Studying Access Control Usability in the Lab: Lessons Learned From Four Studies

Kami Vaniea, Lujo Bauer, Lorrie Faith Cranor
Carnegie Mellon University
Pittsburgh, PA, USA
{kami,lbauer,lorrie}@cmu.edu

Michael K. Reiter
University of North Carolina
Chapel Hill, NC, USA
reiter@cs.unc.edu

ABSTRACT

In a series of studies, we investigated a user interface intended to help users stay aware of their access-control policy even when they are engaged in another activity as their primary task. Methodological issues arose in each study that impacted the scope of the results. We describe the difficulties encountered during each study, and changes to the methodology designed to overcome those difficulties. Through this process, we shed light on the challenges intrinsic to many studies that examine security as a secondary task, and convey a series of lessons that we hope will help other researchers avoid some of the difficulties that we encountered.

Keywords

access control, human factors, methodology, privacy, visualization

1. INTRODUCTION

Websites that allow users to upload and share content often give users the ability to control, via permission settings, who can see their content. However, interfaces and mechanisms for setting and viewing permissions often fall short at providing users with an effective way to detect and correct misconfiguration in their access-control policies [7, 18].

In a series of studies, we investigated an interface intended to help users stay aware of their access-control policy even when they are engaged in another activity as their primary task. More specifically, in the context of a photo-sharing site, we investigate whether making access-control policy visible to users while they are engaged in a non-security-related primary task can improve the users' understanding of, and ability to correctly set, a desired access-control policy.

Our primary hypothesis was that if the current permission settings are shown in close spatial proximity to the resources they control, instead of on a secondary page, users are more likely to notice and fix permission errors. To test our hypothesis we need our participants to interact with the display as as secondary task, where they have a non-security primary task and interacting with permissions is secondary.

Other researchers have studied security as a secondary task using various approaches [6, 13, 16]. One approach, used by Haake et al. [6], is to conduct a long term study where the participant is made aware that security is a part of the study but the study is run for long enough that the user stops focusing on security. Another approach, used by Sunshine et al. [13], is to not make the participants aware of the security nature of the study, but the study design forces participants to engage in a security behavior while trying to complete their primary task. A final approach, used by Wang [16], is to keep the participant unaware that the study is about security and give the participant the option of whether or not to interact with the security functionality.

To test our hypothesis we decided to use the last approach. We conducted a lab study where participants performed various photo management tasks. Depending on condition, participants were shown permission information under the photos, elsewhere on the page, or on a secondary page (control). We endeavored to control for anticipated study issues. However, we stopped the study early when we ran into multiple methodological problems, including outcome measurement and participants not treating security as a secondary task.

When designing the initial study methodology, we wanted to meet the following goals: make security a secondary task (Section 4), give the participant ownership/responsibility for the albums (Section 5), make sure the participants understood the policy they needed to enact (Section 6), and develop clear metrics for measuring the outcomes (Section 7). Despite careful planning we encountered methodological issues on every one of these goals.

In this paper, we discuss this study and three subsequent ones, each of which took into account the methodological issues that arose in the proceeding study. We focus our discussion on aspects of the methodology that tried to accomplish the four goals described above. We describe the difficulties encountered during each study, and changes to the methodology designed to address those difficulties. Through this process, we shed light on the challenges intrinsic to many studies that examine security as a secondary task, and convey a series of lessons that we hope will help other researchers avoid some of the difficulties that we encountered.

2. STUDY GOALS

The purpose of all four studies was to test the hypothesis:

H: Users who see information about access-control permission settings on the main interface notice permission

LASER '12 July 18-19, 2012, Arlington, Virginia USA

errors more often than users who have to proactively open a second interface to view permissions.

When designing study 1 to test H we wanted to create a study environment that met the following four goals:

Secondary permission task Participants should be in an environment where there is little encouragement to engage in security tasks and the benefits, if any, are not immediate. Users treat security as a secondary task because the benefits of security are often hard to envision but the cognitive and time costs of engaging in it are immediate [17].

Other researchers who study security technologies have successfully simulated the secondary task mindset in the lab. Whitten and Tygar's work on email encryption had participants focus on sending and receiving emails while they measured the usability of PGP [19]. Similarly Sunshine et al. asked participants to find information on websites while studying their reactions to SSL errors [13].

Participant responsibility Participants should feel they are sufficiently responsible for the experimental content to be comfortable making changes they deem necessary. Because changing permissions is secondary, the framing of the study should make it clear to participants that they should make changes outside the bounds of their primary task.

When replicating the SSL study described above, Sotirakopoulos et al. experienced issues with participants claiming that the lab was a "safe" environment so they behaved differently [12]. Witten and Tygar overcame this issue in their work [19], but doing so requires careful study design.

Ideal-policy comprehension Participants should be aware of and comprehend the *ideal policy* – the correct set of permissions for the content. The participant needs to have a clear ideal policy associated with the content they are working with. Participants need to be able to consistently decide when a permission setting is "correct" or "wrong."

Effective outcome measurement We need to be able to accurately measure if participants are noticing and fixing errors. In real world environments the presence or absence of an error can be very subjective and dependent on context [2, 3, 8]. To accurately test "noticing" errors the we need to be able to differentiate between environments with no errors, environments where participants are not noticing errors, and environments where errors have been noticed.

2.1 Overall System Design

We decided to use a photo management website as the domain because it is a common type of website where end users might set access-control policy. We chose to use an open source photo management website software, Gallery3 [1], because it was easy to modify and unknown to general users, thereby ensuring minimal bias from prior experience or training.

We built a Gallery module which displays permission information in a small display that appears under the photos/albums (Figure 1), or in other parts of the interface. We also built a new permission modification interface that shows the permissions for every album on a single page. The permission modification interface was designed to be easy to use and comprehend based on prior work [9, 10] and was not the focus of this research. Access-control permissions in Gallery are expressed as four-tuples of (user group, album, action, decision). Permissions cannot be expressed for individual users or photographs.

