I'm Allowing What?

Disclosing the authority applications demand of users as a condition of installation

Jennifer Tam
Carnegie Mellon University
jdtam@cs.cmu.edu

Robert W. Reeder
Microsoft (TUX)
RoReeder@microsoft.com

Stuart Schechter
Microsoft Research
StuS@microsoft.com

ABSTRACT

Computer operating systems, and now websites that serve as application platforms, are increasingly adopting stricter application security models; they restrict the resources applications can access to those authorized by the user. Users are asked to authorize access to these resources either when the application is installed or when previously-unauthorized resources are required. For example, Facebook requires its 400+ million users to make authorization decisions whenever an application first tries to run within a user's account. The Android mobile phone OS requires its millions of users to make application authorization decisions when downloading new applications. While the security of these users' systems and data increasingly rests on their ability to make these authorization decisions, there is little research to guide those designing these application authorization experiences.

We performed a laboratory study to evaluate different designs for disclosing the actions and resources that an application will be authorized to perform once installed. We used a within-participants design to observe thirty-three Facebook users' ability to absorb and search information in seventeen different disclosure designs, all of which were presented in the context of a fictional Facebook application. These designs were chosen to proxy for designs users rely upon today, from platforms including Facebook, Android, OAuth, and HealthVault. Four of these designs conveyed only a set of resources to be authorized, such as the user's *contact information* or *friends*. The other thirteen designs paired resources with different actions that could be performed on them, such as *seeing contact information*, *changing contact information*, or *adding* new *contact information*.

We find that participants overwhelmingly prefer disclosure designs that present resources visually, using icons or pictures, and can search those containing icons most quickly. Surprisingly, we find little variance in participants' performance on our information-absorption tasks over widely varying disclosure designs. We do, however, find that participants perform better when disclosures are organized by actions, and followed by the various resources on which the actions would be authorized, than when information is grouped by the resources.

1. INTRODUCTION

Historically, operating systems and other platforms have implicitly authorized applications to access system and user resources at the time of installation, without conveying the details of this authorization to the user. However, a rapidly growing set of platforms now list the resources and actions to be authorized so that users can make informed decisions to proceed, or not proceed, with installation.

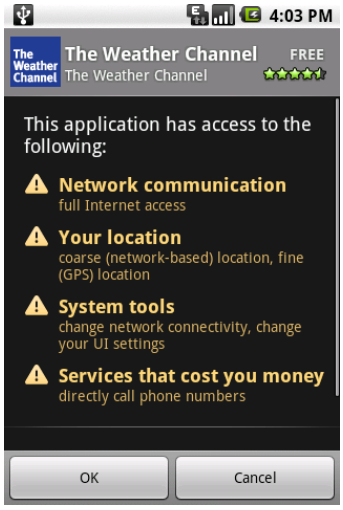
Examples of application installation dialogs that disclose the resources to be authorized are illustrated in Figure 1: the first is for client applications running within Google's Android operating system; the second for web-applications accessing Twitter's implementation of OAuth, an emerging standard authorization of web applications; and the third for applications built on top of Microsoft's HealthVault health data platform. Each dialog discloses the resources that the application will be authorized to access and, in some cases, the specific actions that the application will be authorized to perform on each resource. These resource are specific to the application.

In contrast, Facebook's application installation dialog, illustrated in Figure 1d, discloses a default set of resources authorized for all installed applications. In response to user concerns, which include concern over applications that were "stealing log-in credentials and spamming victims' friends," [16], Facebook announced on August 27, 2009 that it would soon "require applications to specify the categories of information they wish to access" so that it, too, can disclose each application's specific authorization requirements before users choose to install [6].

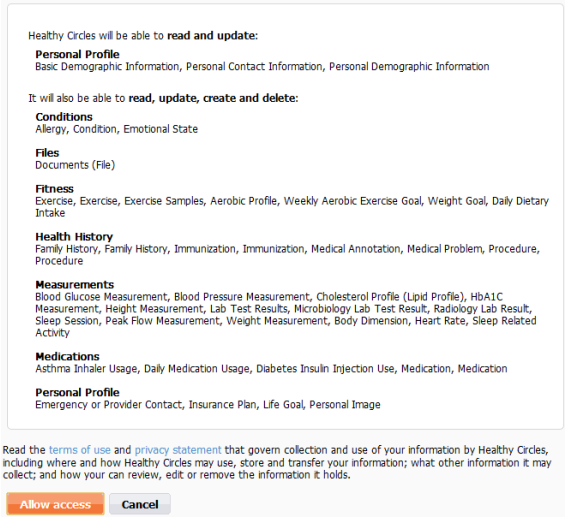
For all of these application installation experiences, authorization is a condition of installation; the user can not run the application without granting access to the requested resources. Authorization disclosures are intended to make users better informed, and ultimately make them better decision makers. Alas, there is little research to guide those designing these disclosures to efficiently convey information about complex authorization decisions.

One challenge facing those researching how users make authorization decisions is to identify successful outcomes. Even if users have perfect information about how applications will behave, their choices of whether to install them are inherently subjective and may stem from considerations outside researchers' control. For example, the user's key decision factor may be guidance from a friend who has experience with the application.

Researchers could learn more by instrumenting application platforms: quizzing users to determine how well they

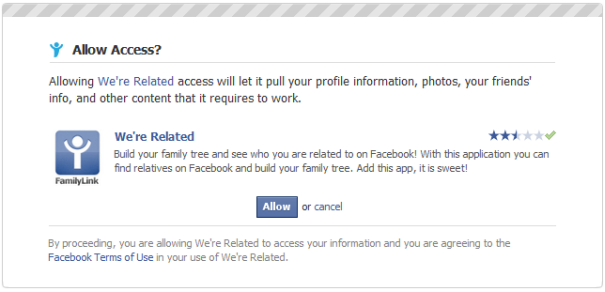


(a) Android



(b) HealthVault

(c) OAuth (Twitter)



(d) Facebook

Figure 1: Application authorization dialogs used by device operating systems (Android) and web application APIs (OAuth, HealthVault, Facebook). The Android, HealthVault, and OAuth dialogs disclose application-specific information about the authority granted upon installation. The Facebook dialog presents the same disclosure for all applications. *Images obtained 9/15/09.*

absorb information in a real-world situation, or asking them to explain the reasons behind a choice. However, exposing users who are making real-world security decisions to untested design options poses risks that should be taken only after other options are exhausted.

As an initial step toward determining the best way to present application authorization disclosures to users, we performed a within-participants laboratory study to examine the efficacy of 17 different disclosure designs. These designs disclose the resources or resource-action pairs that would be authorized if a Facebook application were to be installed. In some trials, participants were shown an authorization disclosure for a short time and then asked to answer a question based on the information they had absorbed. In others, participants were shown a question first and searched through the disclosure to identify whether access to a particular resource was authorized. We created treatments by varying such factors as whether resources were represented by names, as icons, or as images. We also varied the order and grouping of resources and actions. We both tracked participants' performance and, at the completion of the experiment, asked participants to rank the designs. We asked them to rank designs based on how easily they could be absorbed, how easily they could be searched, and their desirability for use when making real-world application installation decisions.

