Effective Risk Communication for Android Apps

Christopher S. Gates, Jing Chen, Ninghui Li, *Senior Member, IEEE*, and Robert W. Proctor

Abstract—The popularity and advanced functionality of mobile devices has made them attractive targets for malicious and intrusive applications (apps). Although strong security measures are in place for most mobile systems, the area where these systems often fail is the reliance on the user to make decisions that impact the security of a device. As our prime example, Android relies on users to understand the permissions that an app is requesting and to base the installation decision on the list of permissions. Previous research has shown that this reliance on users is ineffective, as most users do not understand or consider the permission information. We propose a solution that leverages a method to assign a risk score to each app and display a summary of that information to users. Results from four experiments are reported in which we examine the effects of introducing summary risk information and how best to convey such information to a user. Our results show that the inclusion of risk-score information has significant positive effects in the selection process and can also lead to more curiosity about security-related information.

Index Terms—Risk communication, usability, mobile security

1 Introduction

In recent years smart mobile devices have become pervasive. More than 50 percent of all mobile phones are now smartphones,[1] and this statistic does not account for other devices such as tablet computers that are running similar mobile operating systems. According to Google, more than 400 million Android devices were activated in 2012 alone. Android devices have widespread adoption for both personal and business use. From children to the elderly, novices to experts, and in many different cultures around the world, there is a varied user base for mobile devices.

The ubiquitous usage of these mobile devices poses new privacy and security threats. Our entire digital lives are often stored on the devices, which contain contact lists, email messages, passwords, and access to files stored locally and in the cloud. Possible access to this personal information by unauthorized parties puts users at risk, and this is not where the risks end. These devices include many sensors and are nearly always with us, providing deep insights into not only our digital lives but also our physical lives. The GPS unit can tell exactly where you are, while the microphone can record audio, and the camera can record images. Additionally, mobile devices are often linked directly to some monetary risks, via SMS messages, phone calls, and data plans, which can impact a user's monthly bill, or increasingly, as a means to authenticate to a bank or directly link to a financial account through a 'digital wallet'. This access means that any application (or app) that is allowed to run on the devices potentially has the ability to tap into certain aspects of the information. In the benign case the access is performed to provide useful functionalities, but in other scenarios it may be used to collect a significant amount of personal information and even as a means to have some adverse impact on a user. Furthermore, the line between benign and malicious is often fuzzy, with many apps falling into a gray area where they may be overly invasive but not outright malicious.

Compared to desktop and laptop computers, mobile devices have a different paradigm for installing new applications. For computers, a typical user installs relatively few applications, most of which are from reputable vendors, with niche applications increasingly being replaced by web-based or cloud services. In contrast, for mobile devices, a person often downloads and uses many apps from multiple unknown vendors, with each app providing some limited functionality. Additionally, all of these unknown vendors typically submit their apps to a single or several app stores where many other apps from other vendors may provide similar functionality. This different paradigm requires a different approach to deal with the risks of mobile devices, and offers distinct opportunities.

The present research focuses on the Android platform, because of its openness, its popularity, and the way in which Android handles access to sensitive resources. In Android an app must request a specific permission to be allowed access to a given resource. Android warns the user about permissions that an app requires before it is installed, with the expectation that the user will make an informed decision. The effectiveness of such a defense depends to a large degree on choices made by the users. Indeed whether an app is considered too invasive or not may depend on the user's privacy preference. Therefore, an important aspect of security on mobile devices is to communicate the risk of installing an app to users, and to

1. http://www.nielsen.com/us/en/newswire/2012/smartphones-account-for-half-of-all-mobile-phones-dominate-new-phone-purchases-in-the-us.html.

- *C. S. Gates and N. Li are with the Center for Education and Research in Information Assurance and Security and the Department of Computer Science, Purdue University, West Lafayette, IN. E-mail: {gates2, ninghui}@cs.purdue.edu.*
- *J. Chen and R. W. Proctor are with the Center for Education and Research in Information Assurance and Security and the Department of Psychological Sciences, Purdue University, West Lafayette, IN. E-mail: {chen548, rproctor}@purdue.edu.*

Manuscript received 17 July 2013; revised 21 Nov. 2013; accepted 1 Dec. 2013; date of publication 15 Dec. 2013; date of current version 14 May 2014. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TDSC.2013.58

help them make a good decision about whether to install a given app.

Android's current risk communication mechanism has been shown to be of limited effectiveness. Studies have demonstrated that users tend to ignore the permissions that an app requests [4], [12], [13], [18], and some recent work has attempted to overcome some of these limitations. Felt et al. [13] proposed several improvements, including: modifying permission category headers, emphasizing risk, reducing the number of permissions, enabling customized permission lists, incorporating user reviews and rethinking the timing of when and how permissions are granted. Lin et al. [21] proposed an approach which incorporates crowdsourced (via Amazon Mechanical Turk) expectations of which permissions are considered reasonable, and presents these expectations on the permission page (e.g., "95 percent of users were surprised this app sent their approximate location to mobile ad provider."). Kelley et al. [19] introduced a concept of "privacy facts" which conveys at a high level the types of information an app has access to (e.g., personal information, contacts, location, etc.), and proposed that these facts be displayed these facts on the app's main description page. We consider an alternative approach to this problem which aims to minimize the space that is used to present information while helping a user make installation decisions with a better understanding of the security and privacy implications.

We propose the addition of a summary risk rating for each app. A summary risk rating enables easy risk comparisons among apps that provide similar functionalities. We believe that one reason why current permission information is often ignored by users is that it is presented in a "standalone" fashion and in a way that requires a lot of technical knowledge and time to distill useful information, making comparison across apps difficult. An important feature of the mobile app ecosystem is that users often have choices and alternatives when choosing a mobile app. If a user knows that one app is significantly riskier than another but provides the same or similar functionality, then this fact may cause the user to choose the less risky one. This will in turn provide incentives for developers to better follow the least-privilege principle and request only necessary permissions. Peng et al. [25] presented one possible method for generating a principled metric to rank an app's risk based on the set of permissions it requests. The method can rank the risk of any Android app among all apps available in Google Play, Google's online market for Android apps. Such a risk ranking can be translated into categorical values such as very low, low, medium, and high risk, to provide a summary risk rating.

A summary risk rating also enables proactive risk communication (e.g., when the user searches for apps) so that users can take this information into the decision process. This is in contrast to the current reactive approach, where often times the user sees the permission/risk information of an app as a final warning only after the user has made the decision to choose the app.

Our hypothesis is that when a summary risk rating is presented in a user-friendly fashion, it will encourage users to choose apps with lower risk. In this work, we tested the hypothesis experimentally. Additionally, we explored how to communicate this risk information to an average user effectively and efficiently.

Four experiments were conducted. Experiment 1 was run via Amazon Mechanical Turk (MTurk), with risk rating presented as text. It experimentally tested the extent to which having a risk category affects users' choices. The behavior of users when they were provided with the summary risk was compared to the case when there was no summary risk information. The results confirmed the hypothesis that providing risk summary causes users to prefer apps with lower risk.

Experiments 2 and 3 were designed to investigate how to most effectively communicate this risk information. Instead of presenting the risk category in text, it was presented using symbols, similar to the way in which user ratings of an app are presented using one to five stars. Symbolic depictions in the form of pictures relating to technological and natural hazards have been found to produce higher perceived risk than written text does [36]. When using symbols, a key decision is whether to frame the categories as varying along the dimension of risk, where more symbols represent greater risk, or the dimension of safety, where more symbols represent greater safety. Because user ratings typically are presented as "the more stars the better", we hypothesized that framing the dimension as safety would be more advantageous, in part due to its compatibility with the user-rating dimension. Experiment 2 directly compared ease of processing the rankings in terms of risk or safety with an in-lab choice-reaction task for which response time and accuracy were measured. Experiment 3 was similar but also included user ratings in the display, because risk information would be presented together with user ratings in the naturalistic app selection context.

Experiment 4 returned to the MTurk environment to determine whether symbolic presentation of risk influenced choices in a naturalistic environment much the same way that verbal descriptions did in Experiment 1. It was conducted like Experiment 1 but varying whether risk scores or safety scores were presented to the participants. The results of the four experiments illustrate the value of presenting summary risk information and suggest that there may be an advantage to framing it in terms of safety.

2 Related Work

2.1 Security and Usability

Information security and privacy are issues for users of all types of electronic devices. With regard to smart phones, users are more concerned with privacy on their phones than on computers, and they especially worry about the threat of malicious apps [4]. However, although people are shown the permissions an app requests before it is installed, they do not understand them well [13], [18]. Among the recommendations made by Chin et al. [4] was to provide "new security indicators in smartphone application markets to increase user trust in their selection of applications" (p. 2). The addition of new security indicators not only may decrease the frequency of risky user behaviors, but it may also facilitate the use of smart phones for online transactions by more individuals. Staddon et al. [28] found that users'

engagement and perception of privacy are strongly associated, and people spend more time in social networks when they are less concerned about their privacy. This relation may be true as well for app installation.