3. GENERAL STUDY DESIGN

Our initial study design was intended to test the following hypotheses in addition to our main hypothesis H.

H1 Users who see permission information under photos/albums notice errors more often than users who see permission information in other spatial locations.

H2 When a permission is changed to an error state by a 3rd party, users who see permission information under the photos/albums or on the sidebar notice errors more often than users who see permission information only if they click to a second page.

H3 The type of error, too many permissions or too few, has an effect on the number of errors noticed.

H4 Participants who see permission information under the photos/albums or on the sidebar can recall those permissions better than participants who see permission information only if they click to a second page.

H5 Participants in each of the conditions take the same amount of time to complete each task.

In this work we discuss the methodologies of four similar studies. It is impossible, given space limitations, to due full justice to the methodologies of all four studies. In this section we present the core methodology used in all four studies. In the following sections we detail the unique methodological choices made in each study to meet the goals described in Section 2. We discuss the outcome of the choices and how they informed the methodological choices in the next study.

The first three studies were between-subjects lab studies and the last was a within subject online study. All studies used a round-robin assignment to experimental conditions. Participants in all conditions performed the same tasks. Each study had a slightly different set of conditions, but two conditions were present in every study: the control condition was the default interface, which included a link to the interface for changing permissions; the under-photo condition additionally included a proximity display under photos/albums (Figure 1).

Participants were asked to role play [4, 11, 19] the part of Pat Jones, who manages online photo albums using Gallery. Role playing is a commonly used method of encouraging user engagement. Whitten et al. successfully use it to encourage participants to view security as a secondary task. Tasks were communicated to the participant in email format. In the first three studies the emails were delivered to the participant on paper by the researcher administering the study, in the last study they were shown in an html frame above the website.

Participants started with a training that showed them how to perform several actions on the website including: changing titles, rotating photos, and changing permissions. Participants were asked to perform all actions described in the material to ensure that they understood how to manipulate the interface. In studies 1-3 this training was done on a separate instance of Gallery with fewer albums than the rest of the study. In study 4 the training and the tasks were done on a single Gallery instance.

After the tutorial, participants in study 1 and 2 were given several short warm-up tasks. These tasks were to ensure that the participant had understood the training. It also gave them an opportunity to acclimate to using the interface. Participants in studies 3 and 4 were given 1-2 full task sized warm-up tasks to acclimate to the interface.

The bulk of the studies were composed of a set of tasks presented to the user in sequence. Each task was composed

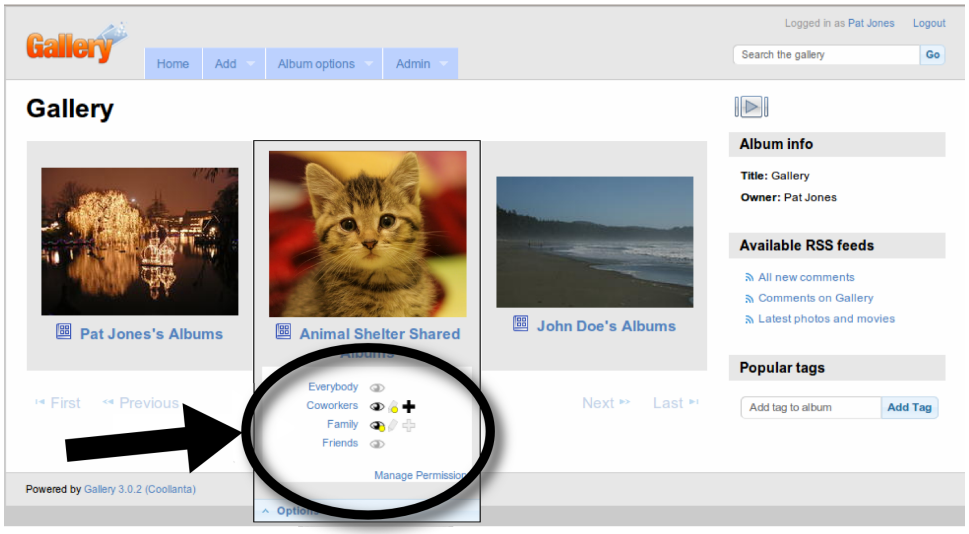

Figure 1: Example of proximity display used in studies 1 and 2. The interface for studies 3 and 4 had a slightly different permission display interface design.

of a set of *subtasks* – issues with the album that the participant is expected to correct to successfully complete a task. A primary subtask was directly expressed in the email and several additional subtasks were implied by observable errors such as rotated photos, misspellings, and incorrect permissions. All tasks contained at least one explicit and one implied title, rotate, delete, or organize subtask intended to distract the participant.

Some tasks were *prompted* in that if the participant failed to correct any subtask, permission related or otherwise, they would be presented with an email pointing out the mistake and asking that it be corrected. *Unprompted* tasks refer either to tasks with no associated prompting or to participant interactions with a task prior to receiving prompting. Participants were unaware of which tasks were prompted until they received a prompt.

Some albums were *changed* mid-way through the study. The participant first interacted with an album and was made aware of the current state, including permission settings. When the participant was distracted by an unrelated task the researcher made changes to the album. The participant was then instructed to interact with the now changed album again.

Finally, participants filled out a survey that asked them to recall permissions for a selection of albums they worked with, as well as non-task albums with correct and incorrect permissions. For each combination of album, group, and permission the participant could answer *True*, *False*, or *Not Sure*. The survey also asked demographic and prior experience questions.

