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, which impacted 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 the currently implemented access-control policy, and ability to correctly set a desired policy.

Our primary hypothesis was that if the current permission settings are shown in close spatial proximity to the resources they affect, instead of on a secondary page, users are more likely to notice and fix permission errors. To test our hypothesis, we needed our participants to interact with the display as a secondary task, while engaged in a non-security-related primary task.

Other researchers have used a variety of approaches to study security as a secondary task. One approach, used by Haake et al., is to conduct a long-term study in which participants are made aware that security is a part of the study but the study is run for long enough that they stop focusing on security [6]. Another approach, used by Sunshine et al., is to hide from participants that the study is about security, but to design the study so that participants engage in a security-relevant behavior while trying to complete their primary task [13]. A final approach, used by Wang, is to keep participants unaware that the study is about security and give participants the option of engaging in security-relevant behavior [16].

We used the last approach to test our hypothesis. We conducted a laboratory study in which participants performed various photo-management tasks. Depending on condition, permission information was displayed under the photos, elsewhere on the page, or on a secondary page (the control condition). We tried to design the study to control for various confounding factors and avoid a range of pitfalls. However, we stopped the study early when we ran into multiple methodological problems, including some that made it difficult to measure study outcomes and others that caused participants not to treat security as a secondary task.

When designing the initial study, we wanted to meet the following goals: make security a secondary task (Section 4), give the participants 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 difficulties in achieving each 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 intended to accomplish the four goals described above. We describe the problems encountered during each study, and changes to the methodology designed to address those problems. We shed light on the challenges intrinsic to many studies that examine security as a secondary task, and identify 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 following hypothesis:

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

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

Permissions as a secondary 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. This is to emulate typical real-world situations, where 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 have studied computer security technologies have successfully simulated the secondary-task context in the lab. In a study on the usability of the PGP email encryption software, Whitten and Tygar had participants focus on sending and receiving emails related to a political campaign [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 content they manipulate during the experiment to be comfortable making changes they deem necessary. Because changing permissions is intended to be a secondary task, the framing of the study should make it clear to participants that they may make changes outside the bounds of their primary task.

When replicating the SSL study described above, Sotirakopoulos et al. observed participants who claimed they had behaved differently in the lab than they would have outside the lab because they considered the lab to be a "safe" environment [12]. Witten and Tygar overcame this issue in their work [19], but doing so requires careful study design.

Ideal-policy comprehension Participants should be able to understand clearly the *ideal policy*—the correct set of permissions for the study scenario. Participants need to be able to figure out when a permission setting is "correct" or "incorrect." If a participant is observed to ignore an error, we need to have confidence that the error was ignored because the participant did not notice the state of the settings rather than because she did not realize it was a violation of the ideal policy.

Effective outcome measurement We need to be able to accurately measure whether participants are noticing and fixing errors. In real-world environments, it may be difficult to determine whether a specific setting constitutes an error; such judgments can be subjective and dependent on context [1, 2, 8]. To accurately test "noticing" errors the we need to be able to distinguish between environments with no errors, environments with errors that participants are not noticing, and environments where errors have been noticed.

2.1 Study Setting

We decided to use a photo-management website as the domain because it is a commonly used environment in which users might set access-control policy. We chose to use an open-source web-based photo-management system, Gallery 3 [4], 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 that displays permission information in a small panel that appears under the thumbnail images of 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, based on prior work [9, 10], was designed to be easy to use and comprehend, but was not the focus of this research. Access-control permissions in Gallery are specified as four-tuples of (*user group*, *album*, *action*, *decision*), where available actions include viewing, editing, and adding to albums, and the decision is to allow or deny access. Permissions cannot be specified for individual users or individual photos.

3. GENERAL STUDY DESIGN

As part of testing our main hypothesis, **H**, our initial study design was intended to test the following specific hypotheses:

H1: Users who see permission information under photo/album thumbnails or on the sidebar notice errors more often than users who see permission information only if they click through to a second page.

H2: When a permission is changed to an error state by a third party, users who see permission information under photo/album thumbnails or on the sidebar notice errors more often than users who see permission information only if they click through 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 photo/album thumbnails or on the sidebar can recall those permissions better than participants who see permission information only if they click through to a second page.

