SPECIAL SECTION: Cybersecurity Decision Making

Influence of Risk/Safety Information Framing on Android App-Installation Decisions

Jing Chen, Department of Psychological Sciences, Purdue University, **Christopher S. Gates** and **Ninghui Li**, Department of Computer Science, Purdue University, and **Robert W. Proctor**, Department of Psychological Sciences, Purdue University

We conducted three experiments with participants recruited on Amazon's Mechanical Turk to examine the influence on app-installation decisions of summary risk information derived from the app permissions. This information can be framed negatively as amount of risk or positively as amount of safety, which was varied in all the experiments. In Experiments 1 and 2, the participants performed tasks in which they selected two Android apps from a list of six; in Experiment 3, the tasks were to reject two apps from the list. This summary information influenced the participants to choose less risky alternatives, particularly when it was framed in terms of safety and the app had high user ratings. Participants in the safety condition reported that they attended more to the summary score than did those in the risk condition. They also showed better comprehension of what the score was conveying, regardless of whether the task was to select or reject. The results imply that development of a valid risk/safety index for apps has the potential to improve users' app-installation decisions, especially if that information is framed as amount of safety.

Keywords: compatibility effects, cyber, cybersecurity, information security, privacy, risk communication, security usability

INTRODUCTION

Assessing risk when making decisions is vital in many areas, including gambling at a casino (Frings, 2012), making healthcare decisions (Schwartz, 2011), and deciding about business strategies and tactics that may affect a business and its customers (Hung & Tangpong, 2010). Because the risks associated with specific actions are often not fully known by the decision maker and may be difficult to comprehend, how best to communicate those risks is a concern (e.g., Brust-Renck, Royer, & Reyna, 2013; McLaughlin & Mayhorn, 2014).

One domain with great risks that is relevant to most people is that of smart mobile devices (Toch, Wang, & Cranor, 2012). In early 2014, for the first time more internet traffic was due to smartphone and tablet devices than to personal computers (O'Toole, 2014). The number of smartphone users world-wide is estimated to be 1.76 billion by the end of 2014 (eMarketer, 2014). Although convenient and pervasive, mobile devices also introduce new dimensions of risk, such as leakage of personal files, physical location, and monetary loss (Shabtai et al., 2010), and raise new privacy and security concerns. However, users often do not have accurate understanding of the risks associated with installing apps on the device (Felt et al., 2012). This lack of understanding leads to the potential for an app to collect data from a user without the user's explicit and intentional consent, which may result in malicious functionality such as intercepting bank authentication messages or sending texts to premium-rate phone numbers.

In our research, we have focused on security of the Android operating system because of its openness and popularity (Mansfield-Devine, 2012) and the fact that 99% of the mobile malware in 2013 targeted Android devices (Cisco Systems, Inc., 2014). In the current Android device, when a user chooses to install an app, a list of the permissions that the app requests is displayed. The defense against malware relies on users comprehending those permissions and

Address correspondence to Jing Chen, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907-2081, USA, chenjing.psy@gmail.com.

Journal of Cognitive Engineering and Decision Making
2015, Volume 9, Number 2, June 2015, pp. 149–168
DOI: 10.1177/1555343415570055

making informed decisions about whether to install this app or select another app that provides similar functionality. However, several studies have provided evidence that users tend to ignore the permissions or fail to accurately comprehend their meanings (Chin, Felt, Sekar, & Wagner, 2012; Felt, Greenwood, & Wagner, 2011; Felt et al., 2012; Kelley et al., 2012).

Several researchers have proposed ways to communicate better the risks associated with installing apps to improve users' app-selection decisions. Felt et al. (2012) suggested changing the wording of permissions, modifying and renaming the permission categories to be more informative, reducing the number of permissions, specifying more clearly the risks associated with a permission, and presenting users only those permissions that are of high risk. Lin et al. (2012) recommended using crowd-sourcing through Amazon Mechanical Turk (MTurk) to discover users' expectations about the permissions an app would need and to signal to users when an app's requested permissions deviate from those expectations. Kelley, Cranor, and Sadeh (2013) proposed that privacy-related information to which an app has access (contacts, personal information, location, etc.) be shown on the app's main description page, visible when the app-selection decision is being made, rather than late in the process when the user has chosen and wants to install an app, as is currently the case. We have also proposed providing risk information early in the decision process, but in the form of summary risk scores that allow easy comparison between apps (Gates, Chen, Li, & Proctor, 2014; Gates, Li et al., 2014). The efficacy of such risk scores may depend on the way in which the information is presented, or framed.

Positive and Negative Framing in Decision Making

It is well-known that people's preferences in risky contexts are influenced by the way in which a problem is framed (the framing effect; Tversky & Kahneman, 1981, 1986). For example, people are risk-averse when the outcomes are presented in terms of potential gains (i.e., in a positive frame) but risk-seeking when they are presented in terms of potential loss (i.e., in a negative frame). This effect has been found to be stable and replicable across time and different age groups (Mayhorn, Fisk, & Whittle, 2002) and in many scenarios (e.g., Barnes, McDermott, Hutchins, & Rothrock, 2011; Gambara & Piñon, 2005; Garcia-Retamero & Dhami, 2013).

In principle, because the alternative options in both frames are logically equivalent, people should choose the same option regardless of the problem framing. This framing effect is of special interest in human decision making because it is counterintuitive and inconsistent with the tenets of rational decision making (e.g., the principle of invariance; Tversky & Kahneman, 1986). Similar to the positive and negative framing of outcomes, the risk information associated with a particular app on mobile devices can be framed positively in terms of the amount of safety or negatively in terms of the amount of risk. Thus, framing the risk information in these two forms may lead to different judgments or preferences by users. We focused on the *safety* and *risk* frames in the current study because safety is a customary antonym of risk (see merriam-webster.com/), and both words are short enough to keep the app interface succinct.

Task Compatibility

The principle of compatibility is that input information is weighted based on its compatibility with the output (response): Inputs that are more compatible with outputs are weighted more and draw more attention than those that are less compatible (Huber, Huber, & Bär, 2014; Rubaltelli, Dickert, & Slovic, 2012; Slovic, Griffin, & Tversky, 1990; Tversky, Sattath, & Slovic, 1988). Compatibility between the nature of the task and the valence of the alternatives has been shown to influence decision making and judgments. Shafir (1993) provided evidence that positive dimensions are weighted more heavily when the task is one of selecting among alternatives, whereas negative dimensions are weighted more heavily when the task is to reject alternatives. For example, Nagpal and Krishnamurthy (2008) found that the combination of task and the valence of the alternatives had an influence on decision difficulty and decision time. When the task and the valence of the alternatives were compatible (e.g., choosing

between two attractive alternatives, or rejecting one of two unattractive alternatives), the decisions were easier than when they were incompatible (e.g., choosing between two unattractive alternatives, or rejecting one of two attractive alternatives). Several other studies have shown that a selection task promotes the decision-maker to focus more on the positive features of the outcome, whereas a rejection task promotes the negative features (e.g., Chernev, 2009; Lai & Hui, 2006).

Installing an app on a mobile device is essentially a selection task. That is, the user selects which app to download among several apps that provide similar functionality. Positive information associated with an app may be weighted more than negative information in this selection task. With regard to an overall risk score, the compatibility principle suggests that when this score is presented as amount of "safety," this information will be weighted more in the app-selection decision than when it is presented as amount of "risk."

Current Study

We previously provided evidence that a summary risk score is beneficial in conveying risks for Android apps (Gates, Chen et al., 2014). The summary score supports easy comparison for risks associated with apps of similar functionality. In that study, we reported experiments that examined the effects of displaying a summary risk score in text or symbol format. The results showed that participants took this summary score into account, and it had a positive effect on their app-selection decisions. Furthermore, in a laboratory experiment in which participants were to decide as fast as possible whether to install an app or not, performance was better when the symbol designated amount of safety rather than risk. However, in an MTurk experiment in which participants selected one app from two presented side-by-side, there was little difference in their decisions as a function of whether the summary score conveyed risk as opposed to safety.

As noted, the latter online experiment of Gates, Chen et al. (2014) involved selection between only two alternative apps, which were presented side-by-side. This differs from the environment in the app stores for mobile devices, where the user is typically confronted with a list of multiple apps with similar functionality. For this reason, in the present study the number of alternative apps to be considered for each decision was increased to six, and they were presented in a list format. Our hypothesis was that in this more realistic scenario with a selection task, framing the summary risk/safety information in terms of safety would lead to less risky decisions than framing the information in terms of risk. Experiments 1 and 2 were designed to test this hypothesis; in Experiment 3, the decision was changed to one of rejecting the same number of apps from the list, to test whether compatibility with the task was a crucial factor.

EXPERIMENT 1

To emulate an app-selection context, we presented lists of alternative apps to users of MTurk, who would have experience installing apps on mobile devices. The display was designed to mimic what one would see when selecting an app from the Google Play store. Of most interest was the inclusion of a summary score, framed as amount of risk for some participants and safety for others, in the form of number of filled circles out of five. The expectancy was that users' choices would be influenced by this score, more so when conveyed as safety than risk.