Study 1 was an hour long between-subjects lab study. Participants were given printed training materials that they worked with for about six minutes. This was followed by five short warm-up tasks which took an average of eight minutes in total. Participants were then given 8 tasks which took an average of two and a half minutes each. Tasks appeared in the same fixed order for all participants. Finally, they filled out the survey. There were five prompted tasks and two changed albums. This study was run on 26 participants and three conditions. It was stopped early because of issues with the methodology.

Study 2 was an 1.5 hours long between subjects lab study. Participants were given printed training materials that they worked with for about five and a half minutes. This was followed by five short warm-up tasks, which took approximately 8 minutes to complete in total. They were then given 12 tasks to perform, which took an average of 3.5 minutes apiece. Tasks appeared in the same fixed order for all participants. Finally, they were asked to fill out the survey. There were five prompted tasks and three changed albums. This study was run with 3 conditions and 34 participants, one participant was excluded, resulting in 11 participants per condition. Further details of this study can be found in Vaniea et al. [15].

Study 3 was a 1.5 hours long between subjects lab study. Participants were given printed training materials that they worked with for about five and a half minutes. This was followed by two large warm-up tasks taking approximately 13 minutes to complete. They were then given 15 tasks in a random order which took an average of 3.5 minutes apiece. Finally, the survey was verbally administered by the researcher, followed by an unstructured debriefing interview. There were three prompted tasks and no changed albums. This study had two independent variables: proximity display and permission modification interface. The proximity display was shown either under the photo (under photo) or not at all (control). The permission modification interface was either a separate page with all permission settings shown or a dialog with only one album's permission settings shown. There were 9 pre-study participants and 33 actual participants in this study.

Study 4 was a hour long within subjects online study conducted on Mechanical Turk. All participants performed training, warm-up, and tasks for both the proximity display condition and the control condition. The order in which participants saw the conditions was assigned round robin. Participants completed a set of training tasks which took an average of four minutes. Then they completed a warm-up

To: Pat Jones <pat@jones.com>
From: Josh Needen <josh@hotmail.com>
Subject: New photos

Yo Pat,

Here are the better photos from the Building Jumping trip last weekend. Could you put them up on your site? Just set it up like any of your other albums. Also could you title the photos with the people in them? I had the red parachute, George had the green one and of course your's was blue.

When you are finished send me back a link so I can forward it to the rest of our friends.

Thanks,
Josh

Figure 2: Email from Pat's friend stating in passive voice that everybody in the Friend's group needs to be able to view the photographs.

task that took an average of three minutes. They were then given 7 tasks, with a maximum of two minutes to complete each of these tasks. Tasks appeared in the same fixed order for all participants. When finished with both conditions they were given a survey to fill out that asked questions about both conditions that the participant worked with. There was one prompted task and one changed album per condition. There were 300 pre-study participants and just over 600 actual participants in this study.

4. SECONDARY PERMISSION TASK

Participants should be in an environment where there is minimal encouragement to engage in security tasks, and the benefits, if any, are not immediate.

4.1 Study 1

We decided to give participants a primary task that would take the majority of their attention while still being sufficiently open ended that they would consider engaging in other subtasks. We communicated the tasks through printed emails because the structure allowed us to give context, such as the ideal policy, to the task without drawing too much attention to it. To prevent users from perceiving permission content as explicit direction, we stated all permission information in passive voice and all primary subtasks in active voice. For example, the email in Figure 2 explicitly asks that the titles be changed, but also implies, that the Friends group needs to be able to view the photos. The ideal policy components, that could not be expressed passively, were embedded in information pages about Pat's friends, family, and co-workers.

We were concerned about giving participants too much permission *priming* – the amount participants are encouraged to engage in permission behaviors. Every time a participant reads or interacts with permission information they are being primed to think about permissions. We compromised by creating three blocks of tasks separated by information pages. Two of the tasks had permission errors and in the third task permissions were never mentioned. This third task was to give the participant time without permission priming.

To test behavior in the absence of prompting, the first two tasks were unprompted. If the participant did not correct permissions on these albums, the researcher did not make them aware of the issue. Participants were first prompted about permissions after the third task. We prompted here to be sure participants knew what the album permissions were before they were changed by the researcher.

Outcome Participants rapidly deduced that this was an error-finding study and tried to find and correct all the errors. However, none of the participants noticed that the study was solely about permissions. While participants may have been biased to look for errors, only 67% of participants noticed any permission errors without prompting and no participant noticed all the errors. For comparison 86% of the title errors were corrected.

Over-priming participants to identify and fix errors in general may have caused a control condition behavior we termed "checklisting." Participants who checklisted would reach the end of a task, pause and appear to go through a mental check list. One participant did this out-loud, listing all the types of errors she had seen in the training material, making sure she had checked all of them before moving on.

Additionally, many participants never obviously consulted the permission display to determine if there was an error before opening the permission modification interface. We hypothesized that since all emails mentioning permissions were associated with albums containing permission errors, participants always needed to open the modification interface and had no need to consult the display.

4.2 Study 2

In study 1 all tasks that expressed permission information in the email had permission errors. Effectively there was no "cost" to checking permissions because participants could determine from the email that there was a permission error. To address this concern we added a new hypothesis:

H6 Participants who see permission information on the main screen are, in the absence of an error, less likely to open the permission modification screen than users who have to proactively open a second interface to view permissions.

New Read-permission tasks We added three new tasks where the email expressed the ideal policy but the current settings matched the ideal policy, so there was no permission error. After this change, 50% of tasks expressed the ideal policy and had permission errors, 25% of tasks expressed the ideal policy but had no error, and 25% of tasks did not express an ideal policy. Two of the new tasks were prompted. If the participant did not obviously check the permissions, the researcher prompted them with an emailed question about the permissions. The new tasks were also intended to test if participants used the displays to determine the lack of an error (H6).

