Real-World Decision Making: Logging Into Secure vs. Insecure Websites

Timothy Kelley
Indiana University, Bloomington
kelleyt@indiana.edu

Bennett I. Bertenthal
Indiana University, Bloomington
bbertent@indiana.edu

***Abstract*—A novel Two-Alternative Forced Choice experiment was used to evaluate the effects of security indicators on participants' decision making when identifying potentially risky websites. Participants recruited from Amazons Mechanical Turk were instructed to visit a series of secure and insecure websites, and decide as quickly and as accurately as possible whether or not it was safe to login. Hierarchical linear regression models were used to identify the importance of the presence of security indicators, security domain knowledge, and familiarity with the presented websites to correctly differentiate between secure and insecure websites. An analysis of participants' mouse trajectories was used to assess how websites were searched before a decision was made. The likelihood to login was modulated by security domain knowledge and familiarity with websites. The mouse tracking data revealed that spoofed websites with security indicators resulted in less search on the website, especially when the browser chrome indicated extended validation. Taken together, these results suggest that participants are aware of security indicators, but their responses are modulated by multiple factors.**

I. Introduction

Users on the Internet must regularly decide whether it is safe to enter their personal information on websites with very different interfaces. While web browsing behaviors are fairly regular, there are plenty of methods for users to reach new websites, or seemingly safe, familiar sites [1]. Mistyped URLs, search engine links, and clicking on links in email can all direct users to potentially malicious websites. Modern web browsers provide security indicators (e.g., the protocol used, the domain name, the SSL/TLS certificate, and visual elements in the browser chrome) to assist users in their decision making, but the technical details of these indicators are arcane for most users.

Studies examining users' attention to web browser security indicators have painted a grim picture. Most users are unaware of available security cues, and those that are aware often choose to ignore the indicators. Users' decision making process is further complicated by the way the indicators display the security information [2]. The design and implementation of websites can also lead to confusion (e.g., a bank that forwards customers to a new domain for logins) [3]. Additionally, users' familiarity with a given website may lead them to trust the website more than the warnings presented by the browser [4].

Empirical studies focus on a simple "yes" or "no" response (i.e., subjects use or do not use the indicators) or some more nuanced version (e.g., how much time subjects spend looking at the security indicators) [5], [6]. One limitation of these paradigms is that correct and incorrect login decisions could be due to more than one source of knowledge or decision process. For example, an incorrect decision might occur because a user is presented with a familiar website that they log into frequently, and thus fails to attend to the absence of security indicators or decides nevertheless to login because it has been safe in past transactions [7]. It is now possible to learn more about the underlying processes responsible for a user's decisions by measuring the mouse movements that occur on the screen prior to a final response.

Another limitation is the difficulty of replicating the experience of risk in an experimental environment. Some studies use roleplaying [5]. Other studies use priming to induce secure-like behavior to avoid exposing participants to real risks [8]. Another category of studies allow participants to engage in a simulation of normal browsing behavior in the lab [6], [8], [9]. Participants playing roles, even when primed, are unlikely to behave as securely as they would when they are personally at risk, particularly in a lab environment [10].

To address these two limitations, a within-subjects two-alternative forced choice paradigm on Amazon's Mechanical Turk was conducted. Participants were presented with simulated versions of real websites to examine the effects of security indicators on participants' decision making processes while they attempted to identify secure and insecure sites. In particular, this study was interested in how security indicators interact with participants' experience and knowledge with regard to their willingness to login to a website and the process participants use when arriving at that decision.

Monetary incentives and penalties were used to create a performance bonus based on both speed and accuracy. These economic incentives were designed to increase participants' motivation and risk taking behavior [11]. To study participants' decision making processes their choice to "login" or not was measured, but, by recording the mouse movements that were produced at each website it was also possible to measure the behavior leading to their responses.