Method

Participants. In total, 295 participants were recruited through MTurk. The experiment took about 10 minutes to complete, and participants were paid $0.50 each. This study, and Experiments 2 and 3, received approval from Purdue University's Institutional Review Board.

Materials and procedure. Participants were randomly assigned to a safety or risk condition. At the beginning of the experiment, an introductory page was displayed that included the purpose and a demonstration of the elements of each app that would be shown in the tasks (see Figure 1). Before performing the app-selection tasks, a pretask questionnaire was conducted for collecting demographic information, the participants' history of use of mobile devices,

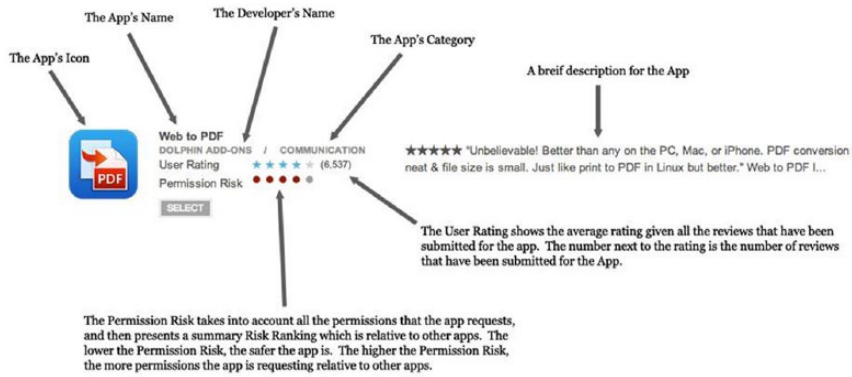

Figure 1. The introductory page demonstrating the elements of each app in the risk condition.

and their app installation activities. Each app-selection task was on one page with a heading, "Which 2 of these Android Apps would you choose?" (see Figure 2). The six apps presented for each task were chosen from the Google Play Store when one of the experimenters did a search for a specific functionality. We skipped the Top 10 apps and chose the 11th to 16th apps to avoid a possible influence of participants' familiarity with the apps.

For each app, the displayed information included icon, app name, developer, user rating (filled stars out of five), user rating count, permission safety/risk (filled circles out of five; more circles indicated increasing safety in the safety condition and increasing risk in the risk condition), and two lines of a brief description that ended with "…" if it was not complete. The user ratings were generated randomly for each participant and each app. The user rating was taken from a distribution based off of approximately 300,000 apps collected from the actual app store. From five stars to one star, the likelihood was [.25, .35, .20, .10, .10], and the actual percentages were [.257, .352, .197, .099, .095]. The permission safety/risk scores were taken uniformly at random for each app. They were displayed as filled red circles for risk and filled green circles for safety, to take advantage of the colors' strong associations with "stop" and "go," respectively (Bergum & Bergum, 1981). Thus, for an app in a specific location in a task, the display was controlled to be identical for all participants except for the user rating and the safety/risk score.

Participants were urged to select two apps out of six by clicking "select" buttons that were positioned under each app. Selection of two apps rather than one was required because users typically do not make their final installation decision based only on the summary information available in the list display. The purpose was to examine what factors influence users' decisions in narrowing their options to a smaller subset of the apps. Upon clicking the button for a particular app, the question, "Why did you pick this app?" was asked, with the following listed options: User Rating Score, User Rating Count, Permission Safety (or Risk), Icon Look and Feel, Description, Familiarity with app or developer, and Other. Participants could indicate as many of these options as they wanted, and they did not have to select any reason before clicking the submit button to continue to the next task. After all six tasks, a posttask questionnaire was conducted regarding the app selection tasks and the participant's security expertise and concern.

Results and Discussion

Demographics. The demographic data are shown in Table 1. Almost two-thirds of the participants were male, with most between 18 and 40 years of age. More than 90% indicated that they used Android devices, with approximately 90% installing apps on at least a monthly basis. Less than 2% responded that they were security experts. Table 1 also includes the demographic information for participants in Experiments 2 and 3, which we will not discuss further since

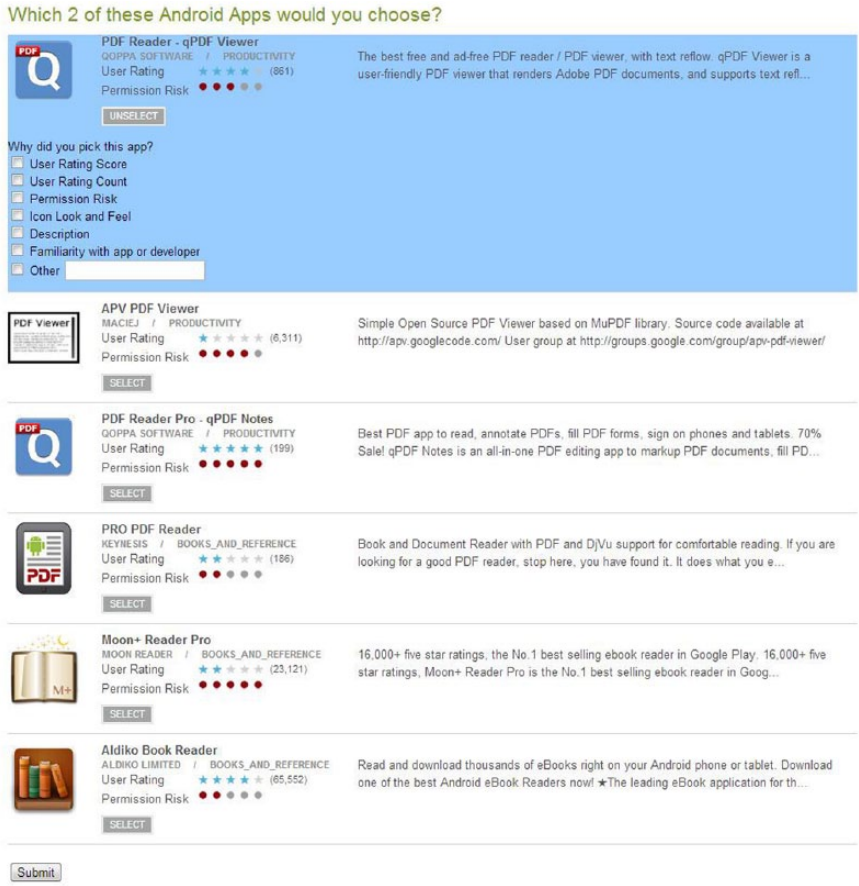

Figure 2. An example of the task: Once an app is selected, its background turns to blue, the "SELECT" button changes to "UNSELECT," and a list of reasons for app selection is shown below the selected app.

their characteristics were similar, except for having somewhat higher percentages of males.

App-selection analysis. The independent variables were the user rating, permission safety/risk score of each app, and the safety/risk condition, and the dependent variable was whether a specific app was selected by the participant. Again, "each app" refers to the app in a specific location in a task. We first conducted correlational analyses between app selection and user rating, and between app selection and safety/risk score, to find out the relation between these variables. These analyses were conducted for each app, which has a fixed icon, position, description, etc. A repeated-measures ANOVA was conducted for the percentage of app selection, by comparing each app's selection percentage under each User Rating × Safety/Risk Score combination. This analysis was performed as if an app was a participant, with user rating, safety/risk score, and safety/risk condition as within-app factors. A binary logistic regression analysis was also performed to examine the effects of user rating, safety/risk score, safety/risk condition, and their interactions. The results were similar to those of the ANOVA but not as detailed, so we report only the ANOVA results in this and the following experiments.

In the safety condition, there was a positive correlation between user rating and app selection for every app (Pearson's rs $\geq .188$, $N = 143$, ps $\leq .025$), and between safety score and app selection for most apps (34 out of 36 apps; rs $\geq .204$, $N = 143$, ps $\leq .015$). In the risk condition,

TABLE 1: Percentage of Participants in Each Demographic Category for All Three Experiments

Variable	Experiment 1 (N = 295)	Experiment 2 (N = 494)	Experiment 3 (N = 398)
Gender			
Male	60.7	66.2	72.6
Female	39.3	33.8	27.4
Age (years)			
18–22	12.2	15.2	18.1
22–30	45.8	46.6	48.7
30–40	27.8	26.1	26.1
40–50	7.8	7.9	5.0
51 and above	6.4	3.8	2.0
Android Device Usage			
Never	6.1	8.9	8.5
Less than 3 months	6.4	8.1	4.8
3 months ~ 1 year	12.5	16.4	15.8
1 year ~ 2 years	27.1	23.7	22.4
More than 2 years	47.8	42.9	48.5
Installing new apps			
Several times a week	19.0	15.8	14.1
About once a week	28.1	24.3	28.4
Several times a month	22.4	25.9	26.4
About once a month	19.7	20.6	18.6
Less than once a month	10.2	12.8	11.3
Security Expertise			
Regular user	69.5	72.5	72.1
Computer novice	5.1	3.8	4.0
Highly skilled	23.7	22.5	22.6
Security experts	1.7	1.2	1.3

there was also a positive correlation between user rating and app selection for every app (rs $\geq$ .220, N = 152, ps $\leq$ .007), but not between risk score and app selection for most apps (29 out of 36 apps; $|r|$s $\leq$ .139, N = 152, ps $\geq$ .091). These correlational results indicate that apps with higher user ratings were selected more often than those with lower ratings in both the safety and risk conditions. More important, apps with higher safety scores were selected more often than those with lower safety scores in the safety condition, but the risk score did not have a similar impact on app selection in the risk condition.

