Retrospective Cued Recall: A method for accurately recalling previous user behaviors

Daniel M. Russell  
Google  
1600 Amphitheatre Parkway  
Mountain View, CA 94043  
drussell@google.com

Mike Oren  
Iowa State University  
1620 Howe Hall, Ames, IA 50011  
moren@iastate.edu

Abstract

A common problem in many user studies is gathering natural user behavior unintrusively over a long period of time. We describe a methodology for conducting passive longitudinal studies, where the participant is able to go about their daily routine without taking any disruptive action such as writing a diary entry or responding to an interruption. Although passive user observations have been done through log analysis, our retrospective cued recall (RCR) method recorded participant screens over a month long period linked to logged trigger events, allowing for a later review that was guided by retrospective recall as cued by the images of screen captures. We find this method of gathering user behaviors to be remarkably accurate, despite fairly lengthy delays between action and recall.

1. INTRODUCTION

Traditional longitudinal study techniques tend to be event driven using event triggered diary entries and interviews or log data analysis. However, log data analysis does not provide insights into user motivations, nor are logs capable of providing context about what a user did before, during or after. Diary studies and interviews, however, are difficult to sustain for a long period of time since they rely on participant motivation, which tends to lag as the study continues. [1] 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 around quiet user behavior is desired and researchers are interested in the context and motivations of participants in a study, the current longitudinal study methods are insufficient.

In this paper, we describe an active logging tool and a methodology for conducting logging-based longitudinal studies that solve many of the problems of current longitudinal study techniques. This method is particularly useful when it comes to the need for consistent user motivation and the ability to observe the context of user activity that is not triggered by an event or task.

We report on the outcome of two studies: the first study reports on the methods used to gather data in a situation that had previously been difficult to track—passive use of a particular web page feature. The second study summarizes our experiences in measuring the accuracy and quality of participant recollections when being retrospectively cued for their past behaviors.

2. CONTRIBUTIONS AND BENEFITS

We present a simple method to support recall of earlier situations and events. This noninvasive method improves recall, helping track and understand activities that are not event triggered by captured image assisted recall.

WHAT DO USERS DO WHEN YOU’RE NOT THERE?

While in-lab usability studies are effective at discovering some kinds of UI use patterns, it is notoriously difficult to track the long-term behavior of users in the wild. [13] Specialized tracking devices, diary or interruption studies [3, 9] can all materially affect user behavior by continually reminding the user that their behavior is being tracked and monitored.

Our approach is to take advantage of human’s well-known ability to visually recognize images, particularly images of an environment (the computer screen) that have been created as part of the normal course of work. [4, 5] This human ability to recall situational information when retrospectively cued by imagery obtained at the time is the key to this method.

These 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.
3. PREVIOUS WORK

There is a long tradition of using photos of key events to cue retrospective memories. Collier [20] is mostly closely associated with the photographic technique in anthropology settings, when photos are used as prompts and foils to elicit memories and context around some circumstance. While van Gog [22] reports on the tradeoffs involved in using concurrent vs. 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 key insight that images can be used for HCI recollection purposes as well can be seen in the work of Van House [19] and Intille [21], although these (and other similar systems) capture images of the world context, and do not provide the internal tracking of events in the user’s experience of the online world.

Many systems have been built to quietly log user events over time for later analysis. These range from the obvious web behavior logging analysis systems to client-side tooling that records user behavior in high resolution from the user’s perspective.

Perhaps the system closest to our approach is LogViewer [1], a tool that logs user events and screen images in web behavior for later analysis. LogViewer also creates a tree analysis 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, how user navigation is organized in terms of landmarks pages, etc. They also interviewed their participants, but with a focus on gathering contextual information to aid in their application redesign purposes.

[17] has a logging system built into a bespoke browser (a modification of Internet Explorer) that allowed the user to label their own behavior as they completed tasks. As with all systems of this kind, 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, 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 [18, 2, 6] log events for later analysis and are intended to support navigation and information browsing behavior analysis. Please see [8] for a complete summary of many such tools developed for tracking and logging web behavior to improve usability analysis.

4. IE Capture: Client-side behavior logging

We developed a plug-in module for Microsoft’s Internet Explorer (IE Capture), to capture every browser click, event and page view. (Figure 1). When the browser’s pageload-complete events fires, IE Capture takes a screenshot of the entire working display space (there are options for capturing the entire screen or just the Internet Explorer window). IE Capture runs quietly in the background, building up a log of all browser actions (page views and click events), each action linked to a matching screen capture. These screenshots are then reviewed with the participant and are often used in conjunction with diary entries to get a complete understanding of user motivations and behaviors.