People will not use security features properly if they fail to understand the purpose of the features or the information on which their decisions should be based. The security features also will not be used if the users find the features intrusive or too difficult to master. Therefore, interactions between users and the systems need to be simple and user-friendly [26]. Despite this need, studies of various security and privacy measures have shown their usability is typically deficient [2], which often leads to user resistance [30], [34]. Studies have also demonstrated that usability can be improved by systematically studying the human information-processing requirements associated with effective use of the measures and incorporating the resulting knowledge into the designs [16], [34], [35].

Usability of security mechanisms has been studied in contexts other than mobile platforms. Biddle et al. [3] laid out some general ground rules concerning the content of security dialogs; e.g., avoid unfamiliar terms, lengthy messages and misleading or confusing wordings. Schwarz and Morris [27] proposed that web search results be augmented with indicators for helping people assess the degree of trustworthiness of web sources. They found that adding such information to search results is useful, but less so when the information is added to web pages, presumably because the content, look, and feel of the page dominate the user's judgment. Cranor et al. [6] developed Privacy Bird specifically with the intent to signal to users whether web sites match their privacy preferences. It provides a red bird icon when visiting a site if the privacy policy does not match the user's preferences and a green bird icon if it does match. They extended this idea to web searches with Privacy Finder, which provides similar information when a search engine returns the results of a query [9]. Studies have found the summary privacy information provided by Privacy Bird (and Privacy Finder) to be effective at improving participants privacy practices [5], [33]. Egelman et al. [10] directly examined the influence of privacy indicators, which showed privacy ratings of online vendors from low to high as one to four green boxes in a row of four that were green), on Internet users' browsing and purchasing decisions. When the privacy indicators were presented alongside the search results, participants who chose to visit only a single website paid more money for a higher level of privacy. However, when this information was provided after a website had been selected, participants did not alter their initial decision to purchase from a cheaper website with lower level of privacy. Finally, Kim et al. [20] proposed the Online Trust Oracle approach for communicating information regarding programs for Windows desktop environment; the interface lists information regarding why a file may be harmful on the left side of the dialogue and why a file may be safe on the right side of the dialog, and it also uses three colors to distinguish programs of different degrees of risk. To summarize, these studies all suggest that presenting high-level summary risk information will be beneficial, particularly if it is displayed early in the selection process.

2.2 Risk Perception and Decision Making

Users make many decisions that affect the overall state of security of any system with which they interact. For security and privacy, most of these decisions relate to the risk to which the individual or system is exposed. Consequently, improving security decisions by users involves taking into consideration factors that influence a user's *risk perception* and *decision making* [8]. Also relevant is *risk communication*, which refers to conveying risk to users in a way that allows accurate risk perception and, hopefully, better choices of actions with regard to the actual risks involved [32]. One factor that has been shown to be critical in risky decisions is the way in which losses and gains are framed. With the exact same scenario, the way in which the information is presented can significantly influence the decision-maker's choice. People are risk-averse when the framing highlights positive outcomes, but risk-seeking when it highlights negative outcomes [23], [31].

It has become customary to conceive of risk-perception judgments and decision making as relying on two distinct modes of thought, or systems: System 1 is automatic and intuitive, and operates outside of awareness, whereas System 2 requires attention and is slower and more logical than System 1 [17]. Because System 2 processing is effortful and time consuming, judgments of uncertainty (including risk perception) and choices among alternatives often rely on the intuitive impressions provided by System 1. Consequently, they violate the rules of probability and normative theories. Because System 1 is automatic, rapid responses are influenced more heavily by it, whereas System 2 processes contribute to more deliberate judgments [14]. Therefore, framing effects would be more apparent under the conditions in which people respond quickly.

3 Motivation

3.1 Overview of Android Security

The Android system's in-place defense against malware consists of two parts: *sandboxing* each app, and *warning* the user about the permissions that the app is requesting. Specifically, each app runs with a separate user ID, as a separate process in a virtual machine of its own, and by default does not have the ability to carry out actions or access resources which might have an adverse effect on the system or on other apps without requesting permission to do so from the user.

The permissions consist of capabilities that an app may require such as accessing geo-location information, sending text messages, receiving text messages, and many more. In total there are around 130 unique permissions in Android depending on the version. Each permission has a name, category, and a high level description of what it allows. An example is the "FULL NETWORK ACCESS" permission in the "NETWORK COMMUNICATION" category with its description as "*Allows the app to create network sockets and use custom network protocols. The browser and other apps provide means to send data to the internet, so this permission is not required to send data to the internet.*"

The risk communication mechanism for permissions relies on the assumption that a user understands and makes an informed decision when presented with a list of permissions requested by an app. For most permissions, risks must be inferred because they are not explicitly stated in the description [13]. When browsing a specific app from the Google Play website, a user is able to see details about the app via a series of tabs at the top of the page. In addition to an overview, user reviews, and 'what's new' section, one of these tabs presents the permission information. When an app has been selected for installation, permissions are displayed before the user confirms installation. When app installation is performed directly on the device, there is a Play Store app which allows users to find and install new apps. The options and information are the same as on the website, with the primary difference being that the screen may be smaller and so when information is displayed, including permissions, a user has to make more of an effort to view that information.

3.2 Risk Communication in Android

Studies have shown that Android users tend to ignore the permissions that an app requests [4], [12], [18], and there are many reasons for ignoring them. Permission descriptions are seen as confusing or difficult to understand by many users [18]. Furthermore, nearly all apps request permissions with some associated risk. Felt et al. [12] analyzed 100 paid and 856 free Android apps, and found that "*Nearly all applications (93% of free and 82% of paid) ask for at least one 'Dangerous' permission, which indicates that users are accustomed to installing applications with Dangerous permissions. The INTERNET permission is so widely requested that users cannot consider its warning anomalous. Security guidelines or anti-virus programs that warn against installing applications with access to both the Internet and personal information are likely to fail because almost all applications with personal information also have INTERNET*"(p. 6). The implication is that since most apps are considered to be benign, and users see very similar warning information for all apps, the users generally ignore the warnings.

Unless a user is highly concerned with security and privacy, and regularly examines the permissions as part of her app selection process, then most likely she has already made the decision to install the app before being presented with the permission information. In Android, a user is able to install the app by clicking a button to 'install' or 'buy' the app. Only then is the user forced to view the permissions that the app is requesting in a final confirmation screen. However, by this point the user has already made the decision to install the app, and this extra warning is often seen as a nuisance and ignored.

There is a parallel between Android's permission warning and Windows' User Account Control (UAC). Both are designed to inform the user of some potentially harmful action that may occur. In UAC's case, this happens when a process is trying to elevate its privileges in some way, and in Android's case, this happens when a user is about to install an app that will have all the requested permissions.

Recent research suggests the ineffectiveness of UAC in enforcing security. Motiee et al. [24] reported that 69 percent of their survey participants ignored the UAC dialog and proceeded directly to use the administrator account. Microsoft itself concedes that about 90 percent of the prompts are answered as "yes", suggesting that "users are responding out of habit due to the large number of prompts rather than focusing on the critical prompts and making confident decisions" [11].

According to Fathi [11], in the first several months after Vista was available for use, people were experiencing a UAC prompt in 50 percent of their "sessions" - a session is everything that happens from logon to logoff or within 24 hours. With Vista SP1, and over time, this number has been reduced to about 30 percent of the sessions. This reduction suggests that UAC has been effective in incentivizing software developers to write programs without elevated privileges unless necessary. The difference between Android and UAC is that UAC encourages the developer to work with fewer privileges since this will lead to a smoother user experience. However, with Android there is no obvious feedback loop to the developer at this point other than the fact that a small fraction of the user reviews may complain about an app being over-privileged.

An effective risk communication approach for Android could provide an incentive for developers to reduce the number of permissions requested by apps, similar to UAC's impact with Windows software developers. By highlighting requested permissions of apps, such risk communication could potentially change user behavior and drive consumption to apps with fewer permissions, thereby creating a feedback loop to developers and having a positive effect on the app ecosystem.

Currently, we are seeing the opposite trend regarding permissions requests. We collected two sets of app data [25], one in February of 2011 containing 155,000 apps and one in February of 2012 containing 325,000 apps. They were collected by crawling as much of the Google Play app store as was discoverable at the time, starting with a seed of the top applications from all categories and branching out based on related apps found on each page. Both data sets were captured using the same methodology, and thus represent comparable snapshots of Google Play at one year apart. We separated out the apps that were common to both datasets by both their name and permission requests, which we call Overlap. We then removed those in Overlap from 2011 and 2012, respectively, obtaining three data sets. Fig. 1 shows the percentage of apps having a certain number of permissions in each set. Apps in Overlap requested the least number of permissions on average, and those in 2012 without Overlap requested the highest number of permissions on average.