The ANOVA showed several findings. First, the percentages of app selection did not differ in the safety and risk conditions on average (Ms = 26.4 % vs. 25.5%), $F(1, 35) = 1.03$, $p = .317$, $\eta_p^2 = .03$, but safety and risk conditions showed distinct patterns across safety rankings (i.e., safety/risk scores; see Figure 3, top left panel), $F(4, 140) = 22.39$, $p < .001$, $\eta_p^2 = .39$. Note that the frequencies with which each app were associated with specific safety/risk score and user rating combinations were not controlled to be strictly equal, and thus, the percentage of app selection was computed as an average weighted by the frequency. As a result, the overall computed percentage can differ from the actual overall selection percentage (33.3%). Figures plotted from the raw data show the same patterns as those from the weighted data. The mean app selection percentage varied with the increased

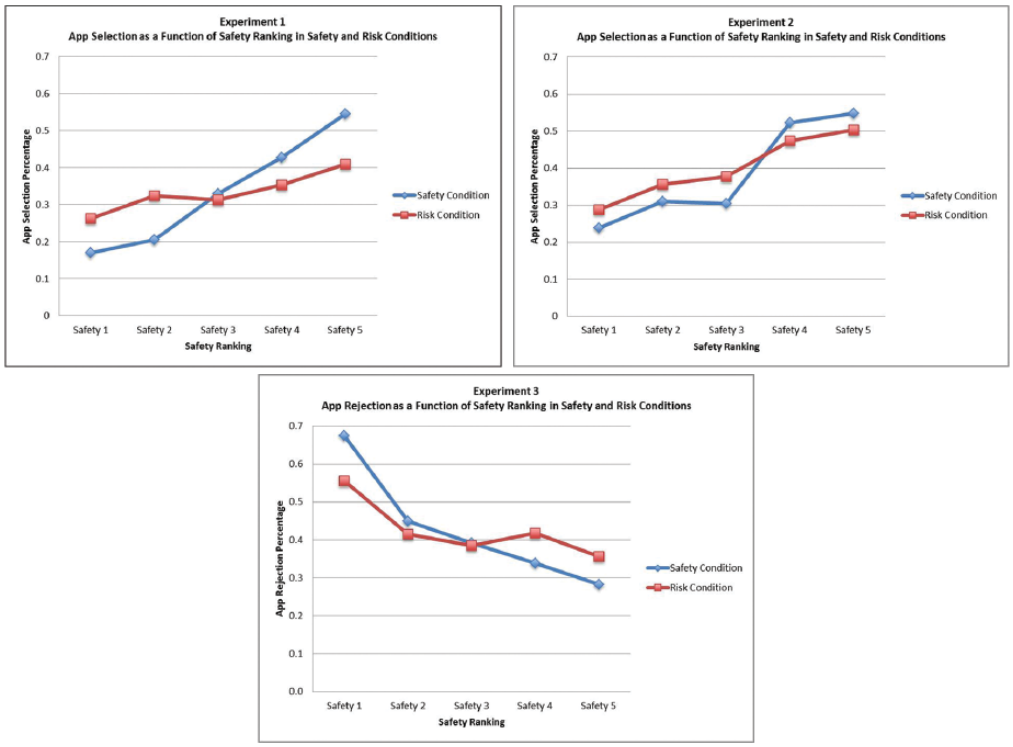

Figure 3. App selection percentage (Experiments 1 and 2) and rejection percentage (Experiment 3) as a function of safety ranking (= safety score or 6 – risk score) in safety and risk conditions.

safety rankings, being 13.4%, 16.4%, 24.9%, 33.1%, 44.0% in the safety condition, and 21.8%, 25.2%, 22.6%, 27.3%, 30.6% in the risk condition. Post hoc pairwise comparisons with Bonferroni adjustment (used for all subsequent pairwise comparisons) showed that the differences between the safety/risk conditions were significant for most of the safety rankings, $ps \leq .007$, except for the middle value, $p = .182$. Compared to the risk condition, participants in the safety condition selected apps with lower safety ranking less often and apps with higher safety ranking more often. In other words, better app selection decisions (safer apps selected more often and riskier apps selected less often) were made in the safety condition than in the risk condition.

Second, overall, the percentage of app selection was higher with increased safety (i.e., increased safety score or decreased risk score) (Ms = 17.6%, 20.8%, 23.7%, 30.2%, and 37.3% for safety rankings 1 through 5), $F(4, 140) = 68.32$, $p < .001$, $\eta_p^2 = .66$, and the percentage was also higher with increased user rating (Ms = 6.7%, 12.1%, 17.5%, 40.4%, and 53.0% for user ratings 1 through 5), $F(4, 140) = 317.78$, $p < .001$, $\eta_p^2 = .90$. Moreover, safety ranking interacted with user rating, $F(16, 560) = 8.61$, $p < .001$, $\eta_p^2 = .20$. Post hoc pairwise comparisons for this interaction showed that for apps with lower user ratings (1, 2, and 3) the safety ranking did not influence app selection much, but for apps with higher user ratings (4 and 5) increased safety rankings led to more app selection (see Table 2). These results show that apps with higher user ratings and higher safety rankings were selected more often than other options and that the percentage of app selection increased as the safety increased or the risk decreased, much more for the apps with high user rating than for those with low user rating (see Figure 4, top row).

Finally, there was a three-way interaction among user rating, safety level, and safety/risk condition, $F(16, 560) = 2.10$, $p = .007$, $\eta_p^2 = .06$. Further analyses showed that this interaction was mainly due to a significant two-way interaction between safety level and safety/risk condition when user rating equaled 2, 3, 4, and 5 ($ps \leq$

TABLE 2: Percentage Difference in App Selection/Rejection Between Different Safety Rankings as a Function of User Rating

User Rating	Safety	Experiment 1					Experiment 2					Experiment 3				
		1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	1	0.0					0.0					0.0				
	2	0.1	0.0				1.6	0.0				6.2	0.0			
	3	1.6	1.5	0.0			−2.2	−3.8	0.0			21.2*	15.1	0.0		
	4	0.6	0.5	−1.0	0.0		5.5	3.9	7.7	0.0		14.1	7.9	−7.2	−15.3	
	5	6.4	6.4	4.9	5.9	0.0	5.3	3.7	7.5	−0.2	0.0	29.4***	23.2***	8.2	15.3	0.0
2	1	0.0					0.0					0.0				
	2	−4.1	0.0				−8.6	0.0				22.5*	0.0			
	3	−6.5	−2.5	0.0			−1.3	7.3	0.0			22.7*	0.2	0.0		
	4	1.8	5.9	8.3	0.0		−0.9	7.7	0.4	0.0		8.6	−14.0	−14.2	0.0	
	5	3.3	7.4	9.9	1.5	0.0	0.1	8.7	1.5	1.0	0.0	36.1***	13.6	13.4	27.6***	0.0
3	1	0.0					0.0					0.0				
	2	3.2	0.0				2.8	0.0				18.8**	0.0			
	3	2.5	−0.7	0.0			10.8	8.0	0.0			27.8***	9.0	0.0		
	4	11.6***	8.3*	9.1***	0.0		14.7*	12.0	4.0	0.0		37.1***	18.3**	9.3	0.0	
	5	19.3***	16.0***	16.8***	7.7	0.0	6.2	3.4	−4.6	−8.6	0.0	25.6***	6.9	2.2	11.5	0.0
4	1	0.0					0.0					0.0				
	2	4.0	0.0				6.1	0.0				24.8***	0.0			
	3	10.9***	7.0*	0.0			5.2	−1.0	0.0			21.3***	3.4	0.0		
	4	20.8***	16.8***	9.9**	0.0		25.2***	19.1***	20.0***	0.0		37.1***	12.3	15.7**	0.0	
	5	33.1***	29.1***	22.2***	12.3***	0.0	27.1***	21.0***	22.0***	2.0	0.0	35.6***	10.8	14.3**	1.5	0.0
5	1	0.0					0.0					0.0				
	2	13.0***	0.0				5.7	0.0				21.6***	0.0			
	3	22.0***	9.0	0.0			14.0**	8.3	0.0			25.0***	3.4	0.0		
	4	28.1***	15.1***	6.1	0.0		14.0**	8.3	0.0	0.0		24.8***	3.2	−0.2	0.0	
	5	36.5***	23.5***	14.5***	8.4	0.0	25.2***	19.5***	11.1	11.2	0.0	27.0***	5.4	2.0	2.2	0.0

*The mean difference was significant at the .05 level. **The mean difference was significant at the .01 level. ***The mean difference was significant at the .001 level.