USEC '16, 21 February 2016, San Diego, CA, USA

http://dx.doi.org/10.14722/usec.2016.23023

The results demonstrate that, rather than no sensitivity, participants' behavior reveal a complicated relationship with security indicators. When encryption appeared to be removed, participants' demonstrated a lack of awareness. When website URLs were manipulated, however, they relied primarily on the presence of locks to guide their responses.

Participants' decision making processes were also more complicated when website URLs were altered. When encryption-based indicators were manipulated, their familiarity with the website was the only predictor of their mouse tracking behavior. When URLs were changed, on the other hand, participants' mouse tracking behavior was affected by multiple factors such as their self-reported use of security indicators, their knowledge of the information security domain, and the presence (or absence) of encryption-based indicators.

II. Background

Modern web browsers use security indicators to provide information about the potential risks (e.g., whether a page is encrypted, users are at the intended location, and third-party vetting of domain ownership) involved in visiting a given web page. Empirical research is beginning to show, however, that users often fail to utilize these security indicators, demonstrating a problem in the way web browsers display security information.

Schecter, Dhamija, Ozment, and Fischer found that role-playing participants were more likely to login to a phished website than participants using their own accounts, but even participants using their own accounts ignored the absence of security indicators [5]. A study of the effectiveness of SSL warnings by Sunshine, Egelman, Almuhimdedi, Atri, and Cranor used survey data as well as simulated bank and library websites to identify participants' willingness to proceed to risky websites [12]. Whalen and Inkpen used eye tracking to identify participants' attention to security indicators, but found no security directed behavior without first priming participants [8].

An additional eye tracking study by Sobey, Biddle, Oorschot, and Patrick demonstrated how changes in design could enhance participants' attention to security indicators on a simulated commerce website [13]. Eye tracking done by Arianezhad, Camp, Kelley, and Stebila showed that task context was more important than security domain knowledge in participants' gazes to security indicators [9]. Alsharnouby, Alaca, and Chiasson found that, despite better designed security indicators, participants generally ignored the indicators, and were still unable to correctly identify phishing websites [6].

Most studies investigating users' decisions are confined to a single outcome variable: whether or not they login to a website or the amount of time they gaze at a particular indicator. Little is known about the decision processes responsible for the responses. An approach to studying these processes is to require mouse movements as the response.

Most traditional theories view the mind's cognitive and motor systems as independent, whereby the motor movement represents the final stage and end-result in a sequence of discrete processes. Recently, more dynamical views of cognitive processing have begun to challenge this view. Substantial evidence has accrued over the past decade to indicate that movements are continually updated by cognitive processing over time.

Early work by Goodale, Pelisson, and Prablanc showed that reaching movements updated continuously, rather than in discrete bursts [14]. Numerous studies, such as work by Freeman, Dale and Farmer, reveal that the process of categorizing a stimulus is shared over time with the motor cortex so that it continuously guides the response [15]. Furthermore, work by Song and Nakayama on human reaching movements suggest that multiple motor plans are prepared in parallel, and that these cascade over time until the human information processor arrives at a final response [16]. Cisek and Kalaska provide converging evidence from monkey studies revealing that motor cortical population codes are coupled to the decision process when the response involves a hand movement [17].

Especially relevant to the current research are studies revealing that high-level decision processes are revealed in the manual dynamics of the hand. McKinstry, Dale, and Spivey found that when participants were instructed to judge the truth of a question by moving their mouse to a "yes" or "no" location, questions of greater uncertainty were answered with a larger curvature and more fluctuations in the mouse trajectory than questions of less uncertainty [18]. Although some researchers might be inclined to use eye tracking rather than mouse tracking, the results will often be quite different because eye movements are rapid and ballistic as the observer saccades from one fixation to the next, whereas hand movements are continuously updated at a much finer level of precision. Thus, hand movements reveal the ongoing dynamics of cognitive processing and can capture the mind in motion with fine-grained temporal sensitivity.