3.3 Risk Score

Previous research proposed one possible scoring mechanism grounded in machine learning and based off of permission requests. That work leverages probabilistic generative models to create a principled way in which to evaluate the risk of an app. Using these models, some parameterized random process is assumed to generate the app data and learn the model parameters based on the data. Then, it is possible to compute the probability that each app was generated by the model. The risk score can be any

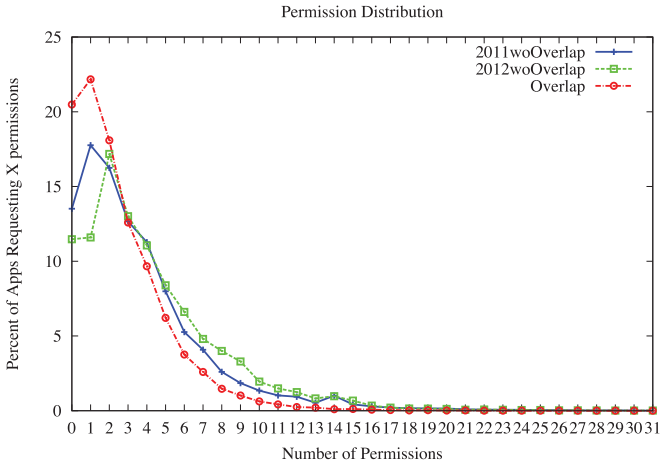


Fig. 1. Percent of apps that request X permissions. Overlap is the data that is common to both the 2011 and 2012 datasets. 2011woOverlap is a data set collected in February of 2011 with the Overlap removed. 2012woOverlap was collected in February of 2012 with the Overlap removed.

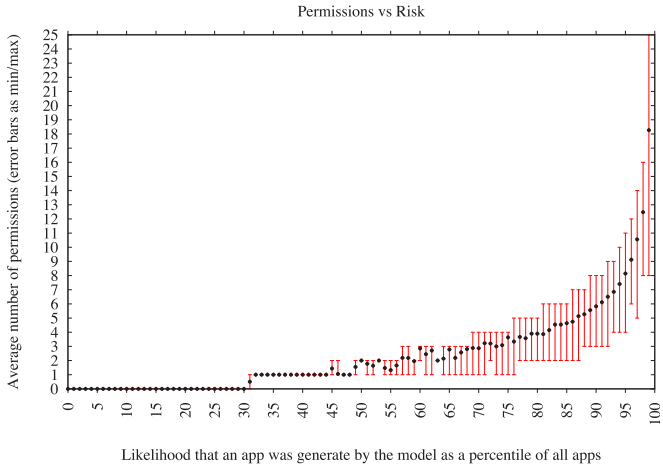


Fig. 2. Average number of permissions for every 1 percent division of apps, sorted in descending order on the basis of likelihood given by the PNB model. The points represent the average number of permissions requested, and the error bars indicate the min and max at that percentile.

function that is inversely related to the probability, so that lower probability translates into a higher score. Several different models were analyzed, and Naive Bayes with an Informative Prior [**P**rior **N**aive **B**ayes (PNB)] was the recommended model. This model balances effectiveness, simplicity and monotonicity, to come up with a principled way to generate a relative risk score among all apps.

Fig. 2 shows the PNB model that was trained on a large set of data and then used to evaluate a separate set of apps. Apps were sorted relative to one another based on their likelihood, and at each percentile we show the min, max, and average number of permissions requested by an app at that percentile. As one moves to the right in the figure, the number of permissions that were requested generally increases, but the the ranges for nearby buckets are overlapping. This is caused by the fact that lower impact permissions may be requested more often, and thus requesting several of these permissions increases the relative risk less than requesting fewer higher-impact permissions. So, while the number of permissions does not strictly increase as the percentile increases, this method does provide the monotonic property in that removing a permission will always make an app less risky.

One possible way to assign one of four risk category values to each app is as follows. The first 60 percent of apps are assigned "low". From Fig. 2, we can see that this includes the 30 percent of apps that requested no permission, approximately 15 percent that requested one permission, and some apps that requested two to three permissions. Apps between 60 percent and 85 percent are assigned "med". Apps between 85 percent and 95 percent are assigned "high"; and apps above 95 percent are assigned "very high". Apps in the 99th percentile requested on average about 18 permissions.

We point out that the concept of risk is a fuzzy one, just as the concept of an app being "malicious" is. There are clear examples of malware, such as banking trojans that intercept text messages, but there are also less clear examples, such as overly invasive ad networks that some users may consider malicious while other users may not. Similarly, it is impossible to come up with a measurement of risk that everyone agrees, and thus there cannot exist a "perfect" metric for risk. Even though a perfect mechanism for computing risk score is lacking, this lack should not prevent one from trying to communicate risk or risk-related information, just as the list of permissions being imperfect does not mean that it should not be presented to users. We see two ways to interpret risk scores. The first is to view a score purely as a summary of the permissions requested by an app, as in [25]. Adding such a score enhances what the current permission-list tries to communicate. The second is to use a risk score to summarize permissions as well as other information such as source code, user comments, and the content of privacy policies, in order to better capture apps that have malicious or problematic activities. The risk communication work in this paper is applicable in both interpretations of risk scores.

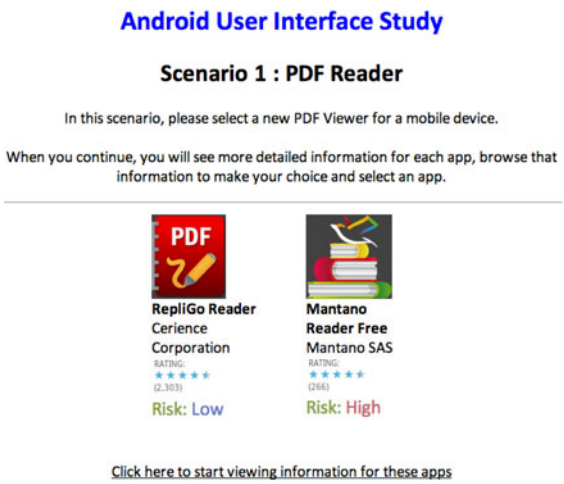
4 Experiment 1: Adding a Risk Metric

The first experiment was designed to identify how useful high-level risk information may be for the app selection process. This was accomplished by having participants select between two alternative apps. We compared their choices and subjective ratings when summary risk information was provided and when it was not.

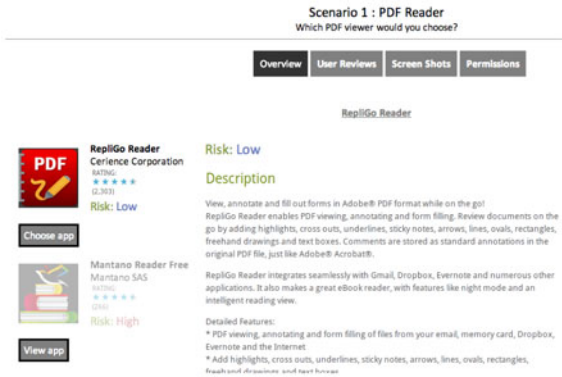
4.1 Method

Two hundred participants were recruited for an online app selection experiment using MTurk. On average the experiment took 12 minutes to complete, and participants' were paid $1.00 each for their efforts. The experiment received approval through Purdue University's IRB process.

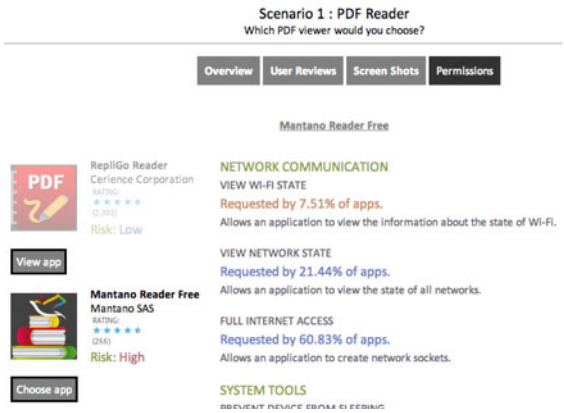
We compared the participants' choices under two interface designs. One, the 'Standard Interface', approximates the official Google Play Store. With the Standard Interface, there are two primary ways to view information about an app. The 'summary view' presents the app's name, publisher, icon and average user rating. Once the app is clicked,



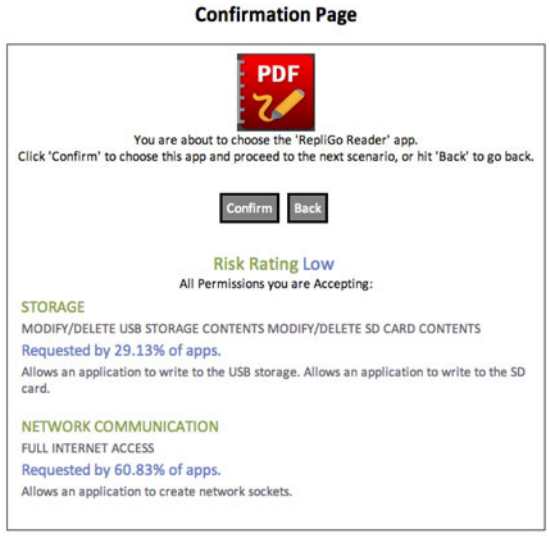
(a) Task Introduction



(b) The Main View when choosing an app



(c) Showing Permissions for the higher risk app



(d) Showing the Confirmation Page when selecting the low risk app

Fig. 3. Selection of Risk Interfaces for use in describing Experiment 1.

then there is a second view with more details available. We call this the 'main view', and it contains a description written by the developer and three user reviews. In the main view there are tabs across the top of the page that can be clicked to browse other information about the app, including a full list of user reviews, screenshots, and permissions.