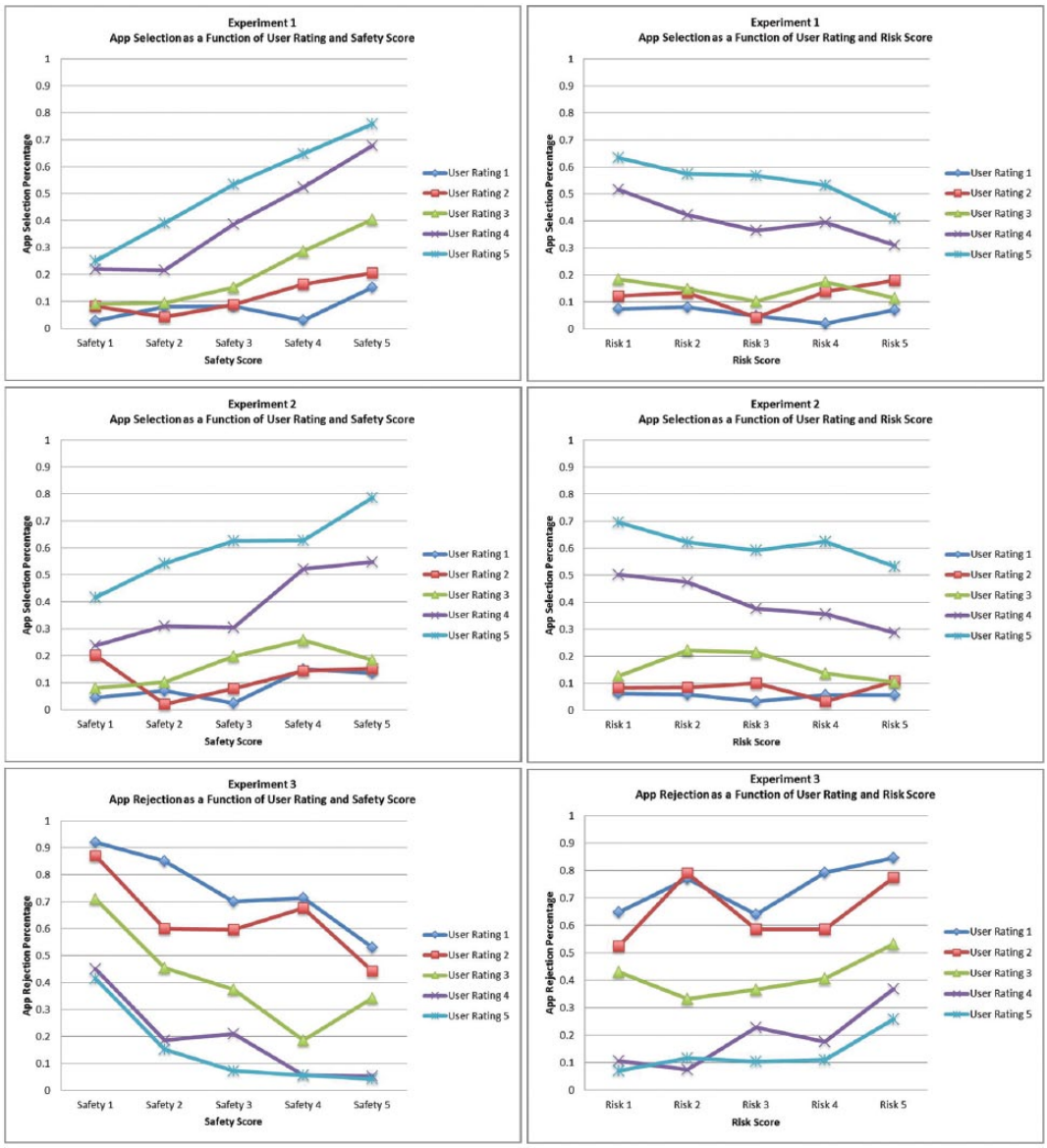

Figure 4. App selection percentage (Experiments 1 and 2) and rejection percentage (Experiment 3) as a function of user ratings and safety or risk score.

.007) but not when user rating equaled 1 (p = .124). An ANOVA excluding the user rating = 1 condition showed no three-way interaction, $F < 1$. There was also an interaction between safety/risk condition and user rating, $F(4, 560) = 4.22$, $p = .003$, $\eta_p^2 = .11$. This interaction did not show up much in the mean data (7.5%, 11.7%, 20.5%, 40.5%, 51.6% for the safety condition, and 5.9%, 12.4%, 14.5%, 40.2%, 54.5% for the risk condition, with increased user rating). It reflected mainly a larger percentage for the safety condition than for the risk condition when paired with an intermediate user rating. Consistent with the mean data, post hoc pairwise comparisons showed that the safety and risk conditions differed when user rating = 3, $p < .001$, but not when it had other values, ps $\geq .159$. There is no rationale for this pattern, and it was not significant

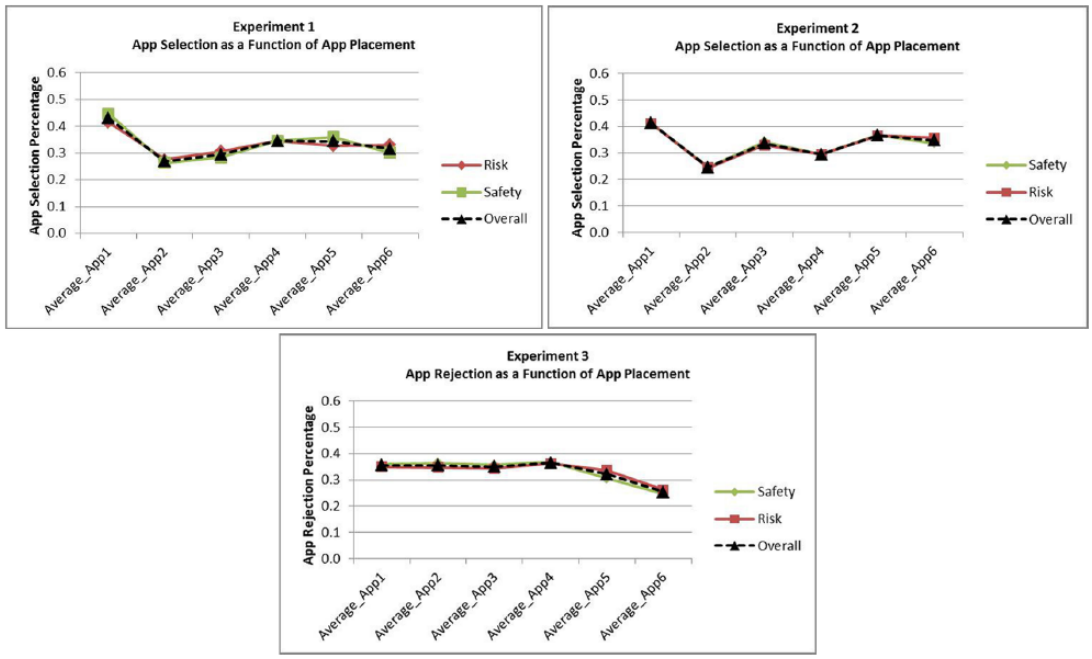

Figure 5. App selection percentage (Experiments 1 and 2) and rejection percentage (Experiment 3) as a function of the placement of the apps.

in Experiments 2 or 3, so we suspect it is a Type 1 error.

We also considered the placement of the app in the list of six in each task. A common pattern for all six tasks was that the first app was selected more often than the second app and, on average, more often than any of the other apps, for which the selection percentage did not differ much (see Figure 5, top left panel). This result is consistent with the take-the-first heuristic in decision making (Johnson & Raab, 2003), and it could also be due to the first app in the list being (perceived as) more popular among the users.

Subjective reasons for app selection. For each app that was selected, participants were to mark the reason(s) why they selected it. A significant positive correlation was found between the safety score of the app and the reason "permission safety" for most of the apps (33 out of 36 apps; Pearson's rs $\geq .350$, Ns = 25~79, ps $\leq .042$), and a significant negative correlation between the risk score and the reason "permission risk" for most of the apps (34 out of 36 apps; $|r|$s $\geq .298$, Ns = 25~84, ps $\leq .026$). Table 3 shows percentage of the reasons marked by the participants in safety and risk conditions. The difference between the percentages of selecting "permission safety" and "permission risk" was significant, $X^2(1, N = 3{,}540) = 68.55$, $p < .001$, as well as the difference for user rating count, $X^2(1, N = 3{,}540) = 13.39$, $p < .001$, and description, $X^2(1, N = 3{,}540) = 6.72$, $p = .010$. No other differences were significant, ps $\geq .215$.

After the participants finished all six tasks, they were asked what factors they considered when selecting apps during the tasks. Table 3 shows the percentage with which each factor was considered. The difference between the percentages of selecting "permission safety" and "permission risk" was significant, $X^2(1, N = 295) = 7.59$, $p = .006$. None of the differences for other reasons in the two conditions was significant, ps $\geq .084$.