While the scope of our experiment could not facilitate a complete solution to the application authorization problem, our concrete findings will be valuable to those designing the next generation of authorization disclosures.

2. CONSTRUCTING DESIGNS TO TEST

In examining each of the different designs used by Facebook, Android, HealthVault, and various OAuth implementations (see again Figures 1 and 1d), we realized that multiple design decisions differentiate each from the others. For example, at the center of Facebook's application installation dialog is an extensive description of the application itself, with the authorization disclosure presented in a paragraph above. In contrast, Android places a small application description at the top of its application authorization dialog, which is dominated by the disclosure presented using an outline format. A direct comparison would not allow us to separate the choice of a paragraph or outline as the central design element from confounding factors such as the type

of application description provided, the fraction of dialog real estate allocated to the authorization disclosure, or the relative locations of the application description and authorization disclosure. Beyond matters of presentation, an additional confounding factor is that these interfaces are used on different architectures, which expose different resources and actions to users.

We sought to create a set of disclosure designs that would allow us to isolate and test key design elements in isolation. We selected many design elements used by real application platforms today and assembled them into a single design framework so that we could minimize confounding variables and ensure a level playing field.

We chose to test different designs for disclosing application authorization requirements in the context of Facebook because most Facebook users have already installed applications and encountered the application authorization disclosure dialog in Figure 1d.

We examined three main design choices when constructing our disclosure designs: the central design element, authorization granularity, and resource-action grouping. From our three variables we generated a total of seventeen different designs.

2.1 Central design elements

All designs were built around one of five central design elements, which dictated the layout of the disclosure information and thus was the design choice that had the most salient visual impact. The *paragraph* element compactly represents an authorization disclosure as a single paragraph; the *outline* element places actions to be authorized in headings and indents their corresponding resources below, or vice versa; the *table* element places actions and resources on different axes of a table; the *image* element contains pictures that resemble the content or appearance of resources; and the *icon* element uses more abstract iconic representations of resources. The paragraph and outline elements are currently in use, and the table, image, and icon elements were derived from what was suggested in related work.

2.2 Authorization granularity

Some platforms, including Facebook, disclose the resources that an application will be authorized to access (e.g. *contact information*) but not the specific actions to be permitted. Others, including HealthVault and OAuth, pair resources with the actions to be authorized (e.g. *see*, *change*, and *add to*). We refer to these two levels of granularity as *resource-only* and *resource-action*, respectively.

The four resource-only designs are illustrated in Figure 2. In these designs, the disclosure stated that applications would simply be able to "access" these resources. No resource-only *table* element design was used, as the design targets two-dimensional data and a resource-list is inherently one-dimensional.

We will next describe how we further subdivided resource-action designs designs by how resource-action pairs are assembled into groups.

2.3 Grouping

When *paragraph*, *outline*, and *icon* design elements are used to present disclosures at the resource-action level of granularity, the resource-action pairs to be disclosed can be presented in groups assembled either by resources or by actions. For example, when *grouping by resource* in the *outline* element one could present a resource name (e.g. contact information) followed by all the actions to be authorized on it (see, change, add to). Alternatively, when *grouping by action*, one could present an action name (see) followed by all the resources that the application will be authorized to perform that action on (contact information, newsfeed, photos).

(a) *icon* (b) *outline* (c) *image* (d) *paragraph*

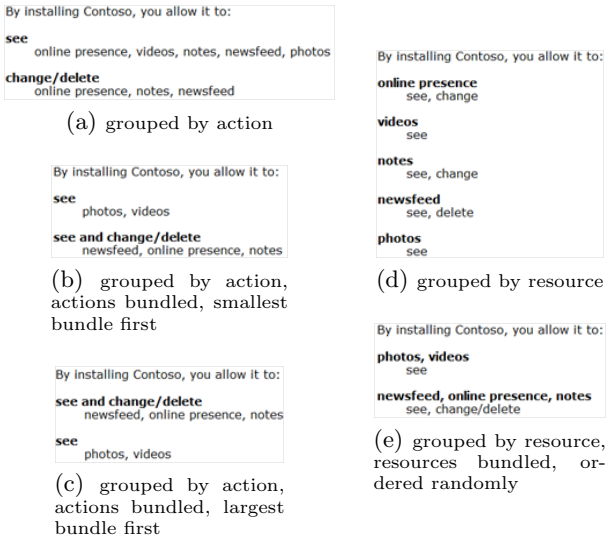
Figure 2: Resource-level designs.

The five *outline* designs are presented in Figure 3. Designs are further bundled by whether more than one action or resource may be clustered into a single heading line.

The two *paragraph* designs are presented in Figure 4. Unlike the other central design elements, the grouping impacted only whether the resource or action was boldfaced; the rules of English made it difficult to put resources in front of actions and make it unreasonably unwieldy to repeat actions for each resource. We could not imagine a workable design that would be grouped by resource.

The lone *table* and *image* designs appear in Figures 5 and 6 respectively. Other variants of *table* and *image* were not possible because we did not want the designs to exceed the available vertical space and therefore require participants to scroll down.

The four *icon* designs are presented in Figure 7. The *icon.a* design is grouped by action and the *icon.d* design is grouped by resource. Two additional designs grouped by action, *icon.b* and *icon.c*, were added mid-study in response to participant enthusiasm for *icon.a*. As a result, we only obtained results for the *icon.b* and *icon.c* designs for the final 13 and 15 participants, respectively.



(a) grouped by action

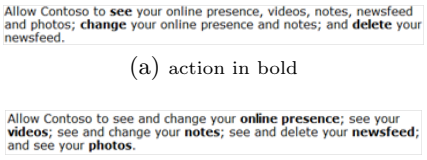
(b) grouped by action, actions bundled, smallest bundle first

(c) grouped by action, actions bundled, largest bundle first

(d) grouped by resource

(e) grouped by resource, resources bundled, ordered randomly

Figure 3: Action-level *outline* designs.



(a) action in bold

(b) resource in bold

Figure 4: Action-level *paragraph* designs.

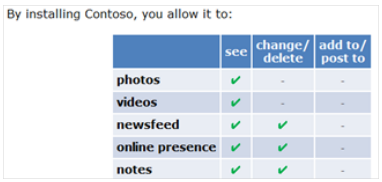


Figure 5: Action-level *table* design.

Figure 6: Action-level *image* design.

3. EXPERIMENTAL DESIGN

We used a within-participant design in which 33 participants were exposed to all of the designs from Section 2 in randomized order. For each disclosure design, we showed each participant a set of 24 consecutive *trials*: disclosures of randomly selected pairs of resources and actions matched with a yes/no question about whether a resource/action pair was among those disclosed. Each question asked: "Would installing this Contoso application allow it to *action* your *resource*," where *action* and *resource* are replaced with names of actions and resources. For resource-only representations only the *access* action was used. Participants could answer with *no*, *probably not*, *not sure*, *probably yes*, or *yes*.

3.1 Trial types

Trials were intended to gauge either a participant's ability to *absorb* the information in a disclosure (*absorption trials*) or to *search* the disclosure for the answer to a question (*search trials*). For each disclosure design, participants were presented with eight long absorption trials, followed by eight short absorption trials, and concluded with eight search trials as described below. Since search trials are the simplest, we start by describing them.

