual information to aid in their application redesign purposes.

Kellar, Watters, and Shepherd (2006) built a logging system is attached to a customized browser (a modification of Internet Explorer) that allowed the user to label their own behavior as they completed tasks. As with all systems that ask for manual labeling in near real time, the presence of the logging system is hard to ignore, and Kellar points out that there is good evidence that users modified their behavior because of its presence. This system was also used as an object of discussion in the retrospective style, but again, the focus was on accurately labeling sequences of behaviors, rather than using the event log as a cue for recollection of overall behavior.

Other loggers such as (Al-Qaimari & McRostie, 1999; Chi, Pirolli, & Pitkow, 2000; Jones, Milic-Frayling, Rodden, & Blackwell, 2007; Siochi & Hid, 1991) log events for later analysis and are intended to support the understanding of user behavior in search and information browsing tasks, often coordinating the log data with other kinds of user data (e.g., field observations). See (Ivory & Hearst, 2001) for a summary of many such tools developed for tracking and logging Web behavior to improve usability analysis.

Video: (Capra, 2002) developed and evaluated a retrospective analysis version of the self-reported critical incident technique. In this study, researchers showed participants a video replay of their entire working session, asking them to detect and describe critical incidents as they observed them in the video.

To speed up the process and simplify the interaction from the participants perspective (Akers, Simpson, Jeffries, & Winograd, 2009), logged critical events in a participant's use of the CAD solid modeling system SketchUp. Each critical event was then automatically extracted from the video 20 s around each incident. After the entire task was completed, the participants answered a series of questions about their performance while watching the video clips of each incident. This approach gave the participants enough visual context, and enough perspective (coming after the task was completed), to be able to explain why this moment was a crucial incident for them.

In these uses of video as the cue stimulus, the events in question were freshly in mind (having just been completed) and the participants were able to comment on their performance in accurate and useful ways.

Eye tracking: Tracking a participants' eyes as they perform a task gives another kind of video trace that can be used to elicit information [See Chap. "Eye Tracking in HCI: A Brief Introduction"]. When gathered shortly after the completion of the task, a participant can narrate what they were spending their attention on, and why. Hyrskykari, Ovaska, Räihä, Majaranta, and Lehtinen (2008) and Guan, Lee, Cuddihy, and Ramey (2006) both use eye tracking video as cues to help participants describe their visual motions in broad-brush terms. By comparing subjects' retrospective narration with their eye movements, they found the post-event accounts to be valid and reliable, providing a useful account of what people paid attention to while completing tasks. They also found that this has a low risk of introducing fabrications, and is unaffected by overall task complexity.

Eye tracking can also be used to have people retrospectively comment on what was noticed, or not, in a user-interface. Muralidharan, Gyongyi, and Chi (2012) describe a retrospective think-aloud protocol (RTA) where, immediately after the tasks, participants were asked to take the researcher through what they were doing while watching a replay of a screen capture (with eye tracks) of their tasks. The researcher would ask probing questions for clarification, and then move on to talk about another task.

If the participant never mentioned the features being tested (even if the feature was always visible in all of the tasks), the researcher would return to a screen capture, point out the UI feature explicitly and ask a series of question: "What is that? Did you notice it? What does this element of the user interface suggest to you? Tell me what you think about this." The goal was to learn what the participant thought the feature was, whether it had been noticed at all, to understand whether or not they perceived it as useful, and why.

Day Reconstruction Method: Another approach that has been used extensively in psychology studies is the "Day Reconstruction Method" (DRM). As introduced by Kahneman, Krueger, Schkade, Schwarz, and Stone (2004), the DRM combines the advantages of an offline method with the accuracy of introspective approaches such as Experience Sampling (described below).

A DRM study asks participants to reconstruct their daily experiences as a continuous series of episodes, writing a brief name for each one. Experiential episodes are recalled in relation to preceding ones, which lets participants draw on episodic memory when reporting on the experience (Schwarz et al., 2009). To minimize retrospection biases, the DRM is typically conducted at the end of a reported day or at the beginning of the next day. Hence, participants are better able to capture the properties of a single experiential episode, avoiding inferences from their global beliefs about the experience. The DRM method works well when the participants understand the nature of what the study is about (for instance, if it is trying to capture their hedonic experience of a particular system), but less well when the research questions are about their experience over multiple days.

Experience Sampling: The "experience sampling method" (ESM) can be thought of as a diary study where the diary entries are driven by some external signal, typically a beeper, phone message, text message, or another way to remind the participant to fill out a questionnaire about their experience at that moment in time (Kuniavsky, 2003; Larson & Csikszentmihalyi, 1983). A modification of the ESM method that makes it much more like a RCR method is "Image-based experience sampling and reflection" (Intille et al., 2002), where a still image (or short video clip) of the participants environment is captured at the sample moment, and then later reflected upon for subsequent analysis.

A Sample Retrospective Analysis Method

Our use of the tool IE-Capture (short for Internet Explorer Capture) (Russell & Oren, 2009) illustrates a retrospective method in HCI. This was a browser add-on that captured not just moments in a user's behavior of a Web browser but also, crucially, complete screen images (the entire screen extent—more than just the Internet Explorer (IE) window being used to work on the Web). Having the complete-screen proved valuable for helping the participants recall their behavior accurately.

In terms of the methodology design dimensions mentioned above, IE-Capture had:

- *Data collection*: complete screen captures, URL, time-stamp; triggered by the completion of the loading of a Web page in the browser.
- *Study duration*: varying from 1 to 6 weeks (mostly 2 weeks in length).
- *Review instruments*: the collected screenshots were reviewed with a custom-built data viewer (see Fig. 3) that allowed the participant to browse forward and backward in time through the collection. Each screen capture occurred whenever a Web page completed loading, so the sampling frequency varied depending on the participant's use of the Web. Typically, this would measure in the many hundreds over the course of the study.
- *Sampling frequency*: samples were collected on-event (at document-load time) whenever the participant was using Internet Explorer as their browser.
- *Delay of review*: 1–6 weeks (most often 2 weeks) after data collected.

IE-Capture was designed to help us to understand how search-engine users would approach questions that required a long effort over time. By their nature, such tasks are difficult to capture in laboratory settings; an unobtrusive data capture method was needed, and hence the creation of the logging system that could be left in place without any intervention on the part of the participant.