Security/computer expertise. To compare whether expertise mediated the influence of safety/risk scores on app selections, the participants were divided into two groups based on their responses to the security expertise

TABLE 3: Percentage of Reasons Chosen as a Function of Safety or Risk Condition When Selecting/Rejecting Apps During the Tasks and in the Posttask Questionnaire

	Experiment 1				Experiment 2				Experiment 3			
	During Task		Posttask		During Task		Posttask		During Task		Posttask	
Reasons	Safety	Risk	Safety	Risk	Safety	Risk	Safety	Risk	Safety	Risk	Safety	Risk
User rating score	79.1	77.4	96.5	96.1	72.3	75.7	92.3	96.4	66.4	66.7	97.9	99.0
User rating count	53.9	47.8	83.2	77.0	46.9	50.1	77.3	81.0	24.0	23.6	63.6	63.5
Permission safety/risk	**47.2**	**33.6**	**75.5**	**60.5**[a]	**34.7**	**32.2**	**55.9**	**57.1**	**48.2**	**44.4**	**81.5**	**83.7**
Icon look and feel	37.8	37.4	66.4	60.5	32.0	31.1	55.5	57.9	16.5	19.2	47.2	50.7
Description	36.7	32.6	62.9	60.5	26.4	32.0	50.2	61.9	8.6	8.4	37.9	33.0
Familiarity with app or developer	9.3	9.0	40.6	30.9	8.0	7.5	28.3	28.3	2.3	3.1	11.8	17.7
Other	1.2	1.6	3.5	3.3	1.2	1.0	0.8	1.6	1.4	1.6	1.0	3.0

[a]Due to a technical issue, some of the participants in the risk condition saw "Permission Safety" for this question. This number was a sum of people who chose Permission Safety and those who chose Permission Risk (17.1+ 43.4).

TABLE 4: Percentage of App Selection (Experiments 1 and 2) and Rejection (Experiment 3) for Experts and Nonexperts in the Safety and Risk Conditions

	Experiment 1				Experiment 2				Experiment 3			
	Experts		Nonexperts		Experts		Nonexperts		Experts		Nonexperts	
Safety Level	Safety	Risk	Safety	Risk	Safety	Risk	Safety	Risk	Safety	Risk	Safety	Risk
1	19.5	18.6	15.9	29.3	17.6	31.0	26.2	32.2	51.5	49.6	59.7	45.8
2	17.1	30.1	21.1	33.2	25.3	28.5	25.0	28.8	38.6	35.4	36.4	33.6
3	38.4	39.2	31.0	29.0	33.1	30.2	31.6	30.1	31.1	31.0	28.2	29.2
4	39.0	36.2	44.4	34.7	38.6	36.4	40.8	35.6	27.1	32.5	25.1	32.8
5	53.7	46.7	55.6	39.0	43.2	36.9	40.2	34.7	23.4	27.8	21.2	29.3

question (see Table 1): One group was experts, participants who indicated highly skilled or security expert, and the other group was nonexperts, those who indicated regular user or computer novice. An ANOVA with expertise, safety/risk score, and safety/risk condition as within-subject (app) factors was conducted on the app-selection percentage. Expertise entered into an interaction with safety/risk score, $F(4, 140) = 5.35$, $p < .001$, $\eta_p^2 = .13$, but this was qualified by a three-way interaction of those two variables with safety/ risk condition, $F(4, 140) = 3.50$, $p = .009$, $\eta_p^2 = .09$. The experts were more sensitive than the nonexperts to the safety level but only in the risk condition (see Table 4), indicating that the experts were less susceptible to the framing of the information as risk or safety. Although the experts' selections were influenced more by the permission safety/risk information, this was not reflected in their marked reasons for app selection: Correlational analyses between the participants' security expertise and the reason "permission safety/risk" showed no significant correlation for 33 of the 36 apps, ps > .05.

Other questionnaire analyses. When asked whether they found the overall permission safety/risk information to be useful, most participants indicated that it was useful. On a 7-point scale, with 1 denoting not useful and 7 extremely useful, 64.3% of participants gave a rating of 5 and above in the safety condition, whereas 59.9% of participants did in the risk condition.

To determine whether the participants understood the permission safety/risk symbols, we showed them four full circles and asked what that symbol stood for. In the safety condition, 86.0% of the participants gave a correct answer, 6.3% gave an opposite answer, and 7.7% indicated that they did not know what it meant; in the risk condition, 63.2% of the participants gave a correct answer, 23.0% gave an opposite answer, and 13.8% marked that they did not know what it meant. Due to a technical issue, some participants saw the opposite color of what they saw during the task. This issue may have led to confusion in answering the question, as suggested by some participants' comments (e.g., "For the last question #4, it is unclear because all the apps had green circles indicating a safe rating but in #4 they are red. So it is confusing.").

EXPERIMENT 2

For the more realistic scenario of Experiment 1 than in our previous study (Gates, Chen et al., 2014), in which choice was between six and two apps, respectively, there was a substantial benefit of the summary score being presented as safety rather than risk. Given the current emphasis on the need for researchers to "integrate replications into their scholarly habits" (Brandt et al., 2014, p. 217), one goal of Experiment 2 was to confirm the reliability of this safety benefit. We intentionally introduced a confound between the safety versus risk variable and symbol color (green vs. red) in Experiment 1, so we designed Experiment 2 to eliminate the

color difference, using a neutral color "blue" for both safety and risk. Also, we employed a more tightly controlled design in which sets of risk and safety conditions had equivalent user rating and risk/safety information, except for whether the latter was specified as risk or safety, to allow for direct comparison between the two conditions. Finally, due to some of the confusion about the risk/safety scores shown in the subjective responses, we made several minor methodological changes to the interface to improve clarity (see the Method section for details).

Method

Participants. A total of 494 participants were recruited through MTurk. On average, the experiment again took 10 minutes to complete, and participants' were paid $0.75 each.

Materials and procedure. The materials and procedure were similar to those of Experiment 1, except as follows: (1) On the introductory page, the last sentence describing the permission safety/risk (i.e., "The higher the permission risk, the more permissions the app is requesting relative to other apps.") was deleted, due to its potential to confuse the participants about the actual meaning of permission safety/risk. (2) The color of the symbols representing permission safety/risk was controlled to be the same (dark blue). (3) To ensure comparability of the risk and safety conditions, 10 sets of randomly generated numbers of user ratings and safety rankings were used in both the safety and risk versions of the tasks. (4) For one of the final questions—"What does a rating of [four full circles] stand for?"—the four circles were changed to five circles to match the display in the tasks. Because the circles in both conditions were blue, the issue of presenting a mismatching color in Experiment 1 was also remedied. (5) The issue of some participants in the risk condition seeing "permission safety" was also eliminated.

Results and Discussion

App-selection analysis. The correlation between app selection, user rating, and safety/risk score for each app was not able to reflect the real relation between them because for each set of tasks the user rating and safety/risk score were fixed. Thus, data from all 36 apps were combined for the correlational analyses. Overall, the results were similar to those in Experiment 1. In the safety condition, there was a positive correlation between user rating and app selection, $r = .367$, $N = 8{,}886$, $p < .001$, and a positive correlation between safety score and app selection, $r = .157$, $N = 8{,}886$, $p < .001$. In the risk condition, there was a positive correlation between user rating and app selection, $r = .410$, $N = 8{,}892$, $p < .001$, and a negative correlation between risk score and app selection, $r = -.059$, $N = 8{,}892$, $p < .001$. Note that this last negative correlation was not significant in Experiment 1, which had fewer participants, even though the value was numerically larger. There is likely a very weak correlation between risk score and app selection that was significant in this experiment due to the larger sample size and power.

Each app was again treated as one "participant" in the ANOVA. A univariate approach was used wherein the user rating and safety/risk score were treated as between-subjects (apps) factors, because each app only underwent some of the combinations of user rating and safety/risk score due to the use of the 10 sets of them. The ANOVA with safety/risk condition (safety vs. risk), user rating (1 through 5), and safety ranking (1 through 5; safety ranking = safety score in the safety condition; safety ranking = 6 – risk score in the risk condition) as between-subjects (apps) factors was conducted for the percentage of app selection.

The same result patterns were found as in Experiment 1. First, the percentage of app selection did not differ between the safety and risk conditions (Ms = 27.3% vs. 26.4%), $F < 1.0$, but the two conditions showed different trends across different safety rankings (see Figure 3, top right panel), $F(4, 540) = 2.37$, $p = .052$, $\eta_p^2 = .02$. Post hoc pairwise comparisons showed that the safety and risk conditions only differed when the safety ranking = 5, $p = .023$, but not when it had other values, $ps \geq .155$, although the data pattern was similar to that in Experiment 1. The result patterns conform to the proposition that users in the safety condition made safer (less risky) decisions than those in the risk condition.