Outcome The addition of the new tasks appears to have reduced permission priming. We observed no participant engage in checklisting type behavior. Additionally, 53% of participants corrected permissions on 3 or less of the 12 tasks before being prompted and no participant corrected all permission errors. In comparison, over 90% of spelling errors were corrected. This suggests that participants were not overly primed to look for permission errors.

The reduction in priming allowed us to observe more subtle issues with our methodology. Participants' permission-checking frequency was impacted by the different tone and wording of the ideal policy in the task emails. Emails with

stronger wording resulted in permissions being checked more frequently by participants in all conditions and emails with weaker wording were checked less. This meant that while we had a valid study-wide result, we couldn't compare the permission identification behavior between tasks. The wording strength added a confounding factor.

4.3 Study 3

Reducing the number of tasks with permission errors to 50% and providing ideal policy information in the absence of errors appeared to cause less checklisting behavior. However, the wording of tasks caused participants to check permissions on some tasks more than others, suggesting that participants did not have consistent priming. In study 3 we wanted the tasks to provide a consistent level of permission priming independent of the presence of a permission error. We also wanted to maintain the "cost" of checking permissions at a 50% chance of there being no error.

One ideal policy We used a single ideal policy that applied to all albums because it 1) better mimicked normal usage where a single user has a consistent set of requirements, 2) was clearer for the participant to understand than getting a new policy with every email, and 3) eliminated wording variability since the participant would only see one policy. To counter differences in participant memory, participants were allowed to look back through any piece of paper the researcher gave them, including the page with the policy.

The ideal policy we ultimately selected had five rules, three of which involved access-control permissions. We were concerned that having a single policy that clearly mentions permissions would overly bias participants to look for permission errors, so we tried the protocol with seven test participants. We found that despite the priming, participants infrequently checked for permission errors but frequently checked for the other types of errors mentioned in the rules.

Consistent task structure Previously the emails were two paragraphs and important information appeared wherever it was most natural based on the email content. For this study the first paragraph was contextual only, indicating how it related to Pat but contained no vital data. The second paragraph clearly explained the primary subtask the participant was to engage in.

Unlike studies 1 and 2, the warm-up tasks in study 3 used the same structure and wording style as the other tasks. Based on observations in the prior studies, the tutorial was sufficient for understanding the system and the warm-up tasks were only necessary for the participant to acclimatise to the system and how tasks were presented.

Randomized tasks We decided, with the exception of the warm-up tasks, to randomize both the order that tasks were presented in and which tasks had permission errors. The goal here was to remove any ordering effects and by removing any effect task wording might have on a participant's inclination to check permissions.

Outcome The use of a single ideal policy allowed us to reduce the number of times we presented the participant with permission information. Only 11 of the 31 participants checked permissions on more than 50% of the tasks suggesting that for the majority of participants permissions remained a secondary task.

Our primary concern was that having explicit permission rules expressed in the beginning of the study would overly prime participants to check permissions regularly. Behavior of practice participants suggested that this would not be the case. However, the results of the full study showed that over priming did impact participants.

Our changes to study 2 appeared to eliminate the checklisting behavior observed in study 1 participants, but the design of study 3 brought it back. A graph of number of tasks where control participants checked permissions shows a non-normal distribution with peaks at 0 and 100. The other conditions showed similar distributions. This suggests that the permission priming effected some participants more than others.

4.4 Study 4

In study 3 we saw no difference between conditions because participants corrected all or none of the permissions with few participants in the middle. Using a single ideal policy worked well in study 3 as did the mix of 50% of tasks having permission errors. Because study 4 was within subjects, we decided to use a fixed permission order for easier comparison.

Time limitation We hypothesized that in study 3 that providing participants with clearer instructions made it easier for them to know what to do, but the only cost to participants for checking permissions was the time required to perform the check. In real life that time would be an opportunity trade off since the user could be doing something else with that time. In study 4 we decided to limit participants to a maximum of 2 minutes per task, forcing them to value their time and make trade offs. The primary researcher, as an expert user who knew where all the errors were, required a minimum 1.5 minutes complete each task, so we tried 2 and 3 minute limits on practice participants. We determined that a limit of 2 minutes created the largest differentiation amongst users.

Compensation variation For our practice participants we were concerned that Mechanical Turk users would not take the tasks seriously and do the minimum to advance through the study. So we offered a bonus based on performance. However, study feedback suggested that participants were deeply concerned that failure to get everything correct meant they would not be paid. They also felt a level of personal responsibility to get all the subtasks correct. So we adjusted compensation to a single rate and explicitly stated that all participants who got more than 25% of the task components correct would be compensated.

Outcome The combination of time limitations and reduction of emphasis on accuracy worked well. Permissions were changed unprompted by 66% of participants. In the underphoto condition only 4 of the 62 participants corrected all permissions. We also saw a reduction in feedback about the number of tasks participants had correctly completed.

5. PARTICIPANT RESPONSIBILITY

The framing of the study should make it clear to participants that they could and should make changes outside the bounds of the subtasks expressed in the emails.

5.1 Study 1

By having participants role play we were able to inform them that they had a responsibility for some albums by telling them it was part of their job or that their mother regularly relied on them for assistance. We wanted participants to be aware of what types of errors (rotations, spelling,

ect.) were within the bounds of the study without overly priming them towards permissions. The tutorial that covered several functionalities of Gallery, included permissions and followed by five prompted warm-up tasks, two of which involved permissions.

Outcome The open-ended nature of the tasks combined with the imparted responsibility made participants uncertain about how to react to tasks and prompts. For example, after a prompt from Pat's mother, in which the mother is panicking about seeing a photo of Pat sky diving, one participant simply responded "Sorry Mom." Another participant asked how old Pat was, then slapped the paper down on the table and declared loudly "I am NOT answering this!"