3.1.1 Search

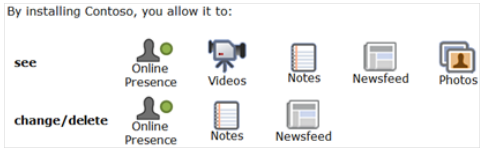
In search trials, participants were asked a question before seeing a disclosure. These trials were designed to represent the behavior of a user who is looking to address a specific concern by inspecting the disclosure. Once a participant had read and understood the question, she could press a key and the disclosure would appear immediately below it, as illustrated in Figures 8 and 9. When the participant pressed a key to answer the question we recorded the amount of time that had passed since the disclosure had been presented.

3.1.2 Absorption

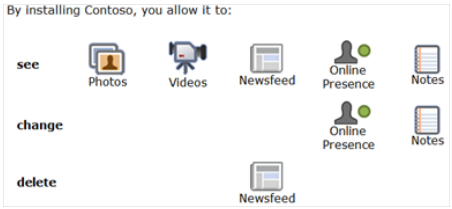
In an absorption trial, participants would be presented a disclosure for a fixed period of time (Figure 10), after which the disclosure would disappear and be replaced by a question about what the participant had seen (Figure 11). Participants had to answer the question using the information they had processed during the absorption period and still remembered after reading the question. These trials were designed to represent the behavior of a user who wants to understand the full content of the disclosure.

We never asked more than one question after presenting a disclosure because processing prior questions could further diminish participants' memories of the disclosure.

We chose to show eight long absorption trials first followed by eight short-absorption trials because this gave participants the opportunity to learn a design before reaching the more challenging short-absorption trials. We determined the length of our absorption periods through a pilot study, tracking scores and user an eye tracker to follow the progress of participants' eyes over each disclosure. Our goal was to ensure that participants could absorb enough information to answer some questions, but would often not be able to absorb all the information. Using the eye tracker, we observed participants read through the bulk of disclosures in eight seconds but fail to traverse the full content in four seconds, especially when the disclosures were long. The optimal timing to statistically differentiate the impact of different design

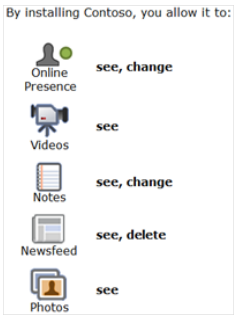


(a) grouped by action, resources shifted left



(b) grouped by action, resources in columns

(c) grouped by action, resources in columns, similar actions (rows) consolidated



(d) grouped by resource

Figure 7: Action-level *icon* designs.

decisions is one for which, on average, participants are able to absorb and recall the information required to answer a question 50% of the time. Since questions were all yes/no, participants had a 50% chance of guessing the correct answer if they did not already know it. Thus, we selected an absorption period that would cause participants to answer questions correctly an average of 75% of the time. The average absorption-trial performance was indeed very close to this goal.

Because resource-only disclosures contain much less information, we reduced the absorption periods by a factor of two (four and two seconds, respectively) for these trials. In retrospect, the reduction factor should have been larger.

3.2 Application platform parameters

We conducted our experiments using sets of resource and action names that were selected to be familiar to Facebook users and make sense for questions posed in the context of a Facebook application.

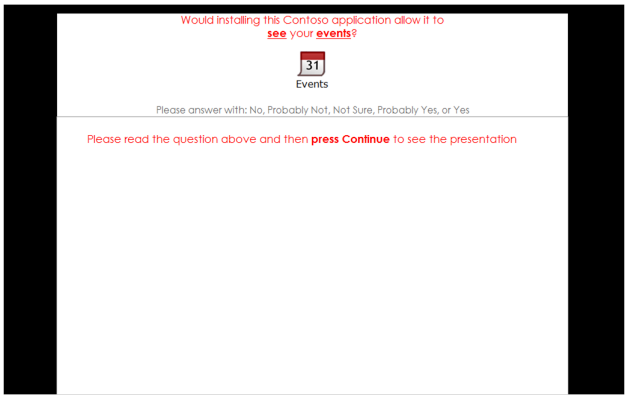


Figure 8: Example search trial before user presses key to see the disclosure.

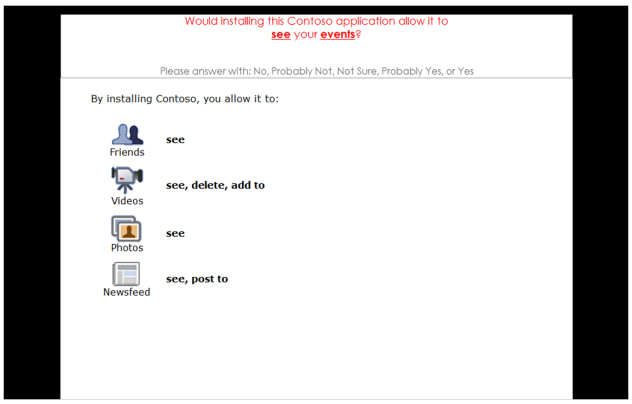


Figure 9: Example search trial where user has pressed key to see the disclosure.

3.2.1 *Resources and actions*

We selected ten resources that we believed Facebook users would be familiar with for use in our application authorization disclosures: *contact information*, *friends*, *photos*, *online presence*, *videos*, *newsfeed*, *education/work history*, *notes*, *events*, and *networks*.

The default actions that could be authorized on each resource were *see*, *change*, and *add to*. Table 1 contains the set of all possible combinations of the three actions, excluding the empty combination and the two combinations for which an application would be authorized to change (or delete) a resource it could not see. While we strove to use only action combinations that were likely to occur in real-world use,

	see	*change* (*delete*)	*add to* (*post to*)
1	✓		
2	✓	✓	
3			✓
4	✓		✓
5	✓	✓	✓

Table 1: The five possible bundles of *actions* that an application could be authorized to perform on a *resource*. Each row represents one bundle of actions. We excluded bundles that would authorize and application to change or delete a resource it could not see.

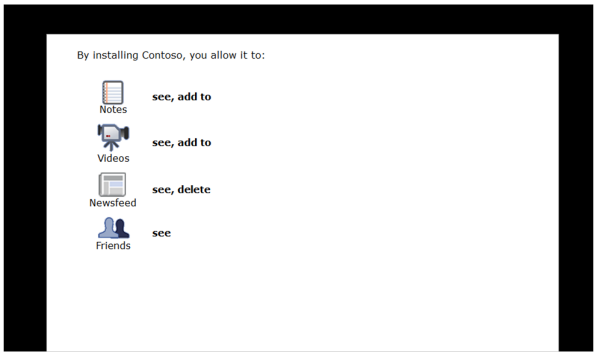


Figure 10: Example absorption trial disclosure.

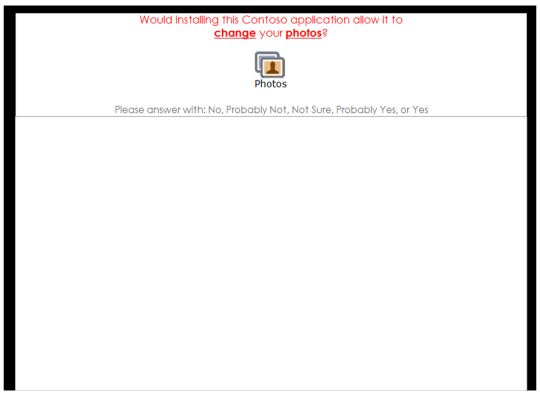


Figure 11: Example absorption trial question.

participants were told that no action implied the presence of any other action so that they would not presume that an application could *see* any resource it could *add to*.