Second, overall, the percentage of app selection was higher with increased safety ranking (*M*s = 20.6%, 22.1%, 25.9%, 32.3%, and 33.4% for safety rankings 1 through 5), $F(4, 540) = 13.83$, $p < .001$, $\eta_p^2 = .09$, and it was also higher with increased user rating (*M*s = 7.1%, 9.9%, 17.6%, 38.6%, and 61.0% for user ratings 1 through 5), $F(4, 540) = 217.47$, $p < .001$, $\eta_p^2 = .62$. Again, safety ranking interacted with user rating, $F(16, 540) = 3.47$, $p < .001$, $\eta_p^2 = .09$, and this interaction did not differ across the safety and risk conditions, $F < 1.0$. Post hoc pairwise comparisons for the interaction between safety ranking and user rating (see Table 2) showed that safety rankings did not influence app selection for apps with lower user ratings (1, 2, and 3), but increased safety ranking led to greater selection percentage for apps with higher user ratings (4 and 5). These results are similar to those in Experiment 1 and indicate that the apps with higher user ratings and higher safety or lower risk scores were selected more than other options (see Figure 4, center row).

The interaction between safety/risk condition and user rating in Experiment 1 did not show up in Experiment 2, $F < 1.0$. Thus, that interaction does not appear to be reliable. When considering the placement of the six apps in each task, the same pattern showed as in Experiment 1: The first app was selected more often than the second one in each task, and on average, the first app was also selected more often than other apps (see Figure 5, top right panel).

Subjective reasons for app selection. For the reasons marked while selecting an app, a significant positive correlation was found between the safety score of the app and the reason "permission safety" for most of the apps (33 out of 36 apps; Pearson's $rs \geq .227$, $Ns = 45$~125, $ps \leq .030$) in the safety condition, and a significant negative correlation between the risk score and the reason "permission risk" for most of the apps (33 out of 36 apps; $|r|s \geq .240$, $Ns = 40$~137, $ps \leq .012$) in the risk condition. Table 3 shows percentage of the reasons marked by the participants in safety and risk conditions. The difference between the percentages of selecting "permission safety" and "permission risk" was significant, $X^2(1, N = 5{,}922) = 4.19$, $p = .041$, as well as the difference for user rating score, $X^2(1, N = 5{,}899) = 9.20$, $p = .002$, user rating count, $X^2(1, N = 5{,}904) = 6.12$, $p = .013$, and icon look and feel, $X^2(1, N = 5{,}912) = 23.33$, $p < .001$.

Regarding the reasons selected in the post-task questionnaire (see Table 3), none of the differences between the percentages of selecting other reasons in the two conditions was significant, $ps \geq .319$, except user rating score, $X^2(1, N = 494) = 3.79$, $p = .052$, and description, $X^2(1, N = 494) = 6.91$, $p = .009$.

Security/computer expertise. An analysis of selections with expertise as a variable similar to that of Experiment 1 was conducted. None of the terms that included expertise even approached being significant, *F*s < 1. However, the tendencies of the mean values were consistent with the result of Experiment 1 that the experts tended to be more sensitive than the nonexperts to the safety level (see Table 4). The mean tendencies did not show any sign that the experts were less affected than the nonexperts by the safety/risk framing. As in Experiment 1, there was no correlation between security concern and expertise for 33 out of 36 apps, *p*s > .05. The main methodological difference from Experiment 1 was use of blue symbols to convey risk and safety, rather than red symbols and green symbols, respectively.

Other questionnaire analyses. When asked whether they found the overall permission safety/risk information to be useful, more than half the participants indicated that it was useful. On a 7-point scale, with 1 denoting not useful and 7 extremely useful, 56.4% participants gave a rating of 5 and above in the safety condition, and 52.9% participants did in the risk condition.

With the technical issues of Experiment 1 corrected, 89.5% of the participants in the safety condition gave a correct answer as to what that display stood for, 3.2% gave an opposite answer, and 7.3% indicated that they did not know what it meant. Participants in the risk condition continued to evidence more confusion, with 71.3% giving the correct answer, 19.0% the opposite answer, and 9.7% marking that they did not know what the symbols meant. Chi-squared analysis showed that participants in the safety condition understood the symbols better than those in the risk condition, $X^2(1, N = 494) = 25.98$, $p = .006$. Thus, without the technical

issues of Experiment 1, participants still showed better understanding of the symbols in the safety condition than in the risk condition.

EXPERIMENT 3

In Experiments 1 and 2, we found that a summary score promoted better app-selection decisions when framed as safety rather than risk. This advantage of safety framing could be due to several factors, including that the safety symbols obey the rule "the more the better," the safety score is more compatible with the user ratings (for which more filled stars indicates better), and the safety score is more compatible than the risk score with the task of choosing apps. We examined this latter possibility in Experiment 3 by changing the task to one of rejecting two of the apps from the list. If the framing of the score as safety rather than risk was better in Experiment 2 due to the compatibility of safety with the task of choosing apps, this benefit should not be evident in Experiment 3 for which the task is one of rejecting apps.

Method

A total of 398 participants were recruited. Experiment 3 was conducted similarly to Experiment 2, except that any wording of "select" or "choose" was changed to "reject."

Results and Discussion

App-rejection analysis. For the same reason as in Experiment 2, data from all 36 apps were combined for the correlational analyses. Overall, the correlational results were similar to those in Experiment 2. In the safety condition, there was a negative correlation between user rating and app rejection, $r = -.443$, $N = 7{,}014$, $p < .001$, and a negative correlation between safety score and app rejection, $r = -.253$, $N = 7{,}014$, $p < .001$. In the risk condition, there was a negative correlation between user rating and app rejection, $r = -.461$, $N = 7{,}308$, $p < .001$, and a positive correlation between risk score and app rejection, $r = .114$, $N = 7{,}308$, $p < .001$.

ANOVAs with user rating (1 through 5), safety ranking (1 through 5; safety ranking = safety score in the safety condition; safety ranking = 6 – risk score in the risk condition), and safety condition (safety vs. risk) as between-subjects (apps) factors were conducted for the percentage of app rejection. The result patterns were similar to those of Experiments 1 and 2. First, the percentage of app rejection did not differ in the safety and risk conditions on average (*M*s = 42.4%), $F < 1.0$, but safety and risk conditions showed distinct patterns across different safety rankings (see Figure 3, bottom), $F(4, 540) = 4.98$, $p = .001$, $\eta_p^2 = .04$. Post hoc pairwise comparisons showed that the differences between the safety/risk conditions were significant for the safety rankings 5, 4, and 1, *p*s = .042, .030, and .001, but not for safety rankings 3 and 2, *p*s = .861 and .404. The result pattern conforms to the proposition that users in the safety condition made better decisions than those in the risk condition.

Second, overall, the percentage of app rejection was lower with increased safety ranking (*M*s = 61.9%, 43.1%, 38.3%, 37.6%, and 31.1% for safety rankings 1 through 5), $F(4, 540) = 41.39$, $p < .001$, $\eta_p^2 = .24$, and it was also lower with increased user rating (*M*s = 74.2%, 63.9%, 40.0%, 20.2%, and 13.7% for user ratings 1 through 5), $F(4, 540) = 192.83$, $p < .001$, $\eta_p^2 = .59$. Again, the safety ranking interacted with user rating, $F(16, 540) = 2.92$, $p < .001$, $\eta_p^2 = .08$, and this interaction did not differ for the safety and risk conditions, $F < 1$. Post hoc pairwise comparisons for the two-way interaction (see Table 2) showed that decreased safety ranking led to more app rejections, but this trend was more significant for apps with higher user ratings (3, 4, and 5) than for apps with lower user ratings (1 and 2). These results are similar to those in Experiments 1 and 2 and indicate that the apps with higher user ratings and higher safety or lower risk scores were rejected less than other options (see Figure 4, bottom row).

The interaction between safety/risk condition and user rating, which was significant in Experiment 1 but not Experiment 2, did not show up in Experiment 3, $F < 1.0$. Thus, that interaction, which was not very evident in the mean data, does not appear to be reliable.

When considering the placement of the six apps in each task, the app-rejection tasks did not show a consistent pattern across all six tasks. On average, the phenomenon that the first app was selected more often than other apps in the selection task was not evident in the rejection task. Rather, the first four apps had a similar rejection

rate, which was higher than that of the last two apps (see Figure 5, bottom panel). This different pattern could be due to two possible reasons: (1) Different processes are involved in the selection and rejection tasks, and (2) the bias of the take-the-first heuristic to reject the first app in the list was canceled out in the present experiment by its being (perceived as) more popular among the users.

Subjective reasons for app rejection. For the reasons marked while rejecting an app, a significant negative correlation was found between the safety score of the app and the reason "permission safety" for most of the apps (30 out of 36 apps; Pearson's $|r|s \geq .258$, Ns = 23~97, $ps \leq .049$) in the safety condition, and a significant positive correlation between the risk score and the reason "permission risk" for most of the apps (30 out of 36 apps; $rs \geq .227$, Ns = 33~86, $ps \leq .037$) in the risk condition. Table 3 shows percentages of the reasons marked in the safety and risk conditions. None of the differences between the percentages of selecting the reasons was significant, $ps \geq .074$, except that for icon look and feel, $X^2(1, N = 4{,}772) = 5.95$, $p = .016$.