Another advantage of mouse tracking is that it is possible to collect this data online through crowd sourcing sites, such as Amazon's Mechanical Turk [19]. Once trajectories are recorded, several measures can be computed, such as maximum deviation toward an unselected response (reflecting the conflict between the correct and incorrect response), area under the curve computed as the difference between the actual mouse trajectory and the straight-line trajectory between the start and terminus of the mouse movement, switches in direction, or movement complexity, all of which reflect a level of uncertainty in the decision making process. Any one or more of these measures can be used as a complement to the more straightforward percent correct measure in order to learn more about the cognitive processes underlying the decision process involved in risky decision making on the web.

III. Methodology

A. Participants

Participants were recruited from Amazon's Mechanical Turk under Internal Review Board (IRB) approval from Indiana University. The sample consisted of 214 participants ranging in age from 18- to 66-years-old ($\mu = 32.6, sd = 9.58$). Before data analysis, 41 participants (19.6%) were excluded due to lack of data, as they closed the web browsing task early, or did not complete it. During data analysis, one participant's data had to be removed from the mouse-tracking portion of the experiment as they used a keyboard or touch-pad to navigate

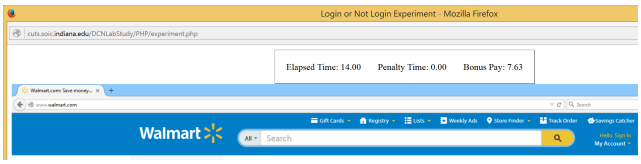


Fig. 1. Screenshot of the top portion of simulated browser chrome sitting in experimental window. The experimental clock is resting on top of the simulated browser chrome displaying the time elapsed (seconds), penalty time accrued (seconds), and the remaining bonus pay (dollars).

Fig. 2. Example website manipulations. On the left a sample eBay spoof (Bottom) compared with the legitimate eBay site (Top). On the right a login site with the encryption removed (Top) compared to the legitimate site with a standard certificate and encryption (Bottom).

between responses. Of the remaining participants there were 100 males and 72 females, primarily Caucasian with the same age distribution as the initial sample ($\mu = 32.6, sd = 9.6$). Most participants listed Firefox ($N = 84$) or Google Chrome ($N = 80$) as their primary browser.

B. Stimuli

This study required that participants use Firefox as their web browser. On each trial, they were presented with the front page of a website, and then had to click on the link (e.g., sign in, my account, login) that brought up the login page or dialog. This standardized the mouse-tracking data collection by presenting the logins on the second page, allowing us to capture the precise mouse location at the start of the decision-making process. All websites were manipulated in a graphical editing program, and presented to participants in a popup window with disabled user interface chrome to minimize confusion between the proxy websites' chrome and their actual browser chrome (Fig. 1.). This also prevented participants from manipulating the experiment by reloading pages or navigating back and forward outside of our simulated website user interface.

All stimulus configurations appearing in the browser chrome were technically feasible. For example, the URL bar could not list a URL beginning with http and display a lock, nor could a website display https and a globe at the same time. Similarly, for the spoof conditions that utilized an extended validation (EV) certificate, that certificate could not point to the originating entity (i.e., ebay.com's EV certificate could not list eBay as the verified entity, but ebuy was a viable validated entity) (Fig. 2).

C. Procedure

Participants were offered $2 to visit 16 websites and decide whether or not they were secure. If secure, then participants were instructed to click on the login instruction (located in the lower center of the screen), but if not secure, then participants were instructed to click on the back button in the upper left hand corner of the screen Participants' bonus pay was dependent on how quickly they proceeded through the websites, incentivizing participants to visit all the websites as quickly as possible. Correct responses (i.e., clicking login on a secure website or back on an insecure website) advanced the experiment to the next trial. Incorrect responses (i.e., clicking back on a secure website or login on an insecure website) resulted in a time penalty. Incorrectly pressing back on a secure website resulted in a penalty screen being displayed for 20 sec and that time was added to the cumulative time, reducing the bonus pay. Incorrectly logging into an insecure website resulted in a 10 sec penalty screen being displayed, adding to the cumulative time and again reducing bonus pay. After the penalty screen was displayed, the experiment advanced to the next trial.