For added realisms, we made some exceptions to the default action names in order to ensure they were semantically meaningful. The *add to* action implies adding new content to a resource that can grow in size, such as a set of notes, friends, or photos. *Online presence* – whether the user is currently online or not – is a single value that can change but not grow, and so the *add to* action was never used for this resource. Newsfeed entries and individual videos are resources that one cannot change but can delete, so the *change* action was replaced by a *delete* action for these two resources. Finally, one is more likely to say that he or she would *post to* a newsfeed than to *add to* it, and so the former action name was substituted for the latter when used in the context of the newsfeed.

3.3 Generating trials from parameters

We generated two pools of disclosure-question pairs used for trials: one for resource-only disclosures and one for the more fine-grained resource-action disclosures. Within granularity categories, the same question pools were used for both search and absorption trials.

3.3.1 Disclosure parameters

We generated each disclosure-question pair starting with the disclosure. First, we determined the number of resources to be disclosed by randomly (and uniformly) picking a number between three and five. We then selected that many resources randomly (and uniformly) from the ten available. This range of three to five resources was selected based on the performance of pilot participants and was intended to ensure that the questions were sufficiently challenging while still being potentially tractable in the time given.

For resource-action disclosures, we matched each resource with a randomly selected bundle of actions from the set of combinations in Table 1. If the resource was *online presence*, we selected randomly from the first two combinations in Table 1 as one cannot *add to* online presence–it is a fixed sized field.

3.3.2 Question parameters

All questions asked whether the disclosure authorized a given action to be performed on a given resource. Thus, the two inputs to each question were the resource and action to ask about.

To ensure that both a *yes* and *no* answer were equally likely, we generated the answer uniformly and at random before generating the question. We then selected randomly from the set of possible questions that would have this answer.

For questions about resource-only disclosures (the single action was "access") we randomly select the resource to ask about from the set authorized (for *yes*) or the set not authorized (for *no*).

For questions about resource-action disclosures, we generated questions for *yes* answers by randomly selecting a resource and action from the set of authorized resource-action pairs. When the answer was to be *no*, we randomly (and uniformly) selected a resource-action pair from the union of two sets: those resource-action pairs not authorized but for which another action is authorized on that resource (the resource will appear in the authorization disclosure) and the set of all possible actions for two resources randomly selected that were not present in any authorized resource-actions pairs (the resources would not appear in the authorization disclosure).

4. EXPERIMENTAL OPERATION

We conducted our study during two-hour sessions between the dates of August 28 to September 1, 2009. The session consisted of an initial study, the experiment over 17 randomly ordered disclosure designs with 21 consecutive trials for each design, and a post-experiment questionnaire. At the conclusion of the study all 33 participants received a gratuity.

4.1 Participant recruitment

We required participants to be Facebook users and to have used at least one Facebook application. We sought to obtain a participant group evenly split along both gender lines and above and below the age of thirty. A recruiting team, organizationally independent of the research team, used a diverse pool of prospective participants from myriad occupations. The pool contained members of the general public who had been recruited via public events, sweepstakes, online advertisements, and a website. The recruiters interviewed each prospective participant to ensure that he or she met the study requirements. Seventeen participants in the full study (7 men, 10 women) were no older than 30 years old and sixteen participants (9 men, 7 women) were at least 31 years of age.

4.2 Preliminaries

Upon arrival, participants were asked to complete a consent form followed by a questionnaire containing demographic questions and questions about their use of Facebook. We asked how recently they had seen the Facebook application installation dialog, the Facebook privacy settings page for all applications, and the application settings page that lists applications the user has installed. We then asked what security concerns they had in installing Facebook applications, and how they would restrict Facebook applications from accessing their account if they could.

4.3 Experimental trials

For each disclosure design, participants completed a set of 24 consecutive trials: first came eight *long absorption* trials, followed by eight *short absorption* trials, and last came eight *search* trials. As this was a within-participants study, all participants were presented with all available designs and the order in which participants encountered each design was randomized.

For each trial, a disclosure-question pair was randomly selected from the set of pairs that the participant had not yet seen.

4.4 Final questionnaire

After completing all trials participants were asked three final questions. In the first and second questions, participants were shown all thirteen resource-action disclosure designs and asked to rank them from easiest to hardest. The first question asked participants to rank the designs based on how difficult they were to absorb while the second question asked participants to rank the designs based on how difficult they were to search. The third and final question asked participants to select the design they would prefer to see if they actually had to make a real-world decision regarding whether or not to install an application.

5. RESULTS

Recall that participants could answer with *no*, *probably not*, *not sure*, *probably yes*, or *yes*. We assigned a score of 1 for each correct answer, −1 for each incorrect answer, and 0 for each *not sure* answer. Answers conditioned with 'probably' were treated as if they had been given with full confidence. The option to provide conditioned answers served its assigned purpose of eliciting answers from participants who were uncertain but believed they had a better than random chance of answering correctly.

The overall average score for long and short absorption trials was 0.576 (0.578 when adjusted for learning effects; see Section 5.2) and the median time for search trials was 2341 milliseconds (2182 milliseconds when adjusted for learning effects). For trials in which the correct answer was *yes*, the average score was 0.606 (still 0.606 when adjusted for learning effects), while for trials in which the correct answer was *no* the average score was 0.548 (0.551 when adjusted for learning effects). This suggests participants had enough time to absorb the correct answer and recall it 77.5% of the time—very close to our goal of 75%.

Sections 5.1 and 5.2 examine two potentially confounding factors – cognitive load and learning effects – that could have affected our results. Our main results follow in sections 5.3, 5.4, and 5.5, which address the performance of designs, whether it is better to group resource-action pairs by action or by resource, and user preferences amongst the designs.

5.1 Cognitive load

We use two metrics to represent the amount of information that participants must absorb to complete absorption-trials: the number of resources presented and the number of resource-action pairs to be authorized. For example, if we were to present an authorization to *see* a user's *notes* then this would be a cognitive load of one resource and one resource-action pair. If now the authorization is to *see* and *change* a user's *notes*, this addition would increase the load to one resource and two resource-action pairs.

We graph mean scores and median times as a function of resources and resource action pairs in Figure **??**. Note that there were very few questions that contained more than 11 pairs and so the resulting figures are likely to have large margins of error. Observe that with one exception, the scores are monotonically decreasing as the load increases from 3 resource-action pairs to 11. The negative correlation between scores and resource-action pairs is statistically significant ($r = -0.737, t = -3.27, df = 9, p < 0.001$) which means cognitive load is negatively impacting user performance on absorption tasks.

Participants' mean scores on the hardest (2s short-absorption) resource-only trials ($M = 0.795, sd = 0.129$) were significantly higher than their mean scores on the easiest (8s long-absorption) resource-action trials ($M = 0.492, sd = 0.129$): Wilcoxon $W = 1025.5, p < 0.001$. Thus, cognitive load increases by much more than a factor of two with the introduction of actions. As resource-action pairs grow as a product of the resources and actions available, the number of pairs is likely the better metric of cognitive load.