However, a key to this research study was to understand *how* participants thought about and framed questions as they went through their research process over hours, days, and weeks. In this study, IE-Capture logged screenshots for a period of 2 weeks (sometimes longer); then a retrospective review was held with the participant in their home or workplace. As seen in Fig. 1, a series of whole-screen captures were collected for the interview. An individual frame of that sequence can be seen

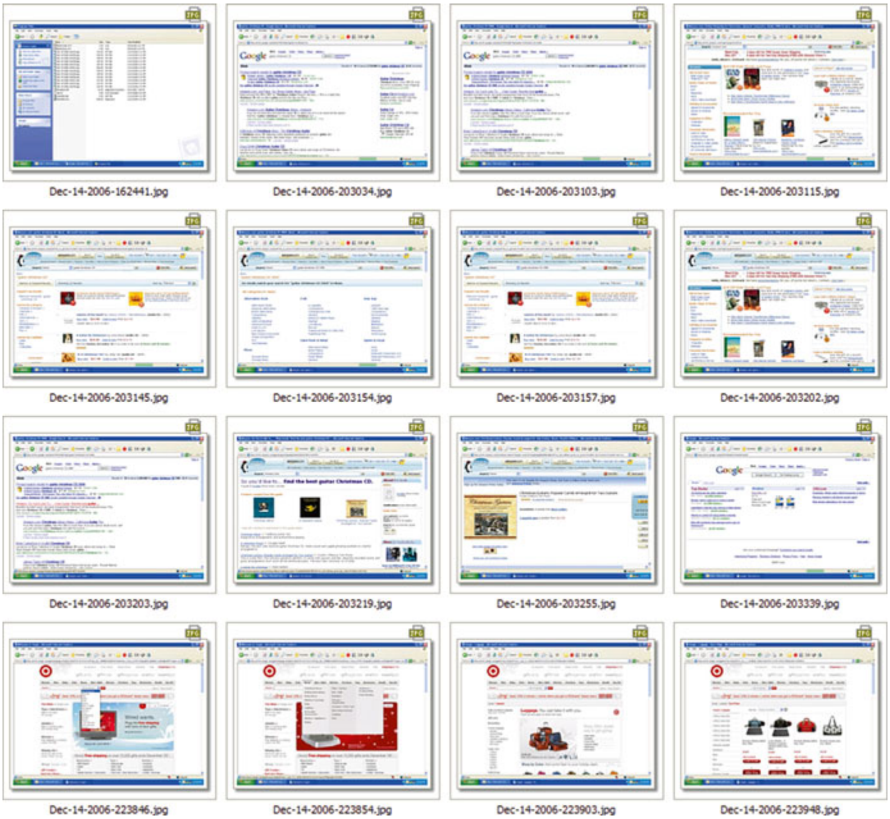
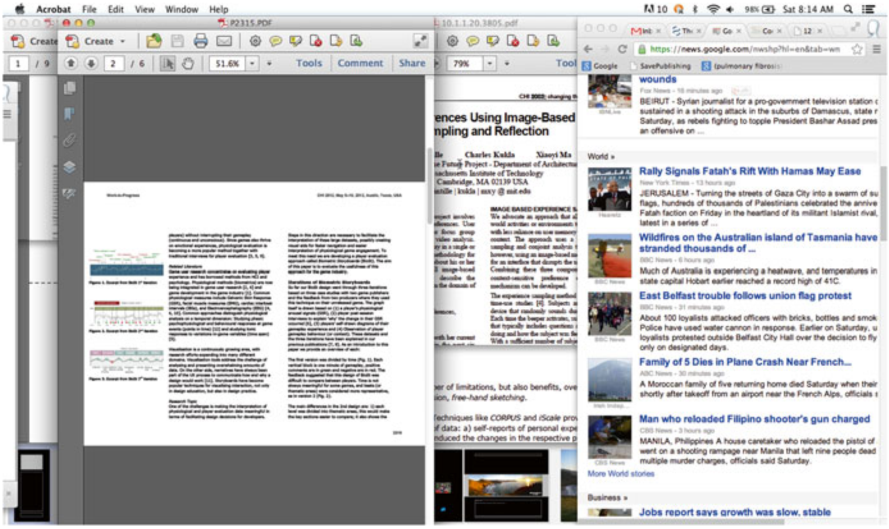


Fig. 1 A series of screen snapshots used in a post-event retrospective interview. These kinds of captured cues support accurate recall by providing a great deal of context to the participant, allowing them to reconstruct what was happening at the time of the events in question (In this case, the browser is maximized to full screen size)

in Fig. 2. With all of the additional visual information available (such as which applications are open, which documents are visible at the same time, the state of work in progress), the participant can recollect what was going on at the time from many different cues.

During the interview, the participant would review the collected series of screen captures, providing the backstory in response to questions asked by the interviewer. Since the number of screen images and logged events could be in the large hundreds, the researcher would select a sample of the events to review in detail.

As is typical for RCR studies, the interviewer began by acquainting the participant with the review instrument operation (how to move forward and backward in the sequence of captured screen images) and then setting the context by reviewing the very earliest data collected in the study. Then, a series of semi-structured review questions were asked, determining the properties of the experience during the study period, elaborated in the next section.



ences Using Image-Based
npling and Reflection

wounds

BEIRUT - Syrian journalist for a pro-government television station
sustained in a shooting attack in the suburbs of Damascus, state
Saturday, as rebels fighting to topple President Bashar Assad pres
an offensive on ...

Rally Signals Fatah's Rift With Hamas May Ease

JERUSALEM - Turning the streets of Gaza City into a swarm of su
flags, hundreds of thousands of Palestinians celebrated the annive
Fatah faction on Friday in the heartland of its militant Islamist rival,
latest in a series of ...

Wildfires on the Australian island of Tasmania have stranded thousands of ...

Much of Australia is experiencing a heatwave, and temperatures in
state capital Hobart earlier reached a record high of 41C.

East Belfast trouble follows union flag protest

About 100 loyalists attacked officers with bricks, bottles and smok
Police have used water cannon in response. Earlier on Saturday, u
loyalists protested outside Belfast City Hall over the decision to fly
only on designated days.