The second design, the 'Risk Interface', contains all the information from the Standard Interface plus additional risk information in both the summary view and the main view. For this experiment the risk information was presented as a text value in the range [low, medium, high, very high] with color differences to add extra emphasis ranging from [blue, brown, light red, bright red]. This risk information was presented at the summary view (Fig. 3a) for the app, below the average user rating, as well as above the developer description on the main view (Fig. 3b). In addition to the high-level risk summary, a more detailed level of information was introduced whenever the permissions were shown, on the permission page (Fig. 3c) as well as the confirmation page (Fig. 3d). Each permission showed the percentage of apps within that category that request that permission, and presented the rare and critical permissions toward the top so that they were seen more easily. For more advanced users, the permission information could be used to learn whether or not these permissions are common, and possibly further refine their decision.

Each participant progressed through six tasks, in each of which two different apps that provide the same general functionality were presented, and the participant was asked to select one (see Fig. 3a). When the participant began a task, one of the two apps at random was presented first. When the participant had viewed information about both apps and made the choice, then the final confirmation page was presented, similar to Fig. 3d, which listed the permissions in a manner similar to Google Play Store. The participant could confirm the selection or go back to the browsing mode to view more information about each app.

Participants were randomly assigned to two groups. Group A performed three tasks using the Standard Interface and then three tasks using the Risk Interface. Group B

TABLE 1
The Percent of Participants Who Chose the Lower Risk App under Different Interface Type (Standard versus Risk)

Apps \ Interface	Standard	Risk	Chi-square Test
Task1_Low/High	59.1%	82%	$\chi^2 = 12.226$, $p < .001$
Task2_Low/Low	51.6%	52%	$\chi^2 = 0.003$, $p = .957$
Task3_Low/Med	60.2%	68%	$\chi^2 = 1.271$, $p = .260$
Task4_Low/High	56%	76.3%	$\chi^2 = 8.863$, $p = .003$
Task5_Med/Med	55%	49.4%	$\chi^2 = 0.592$, $p = .442$
Task6_Med/VHigh	50%	82%	$\chi^2 = 12.226$, $p < .001$
Overall	56.2%	77.2%	$\chi^2 = 38.269$, $p < .001$

When apps have same risk, we designate the 'first' app as the lower risk app. χ^2 (1, N = 193) for all task chi-square tests, χ^2 (1, N = 772) for overall.

performed three tasks using the Risk Interface and then three tasks using the Standard Interface.

To control for other factors in the decision-making process, such as effects from the descriptions, icons, user reviews, familiarity with apps, screenshots, and other features that might skew the results, we alternated which app was assigned which set of permissions within each task. So, one participant would see app1 as a higher risk app for a specific task, and another participant would see everything the same except that app2 would be the higher risk app for the same task. We were actually recording the results for which permission set was selected, and not which app was selected, although for the rest of this paper we will present it as app selection because it is easier to refer to selecting an app than a permission set.

Throughout the experiment participants completed four surveys. An initial survey collected some high-level demographic information as well as experience with mobile devices. After they completed the three tasks from one interface, they are presented a survey related to their decision-making process. A similar survey was answered after the three tasks with the other interface. There was also a final survey which collected some information after they had seen both interfaces.

We collected the apps that were selected for each task, all answers to the survey questions, as well as timing and click data so that we could analyze the participant's browsing behavior.

4.2 Results and Discussion

4.2.1 Demographics

Seven participants did not complete all the task, and 193 valid participants were included in the following analyses. 120 participants were male, and 73 were female. 37 were of age 18-22 years, 74 between 22-30, 51 between 30-40, 22 between 40-50, and 9 were 51 years and above. The majority (88.1 percent) had used an Android device, with 79.3 percent of the participants having used such a device for more than 3 months. 85.5 percent of the participants indicated that they install a new app on their mobile devices at least once a month. Regarding their computer security expertise, 4.1 percent were computer novices, 71.5 percent of the participants were regular users, 22.8 percent were highly skilled engineers, system administrators, etc., and 1.6 percent were security experts.

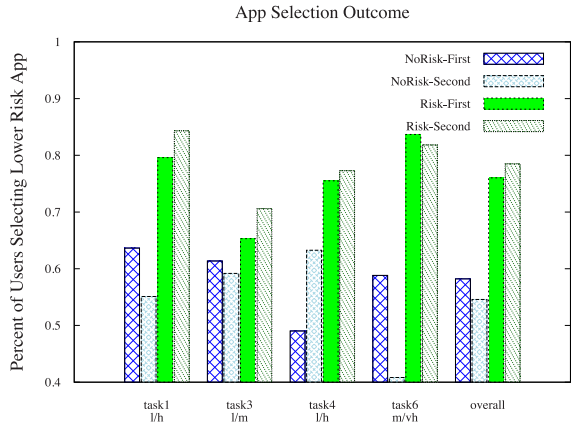


Fig. 4. Percent of participants who chose the lower risk app for the tasks where one app was lower risk.

4.2.2 App Choice Analysis

The main purpose of this analysis was to examine whether the presence of a risk score influenced participants' app-install decision making. Thus, we compared, for the same pair of apps, whether there was any difference when the Standard Interface was used versus when the Risk Interface was used. Chi-square tests were conducted for each pair of apps (see Table 1). A significant result indicates that participants chose the less-risky app more often with the Risk Interface than with the Standard Interface. First, for those pairs with the same risk scores, i.e., Task 2 (Low versus Low) and Task 5 (Med versus Med), there was no difference between the two interfaces. This finding validates that the other factors besides the risk score were controlled well in our experiment, so that participants made similar choices when there were no risk scores and when the presented risk scores were equal. Second, for the four tasks where the two apps had unequal risk scores, there was a pronounced preference for the lower-risk app with the Risk Interface than with the Standard Interface. The difference was statistically significant ($p < 0.001$) for Tasks 1 (Low versus High), 4 (Low versus High), and 6 (Med versus Very High), but not for Task 3 (Low versus Med).

Fig. 4 shows the breakdown of participant choices for these four tasks. Recall that about half of the participants saw the Standard Interface and then the Risk Interface, whereas the other half saw the Risk Interface and then the Standard Interface. Therefore, a task may be presented in four cases: Standard-First, Standard-Second, Risk-First, and Risk-Second. As can be seen from the "overall" columns on the right side of Fig. 4, on average about 56 percent of participants chose the low-risk app without risk information, and about 77 percent chose the lower-risk app with risk information. The results show that the risk scores have a significant impact on participants' app selections, causing them to choose lower-risk apps more often.

Since 11.9 percent of the participants reported never having used Android, we performed separate analyses for the users who had used Android and those who had not. The resulting χ^2 values for both groups showed similar patterns to that of the overall analysis.

TABLE 2
Numbers of Participants Who Specified a Certain Factor as the Most Important One (Left) and Who Considered the Factors When Choosing from the Apps (Right)

	Most Important Factor		Considered Factors	
Factor	Risk	Standard	Risk	Standard
Risk Level	78	N/A	154	N/A
User Ratings	43	70	166	180
User Reviews	32	55	122	136
Descriptions	18	32	87	118
Permissions	15	19	66	58
Screen Shots	4	14	58	78

TABLE 3
Tasks with No Risk Information: Numbers of Participants Who Specified a Certain Factor as the Most Important One (Left) and Who Considered the Factors When Choosing from the Apps (Right)

	Most Important Factor		Considered Factors	
Factor	No Risk First	No Risk Second	No Risk First	No Risk Second
User Ratings	35	35	91	89
User Reviews	28	27	73	63
Descriptions	20	12	66	52
Permissions	2	17	21	37
Screen Shots	8	6	47	31

4.2.3 Questionnaire Analysis

Consistent with the app choice data, the self-report data on the two questions "What did you consider when choosing from the apps?" and "Which one was the most important factor?" also showed that the participants considered risk scores and even regarded risk as the most important factor (see Table 2). When the risk scores were presented with the Risk Interface, the majority (79.8 percent) of participants took risk scores into consideration, and 41.1 percent of the participants took risk scores as the most important factor. This latter value was almost twice as large as the second largest group (22.6 percent), which took the user ratings as the most important factor. We also found that presenting the risk scores led the participants to think that the permission information was more useful. On a 7-point Likert scale, with 1 denoting not useful and 7 extremely useful, 35.2 percent participants gave a rating of 5 and above for the Standard Interface, whereas 45.1 percent participants did for the Risk Interface.

At the end of the session, after the participants had finished the three tasks with each of the two interfaces, we also asked them one question "Which interface did you prefer?" 79.3 percent of the participants preferred the interface with risk information, 12.4 percent thought both interfaces were about the same, and 8.3 percent preferred the interface without the risk information. Although the stated preference for the interface with risk information could be due to its novelty [7], the entire pattern of questionnaire data suggests that the participants were expressing a true desire for risk information to be included.