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

In this paper we discuss the methodologies of four similar studies. It is impossible, given space limitations, to fully describe the methodologies of all four studies. In this section we present the core methodology shared by 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. For each study, we discuss the outcome of the choices and how they informed the methodological choices for the next study.

The first three studies were between-subjects lab studies and the last was a within-subjects 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, which included a link to the interface for changing permissions; and the under-photo condition, which additionally included a proximity display under photo/album thumbnails (Figure 1).

Participants were asked to role play (cf. [3, 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 any Tygar successfully used role playing to encourage participants to view security as a secondary task [19]. Tasks were communicated to the participant in the form of emails. 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 with which the participant was interacting.

Participants started with a training task 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 covered in the training to ensure that they understood how to manipulate the interface. In studies 1, 2, and 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.

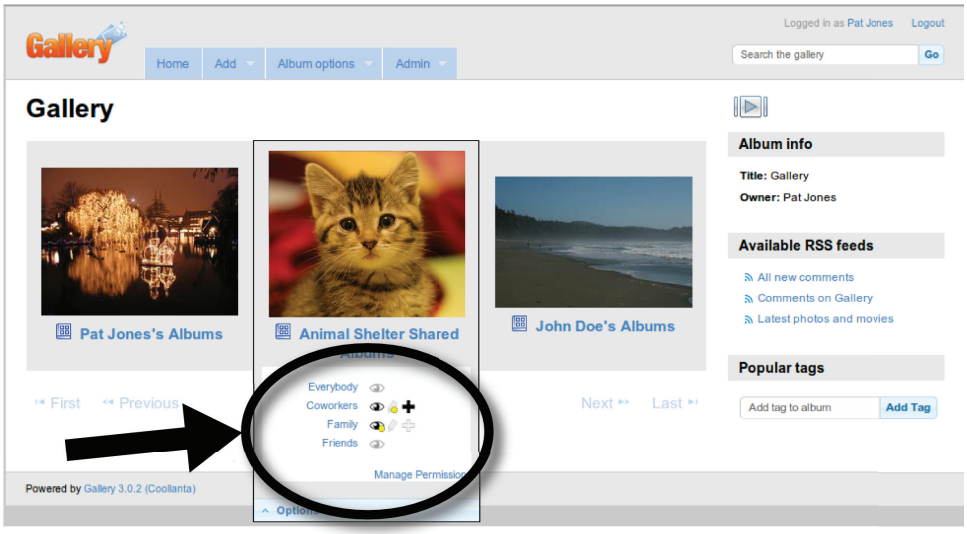

Figure 1: Example of proximity display used in studies 1 and 2. The interface for studies 3 and 4 had a slightly different permission-display design.

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 or 2 warm-up tasks of approximately equal difficulty to the tasks that followed.

The bulk of the studies was composed of a set of tasks presented to the participant in sequence. Each task was composed of a set of *subtasks*—in each subtask the participant was expected or given the opportunity to correct a specific issue with an album. A primary subtask was directly conveyed in the email, and several additional subtasks were implicitly specified by errors that the participant could notice, such as rotated photos, misspellings, and incorrect permissions. All tasks contained at least one explicit and one implicit title, rotate, delete, or organize subtask intended to distract the participant.

Some tasks were *prompted*; 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 are those on which the participant would not have been prompted under any circumstances, as well as those tasks that were completed by a participant without being prompted. Participants were unaware of which tasks were prompted until they received a prompt.

Some albums were *changed* halfway through the study. A participant first interacted with an album and was made aware of the current state, including permission settings. When the participant was distracted by a task the researcher made changes (unrelated to that task) to the album. The participant was then instructed to interact with the now changed album.

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 questions and questions about prior experience.

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 eight tasks, which took an average of two and a half minutes each. Tasks appeared in the same order for all participants. Finally, participants filled out the survey. Five tasks were prompted, and the researcher changed two albums during the study. This study was run on 26 participants and three conditions. It was stopped early because of issues with the methodology.

Study 2 was a 1.5-hour, 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 eight 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 order for all participants. Finally, participants were asked to fill out the survey. Five tasks were prompted, and the researcher changed three albums during the study. 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 [15].