Some participants didn't feel it was their place to change permissions. A couple of participants noticed an error and verbally decided not to correct it because the album belonged to someone else and they expected that the album owner knew what they were doing, even if the permission was odd. Participants were not instructed to talk aloud during the study so we had no way of knowing how many participants noticed an error and chose not to correct it.

5.2 Study 2

Based on observations of participants we theorized that the general uncertainty was caused by a lack of clarity in the task descriptions.

Clearer instructions When observing participants complete the study 1 methodology we noticed numerous small confusion points that together made participants uncertain about what to do in the study. For example, a warm-up task tells participants that a photo of a poster has an incorrect title but doesn't say the correct title. Participants needed to read the title from the photo, but participants became confused. In study 2 we clarified that the titles can be read from the posters in the photos. Another example is from study 1's task 13 where Pat's sister apologizes for messing up Mom's photos and asks Pat to put the photos "back the way you had them." The participant is supposed to undo changes made by the sister so that the album looks like it did at the end of task 11. Some participants tried to change the album back to what it looked like when they first saw it at the beginning of task 11. We clarified the explanation. When running these tasks on practice participants we specifically asked them if these points were clear.

Outcome Participants appear to have taken responsibility for the albums and considered permissions to be in the bounds of the study. We did not observe any participant choosing to not change permissions due to concern about who owned an album. The clarification in wording resulted in less participant uncertainty over how to handle situations.

5.3 Study 3

Directly telling participants that they were responsible for the albums, combined with clear wording, appeared to have caused study 2 participants to sufficiently take responsibility for the albums. In study 3 we tried to keep these themes.

Prompts We initially decided to make only warm-up tasks 1 and 2 prompted tasks to make sure that participants were capable of performing all the actions necessary for the study. As part of the prompting emails, the participant is directly told that it is their responsibility to find and fix these types of errors.

After running the protocol on several practice participants we discovered that around the 5th task, participants would start to become lazy and stop taking responsibility for correcting all the errors. We solved the problem by making task 5 a prompted task. Similar to warm-up tasks 1 and 2, the participant was told in the email that fixing errors is their responsibility.

Outcome Participants took responsibility for the albums and considered permissions to be in the bounds of the study. When asked after the study if they felt they could change permissions, all participants asserted that they felt they were allowed to do so.

Making task 5 a prompted task was very effective in reinforcing participant responsibility. Throughout the study participants would get lazy or careless around this task, receive a strongly worded email from their boss, and immediately start paying more attention. In the debriefing interview we asked participants about their reaction to this email. Participants said that they realized that the boss would be checking their work so they needed to do a good job.

5.4 Study 4

The methodology for study 3 worked well so we made only minor alterations for study 4. We reduced the strength of wording in the prompted warm-up task so that it simply pointed out the error. Because participants only had eight tasks per condition and were limited to 2 minutes we decided to not prompt mid way through.

Outcome Because study 4 is an online study we have limited feedback on participant's feeling of responsibility. Participants who gave study feedback expressed a strong desire to get all the tasks correct. The number of permissions and non-permission subtasks corrected also indicated that participants took responsibility for the albums.

6. IDEAL POLICY COMPREHENSION

Participants should know the ideal policy associated with the content they are working with.

6.1 Study 1

We considered conducting the experiment using participants' own albums and policies but ultimately decided against it. Prior work has shown that participants' ideal policies change over time [8], in reaction to new technology [2], and based on context [3]. Mazurek et al. asked participants to provide ideal policies twice: all at once in a single sitting and by answering the same questions in small batches over the course of a week [8]. They found that the same participants responded with different ideal policies depending upon when asked. We were concerned that participating in our experiment would impact participants answers concerning their ideal policy, negatively impacting our ability to get an accurate ground truth. Instead we decided to create a fictional ideal policy which would be consistent across all participants.

To make the ideal policy appear less like explicit instructions, we expressed it using passive voice in the emails. However, not all permission information, particularly who shouldn't see the albums, could be easily expressed in passive voice so some information was presented in instruction pages that described the people the participant was about to interact with. To make this information simple to internalize, we created characters. For example: Pat's mother was

described as panicking easily, while Pat was described as enjoying dangerous activities. The instruction sheet commented that Pat generally avoided telling his/her mother about the dangerous activities.

We decided to have two permission warm-up tasks to be certain that participants could accurately both read permissions as well as change the permissions. If they were unable to do so the researcher provided guidance. The first permission warm-up task simply asked the participant if a particular album was visible to everybody on the internet or not. The second permission warm-up task asked the participant to change the permissions on a specific album.

Outcome Participants seemed to understand the ideal policy without difficulty and participants who made changes tended to make the correct ones. However, we have no way to determine why participants who did not change permissions chose not to do so.

The warm-up task in which participants were asked to read a permission resulted in participants guessing instead of reading the permission. In the warm-up task, Pat's boss asks if people at other companies can see a particular album. Participants tended to correctly guess that the album was publicly visible and answered the question without even looking at the screen. We had prepared prompting emails in the event of an inaccurate guess, but had not anticipated that the majority of participants would guess accurately. For the non-control conditions there was no way to be certain they had guessed since we could not verify if they had looked at the display.

6.2 Study 2

Participants seemed to understand the ideal policy in study 1 so we made minimal changes to the way it was presented.

Changed permission-read warm-up task In study 1 participants were guessing that anyone on the internet could view the album in the permission reading warm-up task. In study 2 we changed the task so that the correct answer was that anyone on the internet could *not* view the album thereby making it the opposite of the common guess.

Think-aloud protocol For reasons discussed in following sections we made study 2 a think-aloud study. A side effect of this decision was that participants had to read all instruction materials and emails out loud, ensuring that all materials, particularly the ideal policy, were read. We were also able to determine when instructions were confusing.

Outcome In warm-up task 2 (read permission) we observed more participants consulting the display to determine what the permissions were instead of opening the permission modification interface. Participants were still inclined to guess that the album was public but the guesses were now wrong and the researcher was able to prompt them, so every participant understood how to read permissions.