4.2.4 Correlation between App Selection and Self-Report

Each participant performed three tasks with the Risk Interface and three with the Standard Interface. For half of the participants, tasks 1, 2, and 3 were performed with the Risk Interface and for the other half tasks 4, 5, 6 were. Two out of the three tasks in each set had unequal risk levels (e.g., low versus medium), so the number of tasks in which participants selected the less-risky app ranged from 0 to 2. If participants selected the less-risky app at chance, the expected value of this app-selection-index would be 1.

We tested the correlation between the index and participants' subjective reports with Spearman's correlation test, separately for participants who performed tasks 1, 2, and 3 with the Risk Interface (Group 1), and those who performed tasks 4, 5, and 6 with the Risk Interface (Group 2). These tests showed a significant positive correlation, r_s (Spearman's correlation coefficient) $= 0.247$ and 0.518, $p < 0.05$ and $p < 0.01$, for Groups 1 and 2, respectively. On average, participants who considered risk selected less-risky apps more often than those who did not consider risk (1.57 versus 1.24 for Group 1; 1.75 versus 0.94 for Group 2). There was also a significant positive correlation between selecting the less-risky app and rating risk as the most important factor, $r_s = 0.430$ and $0.461, ps < 0.01$, for Groups 1 and 2, respectively. On average, participants who rated risk as the most important factor selected less-risky apps more often (1.79 versus 1.33 for Group 1; 1.90 versus 1.37 for Group 2).

4.2.5 Interface Ordering Analysis

We also examined whether the participants' subjective ratings of risk with the Standard Interface were influenced by first performing with the Risk Interface first, especially their ratings of the usefulness of the permissions, which were presented for both interfaces. Therefore, we compared the self-report data with the Standard Interface when it was used first or second (see Table 3). 37.8 percent of the participants took permissions into consideration when using the Standard Interface after using the Risk Interface, compared to 22.1 percent of those who used the Standard Interface first; 17.3 percent of the participants rated permissions as the most important factor when they used the Standard Interface after the Risk Interface, compared to 2.1 percent of those who used the interfaces in the opposite order. Also, when rating the usefulness of the permissions on the scale of 1 (not useful) to 7 (extremely useful), 38.7 percent participants gave a rating of 5 or above when they used the Standard Interface after the Risk Interface, whereas only 31.6 percent participants gave a rating of 5 or above when they used the interfaces in the opposite order. In addition, when asked whether they considered the permissions on the confirm page, 54.7 percent participants said "yes" when they used the Standard Interface first, compared to 60.2 percent participants who used it second.

We recorded all page views and clicks during the experiment, and looked at how often participants were viewing the permission page prior to their final choice. Fig. 5 illustrates participants who saw the risk score for the first three tasks, and then had the risk score removed for the final three

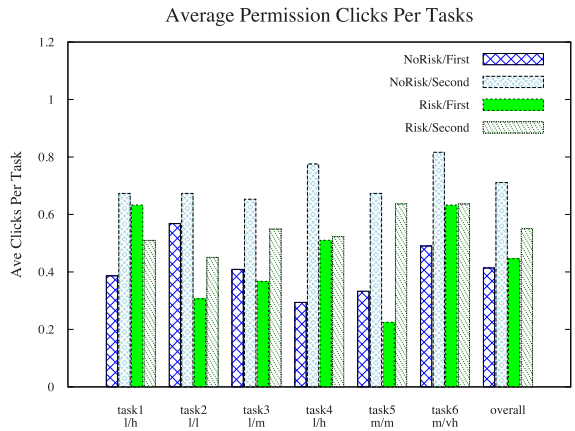


Fig. 5. Average number of clicks on the permissions page per task. Many participants did not click on permissions ever, which puts the average number of clicks per task below 1.

tasks made more clicks on the permission information prior to making a selection. So there is some indication that the risk information has a positive effect on participants' awareness of permissions in practice.

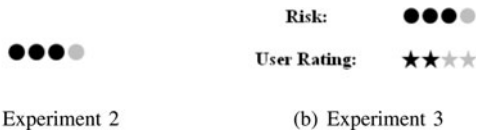
4.2.6 Conclusion

Experiment 1 demonstrates the value of providing summary risk information. When that information was available, participants selected the app labeled less risky more often than they did when the risk information was unavailable. Also, participants subjectively rated risk and permissions as more important when the risk information was included in the interface, and they examined the permission more frequently.

5 Experiment 2: Safety versus Risk

In risk communication, it is insufficient to focus only on the security-relevant information that is communicated. The usability of the security information and how that information is presented to the user is equally important because its effectiveness depends on how the user comprehends and acts on the information. Studies of framing show that people are risk-averse in their decisions when positive outcomes, rather than negative ones, are highlighted [23], [31]. Therefore, in Experiment 2, we presented summary risk information symbolically in terms of number of filled circles, with the scale being one of increasing risk (more filled circles means more risk) for half of the participants and increasing safety (more filled circles means less risk) for the other half.

Because we were interested in people's natural tendencies to react to the risk information, we investigated the difference between the risk and safety conditions in a laboratory experiment that used a speeded, reaction task. Response times obtained under such conditions provide sensitive measures of differences in information-processing efficiency and are widely used in cognitive psychology [22], [29]. In the particular task that was performed, participants were to respond as quickly as possible to the onset of the risk or safety information by making a "yes" or "no" install decision. The time to respond in this task can be assumed to



(a) Experiment 2 (b) Experiment 3

Fig. 6. Stimuli used in Experiments 2 and 3.

reflect mainly System 1 (automatic) processing, and a difference in response time as a function of which scale is used will indicate a framing effect. Specifically, if it is more natural for people to associate larger values with "better", then response times should be shorter on average when the information is presented as amount of safety as opposed to amount of risk.

5.1 Method

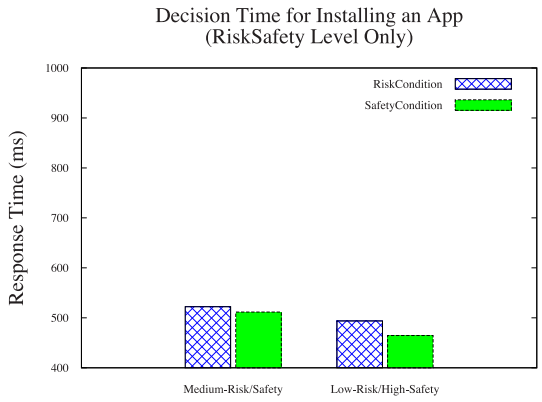
One-hundred and thirty students enrolled in introductory psychology courses in Purdue University took part in this experiment. They received course credits for participation. The participants were invited to our lab and performed the experiment on a computer.

The stimuli were four circles, some of which were black and the rest gray (see Fig. 6a). We used circles rather than more meaningful symbols so that the same symbols would be used for the risk and safety versions. Half of the participants received the "safe" version, and the other half received the "risk" version. A cover story was presented at the beginning to set up the app-install context: "Suppose you are using your smart phone or other electronic device to install an app. You want to check whether the app is safe [or risky] or not, so that it won't do any harm to your device."

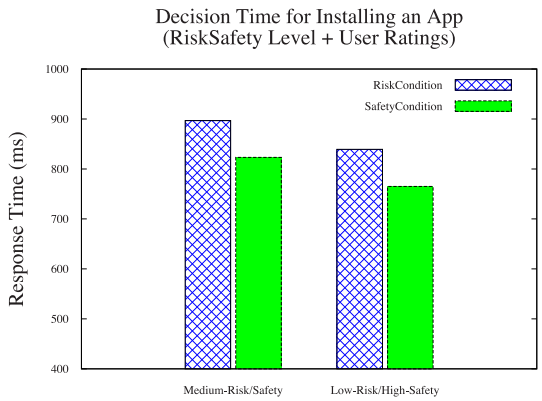
In the "safe" version, participants were told "more black circles mean more safety", with one black circle meaning "the least safety", two meaning "little safety", three meaning "some safety", and four meaning "the most safety". In addition, they were told to install the app only when it had "some safety" or "the most safety" (i.e., three or four filled circles). In the "risk" version, participants were told "more black circles mean more risk", and to install the app only when it has "the least risk" or "little risk" (i.e., one or two filled circles). The word "Correct!" in red was given as visual feedback when the response was correct in accord with the instructions, and "Incorrect!" together with an error tone when it was not.

A Go/Nogo paradigm [15] was used, in which participants were told to press a key for some stimuli ('Go' trial) but not to press any key for other stimuli ('Nogo' trial). This paradigm was adopted to mimic the decision that a participant would make when deciding whether to proceed ('Go') or not ('Nogo') with installing an app. More specifically, participants were asked to press the "ENTER" key if they wanted to install the app and NOT to press any key if they did not want to install it.