An online survey was administered in the original study window after the login tasks were completed. This survey assessed participants' demographic information (e.g., age, gender, education level), applied security knowledge (e.g., self-reported use of security indicators, self-reported password behavior), and technical security knowledge (e.g., definitional knowledge of DDoS, encryption protocols, firewalls). After the survey was complete a final screen was displayed that reported participants' accuracy and total payment amount they could expect to receive. This final screen also displayed a textbox asking for comments on the study.

D. Design

In order to examine both participants' attention to security cues and how those security indicators affect decision making, this study looked at security indicators effects on participants behavior when discerning the safety of encrypted vs. unencrypted websites, and also spoofed vs. not spoofed websites. Participants' ability to use indicators to ascertain encrypted vs. unencrypted websites was tested by manipulating the presence of http or https (https/http manipulation). Their ability to identify spoofed or not spoofed websites was tested by manipulating the domain name of the websites presented (no-spoof/spoof manipulation).

In addition to these basic manipulations, four different levels of encryption information were displayed by the web security indicators:

1) Extended Validation (EV) green lock and https
2) Full Encryption (FE) grey lock and https
3) Partial Encryption (PE) triangle w/exclamation mark & https
4) No Encryption (NE) globe; no encryption

It was not possible to include all encryption levels to test either the https/http or the no-spoof/spoof manipulations. Spoof and no-spoof websites displayed only EV, FE, and PE encryption levels to specifically test whether participants would notice a spoofed website that otherwise appeared secure. Given that unencrypted websites (http) only display a globe (NE), and encrypted websites (https) display the three other security symbols listed above (1-3), nesting all encryption levels under the https/http manipulation was not possible either. Thus, the https/http and no-spoof/spoof manipulations were analyzed separately for this study.

The https/http manipulation contained 8 trials and the no spoof/spoof manipulation contained 6 trials for a total of 14 trials presented to each participant. Within https/http manipulation, there were 4 secure websites (https) and 4 insecure websites (http). The https/http manipulation contained 4 trials under the https condition with each trial corresponding to 1 of the 3 valid levels of encryption information (EV, FE, or PE), and 4 trials with the http condition including only the NE indicator. The spoof/no spoof manipulation, contained 3 secure and 3 insecure trials. Each of those trials consisting of one of three encryption levels (1-3). The secure and insecure websites were counterbalanced between participants and the presentation order of the websites was randomized.

E. Metrics and Data Reduction

Participants' self-reported use of security indicators, their familiarity with the websites they were presented with, and their technical knowledge about security concepts (e.g.,"What type of math is used for RSA?") was collected in the post-task survey as potential predictors of both accuracy and uncertainty. Questions about participants' technical knowledge were taken from Ben-Asher and Gonzalez's work on cyber security and attack detection [20]. In addition to the participants' "back" or "login" responses, mouse-tracking data was also collected automatically without participants' knowledge. Area Under the Curve (AUC), as used by Spivey, Kehoe, and Dale, was used to evaluate participants' mouse tracking behavior [21].

1) Survey Data: The post-login task survey collected participants' self-reported use of security indicators. The survey contained three correct indicators (lock icon in the browser, certificate information, https in the URL) and four incorrect indicators (lock icon on the page, type of website, professional looking website, and website privacy statements).

Participants' Indicator Score was computed as $(\#\text{ correct indicators } + 1)/(\#\text{ incorrect indicators } + 1)$. The collected sample Indicator Score ranged from [0.2, 4.0], with a log-normal distribution $\log N(\mu = 0.14, sd = 0.58)$.

Participants' familiarity with the presented websites was rated on a 5-point Likert scale. The median rating was 3.0, and it ranged from a low of 1.0 to a high of 5.0. Ten technical security knowledge questions were asked and participants were assigned a score from 0 to 1, based on how many questions were answered correctly.