Using a think-aloud protocol forced participants to read all text aloud, thereby ensuring that all materials, including information about the ideal policy, was not skimmed over. Based on the think-aloud statements, participants appear to have understood the ideal policy. However, the protocol had no explicit outcome variable with which to test ideal policy comprehension.

6.3 Study 3

In this study we decided to present one ideal policy to the participant at the beginning instead of presenting the policy in pieces. This was done to provide consistent permission priming (Section 4.3). It was also done to promote participant understanding of the ideal policy and make it easier to test that understanding.

Testing ideal policy comprehension Participants in studies 1 and 2 appear to have understood the ideal policy, but we did not measure their comprehension. Study 3 had a single ideal policy so we were able to perform a pre and post test of participants' ideal policy comprehension. The pre-test was administered after the warm-up tasks, participants were asked by a co-worker if a provided photograph was appropriate for the website and if they should do anything when posting it. The post test is part of the final survey, participants were asked what the permissions for several albums should have been.

Outcome Ideal policy comprehension was provably high in this study. Participants had no problem remembering the ideal policy and were able to apply it to different situations and albums with high accuracy.

In the pre-test 78% of participants correctly mentioned permissions for both comprehension questions and only one participant never mentioned permissions. Participants behaved similarly on non-permission comprehension questions. This means that participants were able to 1) recognize that permissions might need to be set for these photos, and 2) correctly apply the ideal policy. Across conditions participants answered an average of 91% and a minimum of 67% of post-study permission comprehension questions correctly. This shows that the methodology design enabled participants to correctly understand, remember, and apply the ideal permission policy.

6.4 Study 4

As mentioned in Section 4.4 we were concerned that the explicit listing of ideal policy rules in a bulleted list was over priming participants to look for permission errors. With practice participants in study 4 we experimented with several information page designs. We conveyed the ideal policy in paragraph form with varying levels of wording intensity and compared that with providing the policy in bullet point form. We found that presenting the policy in bullet point form lead to the lowest level of variance and the largest difference in permission correction between conditions.

Outcome In study 3 participants could answer "I do not know" to any comprehension question, but it was rare that they did so. In study 4, 50% of participants answered "I do not know" to at least one comprehension question, but only 4% answered all comprehension questions that way. Of the answered questions 90% were answered correctly. Interestingly the design of the information page which conveyed the ideal policy had minimal effect on ideal policy awareness. Participants who saw the ideal policy in paragraph form correctly answered approximately 87% of comprehension questions, with minimal variance between designs.

7. EFFECTIVE OUTCOME MEASUREMENT

We needed to differentiate between environments with no errors, environments where participants are not noticing errors, and environments where errors have been noticed.

7.1 Study 1

We chose a lab study design because it offered us the most amount of control over potential variables. We could control

the task design, types of errors, and when errors would appear. By using a role-playing scenario we could also control participants mindsets when approaching problems.

In order to test our primary hypothesis H we needed to detect when a permission error was "noticed." We anticipated that a participant who noticed an error was very likely to correct it. So for this study we defined "noticed" as "corrected." The number of people correcting a permission error is a strict subset of the number of people noticing errors and we anticipated a large difference in the number of permissions corrected between the conditions. So we were willing to accept that we might not detect a participant that chose not to correct a noticed error.

When designing memory questions we were concerned about participant fatigue leading to questions being guessed at or answered with the fastest answer. To counter this we limited our questions to six albums and only asked about two of the actions. We also required that all memory questions be answered with True, False, or Not Sure. This was to make providing answers the same amount of work as guessing.

Outcome Unfortunately, we did not see a statistically significant difference in the number of permissions corrected between conditions. We also observed participants noticing errors and choosing to not correct them which was not captured by our definition of "noticed." We considered changing our definition but determining if a participant had checked the permissions was impossible for participants in the non-control conditions who might or might not have looked at a proximity display. So, while it may be the case that H is supported if we define "noticed" as "checked permissions," our lack of measurement fidelity prevented us from testing this.

7.2 Study 2

In designing the outcome variables for study 2 we focused on being able to notice when participants checked permissions as well as when they corrected permissions.

Think-aloud and eye tracker Our inability to accurately measure when permissions were noticed but not changed was a major issue with the study 1 methodology. To adjust, we made study 2 a think-aloud study. Study 1 was deliberately not a think-aloud study so we could determine if participants took an equal amount of time to complete tasks (H5). Think-aloud protocols are known for giving inaccurate timing information. In study 2 we felt that accurate timing information was less important than accurately measuring participants' interactions with the displays.

To assist in measuring if and when a participant focuses on a display we decided to use an eye tracker. This data was intended to augment, but not replace, the think-aloud data.

Outcome The think-aloud data enabled us to determine when participants *checked permissions* using the following definition. Control participants were judged to have *checked permissions* if they opened the permission management interface and the permission was visible on the screen. Participants in the other conditions were judged to have *checked permissions* if they (1) opened the permission management interface; or (2) read permission aloud; or (3) clearly indicated through mouse behavior that they were reading the permission display; or (4) pointed at the permission display with their hand while clearly reading the screen. This definition allowed us to measure if a participant paid significant attention to a display.

Data from the eye tracker was less helpful than anticipated. To operate, the eye tracker needed participants' faces to remain in a small area. This is possible for short studies, but our study was 1.5 hours. Participants would shift in their chairs or lean on the desk moving them out of range. We considered prompting participants when they moved outside the required area but decided this would distract participants and alter their behavior. We tried having participants experiment with the eye tracker before the study so they knew where the optimal area was. This helped, but participants still became distracted by the study and started moving outside the optimal area. While incomplete, the eye tracker data did give us a sense of when participants looked at displays.