5.2 Learning

We considered that participants might require a few trials to get up to speed on the absorption and search tasks at the start of the study, and that they might continue to improve over time.

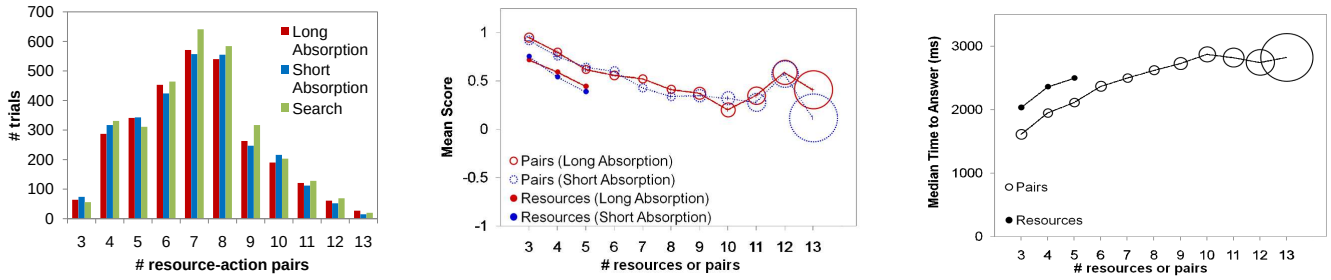
Figures 13a and 13b depict participants' performance on the first 24 resource-action trials they encountered: 8 long-absorption trials, followed by 8 short-absorption trials, and ending with 8 search trials. The mean participant score on the first resource-action trial (a long-absorption trial) was 0.182. This is in contrast to a mean of 0.497 for all other resource-action long-absorption trials. A Wald test shows the difference is significant ($z = -2.00$, $p = 0.023$).[1]

We observed clear improvements in participants' speed on the eight initial search trials that followed. The median time to complete the first search trial is 5.81 seconds and these times drop monotonically to half of this initial value over the following six trials. These speed-ups occurred even though participants should have already been quite familiar with the design–they saw 16 absorption trials before the first search trial.

Figures 13c and 13d illustrate learning over the full sequence of resource-action designs participants saw over the study. We observe that absorption scores are relatively flat, but search times improve over the course of the experiment.

We examine whether participants' performance improves

[1]Since the Wald test assumes scores are normally distributed we ran a Kolmogorov-Smirnoff test of the non-normality hypothesis for these scores, $p = 0.854$.



(a) Distribution of load over all trials

(b) Absorption scores as function of load

(c) Search times as function of load

Figure 12: Number of trials (a) and performance (b,c) as a function of two cognitive load metrics: the number of *resources* presented and the number of resource-action *pairs* presented. Subfigure (a) shows the distribution of the second metric (resource/action pairs) over all trials. Subfigures (b) and (c) show performance as a function of load. To highlight the potential for error from small sample sizes, each data point is surrounded by a circle inversely proportional to the sample size, illustrating that very few trials presented more than 11 resource/action pairs.

within the set of trials for each design in Figures 13e and 13f. We see no discernable improvement in participants' absorption scores and only a small improvement in search time.

Finally, we examine whether learning effects increase participants' performance between the first eight long-absorption trials and the second eight short-absorption trials that follow. If present, learning effects would increase scores in the later short-absorption trials. However, they are countered by cognitive load effects, which should decrease scores in the short-absorption trials. We attempt to isolate the two effects by splitting the trials into those with easier low-load disclosures, those with six or fewer resource/action pairs, and harder high-load disclosures, those with seven or more resource/action pairs. There were a total of 2303 low-load trials (1158 short, 1145 long) and 3527 high-load trials (1754 short, 1773 long).

For the low-load disclosures with at most six resource/action pairs, the mean score on the long-absorption trials (0.61) was actually lower than the mean score for short-absorption trials that followed them (0.67), even though participants only had half as much time in the short-absorption trials. Though the difference was not statistically significant ($V = 199, p = 0.150$), this hints that learning effects trumped cognitive load for this important minority (40%) of trials. This helps explain why the aggregate means of the long- and short-absorption trials are so similar.

For the high-load disclosures with seven or more resource/action pairs, the mean score on the long-absorption trials was higher (0.43) than the mean score for the short-absorption trials that followed them (0.35). This difference was statistically significant ($V = 411, p = 0.019$) and indicates that load effects trumped learning effects for this majority (60%) of trials.

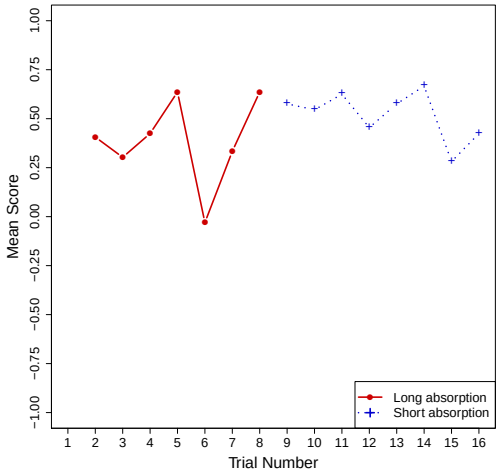
Because we arranged designs in a random order for each participant, most of these learning effects should cancel out. However, to err on the conservative side in our analysis, we eliminated trials that appeared to be most affected by learning: the first long-absorption trial of each participant session; the first set of 8 search trials for each participant session; and the first trial of every set of search trials for the whole study. Learning-adjusted results were presented alongside non-adjusted results above in Section 5; analysis for the remainder of the paper uses learning-adjusted analysis. However, it should be noted that we analyzed results both with and without learning adjustment, and found only negligible differences.

5.3 Performance of designs

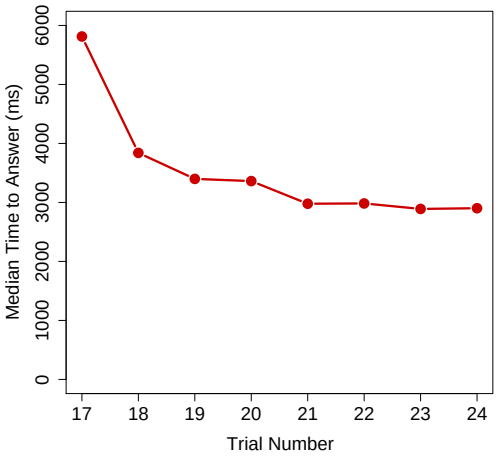
The graphs in Figures 14a and 14b show, for each design, the average participant scores and median search times, respectively. For Figure 14a, error bars represent 95% confidence intervals and are constructed under the assumption that the underlying scores are drawn from a normal distribution. In Figure 14b, boxes represent the inner two quartiles of the distributions with the medians represented as horizontal lines. There were no significant differences between designs in the absorption trials.

Participants answered most quickly for *image* and especially *icon* designs. Users' median search times for *icon* designs were significantly lower than the median times to search all other designs: paired two-sided Wilcoxon signed-rank $V = 99, p < 0.001$. It is important to note that, when asking questions about *icon* and *image* designs we included the icon/image describing the resource as part of the question. Participants knew which image to look for, just as they would have known what text to look for in a non-pictorial representation.