At the beginning of each trial, a fixation point was presented for 1 s, following which the four circles were presented for 2 s or until a response was made. Participants were instructed to respond as accurately and as quickly as possible. Each participant performed 16 practice trials for warm-up plus 208 test trials.



(a) Experiment 2



(b) Experiment 3

Fig. 7. Decision time for installing an app in the risk and safety conditions. Medium-risk stands for risk level 2 in the risk condition, and medium-safe stands for safety level 3 in the safety condition; low-risk stands for risk level 1 in the risk condition, and high-safety stands for safety level 4 in the safety condition. Risk level 3, risk level 4, safety level 1, and safety level 2 have no response time, because participants were not supposed to press the key to install an app.

5.2 Results and Discussion

There were 75 female and 55 male participants. Three participants (2 percent) were replaced by new participants because of mean response times three standard deviations longer than the average. Overall, the response errors were rare (< 1 percent), so we will not present the error data.

We compared the response times for the two different ways of communicating risk (see Fig. 7a). The risk two-black-circle trials and the safety three-black-circle trials were the same in terms of the risk/safety level, as were the risk one-black-circle trials and the safety four-black-circle trials. Thus, for the following analysis, the two- and three-circle trials were treated as a close condition and the one- and four-circle trials were treated as a far condition. We conducted a mixed analysis of variance (ANOVA) with distance (close-23 versus far-14) as a within-subject factor and frame (safety versus risk) as a between-subjects factor. Results showed that responses were faster in the safety condition than in the risk condition ($Mean = 488$ versus 513 ms), $F(1, 128) = 4.82,\ p = 0.030,\ \eta_p^2 = 0.04$, and faster in the far condition than in the close condition ($Mean = 482$ versus 520 ms), $F(1, 128) = 289.68, p < 0.001, \eta_p^2 = 0.69$. The interaction was also significant, $F(1, 128) = 15.77, p < 0.001$, $\eta_p^2 = 0.11$, with the difference between close and far trials being larger in the safety condition ($Mean = 512$ versus 465 ms, $t(64) = 14.93,\ p < 0.001$) than in the risk condition ($Mean = 528$ versus 499 ms, $t(64) = 9.18,\ p < 0.001$). These analyses confirmed that framing the information in terms of "safety" led to faster processing than framing it in terms of "risk". Also, it was easier to make the install decision when the risk/safety level was extreme than when it was medium, especially in the safety condition.

6 Experiment 3: Safety versus Risk with User Ratings

If a risk or safety index were included in the app selection process, it likely would accompany user ratings, for which a high score indicates good. Consequently, the advantage of the safety scale shown in Experiment 2 might be even larger in that situation due to its being compatible (and the risk scale incompatible) with the user rating scale. Therefore, in Experiment 3, we compared the risk and safety versions as in Experiment 2, but with user ratings also displayed.

6.1 Method

One-hundred and thirty Purdue undergraduate students from the same subject pool as those in Experiment 2, but who had not participated in it, took part in Experiment 3.

We aimed at understanding the effect when participants see risk information together with user rating information, which is often presented using stars. Below the four circles representing the risk or safety level, four stars representing user ratings were presented, with some of them being black and the rest of them being gray (see Fig. 6b). A label "User Ratings" was presented to the left of the stars, and "Risk" or "Safety" was presented to the left of the circles. The cover story was slightly different from that in Experiment 2: "Suppose you are using your smart phone or other electronic device to install an app. You want to check whether the app is safe [or risky] or not, and you also want to look at the user ratings."

In each trial, the four stars and the four circles were presented at the center of the screen, and there were two tasks. Task 1 was the same as the task to respond to the risk or safety information in Experiment 2. Following Task 1, the stimuli disappeared and a sentence "How many stars were there?" was presented. Participants were required to recall the number of stars by pressing "1", "2", "3", or "4" key on the number pad. There was no time limit for Task 2, and only visual feedback "Correct!" or "Incorrect!" was used. We used Task 2 to make sure that the participants actually processed the user rating information.

6.2 Results and Discussion

There were 74 female and 56 male participants. Two participants (2 percent) were replaced by new participants because of mean reaction times three standard deviations longer than the average. Overall, the response errors were rare (< 1 percent), so we will not present the error data.

Similar to the analyses in Experiment 2, a mixed ANOVA showed that responses were faster in the safety condition than in the risk condition ($Mean = 794$ versus 868 ms; see Fig. 7b), $F(1, 128) = 5.59, p = 0.020, \eta_p^2 = 0.04$, and in the far

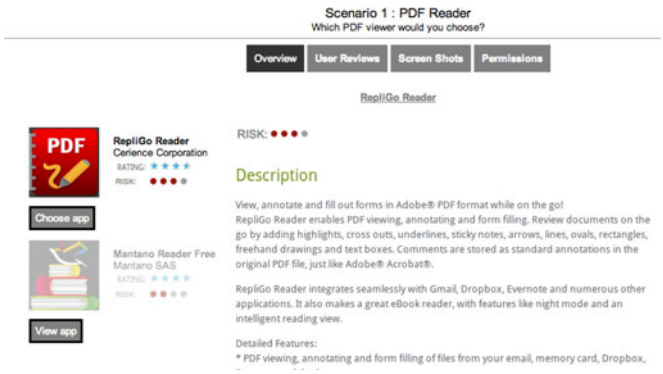


Fig. 8. An example of how the risk/safety information was presented in Experiment 4.

condition than in the close condition ($Mean$ = 802 versus 860 ms), $F(1, 128) = 123.10, p < 0.001$, $\eta_p^2 = 0.49$. However, the interaction was not significant, $F < 1$.

To compare Experiment 2 and Experiment 3, we conducted an ANOVA similar to that just described, but with experiment as an additional between-subjects factor. This ANOVA also showed main effects of distance ($Mean = 690$ versus 642 ms for close and far trials), $F(1, 256) = 285.62$, $p < 0.001$, $\eta_p^2 = 0.53$, and frame ($Mean$ = 641 versus 691 ms for safety and risk conditions), $F(1, 256) = 8.84, p = 0.003$, $\eta_p^2 = 0.03$, as well as of experiment ($Mean = 501$ versus 831 ms for Experiment 2 and Experiment 3), $F(1, 256) = 393.52$, $p < 0.001$, $\eta_p^2 = 0.61$. The only interaction that was significant was that of distance and experiment, $F(1, 256) = 12.23$, $p = 0.001$, $\eta_p^2 = 0.05$, with the close-far difference between smaller in Experiment 2 ($Mean = 520 - 482 = 38$ ms) than in Experiment 3 ($Mean = 860 - 802 = 58$ ms). All other ps > 0.107. That is, although, the advantage for the safety condition was larger numerically in Experiment 3 than in Experiment 2, the difference between studies was not statistically significant.

7 Experiment 4: Safety versus Risk in Context

For Experiment 4, we returned to the more naturalistic setting of Experiment 1. The purpose of this experiment was to test whether using symbols, such as red or green circles, to represent risk and safety levels, respectively, would lead users to select the safer/less-risky apps. We also looked at whether framing the decision as safety or risk would have any impact when making decisions in the context of the app store, where System 2 (deliberative) processes should be more of a factor than in Experiments 2 and 3. An example of how this information is displayed for the task can be seen in Fig. 8.

7.1 Method

The experiment followed a similar format as Experiment 1, presenting a simulated app store and six tasks. We presented the score as "risk" to approximately half the participants and as "safety" to the other half. Because red is associated with stop and warning, and green with go, we used those colors to display the risk and safety information, respectively. All of the tasks were presented with the same information, either red risk circles or green safety circles, and there were four possible values between one and four circles. All possible combinations except ties were represented in the set of tasks, [(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)].

TABLE 4
The Percent of Participants Who Chose the Lower Risk App under Different Interface Types (Safety versus Risk)

Apps \ Interface	Safety	Risk	Chi-square Test
Task2-1v2	63%	71.8%	χ^2 = 1.165, p = .280
Task3-1v3	77%	74.7%	χ^2 = 1.809, p = .179
Task4-1v4	86%	78.6%	χ^2 = 0.139, p = .709
Task1-2v3	73%	66.0%	χ^2 = 1.884, p = .170
Task5-2v4	72%	72.8%	χ^2 = 0.017, p = .897
Task6-3v4	78%	69.9%	χ^2 = 1.724, p = .189
Overall	74.8%	72.3%	χ^2 = 0.981, p = .322

χ^2 (1, N = 203) for all task chi-square tests, (1, N = 1218) for overall. The values 1v2 represent the Risk for an app, so in the Risk interface a participant would see one and two circles, whereas in the safety scenario the participant would see three and four circles.

7.2 Results and Discussion

7.2.1 Demographics

We excluded 17 participants who did not complete the task, and valid data from 203 participants were included in the following analyses. One-hundred and forty-four participants were male, and 59 were female. 52 were of age 18-22 years, 77 were 22-30, 47 were 30-40, 16 were 40-50, and 11 were 51 or above. The majority (82.8 percent) of the participants used an Android device, with 75.4 percent of the participants having used one for more than three months. 86.7 percent of the participants indicated that they install a new app on their mobile devices at least once a month. Regarding their computer security expertise, 3.0 percent were computer novice, 75.9 percent of the participants are regular participants, 20.7 percent were highly skilled engineers, system administrators, etc., and 0.5 percent were security experts.