7.3 Study 3

In study 3 we wanted to get more detailed qualitative data about how and why participants checked permissions. Our definition of "permission checking" from study 2 appeared to be working well so we did not modify it.

Permission modification interface In studies 1 and 2 we observed no difference in memory between the conditions (H4). We hypothesized that this was due to the full sized permission modification interface. Participants who visited the interface frequently changed more than one permission indicating that, even in the control condition, they were looking at other permissions. To address this issue we added the permission modification interface as an independent variable. The permission modification interface was either a separate page with all permission settings shown or a dialog with only one album's permission settings shown. We added the following hypothesis:

H7 Participants who see a comprehensive policy modification interface remember permissions better than participants who see a policy modification interface that displays a single album.

Post-study memory In studies 1 and 2 we asked participants to answer 128 memory questions about 13 albums, 4 groups and 2 actions (view and add) and saw no statistically significant difference between conditions. In this study we wanted more qualitative data to better understand what people remembered. We decided to verbally administer the memory questions and elicit free responses. We felt free form answers would get us a better sense of what the participant remembered. Once all the memory questions had been asked the researcher prompted the participant about anything they had not yet mentioned. For example some participants only answer the questions in terms of the view action so the researcher would ask if they recalled the add or edit action for any of the albums.

When we asked practice participants, who had not checked permissions during the study, memory questions, we found that they started feeling embarrassed that they didn't know the answer, and after a couple questions they started guessing. To discourage guessing we interleaved the memory and comprehension questions. This meant that every participant could, at worst, provide an answer for every other question without having to guess. We found that this discouraged guessing and participants seemed more comfortable admitting that they could not recall the permissions for albums they did not check the permissions on.

Post-study debriefing Once all the questions had been completed we conducted a debriefing interview with the par-

ticipant. In the prior studies participants had occasionally behaved unexpectedly. Initially we thought this was caused by methodology issues but some behaviors persisted through different methodologies. In this study we wanted to get the participant's perspective on why they engaged in these behaviors. However, many of the behaviors were short (1-2 seconds long) and we were concerned that participants would not remember why they had made a comment an hour ago. So we used a contextual interview approach [5] where the participant opened the album they were working with and the researcher explained the context in which the behavior occurred and asked the participant questions concerning what they were thinking or why they had done something.

Outcome This study design allowed us to accurately measure and test all the outcome variables we were initially looking for. The only issue was an unknown confounding variable that caused some participants to check permissions frequently and other participants to check them rarely.

The use of a single ideal policy allowed us to observe natural participant behavior that was inhibited by the design of prior studies. In prior methodologies the participant was unable to choose when to check permissions because they did not know the ideal policy until they started a task. With one ideal policy we observed several participants deciding at a single point in the study to check permissions for every album at once. This behavior was facilitated by the full permission modification interface. We found that participants who saw the full interface performed better across several measurements and were more likely to correct permissions regardless of if they saw the proximity display or not.

The combined use of a single ideal policy, randomized task order, and randomized permission error order allowed us to notice issues with our definition of permission checking. In the control condition we reliably determine when the permissions were shown. In the non-control conditions, we only determine when permissions were checked based on participant behavior. In study 3 non-control participants were statistically more likely to check permissions when there was an error than when there was no error. There was no statistical difference for the control participants. This suggests that participants were able to glance at the display and determine if there was an error fast enough to not vocalize [14]. This is good news for our display but it implies that we can only detect when a participant *focuses on checking permissions* rather than being able to detect every time they check permissions. The eye tracker allowed us to determine when they fixate on a display but similarly did not tell us when they actually checked the permissions.

The use of contextual immersion during the debriefing session was very effective at getting participants to remember their reasoning behind specific actions. In cases where the participant couldn't remember they were still often able to make an educated guess as to why they would have done an action given their behavior up to that point. While a guess is not as good as remembering, participant's guesses as to reasons behind their actions were more accurate than researchers educated guesses.

7.4 Study 4

The prior studies had a small number of participants, and they exhibited a large between-participant variance, making it difficult to detect differences between conditions. In this study we wanted to increase the number of participants and account for the variance.

Within subjects In study 3 we observed that some participants internalized the need to check permissions while others did not. In the debriefing interview the participants who internalized considered it "obvious" and those that did not check permissions appear to have read the ideal policy and then forgot about permissions. To control for the predisposition to pay attention to permissions we decided to make study 4 a within-subjects study where every participant performs the training and tasks on both the control condition and one of the non-control conditions.

Measuring "noticing" Our hypothesis H is that participants in some conditions can "notice" permission errors more frequently than participants in other conditions. In studies 2 and 3 we equated noticing permission errors with checking permissions. However, measuring permission checking requires observation of the participant not possible in an online study. Additionally, we showed in study 3 that our measurement of permission checking was, at best, a lower bound for the number of times permissions were actually checked by participants. In study 4 we returned to our definition of "notice" from study 1 where we equate correcting permissions with checking them. This definition provides only a lower bound but with the larger number of participants and improvements to the methodology we did not anticipate a problem.

Permission modification interface In study 3 we observed that participants who saw the permission modification interface in a dialog had a larger difference in performance between conditions than participants who used the full page permission modification interface. Since our main hypothesis H is concerned with the impact of proximity displays, not permission modification interfaces, we decided to use the dialog for study 4.

Outcome Using the stricter definition of "notice" as "corrected" was effective in that we were able to show statistically significant differences between some of the conditions and control (not all conditions were expected to have a difference). We attribute this to both a larger number of participants and clearer, more tested, study materials.

Similar to study 1 we had a limited ability to measure why participants did or did not make changes to permissions. However, we collected extensive logs which we were able to compare to behaviors observed in prior studies allowing us to imply what users were doing.