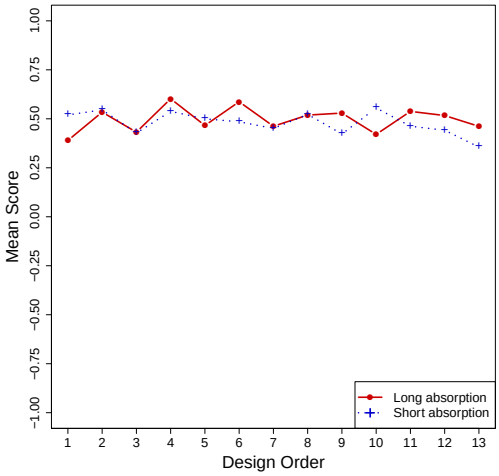
Users' median search times for *image* designs were not significantly lower than the median times to search all non-graphical designs (all others excluding icons): paired two-sided Wilcoxon signed-rank $V = 228, p = 0.83$. Given the ambiguity between the two types of graphical designs, we calculated users' median search times over all graphical designs, both *icon and image*, and compared them to the median search times over all other designs. The difference remains strongly significant: paired two-sided Wilcoxon signed-rank $V = 68.5, p < 0.001$.



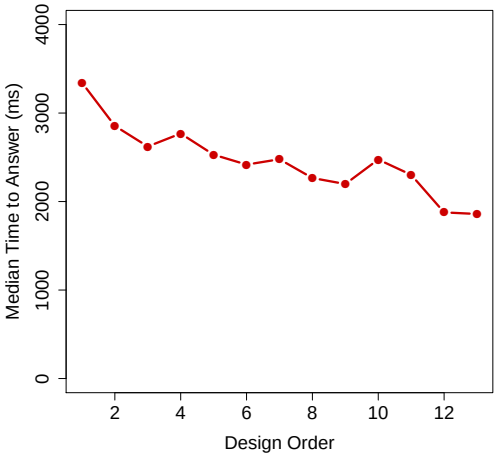
(a) Mean scores over all users over each of the first 16 resource-action absorption trials of the study.



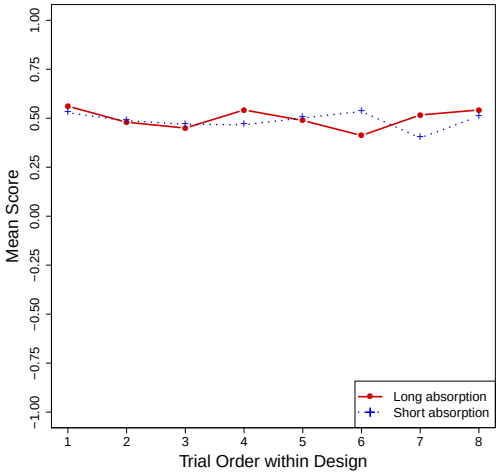
(b) Median times over all users over each of the first eight resource-action search trials of the study, which immediately followed the absorption trials in Figure 13a.



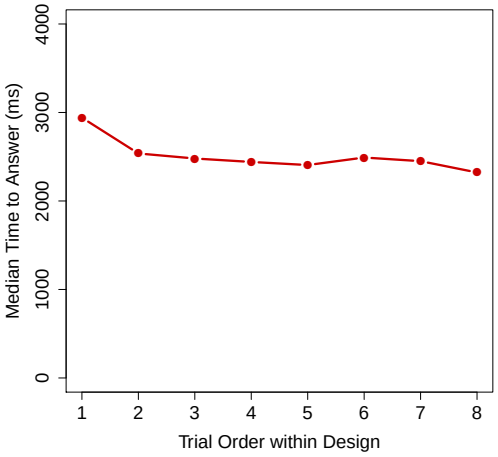
(c) Over the course of the study participants were shown a sequence of resource-action designs. For each step in this sequence we calculate each user's mean short-absorption score for the design at this step. We then calculate the mean short-absorption score over all users for this step in the design sequence. We then do the same for long-absorption trials. Note that different users will see different presentations at each step of the sequence.



(d) Over the course of the study participants were shown a sequence of resource-action presentations. For each step in this sequence, we calculate each user's median search time for the design at this step. We then calculate the median time over all users for this step in the design sequence. Note that different users will see different designs at each step of the sequence.

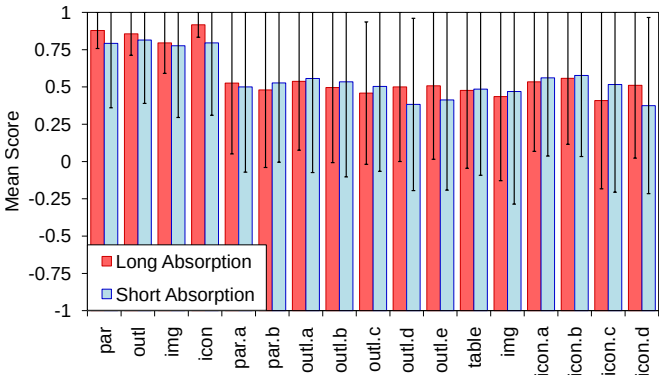


(e) Over the course of trials for each design, participants encountered a sequence of eight long-absorption trials followed by a sequence of eight short-absorption trials. For each of the eight steps in each sequence we calculate the mean short-absorption score and mean long-absorption score over all users and all designs.

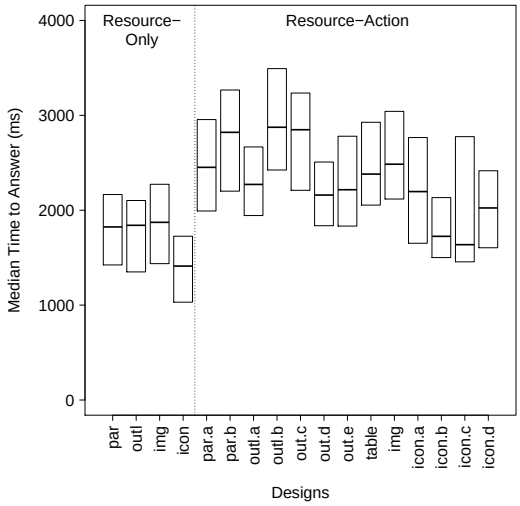


(f) Over the course of trials for each design, participants encountered a sequence of eight search trials. For each user we calculate the median search time for each step through the eight trials over *all* designs . We then take the median over all users for each of the eight steps.

Figure 13: Learning effects



(a) The mean score over all users for each design, with error bars representing the 95% confidence interval.



(b) For each design we calculate each user's median time. Boxes represent the middle two quartiles of those medians with horizontal lines representing the median of medians.