Privacy issues: To address privacy concerns, IE Capture also has a blacklist where participants are free to add any URLs they wish to exclude from being captured. The blacklist works by checking to see if any of the string in the blacklist are contained anywhere in the current URL, if the word appears then no screenshot is captured.

![Figure 1. The IE Capture plug-in control in Internet Explorer provides an interaction point for participants to activate / deactivate logging.](image)

The blacklist works by checking to see if any of the string in the blacklist are contained anywhere in the current URL, if the word appears then no screenshot is captured. For example, if “https” appears in the blacklist and a participant goes to any secured webpage via the HTTPS protocol, then no screenshots of that page are captured. In addition, all screenshots are stored locally on the participant’s machine and they are given the location of the files so they are free to review and edit them before the researcher performs a field study to collect the screenshots and interview the participant. Participants were also given the option of temporarily disabling the screen capture software whenever they want to browse the Internet without generating screen capture logs.

In addition, since there was no need in this study to capture all Internet activity of the participants (especially when capturing based on download complete—which is triggered whenever any single element on a page completes loading), the blacklist was augmented with a whitelist, in hopes that this would also help ease any privacy concerns. The whitelist specifies the substrings of URLs that would be captured, so the whitelist could contain “google.com” while the blacklist would contain “mail.” With this
specification, IE Capture would capture anything with google.com in the URL string except for mail.google.com (GMail).

Figure 2. IE Capture’s privacy options

With these revisions to IE Capture, we were able—with the user’s explicit permission—to track all useful and appropriate web-based browser interaction on the participant’s machine.

In addition, since the study was done as a remote study (meaning that researchers could not visit the home of participants to collect the screenshots and logs) a short Python script was written that zipped up all participant screenshots and e-mailed them to the researchers. The script was converted to an executable file so participants did not have to download and install Python and the files were only zipped up and e-mailed out when the participant manually ran the program (so they were free to review the images before screenshots were sent out).

4. STUDY #1: TRACKING PASSIVE BEHAVIOR

Using IE Capture, one of the authors developed a longitudinal study to explore how users of a personalized homepage product make use of its current capabilities. A second goal was to uncover any discoverability issues that may not be as apparent in a lab study. A personalized homepage is a difficult thing to study, as it is a web page where users leave it up in a kind of background mode, collecting information throughout the day, and only interacting with the page infrequently. As a consequence, traditional longitudinal study techniques such as diaries were ruled out because of their heavyweight nature: manually recording interactions during quickly occurring events that are lightweight by nature seems a recipe for disaster. Log data analysis was conducted as a preliminary first step, but it did not provide information regarding user motivation and was thus very incomplete.

Study #1 Design

For this study, twelve participants were recruited into three different groups. The groups were divided based on how long participants had been active users of the personalized homepage (those who had never visited the site, those who had been active users for over two weeks but less than three months, and active users for more than three months), with each group containing four participants. All participants were remote participants. Participants were divided along these lines to separate users into the user groups of novice, intermediate, and expert users of the personalized homepage. This allowed us to compare the experience between these different user levels.

All participants were scheduled for a total of four interviews over a period of one month. Participants were asked to install and run IE Capture on their computer for the entire month-long duration of the study. One to two days before each weekly participant interview, the participants were sent an e-mail reminding them to run the program to upload the screen capture data. This was a simple operation, usually requiring only a double-click with no more intervention.

Conducting the study

The first interview with participants consisted of a one hour background interview with the participants where they were asked about their current homepage, whether they had ever changed their homepage, and some questions regarding their current use of the personalized homepage and features that they use. If a participant had never used the personalized homepage before, they were asked whether or not they were aware of any personalized homepages, if they knew if the creator of the personalized homepage product had personalized homepage offering, and how they would find it. After discovering the personalized homepage, new users were asked about their initial impressions and asked to make it their own while explaining what they were doing and why.