Family of 5 Dies in Plane Crash Near French...

A Moroccan family of five returning home died Saturday when their
shortly after takeoff from an airport near the French Alps, officials

Man who reloaded Filipino shooter's gun charged

MANILA, Philippines A house caretaker who reloaded the pistol of
went on a shooting rampage near Manila that left nine people dead
multiple murder charges, officials said Saturday.

Jobs report says growth was slow, stable

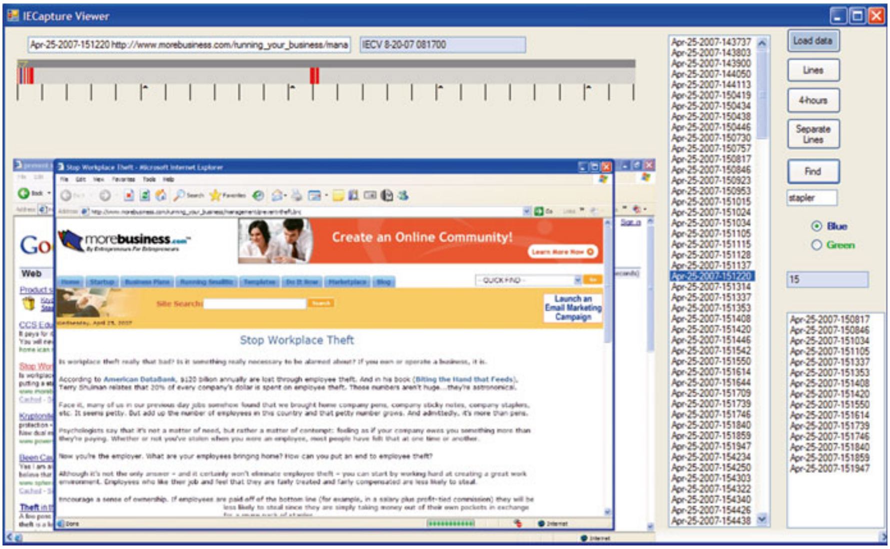
Fig. 2 The choice of cue stimulus is crucial in being able to elicit high quality retrospective recollections. In (Russell & Oren, 2009) the user's entire desktop image was used as a cue to ask questions about overall task intent and search behavior. Cue stimuli with less contextually useful information (e.g., only the browser screen image) were not as successful, and led to lower quality retrospection data

Interview questioning procedure: At the beginning of the interview, after introducing the review tool and reviewing the first day of collected data, the experimenter would view selected visits each day in the log with IECaptureViewer, our tool to view and scroll through the data logs and screenshots (see Fig. 3 for IECaptureViewer). Since this was a study investigating how well people could remember their search tasks, nearly all participants had searches on days that we probed. In the few missing cases, we used the next prior search event (e.g., substituting day 3 for day 4).

For each of the days in question, the experimenter would jump to the first search query made on that day and show it to the participant. (Note that "jumping" to the screen image in question was important, as to avoid showing the participant later screen images that would have shown them the sequence of events.)

The experimenter then asked the participant to describe "what happened next in the search process." The participant was instructed to describe the next event if they felt "reasonably confident" that they knew what happened, in particular, focusing on what search terms were used, and whether or not that particular next search was successful.

While the participant was not prompted for a particular kind of answer, we noted possible variations on their answer. Was the search successful with this query alone? Did the participant have to continue searching after this point in time? If they continued, did they have to continue refining the current query or do something else entirely? This free form question made it easy to assess whether or not the



```
Apr-25-2007-143737
Apr-25-2007-143803
Apr-25-2007-143900
Apr-25-2007-144050
Apr-25-2007-144113
Apr-25-2007-150419
Apr-25-2007-150434
Apr-25-2007-150438
Apr-25-2007-150446
Apr-25-2007-150730
Apr-25-2007-150757
Apr-25-2007-150817
Apr-25-2007-150846
Apr-25-2007-150923
Apr-25-2007-150953
Apr-25-2007-151015
Apr-25-2007-151024
Apr-25-2007-151034
Apr-25-2007-151105
Apr-25-2007-151115
Apr-25-2007-151128
Apr-25-2007-151137
Apr-25-2007-151220
Apr-25-2007-151314
Apr-25-2007-151337
Apr-25-2007-151353
Apr-25-2007-151408
Apr-25-2007-151420
Apr-25-2007-151446
Apr-25-2007-151542
Apr-25-2007-151550
Apr-25-2007-151614
Apr-25-2007-151644
Apr-25-2007-151709
Apr-25-2007-151739
Apr-25-2007-151746
Apr-25-2007-151840
Apr-25-2007-151859
Apr-25-2007-151947
Apr-25-2007-154234
Apr-25-2007-154250
Apr-25-2007-154303
Apr-25-2007-154322
Apr-25-2007-154340
Apr-25-2007-154426
Apr-25-2007-154438
```

```
Apr-25-2007-150817
Apr-25-2007-150846
Apr-25-2007-151034
Apr-25-2007-151105
Apr-25-2007-151337
Apr-25-2007-151353
Apr-25-2007-151408
Apr-25-2007-151420
Apr-25-2007-151550
Apr-25-2007-151614
Apr-25-2007-151739
Apr-25-2007-151746
Apr-25-2007-151840
Apr-25-2007-151859
Apr-25-2007-151947
```

Stop Workplace Theft

Is workplace theft really that bad? Is it something really necessary to be alarmed about? If you own or operate a business, it is.

According to American DataBank, $120 billion annually are lost through employee theft. And in his book (Biting the Hand that Feeds), Terry Shulman relates that 20% of every company's dollar is spent on employee theft. Those numbers aren't huge…they're astronomical.

Face it, many of us in our previous day jobs somehow found that we brought home company pens, company sticky notes, company staplers, etc. It seems petty. But add up the number of employees in this country and that petty number grows. And admittedly, it's more than pens.

Psychologists say that it's not a matter of need, but rather a matter of contempt: feeling as if your company owes you something more than they're paying. Whether or not you've stolen when you were an employee, most people have felt that at one time or another.

Now you're the employer. What are your employees bringing home? How can you put an end to employee theft?