7.2.2 App Choice Analysis

One purpose of this experiment was to validate the finding of Experiment 1, that is, that risk information is useful in participants' decision making. Overall, participants chose the safer/less-risky apps more often with both the Risk Interface (72.3 percent) and the Safety Interface (74.8 percent). This result indicates that the symbols (circles) used in this experiment were comparably effective to the textural risk information used in Experiment 1 (where 77.2 percent of participants chose the less risky app), although we did not strictly control the risk levels and other factors across these two studies. This finding further confirms the benefit of presenting the risk/safety information for Android apps.

The percentage of participants choosing the safer/less-risky apps was slightly higher with the Safety Interface (74.8 percent) than with the Risk Interface (72.3 percent), but this difference was not statistically significant, $\chi^2(1, N = 1218) = 0.981, p = 0.322$. We further analyzed the data by comparing each pair, but still no statistical significance was evident, although for four out of six pairs the Safety Interface yielded better decisions than did the Risk Interface (see Table 4). These results suggest that the safety information

TABLE 5
Numbers of Participants Who Specified a Certain Factor as the Most Important One (Left) and Who Considered the Factors When Choosing from the Apps (Right)

	Most Important Factor		Considered Factors	
Factor	Risk	Safety	Risk	Safety
Risk/Safety Level	28	32	76	74
User Ratings	30	24	80	79
User Reviews	22	19	71	63
Descriptions	4	4	49	47
Permissions	9	13	34	29
Screen Shots	9	7	32	33

worked in a similar way to the risk information when presenting the information within the context of the entire app, at least when distinguished by green ("go") and red ("stop") colors, respectively.

We again separated participants into those who had used Android and those who had not, and performed the app choice analysis on each group. Similar patterns were evident for both groups.

7.2.3 Correlation between App Selection and Self-Report

The self-report data on the two questions "What did you consider when choosing from the apps?" and "Which one was the most important factor?" again confirmed the usefulness of risk/safety information (see Table 5). When the risk levels were presented, 73.8 percent of participants took risk scores into consideration, and 74.0 percent of the participants facing the Safety Interface considered the safety levels when choosing from the apps. With the Risk Interface, 27.2 percent of the participants took risk scores as the most important factor, whereas with the Safety Interface, the percentage was percent.

We computed the number of tasks in which participants selected the safer/less-risky apps out of the total six tasks, which yielded the app-selection-index ranging from 0 to 6 for each participant. If participants selected the less-risky or safer app at chance, the expected value of this app-selection-index would be 3. We then tested the correlation between this app-selection-index and participants' self-reports by using Spearman's correlation test. There was a significant positive correlation between selecting the safer/less-risky app and reports of considering safety/risk in the decision-making process, $r_s = 0.311, p < 0.01$ for Safety Interface, and $r_s = 0.249, p < 0.05$ for Risk Interface. On average, participants who considered safety/risk selected safer/less-risky apps more often than those who did not (4.78 versus 3.65 for Safety Interface, and 4.54 versus 3.78 for Risk Interface). There was also a significant positive correlation between selecting the safer/less-risky app and rating safety/risk as the most important factor, $r_s = 0.483, p < .01$ for Safety Interface, and $r_s = 0.426, p < 0.01$ for Risk Interface. On average, participants who took safety/risk as the most important factor selected safer/less-risky apps more often (5.31 versus 4.07 for Safety Interface, and 5.36 versus 3.93 for Risk Interface).

8 General Discussion

Currently, when purchasing an Android app, a list of permissions required by the app is shown after it has been selected. From prior research on privacy of websites, it is known that the presentation of detailed information late in the selection process is not very effective [1], [9]. Egelman et al. [10] found that providing summary privacy information as part of search results, when multiple options were still available, was more effective at influencing online purchasing decisions. The present research demonstrates a similar value of providing summary risk information early in the process of selecting an app.

Experiment 1 showed that adding a verbal summary risk metric when participants selected between two alternative apps reduced the likelihood of their selecting the riskier app. The majority of participants indicated that they preferred the interface with the summary risk information over the one that did not have that information. Moreover, receiving the Risk Interface prior to the Standard Interface led the participants to rate permission information as more important and to make less risky choices when using the Standard Interface. The Risk Interface also led to an increase in curiosity about risk for the apps, as reflected in the fact that on average more participants checked out the permissions page. This latter finding suggests that adding a summary risk score will have the benefit of raising awareness of the potential risks, beyond the benefit of enabling participants to choose low-risk apps. In the comment section of the final questionnaire, one participant suggested allowing the developer a chance to fill in reasons why an app is requesting a given permission.

In Experiments 2 and 3, participants in a laboratory task made speeded decisions about whether to install an app or not. Responses were faster when visual symbols showed amount of safety rather than amount of risk. In other words, participants were more risk-averse when the indicators conveyed the positive outcome of safety rather than the negative outcome of risk (a framing effect). The difference in response time as a function of safety versus risk was significant when only a risk/safety indicator was displayed and tended to be larger when a User Rating indicator was also shown. The scale of more filled circles designating better safety is compatible with the User Rating scale, whereas that of more filled circles designating higher risk (worse safety) is incompatible with the User Rating scale.

When symbols were used to convey the risk information in a more naturalistic situation in Experiment 4, the percentage of choices of the less-risky app was similar to that for the verbal risk displays in Experiment 1. Thus, summary risk information is beneficial regardless of whether it is conveyed verbally or symbolically. In Experiment 4, there was only slight evidence for a benefit of presenting safety indicators rather than risk indicators. The likely reason why the safety condition was more beneficial in Experiments 2 and 3 than in Experiment 4 is that the former experiments emphasized speed of responding. This speed emphasis would cause System 1, automatic processing, to have a large role in the decision process. In Experiment 4, with no speed emphasis, the relative contribution of System 2, deliberate processing, would increase [14]. An implication is that

users' decisions will tend to be riskier when they are made quickly, without deliberate consideration of the alternatives, and this is where framing is most important.

It is also possible that in a naturalistic setting like that of Experiment 4 an increase in the emotional valance of the indicators would lead to a larger benefit for the Safety Interface over the Risk Interface [14]. We used red circles for risk and green circles for safety in Experiment 4 because those colors are associated with 'stop' and 'go', respectively. However, symbols with more negative affect (e.g., bombs, skull, or crossbones) for risk and more positive affect (e.g., smiling faces) for safety could be used that would produce stronger negative and positive reactions, respectively.

One strength of the present research is the use of two types of experiments, online crowdsourcing under relatively naturalistic conditions and controlled laboratory under reduced and targeted conditions. An obvious limitation is that even the crowdsourcing experiments deviate from what a person would do when choosing to download an app in everyday life. However, experiments like these, with objective measures, as part of a multiple-method approach that includes more qualitative, descriptive studies can provide crucial evidence in helping understand what risk communications are likely to be effective and, especially, why. Finally, it should be noted that the majority of the MTurk participants were Android users, and the duration of the experiments was short. Therefore, the possibility exists that the Risk Interface would lose its effectiveness as the novelty of the risk information wears off due to continued use.

9 Conclusion and Future Work

The results from four user studies validated our hypothesis that when risk ranking is presented in a user-friendly fashion, e.g., translated into categorical values and presented early in the selection process, it will lead users to select apps with lower risk. The majority of participants preferred to have such a risk metric in Google Play Store. We expect that adding a summary risk metric would cause positive changes in the app ecosystem. When users prefer lower-risk apps, developers will have incentives to better follow the least-privilege principle and request only necessary permissions. It is also possible that the introduction of this risk score will cause more users to pay for low risk apps. Thus, this creates an incentive for developers to create lower risk apps that do not contain invasive ad networks and in general over-request permissions.

Our studies are not the last word on the question of how to best present risk information. For example, we have also not examined how the risk score interacts with other factors to affect a users choice, such as user ratings in the natural setting and whether an app is free or not. Also of interest is how users behave when choosing among a list of search results (as opposed to choosing between two options). These topics are important ones for future research.

Acknowledgments

This work was supported by Army Research Office Award 2008-0845-04 through North Carolina State University, and by the National Science Foundation under Grant No. 1314688. Work by C. Gates and N. Li were also supported by Google Award 13033386.

References

[1] A.I. Anton, J.B. Earp, Q. He, W. Stufflebeam, D. Bolchini, and C. Jensen, "Financial Privacy Policies and the Need for Standardization," *IEEE Security and Privacy*, vol. 2, no. 2, pp. 36-45, Mar./Apr. 2004.

[2] D. Balfanz, G. Durfee, D.K. Smetters, and R.E. Grinter, "In Search of Usable Security: Five Lessons from the Field," *IEEE Security and Privacy*, vol. 2, no. 5, pp. 19-24, Sept./Oct. 2004.