In each later interview, after receiving the screenshots from participants we reviewed the log and screenshots, taking notes regarding user activities. We were able to find out when users were on the personalized homepage, when they accessed the page, roughly how long they left the page open (since screenshots were taken on download complete and all participants had at least one gadget on their page that updated every few minutes), which portions of the page they looked at, etc. Often, participants performed other activities on their computer while using the personalized homepage, and we were able to observe these activities since anything placed on top of Internet Explorer was captured by the screenshot taken by IE Capture. Furthermore, even with the whitelist as opposed to the blacklist, we were able to observe a single screenshot of sites participants visited after the personalized homepage when they clicked on links within the personalized homepage. These activities were then compiled and written as interview questions.
so that the participant motivations could be better ascertained.

Many of the activities obtained through the use of IE Capture, such as leaving the personalized homepage open in the background while watching a video, browsing other web sites, etc. could not have been obtained through diary studies where such entries would have been tedious for participants to record. While participants would have been able to inform us that they left the personalized homepage open while performing other tasks on their computer, they would not have been able to give us the same level of details as we were able to obtain through screen captures. Furthermore, much of the personalized homepage activity was passive with participants checking the page periodically to check on updates to news feeds, e-mail, etc. and information about this would have eluded us had we used standard longitudinal study techniques.

**Lessons learned: Pragmatics**

We initially encountered problems when recruiting for this study since potential participants raised privacy concerns due to the use of IE Capture. In the future, we think providing potential participants with a FAQ detailing the steps we have taken to help ensure their privacy, such as the whitelist and the fact that they can review the screenshots before sending them to us.

In addition, we learned that the initial version of the program that uploaded the logs and screenshots had some shortcomings (it didn't split the screenshots into smaller chunks if the zip files were over 20 MB, so the users had to manually split the files up into subfolders). These shortcomings were later fixed but the lack of this feature prevented us from obtaining the screenshots in as timely a manner as we would have liked. In general, it was wise to make the program as smart as possible to prevent the user from needing to take any action beyond running the program to send the log as participant technical abilities varied widely.

Despite stressing the point that participants should use the personalized homepage the way they normally would, several participants (3 of the 12) commented on “feeling bad” if they had not used the personalized homepage much during the week or had not used certain features. This change in usage was problematic as, at times, it seemed to provide an artificial view of the participants personalized homepage usage. Since the participants were contacted every few days, this reminding effect may be countered by having participants go through the study for a longer period of time before talking to a researcher. In such cases, the best outcome is to encourage a low profile and have them "forget" about (or at least not notice) their participation in the experiment. The goal, after all, is to encourage natural use. Without having a longer period of time between interviews, the only way to counter this is to simply be aware of it and try to discover any changes in participant activity by paying attention to their activity patterns and asking them about abnormalities in their usage. However, this is a particular problem for new users as they do not often have a "normal" routine with the personalized homepage, so countering this problem with new users can only be countered by either interviewing them less often or continually stressing that they should just use it like they would if they were not part of the study.

Several participants used the personalized homepage on more than one computer. For some of these participants, we might have been able to get them to install IE Capture on their other computer (but in some cases that was a work computer). An alternative to multiple installations of IE Capture would have been giving users some way of entering in diary entries; however, we felt diary entries would be inappropriate as there is no specific triggering event and participants tended to keep the personalized homepage up—something that would not be captured properly through diary entries.

**Method Benefits**

*Ease of use leads to participant retention:* Out of the twelve participants recruited for this study, only one participant ended up dropping out before the completion of this month long study. This represents a 8.3% dropout rate, which given the length of the study we felt was very good and far below our estimated 25% dropout rate that we had planned into the recruiting. We feel this low dropout rate is in large part due to the ease of participation in the study.

*Richer log content simplifies analysis:* In addition to the low dropout rate of participants, this method also helped the researchers analyze the data. Whereas traditional diary studies require reading through and parsing pages of printed entries that are obtained either on field visits to the participant or through the mail, this technique allowed us to quickly go through visual entries which could quickly be sifted through and analyzed for trends. This visual analysis was much faster and easier than reading through pages of diary entries, plus because the screenshots were what the participant actually saw we did not have to worry about inaccuracies due to participant memories.

While other new techniques such as online and mobile diary entry provide researchers with ways of actively monitoring diary entries these techniques still rely on participants to report their activity, and participants may not always report all activity [3]. Using IE Capture researchers do not have to rely on participants to determine and report the activities they deem to be important. Instead, all activity is recorded and the researcher can go through it and discuss the motivations with the participant. This allows for more
accurate and more complete record of participant activity than is available from other longitudinal study techniques.