Although it's not the only answer – and it certainly won't eliminate employee theft – you can start by working hard at creating a great work environment. Employees who like their job and feel that they are fairly treated and fairly compensated are less likely to steal.

Encourage a sense of ownership. If employees are paid off of the bottom line (for example, in a salary plus profit-tied commission) they will be less likely to steal since they are simply taking money out of their own pockets in exchange for a mere pack of staples.

Fig. 3 IE-Capture Viewer—a tool for reviewing the participant's log and screen images for discussion and retrospective cueing. The participant's screen image is visible in the center of the display, with the stack of windows present at the time of screen capture, an essential part of cueing for long-term recall. The lists on the *right hand side* are for quickly moving among the log events and captured images for discussion purposes with the participant

participant could recall the situation at all since we had data on what actually did happen next.

If the participant could not recall, then the experimenter would go forward in time, showing them one event image after another, pushing forward in time, until the participant could recollect what was going on and was able to predict what the next search event would be.

We were interested in measuring the participant's ability to speak accurately about the next major event in their search process. That is, having cued their memory of an event, we measured their ability to recall the next step in the process. (For example: looking at a screen image from 6 days ago, the participant would be asked "What's the next search you did after this point?") In nearly all cases, the assessment by the researcher of the participant's memory was clear and evident: either the participant could accurately predict what was coming up in the log, or they just could not say. Only rarely did a participant guess and feel confident; when they guessed, they would say so and express a lack of confidence in their prediction.

Results: For each participant we had two measures—the number of correct predictions based on a cued recall, and the number of times they had to go to a previous page before they could recollect what was going on in the search (see Fig. 5). A good recollection happens when the participant can accurately recall the next search event after just one or two "cue" screen images.

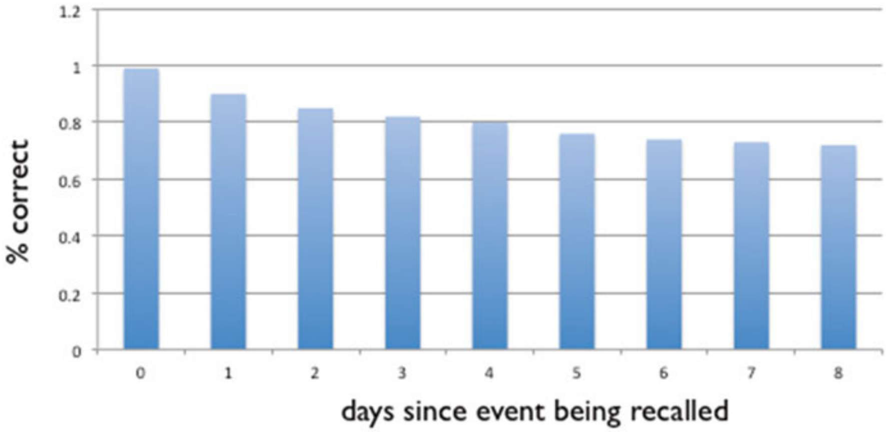


Fig. 4 The number of correct next event predictions drops below 86 % for events 2 days in the past, but is still at 75 % correct for events that are 7 days in the past

As can be seen in Fig. 4, the majority of participants could accurately recall the next search event after the probe within the past 4 days. This is not terribly surprising, given that searches are relatively infrequent. In this participant pool, the average number of searches per week was 11. A search done 3 or 4 days ago is relatively recent and is memorable (and recallable) by its relative rarity among the total number of Web interactions in that time, and its distinctness (a search is an event that requires explicit interaction to achieve some goal).

However, as we tested farther into the past (6 days and 7 days out), participant recall was still quite good. Even after nearly a week, participants were able to recall the next search event correctly around 75 % of the time. With the additional prompting of advancing to the next page in their cached screenshots using the IE-Capture viewer, participants could recall their next search event accurately after seeing only three additional screen images taken from the log/screen files. (Remember that the cue screen images are usually *not* search events, but usually just the next Web page that the participant visited.)

It was clear during the interviews that participants really could recollect not just the next event, but also how this search fit into the larger story of what was going on at the time. Even for events 7 days in the past, participants were able to not just make a prediction about the next event, but also complete the story and say whether or not the entire task (of which search was just a part) was successful or not. We viewed this “story ability” as suggestive that the entire sequence of behaviors was being recollected, and not just the single search event in isolation. The relatively high accurate recall rate after cueing also suggested that more than just one frame of the sequence was necessary for context restoration—a few images in sequence seemed to work the best for accurate recall. As is shown in the context capture method of (Akers et al., 2009), truly effective retrospective analysis means capturing the *context* of use over time, as well as memorable instances with visual context.

Intriguingly, during the interviews, participants seemed able to speak with assurance about what had happened even quite a while ago. But questions about the accuracy of the recalled memories worried us. As we can see in Fig. 4, while accuracy drops off as the events become more distant in the past, within the weeklong period we tested, accuracy rates were quite reasonable.

It became clear to us that some kinds of questions are more easily answered than others. In general, broad descriptions of what kind of thing happened next (e.g., "And then what did you do?") were more effective than asking highly detailed questions (e.g., "What was the next query you performed?"). It also quickly became evident that participants were not only able to make accurate recollections about particular events for which they had not been preconditioned to attend, but that it was the presence of the cueing screen images that was causing the effect. More than one participant commented on the how simple it was to remember what had happened then. Because they could often see other windows in the background (the corner of the Excel spreadsheet, say) those small peripheral cues would give them a distinct sense of time, activity, and place (Wilson, Evans, Emslie, & Malinek, 1997).

Retrospective HCI Methods: Three Time Spans

In HCI, retrospective studies have been used to elicit reflections from study participants on time scales varying from minutes to weeks. Because retrospective memories (and the reflections elicited) vary so much by the amount of time from the original event, it is useful to divide retrospective studies into three categories: Each time period has a distinctive character, with particular challenges and properties.