Figure 14: Performance of different designs

5.4 Ordering actions and resources

To test the hypothesis that users perform better when resource-action pairs were grouped by action rather than when they were grouped by resource, we examined pairs of designs that differed primarily by grouping. The first two pairs, *icon.d*/*icon.a* and *outl.d*/*outl.a* differ only by grouping. The third pair pits *outl.e* against the mean of *outl.b* and *outl.c*, as the latter two are similar to *outl.e* but take different approaches to ordering clustered actions. For each user we computed the mean of the three scores from action-grouped designs and for the resource-grouped designs. Mean short-absorption scores were higher on designs that grouped by action ($M = 0.545$, $sd = 0.181$) than those that grouped by resource ($M = 0.390$, $sd = 0.180$). We compare the two sets of mean scores using a Wilcoxon signed-rank test and find the difference is significant ($V = 446.5$, $p = 0.003$).

Because our *icon* designs only used icons to represent resources (not actions) it is possible that improved performance when grouping icons by action is actually the result of participants being better at grouping by textual elements rather than picture elements. However, even removing the *icon* designs, the remaining designs grouped by action outperform those grouped by resource ($V = 396$, $p = 0.014$).

In interpreting the higher scores for disclosure designs grouped by action, it is important to remember that, like these designs, the questions posed in each trial began with an action and the resource followed (would you allow Contoso to *action* this *resource*). We considered posing questions with resources first but found it to be too linguistically awkward. It may be the case that the English language favors the placement of actions before resources, and so these results might not hold for other languages.

Rank	*Paragraph*	*Outline*	*Table*	*Image*	*Icon*
best for absorption					
1	1	0	2	12	18
2	0	5	4	12	12
3	4	16	10	3	0
4	4	12	11	4	2
5	24	1	5	2	1
best for search					
1	1	1	4	8	19
2	0	2	1	19	11
3	3	12	14	2	2
4	5	17	9	2	0
5	24	1	5	2	1
desired for real-world use					
	0	0	7	9	17

Table 2: Participants ranked each design from those they considered easiest to hardest for use in *absorption* and *search* trials, then identified the format they would most like to use for real-world decisions.

5.5 Participant preferences

Having examined our data to determine whether participants perform better on some designs than others, we next examined the post-experiment rankings participants had assigned to each design. They had ranked the designs from easiest to the hardest to use for the absorption trials and again for the search trials. We consolidated these rankings by taking the highest score for each of the central design element: *paragraphs*, *outlines*, *tables*, *images*, and *icons*. The resulting rankings are shown in Table 2, along with tallies of participants' preferred designs for real-world decisions. Icons are the favorite over all metrics and images an uncontested runner up. However, our methodology may have favored icons over images, as participants saw multiple icon designs and thus had more opportunities to rank one of them above the lone image design. Paragraphs and outlines are almost universally disliked.

We examined the rankings to determine if participants expressed clear preferences for designs grouped by action or by resource. We did not find a clear preference, though in our small sample participants were more likely to rank designs grouped by action higher for absorption than for search tasks.

6. RELATED WORK

Prior research with bearing on application authorization disclosures includes the design of systems to restrict application authority, cognitive science studies that present information to people in different graphical formats, and HCI studies of consumers' privacy- and security-related decision making.

6.1 Systems for restricting authority

While the past few years have seen rapid growth in platforms that restrict what applications are authorized to do, the underlying concepts are not new. When Java was introduced in 1995, one of the features of the Java run time system was that it could restrict applets with "subversive code" from accessing restricted resources [9]. At roughly the same time, Goldberg *et al.* were developing sandboxes to restrict native code applications and limit their access to system APIs [7].

Further efforts to confine applications and impose policies that restrict access to system APIs included Cowan *et al.*'s AppArmor (then called SubDomain) [3] and Provos's SysTrace [18]. Both enable users and administrators to share policies, and SysTrace also allows users to grant additional authority on an as-needed basis.

The BitFrost security system, developed as part of the One Laptop Per Child (OLPC) initiative, provided an operating system in which the authority to access resources "is only provided... if the capability is required when the application is installed." [13]. Google's Android operating system, currently deployed in mobile phones, also restricts applications to the resources and actions authorized by the user at installation time. Android applications must provide manifests specifying the actions and resources to be authorized, such as *reading* and/or *writing* to the user's *contacts* [1].

The emerging OAuth standard is a mechanism for authorizing access to APIs of conventional OS applications and web applications [17]. Alas, the existing research by the institutions pioneering OAuth has yet to be peer reviewed and the research that is available, presented at SOUPS 2009 [20], focuses only on maximizing the percent of users who complete the authorization process.

6.2 Cognitive science

Since the capacity of human working memory is limited [15], external representations of data can improve cognitive performance by reducing memory overload. In particular, graphical representations of data, such as tables and graphs, have been shown to improve performance on decision-making tasks by leveraging the pattern-detecting capabilities of the human perceptual system [2, 5, 12, 14]. Some work has attempted to explain which graphical presentation formats (namely graphs or tables) are most useful for a given type of task. Vessey's cognitive fit theory suggests that tables tend to lead to better performance for symbolic (perceptual) tasks, while graphs lead to better performance for spatial (analytic) tasks [22]. However, Speier's later work suggests there may be a crossover point at which even symbolic tasks become sufficiently complex that spatial presentations (graphs) are superior [21]. The task of reviewing an authorization disclosure to answer a comprehension question (the task given to participants in our work) is a symbolic task, with complexity varying as a function of the amount of information presented in the disclosure [23]. Since the complexity of the application disclosure can vary, Speier's work suggests both spatial and symbolic representations may be appropriate for authorization disclosures.

6.3 Usability of privacy presentation formats

Several researchers have studied the comprehensibility and utility of various formats for presenting privacy and authorization information to consumers.

Besmer *et al.* created a Facebook application model and a corresponding user interface that allowed users to authorize applications to access users' personal data. Their user interface lists data categories and samples of the actual data to be shared with an application if the user allows access. They conducted a user study, but were primarily concerned with how users responded to different requests for data from applications. They did not compare multiple representations, as we do in the present work.

Good *et al.* developed a short summary format for showing the contents of End User License Agreements (EULAs) [8]. Their user study showed that these short summaries helped users recognize unwanted software more readily than full-length EULAs did.

A report by the Kleimann Group used an extensive iterative design and testing process to design a format for displaying financial privacy notices to consumers. The work ultimately concluded that a tabular format was best [19].

Kelley *et al.* designed a tabular layout with icons to display website privacy policy data to consumers. Their ultimate design was the result of an iterative process and its efficacy was demonstrated using a lab-based user study [11] and a subsequent large-scale online study [10].

Cranor *et al.* introduced Privacy Bird, a browser attachment that showed simple iconic evaluations of a website privacy policy relative to pre-specified user preferences: a green icon for a policy that did match user preferences, a yellow bird for a website with no P3P privacy policy, and a red icon for a policy that did not match user preferences [4]. Privacy Bird was designed to minimize the amount of information in each disclosure: it displayed one of three colors (red, yellow, or green) depending on the site's privacy policy. Users could request more information to determine why a website's privacy policy did not match their policy preferences.

7. DISCUSSION

7.1 Limitations

When interpreting our results, it is imperative to keep in mind the limitations already disclosed above and to explore further limitations inherent in our study design. Some of these limitations result from our choice of Facebook as the context of our experiment, and our choice of Facebook-specific resources and actions. For example, grouping by action may only be beneficial when the number of actions is smaller than the number of resources, which was true in our study but may not be true for all systems. Similarly, representing resources graphically may be less effective when a platform's resources do not map as well to graphical designs.