Cueing works well: To learn about user motivations, we asked the participants what made them decide to take certain actions (while reminding them of the context of the action, based on the logs), such as before looking for additional content to add. We noticed that looking at the screen image while they were answering the question made them much more confident about their answers. When asked about why they chose to add particular content to the page after browsing or searching for it as opposed to other options they looked at, participants would often comment “I remember this...” while looking at the screen image. Participants were also asked if content they added had met their expectations and met the needs or desires they had before looking for new content. If a participant had looked at content, but ultimately did not add any content or did not keep the content they had added, she was asked what was missing from the content options available. Again, the visual cueing worked very well.

Tracking tacit behaviors and the lack thereof: One piece of crucial information we were able to obtain from this study was information about user abandonment. Two of the participants in the “new user” group ended up abandoning the personalized homepage during the course of the study. We were able to determine this since the log data indicated that they were only on the personalized homepage for a few minutes on the day we sent them an e-mail asking them to send the screenshots of their usage. Participants were reluctant to discuss their abandonment of the personalized homepage and we feel it is unlikely we would have caught this abandonment had we done a traditional longitudinal study where participants report their own usage since we feel they would have been unlikely to report their abandonment of the product. In addition, by discovering this abandonment before the end of the study, we were able to develop questions for the final interviews in order to get a better understanding of their reason for abandoning the product. This would have been extremely difficult to do if we were looking through diary entries in their presence and the trends may not have been as obvious.

Furthermore, with most diary studies lasting 2-3 weeks on average due to high drop out rates beyond that period, the abandonment of the personalized homepage would likely have been imperceptible in a traditional diary study. We were only able to begin to observe the abandonment on the third week and if we only had the data collected from the first three weeks, we would have assumed the participant had just been busy during the third week. This critical information was only attainable because we were able to observe first hand through screenshots the participants activity on the page over an extended period of time.

Study #1 Summary
This study was rewarding. We found that month-long longitudinal studies are not only possible, but also fairly straightforward, even when the participants are (as in this study) remote. It became clear that the visual examination of the screenshots (particularly when seen in their original event sequence) were acting as powerful retrospective cues for the participants. Furthermore, because the data collection process was essentially invisible, we had reason to believe that much of the behavior was genuinely uninfluenced by the presence of our instrumentation.

But we were left with a question: How accurately were participant able to recall their behaviors? Were the stories and memories accurate? Or were they generated in response to our questions, with the potential for falseness? [14]

5. STUDY #2: HOW GOOD ARE THE MEMORIES?
An important question is whether or not our participants could actually accurately recall the circumstances of their use of the individualized home page.

To find out how well people could remember their past behaviors from a retrospective-cue such as that provided by IE Capture, we performed a second study.

Study #2 Design
In this study, we recruited 8 participants, (5 women, 3 men; ages 25 to 46; all with advanced education, but not computer scientists or HCI professionals). All were located nearby so we could interview them in person at the end of data collection period. At the beginning of the 7-day experimental run, an experimenter installed IE Capture on their personal computer (no shared computers allowed) and let it run for 7 days. At the end of that time, the experimenter would return to hold a structured interview in the location and context in which the web browsing behaviors took place. The participants were not told the object of the study, but simply that we were studying “home computer use.” All of the participants were screened to be web search users (of at least 3 searches per week).

The interview was structured to ask a specific set of questions about the participant’s web searching behavior. In all cases, participants had many more than 3 searches in the week of study (range: 6 – 37 searches in the week of study).

Interview questioning procedure: At the beginning of the interview, the experimenter would choose 3 evenly spaced queries in the session log—2 days back before the interview, 4 days back and 6 days
Then visit those pages in the log with IECaptureViewer, our tool to view and scrub through the data logs and screenshots. (See Figure 5 for the log viewing tool used.) As it happened, all participants but one had searches on days 2, 4 and 6. In the one missing case, the next prior search event was used (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.

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 participant could recall the situation at all.

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

We were measuring the participant’s ability to speak accurately about the next major event in their search process. That is, having cued their retrospective memory of an event, we measured their ability to recall the next step in the process. In nearly all cases, the judgment by the research was clear and evident: either the participant could accurately predict what was coming up in the log, or they just couldn’t say. Rarely, did a participant guess and feel confident.

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 the next-page before they could recollect what was going on in the search.

As can be seen in Figure 3, seven of the eight participants could accurately predict the next search event after two days. That’s not terribly surprising, given that searches are relatively infrequent (in this participant pool, the average number of searches / week was 11). A search done two days ago is relatively recent and stands out by its relative rarity among the total number of web events.