Study 3 was a 1.5-hour, 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: location of proximity display and type of 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 an 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. For each condition, participants

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 yours 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 implying that everybody in the Friends group needs to be able to view the photographs.

first completed a set of training tasks, which took an average of four minutes. Then they completed a warm-up task, which took an average of three minutes. They were then given seven tasks, with a maximum of two minutes to complete each. Tasks appeared in the same order for all participants. This process was then repeated for the second condition. When finished with both conditions, participants were asked to fill out a survey that asked questions about both conditions. 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. PERMISSIONS AS A SECONDARY TASK

The first goal we wanted to accomplish with our study designs was to put participants in an environment where there is minimal encouragement to engage in security tasks, and the benefits, if any, are not immediate. We explain how we attempted to accomplish this goal in each study, taking into account any problems encountered in previous studies.

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 this allowed us to provide context for the task, such as the ideal policy, without drawing too much attention to it. To prevent users from perceiving permission content in emails as explicit direction, permission subtasks were described only through implied requests while primary subtasks were described through explicit requests. For example, the email in Figure 2 explicitly asks that the titles be changed, but also implies that the Friends group should have permission to view the photos. The ideal policy components that could not be implied were embedded in information pages about Pat's friends, family, and co-workers. These information pages were handed to participants as needed during the course of the study.

We were concerned about giving participants too much permission *priming*—causing them to be more aware of permissions than they would be in a more realistic setting. Every time a participant reads or interacts with permission information, they are being primed to think about permissions. We attempted to avoid over-priming participants 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 included to give participants 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 point this out. Participants were first prompted about permissions, if needed, after the third task. We prompted here to be sure participants knew what the album's permissions were before they were changed by the researcher, since the album used for the third task was one of those that was changed by the researcher during the study.

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 checklist. One participant did this out loud, listing all the types of errors she had seen in the training material, and making sure she had checked all of them before moving on.

Additionally, many participants never obviously consulted the proximity display to determine if there was an error, even though they opened 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 mentioned permission information in emails contained permission errors. Thus there was no reason to check permissions using the permission-modification interface unless permissions were mentioned in the email. To address this concern we added tasks that mentioned permissions but had no permission errors. 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 participants who have to proactively open a second interface in order to view permissions.

New read-permission tasks We added three new tasks for which emails expressed the ideal policy, but the current settings matched the ideal policy, i.e., 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 permission 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 appeared to reduce permission priming. We observed no participant engage in checklisting behavior. Additionally, 53% of participants corrected permissions on 3 or fewer 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 revealed subtler problems 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 than did emails with weaker wording. This meant that while we had found a valid study-wide result (that the proximity display helped users notice permission errors), we could not compare the permission-identification behavior between tasks. The wording differences between conditions 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, independently of the presence of a permission error. We also wanted to maintain the "cost" of checking permissions by retaining a 50% probability that an album would have an error.

One ideal policy We used a single ideal policy that applied to all albums (rather than different ideal policies for different albums) because this (1) better mimicked normal usage, where a single user has a consistent set of requirements that spans albums, (2) was easier 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 recall for tasks that took place towards the end of the study, 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 permissions. We were concerned that having a single policy that clearly mentions permissions would overly prime 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 long, and information important for the task appeared in the email wherever it was most natural based on the email content. For this study, the first paragraph of emails always provided only contextual information, indicating how it related to Pat. 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 other tasks. Based on observations in the prior studies, the tutorial was sufficient for participants to understand Gallery and the warm-up tasks were only necessary for the participants to acclimatize to the system and to how tasks were presented.

Randomized tasks We decided, with the exception of the warm-up tasks, to randomize both the order in which tasks were presented and which tasks had permission errors. The goal here was to remove any ordering effects, as well as any effects of task wording on participants' 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 with the design of study 3 was that showing explicit permission rules to participants at the beginning of the study would overly prime participants to check permissions regularly. Behavior of pilot participants suggested that this would not be the case. However, the results of the full study suggested that over priming did occur, at least for some participants. Our changes for study 2 appeared to eliminate the checklisting behavior observed in study 1 participants, but the design of study 3 brought it back. The incidence of control-condition participants checking permissions followed a non-normal distribution with peaks at 0 and 100% of permissions checked. Other conditions exhibited 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 in permission-error correction between conditions, because many participants corrected all or none of the permissions, with few participants in the middle.