2) Mouse Tracking Data: In order to ascertain participants' decision-making uncertainty their mouse-tracking behavior was recorded on each trial. Each trial presented the front page of a website, requiring participants to click on the relevant account-login link. Clicking on the account-login link on the front page displayed the site's login page and began the portion of the trial during which the mouse movements were sampled. The click on the account-login link served to advance the trial as well as set a known start value for recording the movements to one of two known locations: login vs. no-login (i.e., back to previous page).

Participants' mouse position was sampled at 100Hz. Participants' click position was sampled on click and again on click-release. The mouse-tracking data was analyzed from the start of the click on the account-login link on the front page until participants clicked either the simulated back button (located in the upper left hand corner of the page), or the simulated website's login button (located in the lower center of the page). Despite having a known start and end point, differences in monitor size and resolution required a normalization process to allow us to compare mouse-tracking trajectories.

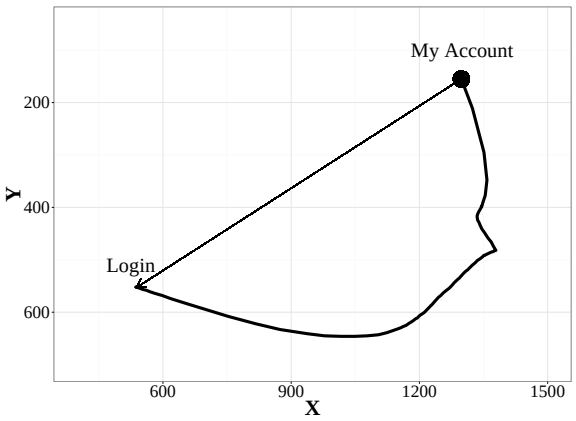
Normalization of the mouse trajectories was achieved by an affine transform that placed each trajectory's start at the origin in order to create a unit vector (and consequently scale the remaining trajectory) from the start to the end points, and then rotate the trajectory such that the start and end points are at a 45 angles to the horizontal line running through the origin (i.e., $(0, 0)$) (Fig. 3.). This results in trajectories that can be displayed and analyzed in a manner similar to other mouse tracking studies using a two-alternative forced choice design such as Spivey and Dale's work on the continuous dynamics of cognition [22], or Koop and Johnson's work on response dynamics of preferential choice [23]. After the mouse trajectories were normalized, Area Under the Curve (AUC) was calculated by taking the integral of the distance of the trajectory from the straight line vector through the start and end points as used by Spivey, Kehoe, and Dale and described by Hehman, Stolier, Freeman [21], [24]. For analysis, the log-transform of AUC was used, since the data could be described by a log-normal distribution $\log N(\mu = 1.33, sd = 1.24)$, and working with a normal distribution simplifies the statistical analysis

F. Statistical Analysis

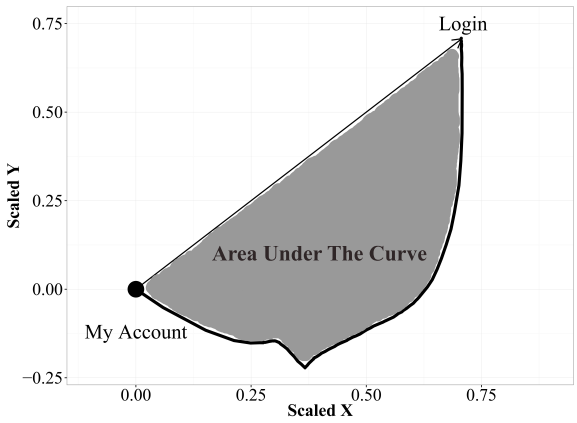
1) Likelihood to login: For both manipulations (https/http and no spoof/spoof), a hierarchical general mixed effects model with likelihood to login (login model) as the predicted variable and manipulation as the independent variable was tested. Participants' familiarity with a presented website (familiarity) and their security domain knowledge were included as covariate predictors. The balanced design of the no-spoof/spoof manipulations also allowed the use of lock presence (lock vs. no-lock) as a predictor. Due to a significant correlation between participants' indicator score and security knowledge $(r(170) = 0.36, p < 0.0001)$, and the fact that correlations between security knowledge and $\log(AUC)$ were stronger than indicator score in both manipulations, indicator score was not included as a covariate. Neither security knowledge, nor indicator score were correlated with likelihood to login, but security knowledge was kept as a predictor to allow comparisons between the model for likelihood to login and area under the curve easier.