However, as we tested farther into the past (4 days and 6 days out), participant recall was still quite good. Even after nearly a week, participants were able to recall the next search event correctly 75% of the time. With the additional prompting of advancing to the next page in their cached screenshots using the IE Capture Viewer, participants can recall accurately after seeing only 3 additional screen images taken from the log/screen files.

![Figure 3. The number of correct next event predictions drops after 2 days, but is still at 75% correct.](image)

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 after 6 days, 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.

6. DISCUSSION

In Study #1 we discovered that the use of IE Capture enabled our participants to richly and confidently talk about events that happened in the course of their web use. Even the absence of certain kinds of events, such as the diminution in the use of a particular feature, was available for discussion and introspection.

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 kept worrying us. That concern led to Study #2 and an attempt to measure the accuracy of the recollections.
As became quickly clear, participants were not only able to make accurate predictions about particular events for which they had not been pre-conditioned 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 inclusiveness of the screen images: because they could often see other windows in the background (the corner of the Excel spreadsheet, say) those tiny peripheral context cues would give them a distinct sense of time, activity and place. [16].

A particularly difficult part of the protocol described here is that the recollection process begins on the first query in a search session. While beginning on the initial query makes the questions unambiguous, it also requires that the participant remember the context that leads up to the query as cued by that single instance. In our next trials, we plan on backing up two or three pages in order to give our participants a bit more of the temporal context as a way of improving recall even more.

One possible confound that we noted (but were luckily able to avoid in this study) was the potential for confusion over repeated searches. As others have noted [15], repeated searches are fairly common. When we were selecting queries from the participant’s log, we took care to avoid queries that had been repeated, as it would be difficult for the participant to reconstruct what happened on that particular instance of the search (as opposed to other instances within the past few days).

In the course of our studies, we have actually run many more participants than we tested for this paper. To date, we now have experience with more than 30 participants, and have been struck both by how well people remember their previous search history—when retrospectively cued. Without cueing or just by cueing with web log history entries, recollection is very low after one day.

We have also been impressed by the discovery in Study #2 that participants seem to forget about the presence of the IE Capture logger after about four days. As we noted in Study #1, people do seem to slightly modify their behavior within the first 3 to 4 days after installation. There is also compensatory behavior after being reminded of its presence, as when the participant manually causes an upload event to occur.

Few people disabled IE Capture after it was initially turned on. This suggests both that participants did not find it to be invasive and that its performance did not affect machine behavior substantially. In fact, in one memorable instance, we discovered IE Capture was still running in the background of one participant’s computer 90 days after the experiment. Before turning it off, we took advantage of the situation to do a study of 30, 60, and 90 day post recollections and then going through the same protocol as before. We repeated the experimental design from above, locating searches done 30, 60 and 90 days in the past. Surprisingly, we found that our participant could recall many of the search events from as long as 90 days ago, albeit with many more next-page looks at the surrounding context. In each case, after looking at a few surrounding pages, the participant would exclaim “oh yes…that search,” and then proceed to describe the rest of the session fairly accurately.

This unexpected anecdote gives us confidence that memories for the relatively recent past, as measured in our studies, can be relied upon as a source of correct behavioral information.

7. CONCLUSION

Using a lightweight, passive longitudinal study method for capturing screenshots with log data, then conducting later interviews using the screenshots as retrospective cues, we were able to collect accurate contextual information from participants with little effort on their part. Participant dropout rates also seem to be lower than traditional longitudinal studies. From a researcher perspective, the data analysis can be completed more quickly and the recollections of the participants are fairly certain to be accurate.

While the results of this logging method are promising, privacy concerns for participants make it a bit more difficult to recruit for studies. While we have had reasonably good experience in recruiting participants, we have had to be very clear in explaining what is happening and the data retention policies. These privacy concerns must be kept in consideration when recruiting participants to ensure proper lead team for recruitment and that any potential participants are properly briefed on the data being collected.
Once explained clearly, the use of the screenshot logger and the retrospective cued recall method seems promising as a way of quietly collecting longitudinal behavioral data accurately.

8. ACKNOWLEDGMENTS

We would like to thank Cindy Yepez for her help and efforts in recruiting participants for the study. In addition, we would like to thank the participants of this study.

9. REFERENCES


Figure 5. IE Capture Viewer—a tool for reviewing the participant’s log and screen images for discussion and retrospective cuing. 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.