Because we saw very different behavior between participants in study 3, we decided to make study 4 a within-subjects study, where each participant would experience both a control condition and an experimental condition. We continued using a single ideal policy, as in study 3, as well as the same ratio of tasks that had errors to those that did not. Because study 4 was within subjects, we decided to use a fixed task order for easier comparison. We also introduced two more factors: a time limit, and variable compensation.

Time limit We hypothesized that, in study 3, 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. Unlike in real life, participants were not making a choice between something more interesting (e.g., browsing YouTube videos) and checking permissions; in the study, even if they chose not to check permissions, they would only get to move on to another study task. In study 4 we decided to limit participants to a maximum of 2 minutes per task, increasing the relative cost of unnecessarily checking permissions (since it would use time needed to complete other subtasks). The primary researcher, who was familiar with all the errors in the albums, needed 1.5 minutes to complete each task. We experimented in pilots with time limits between 2 and 3 minutes. We determined that a limit of 2 minutes was most effective at preventing participants from always checking permissions, without completely discouraging them from checking.

Compensation variation In pilots of the online study we were concerned that Mechanical Turk users would not take the tasks seriously and would do the minimum needed to advance through the study. Hence, we offered a bonus based on performance. However, study feedback suggested that participants were deeply concerned that they would not get paid if they did not correct all errors, and were in general strongly motivated to correctly carry out each subtask. To induce more realistic behavior, 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 time limits and reduction of emphasis on accuracy—combined with a single ideal policy and a within-subjects design—worked well. Permissions were changed unprompted by 66% of participants, and we observed few instances of checklisting behavior. Variances in permission-checking behavior due to wording differences between tasks were minimized. In the under-photo condition, only 4 of the 62 participants corrected all permissions.

5. PARTICIPANT RESPONSIBILITY

A second goal of our study designs was to make it clear to participants that they could and should make changes outside the bounds of the explicit subtasks expressed in the emails.

5.1 Study 1

Role playing allowed us to convey to participants that they were fully responsible for some albums: instructions informed them that it was part of their job or that their mother regularly relied on them for assistance. We wanted participants to be aware of the types of errors (e.g., rotations, spelling) that were within the bounds of the study without overly priming them towards permissions. The tutorial that described the features of Gallery mentioned permissions as well as other features, and was 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 skydiving, 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 did not feel it was their place to change permissions. A couple of participants noticed an error and verbally conveyed their decision 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 our observation of participants in study 1, we theorized that the general uncertainty was caused by a lack of clarity in the task descriptions.

Clearer instructions When observing participants in study 1, we noticed numerous points that caused minor confusion in participants, and we hypothesized that these together made participants uncertain about what action to take at various points in the study (e.g., when they noticed specific permission problems). For example, a warm-up task told participants that a photo of a poster has an incorrect title but did not convey the correct title. Participants needed to read the title from the photo to recognize the error and the obvious way to correct it, but often became confused. In study 2, we additionally explained that the titles can be read from the posters in the photos. Another example is from task 13 in study 1, 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 intended to undo changes made by the sister (in reality, by the researcher) 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 changing permissions to be part of their responsibility. 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 various simulated problems that arose during the study.

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 build on this by fine-tuning the incidence of prompting to ensure that participants did not forget which behaviors were in scope.

Prompts We initially decided to make only warm-up tasks 1 and 2 prompted tasks, as we wanted to make sure that participants were capable of performing all the actions necessary for the study. As part of the prompting emails, the participant was 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 within the bounds of the study. When asked after the study whether 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. Those participants who became lazy or careless around this task received a strongly worded email from their boss, and immediately started paying more attention. In the debriefing 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 with respect to participant responsibility, and so we made only minor alterations for study 4. We reduced the strength of the 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 halfway through the tasks.

Outcome Because study 4 was an online study, we have limited feedback on participants' feelings of responsibility. Participants who gave study feedback expressed a strong desire to get all the tasks correct. The number of errors corrected throughout the study also indicated that participants took responsibility for the albums.

6. IDEAL POLICY COMPREHENSION

The third goal we wanted to achieve in our studies was that 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 [1], and based on context [2]. 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 they were 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 un-

derstanding of ground truth—what their ideal policy really was. Instead, we decided to create a fictional, static ideal policy that would be consistent across all participants.