1. *Short-term studies* (study period <2 h; the retrospective is gathered immediately after task) are typically performed in usability labs, often with the retrospective gathered by a think-aloud protocol as the participant observes a playback of the actual study as captured by video recording of the participant, their screen behavior, or their eye movement behavior (Guan et al., 2006; Hyrskykari et al., 2008). While such studies can be valuable for understanding the instantaneous motivations and reasons for making the choices they do, the temptation is to ask the participant to tell "more than they can know" about their performance. By asking for motivational responses to behavior that might be not open to conscious understanding (such as "why did you choose to read that particular passage"), the participant might easily fall into rationalization about prior behavior that is actually only inadequately remembered. On the other hand, different attributes of the interaction (e.g., why a particular behavior strategy was followed) that are explicitly informational (rather than motivational) can still be commented on accurately (Kuusela & Paul, 2000).
2. *Intermediate-term studies* (study period ≥2 h, <2 days; retrospective gathered 1 or 2 days after completion). These studies are currently somewhat rare in the HCI literature, but strike a nice balance between the accuracy of immediate, short-term labs studies versus the long term studies required to gather enough

rare events, like errors. Studies that run for 1 or 2 days can be naturalistic in ways that the short-term studies are not, because they are conducted in lab settings under tight time constraints.

3. *Very long term* studies (study period >1 day; retrospective gathered 1 or 2 days after the end of the study). For retrospective studies over a long term, the participant cannot ignore the memories and experiences that have happened since the study period. The participant knows how it all turns out, so every recollection will be informed by that knowledge. This effect can be useful (by giving a report about the outcome of actions taken and decisions made), but it can also lead to the "irresistible tendency for subjects to clean up their act and to describe a more coherent and well-thought-out strategy than is normal" (Kuusela & Paul, 2000). Longer-term studies often use daily debriefs of the participant by the researcher. Remote-usability studies sometimes follow this daily check-in protocol as a way to keep in touch with their participants, growing a rapport with them and learning additional information that is nominally outside the scope of the study (Brush, Ames, & Davis, 2004). Furthermore, retrospection from several days in the past is also subject to bias effects—forgetting of the options *not* explored (and in particular, options that we considered at the time, but that left no trace in the cueing record), current mood, and beliefs acquired since the study period (Schacter, 1999).

Evaluating Retrospective Methods in HCI

Many HCI studies have a retrospective aspect to them. At any time a study that has a performance component followed by an evaluation that occurs a significant amount of time after the performance is effectively a retrospective study, even though it may not be labeled as such. (And, in particular, most studies of this sort are not *cued* retrospective studies.) However, many research works in HCI have some aspect of retrospection, for example, when a survey is given to a user population that asks them to reflect on their experiences with experimental software, or when a longitudinal study asks questions about earlier uses of the system under study (both of these are retrospective analyses) (Jain & Boyce 2012).

What about retrospective studies is broadly useful to know for HCI practitioners? We believe that there are two answers. First, what kinds of biases and response effects occur as a result of the passage of time over the course of a study? And second, how do the experimental methods used influence the validity of the retrospection?

It is useful to consider a retrospective study in terms of the important experimental design choices (briefly described in Sect. "Earlier Retrospective Work", above),

- *Data collection:* how will the data be collected and what kinds of data will be collected? Automatically? Or will it be collected by manual intervention (as in the DRM and manual labeling methods)? To what extent will manual annotation interfere with the actual behaviors under study?

- *Study duration*: how long will the study run? Longer runs have the advantage of collecting larger amounts of data, and thus have a higher chance of observing events of real interest, but this interacts with longer term biasing effects.
- *Review instruments*: how will the participant and the researcher review the data? Usually some kind of playback system is needed to select salient episodes or events from the data stream. Such playback systems need to have the ability to "jump to" the prompt of interest without revealing any of the interstitial events. (This avoids giving the participant unanticipated cues which would then degrade the value of the recall).
- *Sampling frequency*: what data sampling rate should be expected and what events cause the data to be collected? Will it be random time sampling (à la ESM), event driven (e.g., by a user action being taken), or periodic (e.g., every hour or at the end of the day).
- *Delay of review*: when will the data be reviewed with the participant? Periodic reviews are useful for longer term experiments, but it becomes difficult to avoid giving the participant subtle clues about what kind of behaviors are the "right" ones (or the opposite—it is difficult to not reveal with responses are surprising from the researcher's perspective).

Can retrospection bias be useful? One may wonder how much retrospection biases influence the accuracy of recall, even during RCR studies. As we have seen, bias is inevitable over the passage of time. But there is an important way to consider this bias: *The memory is what matters*. The veridicality of reconstructed experiences can be of minimal importance as these memories will guide future behavior of the individual and will be communicated to others (Karapanos, Zimmerman, Forlizzi, & Martens, 2009; Norman, 2009). In other words, while what participants remember might be different from what they experienced at the time, memories that are consistent over multiple recalls provide valuable information about future actions. In effect, the memory (no matter how inaccurate), and not the actuality of what happened, becomes the basis on which future decisions are made (Karapanos et al. 2009, 2010).

The influence of retrospective experimental methods. As is true with most psychological or HCI experiment designs, the experimental methods used during retrospective studies can have a profound influence on the results. Even time-honored experimental design patterns can be influential. It is well-known that even something as simple as *assigning* tasks to participants (versus having them use their own, ecologically valid and personally important tasks) can heavily influence outcomes (Russell & Grimes, 2007). Likewise, choices in retrospective experiment designs can be highly influential as well.

We found, for instance, that the choice of cues gathered for recall purposes can spell the difference between no useful results and highly reliable results (Russell & Oren, 2010; van den Haak, De Jong, & Schellens, 2003). In an early (and naïve) version of our study, we tried cueing previous behavior recollection with the search queries presented as strings and associated dates (e.g., "You searched for {Vancouver hotel OR B&B} on Nov 7, 2007. What was your next search for?"). We quickly