Regarding the reasons selected in the posttask questionnaire (Table 1), none of the differences between the percentages of selecting other reasons in the two conditions was significant, ps $\geq .346$.

Security/computer expertise. Similar to Experiment 2, an ANOVA of rejection percentage with expertise as a factor showed no significant terms that included expertise, Fs < 1. The mean values showed no sign of the experts being more sensitive than the nonexperts to the safety/risk information (see Table 4), but there was a tendency for their rejections to be affected less by the safety/risk framing. Again, there was no correlation between security concern and expertise for 32 of the 36 apps, ps > .05.

Other questionnaire analyses. More than half of the participants indicated they found the overall permission safety/risk information to be useful. On a 7-point scale, with 1 denoting not useful and 7 extremely useful, 68.2% participants gave a rating of 5 and above in the safety condition, and 71.9% participants did in the risk condition.

Regarding the question of whether the participants understood the safety/risk symbols, in the safety condition, 89.7% of the participants gave a correct answer, 3.1% gave an opposite answer, and 7.2% indicated that they were not sure what it meant; in the risk condition, 75.9% of the participants gave a correct answer, 19.7% gave an opposite answer, and 4.4% marked that they were not sure. Chi-squared analysis showed that participants in the safety condition understood the symbols better than those in the risk condition, $X^2(1, N = 398) = 13.37$, $p < .001$.

GENERAL DISCUSSION

Progress is being made toward development of summary risk scores for improving the security of mobile applications. Methods have been developed to generate risk scores based on machine learning techniques that can identify certain apps as risky and others as less risky (Gates, Li et al., 2014). To be effective, though, this summary risk information must be presented to users in a way that they can comprehend and that will cause their app-installation decisions to be less risky. The present study demonstrates, in a relatively lifelike scenario, that people's app-installation decisions can be affected by summary risk/safety scores. In all three experiments, less risky apps tended to be chosen over more risky ones, more so when the score was framed as amount of safety rather than amount of risk. Although the risk/safety score influenced app installation, it did so less than the user ratings. When the user rating was high, the risk/safety score exerted the most effect. But when the user rating for an app was low, having a low risk or high safety score did not typically lead to selection of that app. Consistent with the decision data, the risk/safety score was reported as a reason for selecting or rejecting an app by participants less often than user ratings, but more often than other app elements such as icon and description.

There are two general types of reasons why framing the decision as one of safety was more effective than framing it as one of risk. The first type is one of the safety score being more compatible than the risk score with some aspect of the decision context (Shafir, 1995). The tasks of Experiments 1 and 2 required selections to be made, which is a positive decision in that two apps were chosen as being the most desirable of the six alternatives. The safety framing of the information, for which "more" means "better,"

is more compatible with the task goal of determining the two best apps in a selection task. However, the benefit for the safety score in Experiment 3, which required rejection of two apps, suggests that task compatibility was not a significant factor in the present context. Another possibility is that the safety frame is more compatible with the population stereotype for scores, for which a higher number most often indicates better. Furthermore, the safety frame is more compatible with the user ratings than is the risk frame, for which "more" equals "worse." These last two compatibility relations could contribute to the benefit for safety framing in Experiment 3 as well as in Experiments 1 and 2.

The second type of explanation is that, safety differs from risk in more than valence, although dictionaries and thesauruses identify them as antonyms. Safety seems to be a holistic concept in that people rarely talk about dimensions of safety. In contrast, risk seems to be more multidimensional in that it is customary to decompose overall risk into distinct risks. This difference is illustrated by article titles using "safety" singular but "risks" plural (e.g., see Livingstone, Haddon, Görzig, & Ólafsson, 2011). Thus, people may tend to think of overall safety but components of risk.

In agreement with the app choice data, participants showed more confusion about the risk score than the safety score. When asked whether a symbol with all circles filled indicated high safety or high risk, more participants answered incorrectly in the risk condition than in the safety condition. This confusion about the meaning of the risk symbol was not a result of the task requiring selection of apps, as in Experiments 1 and 2, because the confusion was also evident in Experiment 3 when the task was to reject apps. In addition, in Experiment 1, compared to the safety condition, participants in the risk condition stated after their individual choices and in the final questionnaire that they did not rely on that information as much. However, this result was not replicated in Experiments 2 and 3. Regardless of the reliability of these subjective judgments, the more objective symbol-identification data show the greater confusion for the risk symbols that would be expected from risk being a multidimensional concept.

To examine whether the greater confusion for the risk symbols accounted for their disadvantage in app choices, we conducted follow-up analyses of the app-choice data for each experiment comparing the original data based on all participants to data based only on those participants who identified the symbols correctly. Of interest was whether the Safety/Risk Condition × Safety Level interaction (indicative of the advantage for safety symbols) differed across the two data sets. For Experiment 1, that interaction was smaller for the participants who correctly identified the symbols than for all participants, $F(4, 280) = 2.69$, $p = .031$, $\eta_p^2 = .04$, but it was still significant, $F(4, 140) = 10.17$, $p < .001$, $\eta_p^2 = .23$. For Experiment 2, there was no significant difference in the Safety/Risk Condition × Safety Level interaction between the two data sets, $F < 1$, although that interaction only approached significance for the correct-identification data set, $F(4, 140) = 2.30$, $p = .061$, $\eta_p^2 = .06$. For Experiment 3, the Safety/Risk Condition × Safety Level interaction tended to be smaller for the correct-identification data set than for all participants, $F(4, 280) = 2.05$, $p = .088$, $\eta_p^2 = .03$, and was still significant for the former data set alone, $F(4, 140) = 5.50$, $p < .001$, $\eta_p^2 = .14$. In summary, when participants who did not identify the symbols correctly were omitted, the safety score still led to more secure (less risky) app-installation decisions than did the risk score, although the difference between the two framings tended to decrease. Thus, the disadvantage of the risk score for app-installation decisions cannot be attributed solely to the greater confusion regarding the meaning of the risk symbols.

We previously reported an MTurk experiment in which participants had to select which one of two apps to install (Gates, Chen et al., 2014). The results showed only a slight tendency for a safety framing to be better than a risk framing. In the present study, for which participants were required to select two out of six apps, safety scores influenced performance considerably more than risk scores. Because the conditions of the prior experiment and the present ones differed in several ways, including the amount of information presented, we are not able to determine definitively the basis for the difference in results. It likely is due in part to the greater information-processing demands of comparing multiple alternatives than for making binary comparisons (Luce, 1986). Because consideration of multiple alternatives is part of most

app decisions, and the format used to present the alternative apps in the present study is more similar to what would be seen when actually installing an app, the present results can be regarded as more ecologically valid to the mobile computing environment. Thus, an applied implication of our findings is that when risk/safety summary information about apps is provided, this information should be in the form of a safety score, though further research is needed to develop the methods of generating proper safety scores (e.g., Gates, Li et al., 2014; Peng et al., 2012).

Although the general patterns of results in Experiments 1 and 2 were similar, the influence of the safety/risk scores was larger in Experiment 1. A between-experiment ANOVA showed that this difference was statistically significant, yielding an Experiment × Safety/Risk Score interaction, $F(4, 280) = 3.17$, $p = .014$, $\eta_p^2 = .04$. Additionally, those two variables interacted with safety/risk condition, $F(4, 280) = 5.08$, $p = .001$, $\eta_p^2 = .07$, reflecting that the benefit of the safety frame was larger in Experiment 1 than in Experiment 2. The larger effect of the safety/risk scores in Experiment 1 suggests that red and green colors were more effective at signaling risk and safety, respectively, than was the dark blue color. Whether the difference in effectiveness is replicable and due to the stereotypic mapping of colors to risk and safety in Experiment 1, or to red and green being more salient than dark blue, remains to be determined.

The performance data for rejecting apps in Experiment 3 were similar to those for selecting apps in Experiments 1 and 2 in that the safety score had more influence on performance than the risk score. This finding suggests the possibility that participants converted the rejection task to one of selecting which apps not to reject (e.g., Meloy & Russo, 2004). However, Shafir (1993) found that such conversion was used to reduce the number of decisions that had to be made, whereas in Experiment 3, altering the task to one of selection would increase the number of decisions from two to four. Moreover, other aspects of the data differed from those of Experiments 1 and 2. Those experiments showed a primacy effect, that is, a bias to choose the app in the first position (the take-the-first heuristic), but Experiment 3 did not. Also, participants in Experiment 3 indicated that they were placing greater reliance on the permission risk/safety than did those in Experiments 1 and 2. Thus, the rejection decisions of Experiment 3 apparently involved somewhat different strategies than the selection decisions of Experiments 1 and 2.