2) Area under the curve: A hierarchical linear mixed effects model was also used to examine the factors influencing participants' behavior leading to their response. The log-transformed AUC was the predicted variable (AUC model). In addition to the predictors in the login model, correct responses (incorrect/correct) were added as additional categorical predictor. As in the login model, the AUC model was run separately for both the https/http manipulations and the no spoof/spoof manipulations, with lock presence was included as a predictor in the analysis of the no spoof/spoof experiment.

Both the login models and the AUC models utilized the maximal random effects structure justified by the data as



(a) Raw mouse path trajectory. The trajectory begins when the user clicks on "My Account" at the top right corner and ends at the "login" at the bottom left. For comparison purposes, all mouse trajectories were rotated so that the origins of all trajectories began in the lower left corner (see right panel).



(b) In order to compare trajectories from different sized monitors and different resolutions, the trajectory is scaled such that its start and end points are a unit vector, translated to the origin, and placed at a $45°angle$.

Fig. 3. Example transformation of raw mouse tracking path to a standardized coordinate space.

described by Barr, Levy, Scheepers, and Tily [25]. Models were analyzed in R using the `arm`, `lme4`, `lsmeans`, `car`, and `psych` packages [26]–[31]. Plotting was done with the `ggplot2` package [32], and data manipulation was assisted by the `plyr` package [33].

Multiple comparisons between categorical predictors were performed with Tukey's Honest Significant Difference test (THSD) [34]. Continuous predictors were centered to improve analysis according to the techniques detailed by Gelman and Hill [35]. Coefficient estimates are reported with standard error and p-values where appropriate. P-values for model coefficients were calculated using Type III Wald chi-square tests.

IV. Results

The primary research questions were concerned with how security indicators affect not only participants likelihood to log into secure and insecure websites (Login Model), but also their behaviors leading up to their decision (Area Under the Curve Model). These models were analyzed for both the https/http manipulation and the no spoof/spoof manipulation and the effects of the different predictors as well as their interactions were assessed.

A. https/http Manipulation

1) Likelihood to Login: Four of the eight websites involved in this manipulation were secure. If participants were perfectly accurate, the mean likelihood to login was 50% (100% in the https condition and 0% in the https condition). In reality, the overall likelihood to login was higher than 50%, but not significantly so ($\mu = 0.72, 95\%CI = 0.38, 1.0$). Participants made errors in both not logging into secure sites as well as logging into inseure sites. Overall, they did not login to all https sites ($\mu_{\text{https}} = 0.84, SE = 0.03$), and they logged into a significant number of insecure sites ($\mu_{\text{http}} = 0.61, SE = 0.04$).

As summarized in Table I, the probability of logging in was greater for http than https. This login probability was modulated by familiarity of the website, as well as security domain knowledge. Both of these covariates increased participants' likelihood to login. Although the familiarity effect is consistent with previous findings [4], the increased bias to login among those with greater security knowledge is a counter-intuitive finding. One possible explanation for this result is that knowledge and behavior do not always correspond. We will return to this issue in the Discussion.

TABLE I. Coefficient estimates for the omnibus analysis for the https/http manipulation for participants' probability to login.