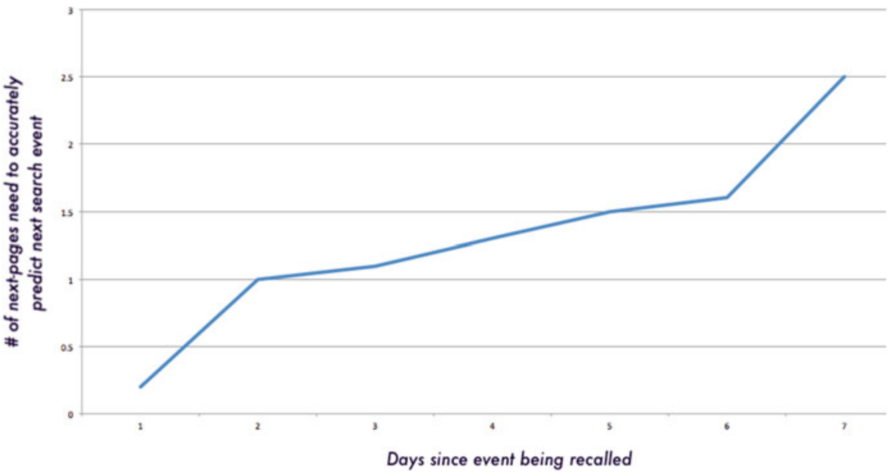


Fig. 5 Events farther in the past required more and more pages to accurately recall the next search event. After about 6 days out, a participant usually needed around 2 or 3 pages to remember what happened next (The numbers are non-integers as they represent the average number of pages required by all study participants)

found that such cues are effectively useless as memory prompts—people simply cannot remember their Internet searching behaviors with probes of this kind.

When we switched to capturing just the browser window (as seen in Fig. 5), recall clearly improved, but the recall error rates were still fairly high. The fairly small switch of capturing the *entire screen* (rather than just the browser window) ended up also triggering memory for a good deal of additional task context information.

The experimental protocol must also include methods to validate that the behavior recalled by the participant is actually the behavior that was performed. As is well known from the design of surveys [cf. Chap. "Survey Research in HCI"] (Holbrook, Green, & Krosnick, 2003), biases in recollection also occur as a consequence of trying to conform to social expectations, simple satisficing, pleasing the experimenter, or to rationalize behavior that seems awkward in after-the-fact review.

Pragmatics of Using Retrospective Methods

While an entire book could be written just on good experimental practices for retrospective methods, a few pragmatic guidelines will be useful for the practitioner.

Choosing good cues: When designing a retrospective study, it is important to capture data that will provide useful memory cues. In general, memories are cued by images or data that encapsulate a good deal of rapidly recognizable context. Thus, images such as screen captures or videos of user performance can be used as cues. Reconstructions of a situation (for example a simulation whose state can be captured in a few variables) that lose memorable or recognizable contextual details are not as promising for cued recall. For example, in an un-cued memorability study of search results (Teevan & Karger, 2005), many of the features of the search results page were forgotten within 60–90 min after the query was run. The only memorable results were ones that had been highly ranked or clicked-on by the user as these were salient to the user's goals. Otherwise, without a good recognizable cue to provide surrounding context, memory is difficult.

Walkthrough methods of interviewing: When interviewing the participants, it is often useful to present earlier data and maintain the chronology of events as they happened in the course of the interview. That is, jumping around from near-past to distant-past (and back) and asking questions about each segment out of order only invites confusion on the participant's part. Just as important, when skipping from one segment of the retrospective data to another, the participant should not be aware of any of the intervening data. Be aware that the cueing stimuli used are the *only* stimuli being tested for recall. Seeing additional cues may significantly alter the answers (and improve!) to later interview questions.

Asking for predictions: While there are many ways to validate the accuracy of the recollections being elicited by the cues, one particularly useful approach is to ask for predictions from the participant. (Roughly, "After you saw this screen, what was your next action?") While not always applicable in exactly this form, the general idea of looking for inter-response consistency in retrospective reports is a valuable thing to measure. This is similar to a method often used in survey design is to ask slight variations on the same question at different points in the survey, testing for consistent replies in responses across variations (Weisberg et al., 1996).

Face-to-face interviews: For retrospective interviews, a face-to-face connection between the participant and the researcher is more effective than distance methods (e.g., telephone surveys). Holbrook et al. (2003) have shown that satisficing and social desirability response biases are more likely to take place in telephone interviews than in face-to-face interviews.

Ways to avoid the false memories effect: As is well-known (Loftus, 2005), interviewers can easily (and often accidentally) introduce false memories by the way they frame their questions. While there is an entire literature on interviewing techniques to avoid introducing spurious information (Loftus, 2005; Memon, Wark, Holley, Bull, & Koehnken, 1997), in an HCI context the challenges are far simpler. Typically, the behavior under question is not emotionally laden (thereby avoiding the effects of eye-witness testimony when charged events occur), and the cueing stimuli are usually data gathered from the participant's own behavior. Good advice to follow when asking questions of past behaviors include:

1. *Avoid direction* about what parts of the behavior should be noticed. That is, avoid cueing the participant to pay special attention to the behaviors that are the focus of the study. If they skip over the important parts, the researcher can ask follow-up questions, noting that they are replies to direct questions.

2. *Avoid value statements* about the behaviors in question, e.g., “When did it stop acting badly?” or “When did you start liking that awesome new interaction widget?” Introducing affectively laden terms can easily alter people’s responses.
3. *Avoid asking for global affective responses from experiences in the past*. As (Schwarz et al., 2009) shows, asking for accurate evaluations of emotional perceptions from the past *cannot help* but be influenced by subsequent events and especially the perception of the entire experience at the end. No amount of rationality can apparently overcome this strong cognitive self-perception bias effect. The participant might intellectually understand that they enjoyed using a system at the beginning of their use experience. But if a later experience turned out to be highly negative, it is tremendously difficult to evaluate the entire experience as positive, even though the average experience might be highly positive. Although factual information about specific events in the past can be accurate, the reconstructive nature of emotional memories makes accuracy difficult.

Avoiding testing children: The age of the participants can be another factor for caution in using RCR. van Kesteren, Bekker, Vermeeren, and Lloyd (2003) found that children between the ages of 6 and 7 often have difficulty holding multiple concepts in memory at once, limiting their ability to both watch a retrospective video of their behavior *and* comment on what they were doing at that time, although it was clear that they could correctly report on changes in their understanding that occurred during the study. (See also (Höysniemi, Hämäläinen, & Turkki, 2003) who found similar cognitive limits on younger children’s ability to reflect on previous performances.) However, Baauw and Markopoulos (2004) found that post-task interviews for usability problems worked about as well as in-lab, real-time usability analysis for children between the ages of 9 and 11.