[3] R. Biddle, P.C. van Oorschot, A.S. Patrick, J. Sobey, and T. Whalen, "Browser Interfaces and Extended Validation SSL Certificates: An Empirical Study," *Proc. ACM Workshop Cloud Computing Security*, pp. 19-30, 2009.

[4] E. Chin, A.P. Felt, V. Sekar, and D. Wagner, "Measuring User Confidence in Smartphone Security and Privacy," *Proc. Eighth Symp. Usable Privacy and Security (SOUPS '12)*, pp. 1-16, 2012.

[5] L.F. Cranor, M. Arjula, and P. Guduru, "Use of a P3P User Agent by Early Adopters," *Proc. ACM Workshop Privacy in the Electronic Soc.*, pp. 1-10, 2002.

[6] L.F. Cranor, P. Guduru, and M. Arjula, "User Interfaces for Privacy Agents," *ACM Trans. Computer-Human Interaction (TOCHI '06)*, vol. 13, no. 2, pp. 135-178, 2006.

[7] N. Dell, V. Vaidyanathan, I. Medhi, E. Cutrell, and W. Thies, "Yours is Better!: Participant Response Bias in HCI," *Proc. Conf. Human Factors in Computing Systems*, pp. 1321-1330, 2012.

[8] A. Diederich and J.R. Busemeyer, "Judgment and Decision Making," *Experimental Psychology*, A.F. Healy and R.W. Proctor, eds., second ed., pp. 295-319, John Wiley & Sons, 2013.

[9] S. Egelman, L.F. Cranor, and A. Chowdhury, "An Analysis of P3P -Enabled Web Sites among Top-20 Search Results," *Proc. Eighth Int'l Conf. Electronic Commerce*, pp. 197-207, 2006.

[10] S. Egelman, J. Tsai, L.F. Cranor, and A. Acquisti, "Timing Is Everything?: The Effects of Timing and Placement of Online Privacy Indicators," *Proc. 27th Int'l Conf. Human Factors in Computing Systems*, pp. 319-328, 2009.

[11] B. Fathi, *Engineering Windows 7 : User Account Control*, MSDN blog on User Account Control, http://blogs.msdn.com/b/e7/archive/2008/10/08/user-account-control.aspx, Oct. 2008.

[12] A.P. Felt, K. Greenwood, and D. Wagner, "The Effectiveness of Application Permissions," *Proc. Second USENIX Conf. Web Application Development (WebApps '11)*, 2011.

[13] A.P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner, "Android Permissions: User Attention, Comprehension, and Behavior," *Proc. Eighth Symp. Usable Privacy and Security*, 2012.

[14] M.L. Finucane, A. Alhakami, P. Slovic, and S.M. Johnson, "The Affect Heuristic in Judgments of Risks and Benefits," *J. Behavioral Decision Making*, vol. 13, no. 1, pp. 1-17, 2000.

[15] M. Gondan, C. Götze, and M.W. Greenlee, "Redundancy Gains in Simple Responses and Go/no-Go Tasks," *Attention, Perception, & Psychophysics*, vol. 72, no. 6, pp. 1692-1709, 2010.

[16] K.A. Juang, S. Ranganayakulu, and J.S. Greenstein, "Using System-Generated Mnemonics to Improve the Usability and Security of Password Authentication," *Proc. Human Factors and Ergonomics Soc. Ann. Meeting*, vol. 56, no. 1, pp. 506-510, 2012.

[17] D. Kahneman, *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[18] P.G. Kelley, S. Consolvo, L.F. Cranor, J. Jung, N. Sadeh, and D. Wetherall, "A Conundrum of Permissions: Installing Applications on an Android Smartphone," *Proc. Workshop Usable Security (USEC '12)*, Feb. 2012.

[19] P.G. Kelley, L.F. Cranor, and N. Sadeh, "Privacy as Part of the App Decision-Making Process," *Proc. Conf. Human Factors in Computing Systems (CHI '13)*, pp. 3393-3402, 2013.

[20] T. H.-J. Kim, P. Gupta, J. Han, E. Owusu, J. Hong, A. Perrig, and D. Gao, "OTO: Online Trust Oracle for User-Centric Trust Establishment," *Proc. ACM Conf. Computer and Comm. Security*, pp. 391-403, 2012.

[21] J. Lin, S. Amini, J.I. Hong, N. Sadeh, J. Lindqvist, and J. Zhang, "Expectation and Purpose: Understanding Users' Mental Models of Mobile App Privacy Through Crowdsourcing," *Proc. ACM Conf. Ubiquitous Computing (UbiComp '12)*, pp. 501-510, 2012.

[22] R.D. Luce, *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford Univ. Press, 1986.
[23] S. Mishra, M. Gregson, and M.L. Lalumière, "Framing Effects and Risk-Sensitive Decision Making," *British J. Psychology*, vol. 103, no. 1, pp. 83-97, Feb. 2012.
[24] S. Motiee, K. Hawkey, and K. Beznosov, "Do Windows Users Follow the Principle of Least Privilege?: Investigating User Account Control Practices," *Proc. Sixth Symp. Usable Privacy and Security*, 2010.
[25] H. Peng, C.S. Gates, B.P. Sarma, N. Li, Y. Qi, R. Potharaju, C. Nita-Rotaru, and I. Molloy, "Using Probabilistic Generative Models for Ranking Risks of Android Apps," *Proc. ACM Conf. Computer and Comm. Security*, pp. 241-252, 2012.
[26] E.E. Schultz, "Web Security, Privacy, and Usability," *Handbook of Human Factors in Web Design*, K.-P.L. Vu and R.W. Proctor, eds., pp. 663-677, CRC Press, 2011.
[27] J. Schwarz and M. Morris, "Augmenting Web Pages and Search Results to Support Credibility Assessment," *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pp. 1245-1254, 2011.
[28] J. Staddon, D. Huffaker, L. Brown, and A. Sedley, "Are Privacy Concerns a Turn-Off?: Engagement and Privacy in Social Networks," *Proc. Eighth Symp. Usable Privacy and Security (SOUPS '12)*, pp. 1-13, 2012.
[29] S. Sternberg, "Inferring Mental Operations from Reaction-Time Data: How We Compare Objects," *An Invitation to Cognitive Science: Methods, Models and Conceptual Results*, D. Scarborough and S. Sternberg, eds., pp. 703-863, MIT Press, 1998.
[30] J. Sun, P. Ahluwalia, and K.S. Koong, "The More Secure the Better? A Study of Information Security Readiness," *Industrial Management and Data Systems*, vol. 111, no. 4, pp. 570-588, 2011.
[31] A. Tversky and D. Kahneman, "The Framing of Decisions and the Psychology of Choice," *Science*, vol. 211, no. 4481, pp. 453-458, 1981.
[32] W. Van Wassenhove, K. Dressel, A. Perazzini, and G. Ru, "A Comparative Study of Stakeholder Risk Perception and Risk Communication in Europe: A Bovine Spongiform Encephalopathy Case Study," *J. Risk Research*, vol. 15, no. 6, pp. 565-582, 2012.
[33] K.-P.L. Vu, V. Chambers, B. Creekmur, D. Cho, and R.W. Proctor, "Influence of the Privacy Bird User Agent on User Trust of Different Web Sites," *Computers in industry*, vol. 61, no. 4, pp. 311-317, 2010.
[34] K.-P.L. Vu, R.W. Proctor, A. Bhargav-Spantzel, B. Tai, J. Cook, and E.Eugene Schultz, "Improving Password Security and Memorability to Protect Personal and Organizational Information," *Int'l J. Human-Computer Studies*, vol. 65, no. 8, pp. 744-757, 2007.
[35] S. Werner and C. Hoover, "Cognitive Approaches to Password Memorability—The Possible Role of Story-Based Passwords," *Proc. Human Factors and Ergonomics Society Ann. Meeting*, vol. 56, pp. 1243-1247, 2012.
[36] XF. Xie, M. Wang, R. Zhang, J. Li, and QY. Yu, "The Role of Emotions in Risk Communication," *Risk Analysis*, vol. 31, no. 3, pp. 450-465, 2011.

Christopher S. Gates received the BS degree in computer science as well as in mathematics and the MS degree in computer science both from the Rutgers University in 2002 and 2005, respectively. After this, he worked in industry for several years until 2009 when he returned to academia to pursue the PhD degree in computer sScience at Purdue University.

Jing Chen received the bachelor's and master's degrees in psychology from Zhejiang University in China. She is currently working toward the PhD degree in cognitive psychology and the master's degree in industrial engineering at the Purdue University.

Ninghui Li received the BEng degree in computer science from the University of Science and Technology of China in 1993, and the MSc and PhD degrees in computer science from the New York University, in 1998 and 2000, respectively. He is currently an associate professor in computer science at Purdue University. His research interests include security and privacy in information systems. He is a senior member of the IEEE, and an ACM distinguished scientist.

Robert W. Proctor received the masters degree and PhD degree in experimental psychology from the University of Texas at Arlington. He is a distinguished professor in the Department of Psychological Sciences and a fellow of the Center for Education and Research in Information Assurance and Security at Purdue University. His research interests include basic and applied aspects of human performance in a variety of tasks and settings.