Performance on the absorption tasks may have been confounded by factors that would not be of concern to real-world users when reading actual authorization disclosures. For example, we believe that users who examine authorization disclosures in real-world situations are more likely to care about processing the information, and identifying items of concern, than remembering the details. One confounding factor affecting performance on absorption tasks was memory; participants had to remember information between the time the authorization disclosure of a given design appeared and the time they finished processing the question.

Our study design did not allow us to examine the efficacy of different ways to explain the purpose of authorization disclosures or the underlying decision. If certain disclosure

designs help users to grasp the underlying task, we would not be able to discover this in our study. This question cannot be studied in a within-participants design, as once a participant grasps a concept one cannot test whether another treatment can help the participant grasp it again.

Furthermore, we did not design our experiment to reproduce the full experience of installing an application, let alone do so in an ecologically valid way. Rather, we only examined participants' general capacity to absorb and search information in a laboratory context. Unlike real-world users, our participants were briefed on the purpose of application authorization dialogs before they encountered them in our experiment. Users in real-world situations may not pay any attention to an authorization disclosure; they will not have seen hundreds of instances of authorizations presented within a short block of time as our participants did; they may be more focused due to the potential of real harm from rogue applications; or they may be less focused as a result of a desire to complete the installation process and start using an application.

The scope of our study prevented us from comparing the efficacy of installation-time authorizations to those that appear only after applications require access to previously unauthorized resources. Nor could we compare the efficacy of consent approaches that disclose authorization policy to approaches such as presenting third-party evaluations of these policies or other reputation data.

7.2 Guidance

While we found that designs of the same type performed better when grouped by action, no single design appeared significantly easier to absorb than the others. Representations in which resources were represented as icons could be searched more quickly, though this may not be the case in real-world situations in which users are not shown the icon representing the concept they are to seek immediately before searching for it. While the lack of a clear winner amongst the designs limits the design guidance we can infer from our results, it does give researchers a great deal of leeway in designing in-the-wild experiments—it would be hard to argue that one would be harming users by exposing them to any one of these designs instead of another.

Regardless of actual performance, we found an overwhelming majority of participants perceived absorption and search tasks to be easier when resources are presented graphically, especially when using icons. An overwhelming majority also expressed a preference for graphical designs in real-world authorization disclosures, again favoring icons. It is possible that, given the option, participants would have also liked to have seen actions or other information represented graphically. It is possible that these perceptions and preferences may be true of graphical presentations of other information. Given the similarity of performance among designs, the most useful guidance may be to select designs that users *perceive* to be the easiest or most effective and prefer; the likelihood that they will read a disclosure will presumably increase as the perceived effort decreases. Future work may well find that, regardless of how well authorization disclosures are represented, the greatest challenge is to convince users that reading them is worth their effort.

8. CONCLUSION

We have provided evidence of a growing trend among application platforms to disclose, via application installation consent dialogs, the resources and actions that applications will be authorized to perform if installed. To improve the design of these disclosures, we have have taken an important first step of testing key design elements. We hope these findings will assist future researchers in creating experiences that leave users feeling better informed and more confident in their installation decisions.

Within the admittedly constrained context of our laboratory study, disclosure design had surprisingly little effect on participants' ability to absorb and search information. However, the great majority of participants preferred designs that used images or icons to represent resources. This great majority of participants also disliked designs that used paragraphs, the central design element of Facebook's disclosures, and outlines, the central design element of Android's disclosures.

9. REFERENCES

[1] Android development team. Manifest.permission, July 2009. http://developer.android.com/reference/android/Manifest.permission.html.

[2] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, San Francisco, CA, January 1999.

[3] C. Cowan, S. Beattie, C. Pu, P. Wagle, and V. Gligor. SubDomain: Parsimonious security for server appliances. In *In Proceedings of the 14th USENIX System Administration Conference (LISA 2000)*, 2000.

[4] L. F. Cranor, P. Guduru, and M. Arjula. User interfaces for privacy agents. *ACM Transactions on Computer-Human Interaction*, 13(2):135–178, June 2006.

[5] G. DeSanctis. Computer graphics as decision aids: Directions for research. *Decision Sciences*, 15(4):463–487, 1984.

[6] Facebook Press Release. Facebook announces privacy improvements in response to recommendations by canadian privacy commissioner. http://www.facebook.com/press/releases.php?p=118816, Aug. 27 2009.

[7] I. Goldberg, D. Wagner, R. Thomas, and E. A. Brewer. A secure environment for untrusted helper applications: Confining the wily hacker. In *In Proceedings of the 6th Usenix Security Symposium*, 1996.

[8] N. S. Good, J. Grossklags, D. K. Mulligan, and J. A. Konstan. Noticing notice: a large-scale experiment on the timing of software license agreements. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 607–616, New York, NY, USA, 2007. ACM.

[9] J. Gosling and H. McGilton. The java language environment: A white paper. Sun Microsystems, May 1995, http://java.sun.com/docs/white/langenv/.

[10] P. Kelley, L. Cesca, J. Bresee, and L. Cranor. Standardizing privacy notices: An online study of the nutrition label approach. In *CHI '10: Proceedings of the SIGCHI conference on Human factors in computing systems*, 2010.

[11] P. G. Kelley, J. Bresee, L. F. Cranor, and R. W. Reeder. A "nutrition label" for privacy. In *SOUPS '09: Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–12, New York, NY, USA, July 2009. ACM.

[12] D. N. Kleinmuntz and D. A. Schkade. Information displays and decision processes. *Psychological Science*, 4(4):221–227, 1993.

[13] I. Krstić and S. L. Garfinkel. Bitfrost: The one laptop per child security model. In *SOUPS '07: Proceedings of the Third Symposium on Usable Privacy and Security*, pages 132–142, New York, NY, USA, 2007. ACM.

[14] G. L. Lohse. The role of working memory on graphical information processing. *Behaviour & Information Technology*, 16(6):297–308, Nov-Dec 1997.

[15] G. A. Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.

[16] E. Mills. Rogue facebook apps steal log-in data, send spam, Aug. 2009. http://news.cnet.com/8301-27080_3-10313618-245.html. Accessed on September 24,2009.

[17] O. Project. Oauth. http://oauth.net/.

[18] N. Provos. Improving host security with system call policies. In *In Proceedings of the 12th Usenix Security Symposium*, Aug. 4–8 2003.

[19] Report by Kleimann Communication Group for the FTC. Evolution of a prototype financial privacy notice, 2006. Available at http://www.ftc.gov/privacy/privacyinitiatives/ftcfinalreport060228.pdf. Accessed on September 11, 2009.

[20] E. Sachs. Invited presentation to the 2009 Symposium On Usable Privacy and Security (SOUPS 2009). http://docs.google.com/Present?docid=ajkhp5hpp3tt_63gr8gsvhq&skipauth=true, July 16 2009.

[21] C. Speier. The influence of information presentation formats on complex task decision-making performance. *Int. J. Hum.-Comput. Stud.*, 64(11):1115–1131, 2006.

[22] I. Vessey. Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences*, 22(2):219–240, 1991.

[23] R. E. Wood. Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes*, 37(1):60–82, February 1986.