Another age-related issue that appears with younger children is that reviewing videos is not always an exciting prospect, leading to a certain amount of attentional drift during the retrospective review part of the study. Retrospective interviewing of children is often a researcher’s most challenging task.

Summary

With all this in mind, retrospective studies are a set of methods to gain insight into behavior that is otherwise very difficult to learn. As we have seen, RCR methods can be used to reconstruct participants’ behaviors, rationales, affective reactions, and responses for events that have been recorded. However, there are many challenges to creating a carefully design retrospective study. Such studies must be designed with care, paying particular attention to capturing cues that are useful and engaging for recall, asking questions that do not ask the participant to over-infer what they can accurately recall, and continually validating the responses with the record of actual behavior.

We find this method of gathering user behaviors to be remarkably accurate when recollection cues and interview methods are well-designed, even when there are fairly lengthy delays between original action and recall.

Further Reading and Resources

For us, the development of the RCR technique grew out of a frustration with not being able to see normal user behavior over an extended period of time. Logs analysis is a splendid technique [See Chap "Understanding Behavior Through Log Data and Analysis"], but it does not allow for any particular insight into attitudinal data or an understanding of individual responses over a longer period of time.

To deal with this issue, we built IE-Capture (see above) as a tool to allow our users to "tell their own story" and give us those additional insights into their use of our system. As we interviewed more and more participants, it became clear that the RCR method was both powerful and sensitive. The concern for not over-interpreting the data became evident when we found our participants rephrasing things they had said earlier in our interviews. This in turn led us to study the accuracy of recalled behavior, and to develop our own skills in asking questions that would not bias the participant.

For additional information about the pragmatics of asking questions in retrospective interview settings, please see (Beatty & Willis, 2007) and (Willis, 2005).

For guidance in using the Experience Sampling Method (reconstructing the day's events at the end of each day), see (Hektner et al., 2007).

Exercises

1. Which of the other methods in this book work well with retrospective study methods?
2. What kinds of reports are not generally accurate when people are reviewing a record and/or visualization of their past behavior?

References

Akers, D., Simpson, M., Jeffries, R., & Winograd, T. (2009, April). Undo and erase events as indicators of usability problems. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems* (pp. 659–668). New York, NY: ACM.

Al-Qaimari, G., & McRostie, D. (1999). KALDI: A computer-aided usability engineering tool for supporting testing and analysis of human computer interaction. In J. Vanderdonckt & A. Puerta (Eds.), *Proceedings of the 3rd International Conference on Computer-Aided Design of User Interfaces (CADUI'99), Dordrecht, October, 1999*. Louvain-la-Neuve: Kluwer.

Baauw, E., & Markopoulous, P. (2004, June). A comparison of think-aloud and post-task interview for usability testing with children. In *Proceedings of the 2004 Conference on Interaction Design and Children: Building a Community* (pp. 115–116). New York, NY: ACM.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*(2), 287–311.

Blackwell, A., Jones, R., Milic-Frayling, N., & Rodden, K. (2005, April). *Combining logging with interviews to investigate web browser usage in the workplace*. Position paper for Workshop Usage Analysis: Combining Logging and Qualitative Methods, ACM Conference on Human Factors in Computing Systems (ACM CHI 2005).

Brandt, J., Weiss, H., & Klemmer, S. (2007). txt 4l8r: Lowering the burden for diary studies under mobile conditions. In *CHI '07*, April 23–May 3, 2007, San Jose, California.

Brewer, W. F. (1986). What is autobiographical memory? In D. Rubin (Ed.), *Autobiographical memory* (pp. 25–49). Cambridge, UK: Cambridge University Press.

Brewer, W. F. (1988). Qualitative analysis of the recalls of randomly sampled autobiographical events. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 263–268). Chichester, UK: Wiley.

Brush, A. J., Ames, M., & Davis, J. (2004, April). A comparison of synchronous remote and local usability studies for an expert interface. In *CHI'04 Extended Abstracts on Human Factors in Computing Systems* (pp. 1179–1182). New York, NY: ACM.

Capra, M. (2002). Contemporaneous versus retrospective user-reported critical incidents in usability evaluation. In *Proc. Human Factors 2002,* HFES, 1973–1977.

Chi, E. H., Pirolli, P., & Pitkow, J. (2000). The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site. In *Proc. CHI 2000* (pp. 161–167). New York, NY: ACM Press.

Collier, J. (1967). *Visual anthropology: Photography as a research method*. New York, NY: Holt, Rinehart and Winston.

Czerwinski, M., Horvitz, E., & Wilhite, S. (2004). A diary study of task switching and interruptions. In *CHI* (pp. 175–182). New York, NY: ACM Press.

Dickson, J., McLennan, J., & Omodei, M. M. (2000). Effects of concurrent verbalization on a time-critical, dynamic decision-making task. *The Journal of General Psychology, 127*(2), 217–228.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P. Feltovich, & R. R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 223–242). Cambridge, UK: Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1985). *Protocol analysis*. Cambridge, MA: MIT Press.

Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proc. CHI 2006* (pp. 1253–1262). New York, NY: ACM Press.

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. Thousand Oaks, CA: Sage Publications.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly, 67*(1), 79–125.

Höysniemi, J., Hämäläinen, P., & Turkki, L. (2003). Using peer tutoring in evaluating the usability of a physically interactive computer game with children. *Interacting with Computers, 15*(2), 203–225.

Hyrskykari, A., Ovaska, S., Räihä, K., Majaranta, P., & Lehtinen, M. (2008). Gaze path stimulation in retrospective think-aloud. *Journal of Eye Movement Research, 2*(4), 1–18.

Intille, S. S., Kukla, C., & Ma, X. (2002). Eliciting user preferences using image-based experience sampling and reflection. In *Proc. CHI '02 Extended Abstracts on Human Factors in Computing Systems* (pp. 738–739). New York, NY: ACM Press.

Ivory, M. Y., & Hearst, M. A. (2001). The state of the art in automated usability evaluation of user interfaces. *ACM Computing Surveys, 33*(4), 470–516.

Jain, J., & Boyce, S. (2012). Case study: Longitudinal comparative analysis for analyzing user behavior. In *Proceedings of the 2012 ACM Annual Conference Extended Abstracts on Human Factors in Computing Systems Extended Abstracts*. New York, NY: ACM.