To make the ideal policy appear less like explicit instructions, we expressed it through implied requests in the emails given to participants. However, not all permission information, particularly information about who should not see certain albums, could be easily expressed implicitly; hence, we also conveyed information important for understanding the ideal policy in the instruction pages that described the people the participant, who was role-playing the part of Pat Jones, 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 verify that participants could accurately both read and change permissions. If they were unable to do so, the researcher provided guidance. The first permission warm-up task simply asked the participant whether a particular album was visible to everybody on the internet. 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 to permissions tended to make the correct ones. However, we were not able 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 many 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 often 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 or read the permission, since we could not determine whether 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 the correct answer the opposite of what was most frequently guessed.

Think-aloud protocol For reasons discussed in Section 7, 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 which 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 that after that task 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, were fully read, instead of just skimmed. Based on the think-aloud comments made by participants, they appear to have understood the ideal policy. However, the protocol had no explicit outcome that allowed testing 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 better 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 test ideal-policy comprehension both early and late in the study. The first test was administered after the warm-up tasks: participants were asked by a co-worker whether a particular photograph was appropriate for the website and whether they should do anything when posting it. The second test is part of the final survey: participants were asked what permissions should have been set on several albums.

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 first 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 correctly an average of 91% and a minimum of 67% of the permission-comprehension questions asked during the survey at the end of the study. This shows that revising the methodology allowed participants to correctly understand, remember, and apply the ideal policy.

6.4 Study 4

As mentioned in Section 4.4, we were concerned that conveying the ideal policy as an explicit, bulleted list of rules was over priming participants to look for permission errors. In pilots of 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 bulleted lists. We found that presenting the policy in bulleted lists 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

The final goal of our study designs was to allow us to accurately measure participants' ability to recognize errors in the permissions set on albums. In order to do so, we needed to distinguish between environments that had no errors, environments that had errors that participants did not notice, and environments where errors were noticed.

7.1 Study 1

We chose to carry out the study in the lab because this offered us the most 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, to a degree, how participants would approach 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. Hence, for this study we measured "noticing" by counting the number of errors corrected. The number of permission errors corrected is a strict subset of the number of errors noticed, and we anticipated a large difference in the number of permissions corrected between the conditions. Because of this, we were willing to accept that we would not detect that a participant recognized an error if she chose not to correct it.

When designing questions to test recall we were concerned about participant fatigue leading to questions being guessed or answered arbitrarily. To counter this, we limited our questions to six albums and only asked about two of the actions actions that could be performed on albums. All recall questions could be answered with "Not sure" to make providing valid answers no more difficult than 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; such behavior led to undercounting the number of times errors were "noticed" (since we actually counted only corrected errors). We considered changing our measurement methodology, but determining whether a participant had checked the permissions was impossible for participants in the non-control conditions, who may or may not have looked at a proximity display.

7.2 Study 2

In designing study 2 we focused on being able to observe 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 problem with the methodology of study 1. To adjust, we made study 2 a think-aloud study. Study 1 was deliberately not a think-aloud study to allow us to measure whether participants took an equal amount of time to complete tasks in different conditions (**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 other aspects of 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 permissions 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 permissions 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 relative to the screen. This is possible for short studies, but our study took 1.5 hours on average. Participants would shift in their chairs or lean on the desk, moving them out of range of the eye tracker. We considered prompting participants when they moved outside the required area, but decided this would distract participants and alter their behavior in other ways as well. We tried having participants experiment with the eye tracker before the study so that they knew where the optimal area was. This helped to a degree, but participants still became distracted by the study and started moving outside the optimal area. The eye-tracker data did give us a sense of when participants looked at permissions displays, but was insufficiently complete for reliable, accurate measurement.