That presenting a summary risk/safety score facilitates users' secure app-installation decisions has theoretical and practical implications. Theoretically, it is consistent with fuzzy trace theory (Reyna, 2008; Reyna & Brainerd, 1995), according to which people have two types of mental representations, gist and verbatim. For reasoning and decision-making, most people rely on the gist representations to make decisions (Brainerd & Renya, 2002). The summary score serves as a basis for users to perform the gist processing of the overall risk associated with an app, and our results suggest that framing the summary score as one of safety may be especially effective. Also, in practice, this idea of presenting a summary risk/safety score fits with a general design principle that an effective interface should be direct and not overburden users with too much cognitive processing (e.g., Krug, 2000). We also found that the experts were influenced at least as much as the nonexperts by the summary scores in all three experiments. Overall, our results fit with Brust-Renck et al.'s (2013, p. 244) conclusion that in a variety of contexts "risk communication should convey the bottom-line (gist) message of risk rather than only the facts to help people make informed decisions."

ACKNOWLEDGMENTS

A preliminary version of this work including only Experiment 1, analyzed differently, was presented at the 2014 annual meeting of the Human Factors and Ergonomics Society and is included in the proceedings. 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.

REFERENCES

Barnes, M. J., McDermott, P. L., Hutchins, S., & Rothrock, L. (2011). Framing, loss aversion, and visualization of risk for a dynamic simulation environment. *Journal of Cognitive Engineering and Decision Making*, *5*, 294-308.

Bergum, B. O., & Bergum, J. A. (1981). Population stereotypes: An attempt to measure and define. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *25*, 662-665.

Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. *Advances in Child Development and Behavior*, *28*, 41-100.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., … van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217-224.

Brust-Renck, P. G., Royer, C. E., & Reyna, V. F. (2013). Communicating numerical risk: Human factors that aid understanding in health care. *Reviews of Human Factors and Ergonomics*, *8*, 235-276.

Chernev, A. (2009). Choosing versus rejecting: The impact of goal-task compatibility on decision confidence. *Social Cognition*, *27*, 249-260.

Chin, E., Felt, A. P., Sekar, V., & Wagner, D. (2012). Measuring user confidence in smartphone security and privacy. In *Proceedings of the Eighth Symposium on Usable Privacy and Security* (pp. 1-16). New York: ACM.

Cisco Systems, Inc. (2014, Jan 16). *Cisco 2014 annual security report*. Retrieved February 18, 2014, from https://www.cisco.com/web/offer/gist_ty2_asset/Cisco_2014_ASR.pdf.

eMarketer. (2014). *Worldwide smartphone usage to grow 25% in 2014*. Retrieved from http://www.emarketer.com/Article/Worldwide-Smartphone-Usage-Grow-25-2014/1010920

Felt, A. P., Greenwood, K., & Wagner, D. (2011). The effectiveness of application permissions. In *Proceedings of the 2nd USENIX conference on Web application development* (pp. 75-86). Berkeley, CA: USENIX Association.

Felt, A. P., Ha, E., Egelman, S., Haney, A., Chin, E., & Wagner, D. (2012). Android permissions: User attention, comprehension, and behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security* (pp. 1-14). New York: ACM.

Frings, D. (2012). The effects of sleep debt on risk perception, risk attraction and betting behavior during a blackjack style gambling task. *Journal of Gambling Studies*, *28*, 393-403.

Gambara, H., & Piñon, A. (2005). A meta-analytic review of framing effect: Risky, attribute and goal framing. *Psicothema*, *17*, 325-331.

Garcia-Retamero, R., & Dhami, M. K. (2013). On avoiding framing effects in experienced decision makers. *The Quarterly Journal of Experimental Psychology*, *66*, 829-842.

Gates, C., Chen, J., Li, N., & Proctor, R. W. (2014). Effective risk communication for Android Apps. *IEEE Transactions on Dependable and Secure Computing*, *11*, 252-265.

Gates, C., Li, N., Peng, H., Sarma, B., Qi, Y., Potharaju, R., Nita-Rotaru, C., & Mollay, I. (2014). Generating summary risk scores for mobile applications. *IEEE Transactions on Dependable and Secure Computing*, *11*, 238-251.

Huber, O., Huber, O. W., & Bär, A. S. (2014). Framing of decisions: Effect on active and passive risk avoidance. *Journal of Behavioral Decision Making*, *27*, 444-453.

Hung, K., & Tangpong, C. (2010). General risk propensity in multifaceted business decisions: Scale development. *Journal of Managerial Issues*, *22*, 88-106.

Johnson, J. G., & Raab, M. (2003). Take the first: Option-generation and resulting choices. *Organizational Behavior and Human Decision Processes*, *91*, 215-229.

Kelley, P. G., Consolvo, S., Cranor, L. F., Jung, J., Sadeh, N., & Wetherall, D. (2012). A conundrum of permissions: Installing applications on an android smartphone. In *Financial cryptography and data security* (pp. 68-79). Berlin: Springer.

Kelley, P. G., Cranor, L. F., & Sadeh, N. (2013). Privacy as part of the app decision-making process. In *Proceedings of the 2013 ACM annual conference on Human Factors in Computing Systems* (pp. 3393-3402). New York: ACM.

Krug, S. (2000). *Don't make me think: A common sense guide to Web usability*. Berkeley, CA: New Riders.

Lai, Y. L., & Hui, K. L. (2006). Internet opt-in and opt-out: Investigating the roles of frames, defaults and privacy concerns. In *Proceedings of the 2006 ACM SIGMIS CPR conference on computer personnel research: Forty four years of computer personnel research: achievements, challenges & the future* (pp. 253-263). New York: ACM.

Lin, J., Amini, S., Hong, J. I., Sadeh, N., Lindqvist, J., & Zhang, J. (2012). Expectation and purpose: Understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 501-510). New York: ACM.

Livingstone, S., Haddon, L., Görzig, A., & Ólafsson, K. (2011). *Risks and safety on the internet: The UK report*. London: EU Kids Online.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Mansfield-Devine, S. (2012). Paranoid Android: Just how insecure is the most popular mobile platform*? Network Security*, *2012*(9), 5-10.

Mayhorn, C. B., Fisk, A. D., & Whittle, J. D. (2002). Decisions, decisions: Analysis of age, cohort, and time of testing on framing of risky decision options. *Human Factors*, *44*, 515-521.

McLaughlin, A., & Mayhorn, C. B. (2014). Designing effective risk communications for older adults. *Safety Science*, *61*, 59-65.

Meloy, M. G., & Russo, J. E. (2004). Binary choice under instructions to select versus reject. *Organizational Behavior and Human Decision Processes*, *93*, 114-128.

Nagpal, A., & Krishnamurthy, P. (2008). Attribute conflict in consumer decision making: The role of task compatibility. *Journal of Consumer Research*, *34*, 696-705.

O'Toole, J. (2014, February). Mobile apps overtake PC Internet usage in U.S. *CNN Money*. Retrieved from http://money.cnn.com/2014/02/28/technology/mobile/mobile-apps-internet/.

Peng, H., Gates, C., Sarma, B., Li, N., Qi, A., Potharaju, R., Nita-Rotaru, C., & Molloy, I. (2012). Using probabilistic generative models for ranking risks of Android apps. In *Proceedings of the 2012 ACM conference on computer and communications security* (pp. 241-252). New York: ACM.

Reyna, V. F. (2008). A theory of medical decision making and health: Fuzzy trace theory. *Medical Decision Making*, *28*, 850-865.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*, 1-75.

Rubaltelli, E., Dickert, S., & Slovic, P. (2012). Response mode, compatibility, and dual-processes in the evaluation of simple gambles: An eye-tracking investigation. *Judgment & Decision Making*, *7*, 427-440.

Schwartz, A. (2011). Deceiving and informing: The risky business of risk perception. *Medical Decision Making*, *31*, 378-379.

Shabtai, A., Fledel, Y., Kanonov, U., Elovici, Y., Dolev, S., & Glezer, C. (2010). Google android: A comprehensive security assessment. *IEEE Security & Privacy*, *8*(2), 35-44.

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, *21*, 546-556.

Shafir, E. (1995). Compatibility in cognition and decision. In J. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *The psychology of learning and motivation* (Vol. 32, pp. 247-274). San Diego, CA: Academic Press.

Slovic, P., Griffin, D., & Tversky, A. (1990). Compatibility effects in judgment and choice. In R. M. Hogarth (Ed.), *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 5-27). Chicago: University of Chicago Press.

Toch, E., Wang, Y., & Cranor, L. (2012). Personalization and privacy: A survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction*, *22*, 203-220.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453-458.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, *59*, S251-S278.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review*, *95*, 371-384.

Jing Chen received her BS and MEd degrees in cognitive psychology from Zhejiang University in China, in 2007 and 2010, respectively. She is currently working toward her PhD degree in cognitive psychology and her MS degree in industrial engineering at Purdue University.

Christopher S. Gates received his BS degree in computer science as well as in mathematics and his MS degree in computer science, both from Rutgers University in 2002 and 2005, respectively. After this, he worked in industry for several years until 2009, when he returned to academia. He received his PhD degree in computer science from Purdue University in 2014.

Ninghui Li received a BEng degree in computer science from the University of Science and Technology of China in 1993 and MSc and PhD degrees in computer science from New York University, in 1998 and 2000, respectively. He is currently a 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 an MA degree and PhD degree in experimental psychology from the University of Texas at Arlington, in 1972 and 1975, respectively. 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.