8. DISCUSSION

We discussed the methodologies of four studies designed to test our hypothesis. When designing our initial study we tried to account for anticipated methodology issues. Our initial design succeeded in some aspects and was lacking in others. Subsequent studies were adjusted to account for observed issues.

Secondary permission task Users treat security as a secondary task because the benefits of security are hard to envision but the costs of engaging in it are immediate [17]. In our studies we did not want to incentivize the participant to check permissions so we tried to balance the amount of priming with the cost of checking. We successfully managed priming on study 2 and 4, but in studies 1 and 3 we over-primed, first by mentioning permissions too frequently and then by using strong wording to express the ideal policy without forcing participants to consider trade-offs. We in-

creased the immediate cost of checking permissions in studies 2 and 3 by adding tasks where the permissions were already correct and checking them cost time and effort. We further increased the cost in study 4 by adding a time limitation which forced the participant to make trade-offs. We found that at least 50% of the tasks needed to have no permission error in order to give checking a high cost compared to the benefit.

Participant responsibility Role playing was very effective in making participants feel responsible for albums that belonged to Pat. Our main issue was when we asked participants to be responsible for albums that belonged to people such as Pat's mother. We countered this issue in the second study by making it clearer that others trusted Pat to make changes.

Ideal policy comprehension We tried two methods of expressing the ideal policy to participants. The first was to have a different policy for each album. The policy was expressed using passive voice in the emails (studies 1 and 2). The second way was to have a policy that applied to all the albums. The policy was expressed using direct wording at the beginning of the study (study 3 and 4). Both methods sufficiently communicated the policy to the participant. The per-album policy gave participants less priming towards fixing permissions but was difficult to make consistent. The study-wide policy over-primed some participants to look for permission errors, but provided consistent priming to all participants on all tasks.

Effective outcome measurement Our primary issue with measuring the study outcome was defining and testing participants' ability to "notice" permission errors. In the first study we defined "notice" as changing permissions, but this definition was insufficiently precise to measure the difference between conditions. In later studies we changed our definition of "notice" to checking the permissions for errors. This definition allowed us to observe if participants were looking for errors independently of whether they found the error or decided to fix it.

In conclusion we presented the methodologies of four studies and discussed the decisions and outcomes of each study. We were able to describe our methodological successes and difficulties in terms of our four goals: 1) secondary permission task, 2) participant responsibility, 3) ideal policy comprehension, and 4) effective outcome measurement. Through this process, we have shed light on the challenges intrinsic to many studies that examine security as a secondary task.

9. REFERENCES

[1] Gallery 3. Website. http://gallery.menalto.com/.

[2] Lujo Bauer, Lorrie Cranor, Robert W. Reeder, Michael K. Reiter, and Kami Vaniea. Comparing access-control technologies: A study of keys and smartphones. Technical Report CMU-CYLAB-07-005, Carnegie Mellon University, 2007.

[3] Sunny Consolvo, Ian E. Smith, Tara Matthews, Anthony LaMarca, Jason Tabert, and Pauline Powledge. Location disclosure to social relations: Why, when, & what people want to share. In *Proc. CHI*, 2005.

[4] S. Egelman, A. Oates, and S. Krishnamurthi. Oops, i did it again: Mitigating repeated access control errors on facebook. In *Proc. CHI*, 2011.

[5] D Godden and A.D Baddeley. Context-dependent memory in two natural experiments: on land and under water. *British Journal of Psychology*, 66:325–331, 1975.

[6] Joerg M. Haake, Anja Haake, Till Schümmer, Mohamed Bourimi, and Britta Landgraf. End-user controlled group formation and access rights management in a shared workspace system. In *Proc. CSCW*, 2004.

[7] M. Madejski, M. Johnson, and S. M. Bellovin. The failure of online social network privacy settings. Technical Report CUCS-010-11, Department of Computer Science, Columbia University, 2011.

[8] M. L. Mazurek, P. F. Klemperer, R. Shay, H. Takabi, L. Bauer, and L. F. Cranor. Exploring reactive access control. In *Proc CHI*, 2011.

[9] R. W. Reeder, L. Bauer, L. F. Cranor, M. K. Reiter, K. Bacon, Ke. How, and H. Strong. Expandable grids for visualizing and authoring computer security policies. In *Proc. CHI*, 2008.

[10] R. W. Reeder, L. Bauer, L. F. Cranor, M. K. Reiter, and K. Vaniea. More than skin deep: Measuring effects of the underlying model on access-control system usability. In *Proc. CHI*, 2011.

[11] S. Sheng, M. Holbrook, P. Kumaraguru, L. Cranor, and J. Downs. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proc CHI*, 2010.

[12] Andreas Sotirakopoulos, Kirstie Hawkey, and Konstantin Beznosov. On the challenges in usable security lab studies: Lessons learned from replicating a study on ssl warnings. In *Proc. SOUPS*, 2011.

[13] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying wolf: An empirical study of SSL warning effectiveness. In *Proc. USENIX Security Symposium*, 2009.

[14] Maarten W. van Someren, Yvonne F. Barnard, and Jacobijn A.C. Sandberg. *The Think Aloud Method: A practical guide to modelling cognitive processes.* Academic Press, London, 1994. hu, R.

[15] K. Vaniea, L. Bauer, L. F. Cranor, and M. K. Reiter. Out of sight, out of mind: Effects of displaying access-control information near the item it controls. In *Proc PST*, 2012.

[16] Y. Wang. *A Framework for Privacy-Enhanced Personalization.* Ph.D. dissertation, University of California, Irvine, 2010.

[17] Ryan West. The psychology of security. *Communications of the ACM*, 51:34–40, 2008.

[18] T. Whalen, D. Smetters, and E. F. Churchill. User experiences with sharing and access control. In *Proc. CHI*, 2006.

[19] A. Whitten and J. D. Tygar. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In *Proc. USENIX Security Symposium*, 1999.