7.3 Study 3

In study 3 we wanted to obtain 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 permission recall between conditions (**H4**). We hypothesized that this was due to the full-sized permission modification interface. Participants who visited the interface frequently changed permissions for more than one album, indicating that, even in the control condition, these participants were looking at other permissions. To address this issue, we adjusted our methodology to make the permission-modification interface 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 recall In studies 1 and 2 we asked participants to answer 128 questions to test their recall of the permission settings related to 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 recall questions and elicit free responses. We felt free-form answers would get us a better sense of what participants 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 answered 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 posed recall questions to practice participants, who had not checked permissions during the study, we found that they

became embarrassed that they did not know the answers, and after a couple questions they started guessing. To discourage guessing, we interleaved the recall and comprehension questions, which we expected a much larger fraction of participants to answer correctly. We found that this discouraged guessing and participants seemed more comfortable admitting that they could not recall the permissions for albums for which the did not check the permissions.

Post-study debriefing At the end of the session we debriefed the participant. 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 in duration (1–2 seconds) and we were concerned that participants would not remember why they had engaged in a particular action or made a comment an hour ago. Hence, we used a contextual interview approach [5], where the participant opened the album they had been 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 a single 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 by several metrics than those that saw the partial permission-modification interface, and were more likely to correct permissions regardless of whether they saw the proximity display.

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, it was evident when permissions were checked, because this involved opening a new interface. In the non-control conditions, we could not as reliably determine whether permissions were checked. 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 quickly enough to not vocalize that they had checked [14]. This indicates that our proximity displays are effective, but 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 was very effective at helping participants remember their reasoning behind specific actions. In cases where the participant could not remember, they were still often able to make an educated guess as to why they would have performed an action given their behavior up to that point. While a guess is not as good as remembering, participants' guesses as to reasons behind their actions were likely more accurate than researchers' educated guesses.

7.4 Study 4

Studies 1 through 3 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, the participants who internalized the policy 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 anticipated it to be sufficiently precise to detect differences in behavior between conditions.

Permission-modification interface In study 3 we observed that participants who saw the permission-modification interface in a dialog window experienced 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 "noticed" as "corrected" was effective in that we were able to show statistically significant differences between some experimental conditions and control conditions (not all conditions were expected to perform differently from the control conditions). We attribute this to both a larger number of participants and clearer, more tested, study materials.

Similarly to study 1, we had a limited ability to measure why participants changed or did change permissions. However, we collected extensive logs, which we compared to behaviors observed in prior studies to infer what users were doing and why.

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.

Permissions as a secondary 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 overly incentivize participants to check permissions so we tried to balance the amount of priming with the cost of checking. We successfully managed priming on studies 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. In studies 2 and 3 we added tasks that required time and effort to determine there were no permission errors. We increased the cost of checking permissions in study 4 by adding a time limitation, which forced participants to make trade-offs. We found that at least 50% of the tasks needed to have no permission errors 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. However, we encountered some problems when we asked participants to be responsible for albums that "belonged" to other people, such as Pat's mother. We countered this issue in the second study by making it clearer that these other people 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, and to express the policy implicitly in the emails (studies 1 and 2). The second way was to have a single, concise policy that applied to all the albums, and to express it using direct wording at the beginning of the study (studies 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 across tasks. 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 measurement challenge was defining and testing participants' ability to "notice" permission errors. In the first study we measured the rate at which participants corrected permission errors, but this approach was insufficiently precise to measure the difference between conditions. In later studies we measured the rate at which participants checked for permission errors. This definition allowed us to observe whether 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 described our successes and difficulties in terms of our four methodological goals: 1) permission as a secondary 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.

Acknowledgments

This work was supported in part by Carnegie Mellon CyLab under Army Research Office grant DAAD19-02-1-0389; by ONR grants N000141010155 and N000141010343; and by NSF grant CNS-1018211.

9. REFERENCES

[1] L. Bauer, L. Cranor, R. W. Reeder, M. K. Reiter, and K. Vaniea. Comparing access-control technologies: A study of keys and smartphones. Technical Report CMU-CYLAB-07-005, Carnegie Mellon University, 2007.

[2] S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge. Location disclosure to social relations: Why, when, & what people want to share. In *Proc. CHI*, 2005.

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

[4] Gallery 3. `http://gallery.menalto.com/` [accessed 23 Sep 2012].

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

[6] J. M. Haake, A. Haake, T. Schümmer, M. Bourimi, and B. 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, K. 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] A. Sotirakopoulos, K. Hawkey, and K. 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] M. W. van Someren, Y. F. Barnard, and J. A. 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] R. 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.