Jones, R., Milic-Frayling, N., Rodden, K., & Blackwell, A. (2007). Contextual method for the redesign of existing software products. *International Journal of Human-Computer Interaction, 22*(1–2), 81–101.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science, 306*(5702), 1776.

Karapanos, E., Martens, J.-B., & Hassenzahl, M. (2009). Reconstructing experiences through sketching. Arxiv preprint, arXiv:0912.5343.

Karapanos, E., Martens, J., & Hassenzahl, M. (2010). On the retrospective assessment of users' experiences over time: Memory or actuality?. In *Proceedings of the 28th of the International Conference Extended Abstracts on Human Factors in Computing systems*. New York, NY: ACM.

Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J. B. (2009). User experience over time: An initial framework. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems* (pp. 729–738). New York, NY: ACM.

Kellar, M., Watters, C., & Shepherd, M. (2006). A goal-based classification of web information tasks. In *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*, Austin, TX (ASIS&T).

Kuniavsky, M. (2003). *Observing the user experience: A practioner's guide to user research*. New York, NY: Morgan Kaufman.

Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology, 113*(3), 387–404.

Lamming, M., Brown, P., Carter, K., Eldridge, M., Flynn, M., Louie, G., et al. (1994). The design of a human memory prosthesis. *Computer Journal, 37*(3), 153–163.

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. In H. T. Reis (Ed.), *Naturalistic approaches to studying social interaction: New directions for methodology of social and behavioral science*. San Francisco, CA: Jossey-Bass.

Loftus, E. F. (1996). Memory distortion and false memory creation. *Journal of the American Academy of Psychiatry and the Law Online, 24*(3), 281–295.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*(4), 361–366.

McCarney, R., Warner, J., Iliffe, S., van Haselen, R., Griffin, M., & Fisher, P. (2007). The Hawthorne effect: A randomised, controlled trial. *BMC Medical Research Methodology, 7*, 30.

Memon, A., Wark, L., Holley, A., Bull, R., & Koehnken, G. (1997). Eyewitness performance in cognitive and structured interviews. *Memory, 5*(5), 639.

Muralidharan, A., Gyongyi, Z., & Chi, E. (2012). Social annotations in web search. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems*. New York, NY: ACM.

Norman, D. (2009). Memory is more important that actuality. *Interactions, 16*(2), 24–26.

Novick, D. G., Santaella, B., Cervantes, A., & Andrade, C. (2012, October). Short-term methodology for long-term usability. In *Proceedings of the 30th ACM International Conference on Design of Communication* (pp. 205–212). New York, NY: ACM.

Rieman, J. (1993). The diary study: A workplace-oriented research tool to guide laboratory efforts. In *Proceedings of CHI: ACM Conference on Human Factors in Computing Systems* (pp. 321–326).

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words that were not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803–814.

Russell, D. M., & Grimes, C. (2007). Assigned tasks are not the same as self-chosen Web search tasks. In System Sciences, 2007. *HICSS 2007. 40th Annual Hawaii International Conference on* (pp. 83-83). IEEE.

Russell, D. M., & Oren, M. (2009, January). Retrospective cued recall: A method for accurately recalling previous user behaviors. In *System Sciences, 2009, HICSS'09. 42nd Hawaii International Conference on* (pp. 1–9). IEEE.

Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist, 54*(3), 182–203.

Schacter, D. L. (2001). *The seven sins of memory: How the mind forgets and remembers*. Boston, MA: Houghton Mifflin.

Schwarz, N., Kahneman, D., Xu, J., Belli, R., Stafford, F., & Alwin, D. (2009). Global and episodic reports of hedonic experience. In R. Belli, D. Alwin, & F. Stafford (Eds.), *Using calendar and diary methods in life events research* (pp. 157–174). Newbury Park, CA: Sage Publishing.

Shiffman, S., Hufford, M., Hickcox, M., Paty, J. A., Gnys, M., & Kassel, J. D. (1997). Remember that? A comparison of real-time versus retrospective recall of smoking lapses. *Journal of Consulting and Clinical Psychology, 65*, 292.

Siochi, A. C., & Hid, D. (1991). A study of computer-supported user interface evaluation using maximal repeating pattern analysis. In *Proceedings of ACM CHI '91* (pp. 301–305).

Steele-Johnson, D. (2000). Goal orientation and task demand effects on motivation, affect, and performance. *The Journal of Applied Psychology, 85*(5), 724–738.

Teevan, J., & Karger, D. (2005). *The research engine: Helping people return to information on the Web*. Paper presented at the Proceedings of the ACM Symposium on User Interface Software and Technology (UIST'05), Seattle, WA.

Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology, 70*, 122–129.

Van Boven, L., Kane, J., & McGraw, A. P. (2009). Temporally asymmetric constraints on mental simulation: Retrospection is more constrained than prospection. *The handbook of imagination and mental simulation* (131–147).

van den Haak, M., De Jong, M., & Schellens, P. J. (2003). Retrospective versus concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology, 22*, 339–351.

Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied, 11*(4), 237–244.

Van House, N. (2006). Interview viz: Visualization-assisted photo elicitation. *Ext. Abstracts CHI 2006* (pp. 1463–1468). New York, NY: ACM Press.

van Kesteren, I. E., Bekker, M. M., Vermeeren, A. P., & Lloyd, P. A. (2003). Assessing usability evaluation methods on their effectiveness to elicit verbal comments from children subjects. In *Proceedings of the 2003 Conference on Interaction Design and Children* (pp. 41–49). New York, NY: ACM.

Weisberg, H., Krosnick, J. A., & Bowen, B. D. (1996). *An introduction to survey research, polling, and data analysis. Thousand Oaks*, CA: Sage Publications.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage Publications.

Wilson, B. A., Evans, J. J., Emslie, H., & Malinek, V. (1997). Evaluation of NeuroPage: A new memory aid. *Journal of Neurology, Neurosurgery and Psychiatry, 63*, 113–115.

Wilson, T. D. (1994). The Proper Protocol: Validity and Completeness of Verbal Reports. Psychological Science, 5(5), 249–252.