	Est. (β)	**Std. Err**	χ^2	**df**	**P**($>\chi^2$)
$(Intercept)$	1.98	0.15	180.55	1	< 0.0001
http	-1.48	0.15	90.84	1	< 0.0001
Familiarity	1.04	0.26	16.18	1	< 0.0001
Knowledge	0.86	0.28	9.17	1	0.002
$\text{http}_{\text{Familiarity}}$	-0.67	0.32	4.29	1	0.04
$\text{http}_{\text{Knowledge}}$	-0.98	0.30	10.79	1	0.001

The omnibus analysis also revealed a two-way interaction between the manipulation and security knowledge (Figure 4). On https sites, the likelihood to login increased with higher security domain knowledge, whereas security knowledge did not affect responses on http sites ($\beta = -13, SE = 0.22$). This finding suggests an asymmetry in how security knowledge affects logging in. Even though greater security knowledge increased the likelihood to login to secure https sites, this same increase in knowledge had a negligible effect on decreasing logins to insecure http sites conceivably because the bias to login was so strong.

Familiarity also interacted with the manipulation ($\chi^2(1) = 4.29, p = 0.04$). Unlike security knowledge, familiarity strictly increased the likelihood to login. In the https condition, however, the effect of familiarity was higher, than in the http condition ($\beta_{\text{Familiarity}_{\text{https}-\text{http}}} = 0.67, SE = 0.32, z = 2.07, p = 0.04$), (Figure 5).

2) Area Under the Curve: When $\log(AUC)$ was the dependent variable, the analyses revealed a single main effect

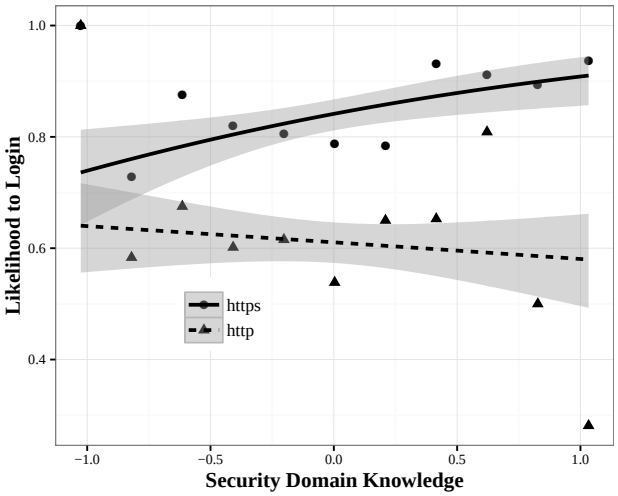


Fig. 4. Effects of participants' security domain knowledge on their likelihood to login. Higher knowledge leads to higher likelihood to login in the "https" condition, but knowledge has no effect in the "http" condition.

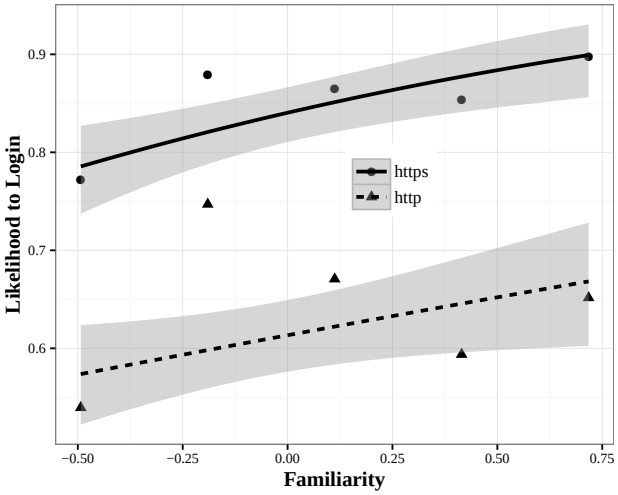


Fig. 5. Effects of participants'familiarity on their likelihood to login. Higher familiarity leads to higher likelihood to login in both the "https" and "http" conditions, but there is a reduced effect in the "http" condition.

for security domain knowledge. (Table II). Higher security domain resulted in higher $\log(AUC)$. This result suggests that more knowledgeable participants were more likely to login, because they considered more features on the website which accounted for a larger AUC. As such, this finding suggests that the decisions reached by high-knowledge participants were based on more information, though not necessarily the critical information.

TABLE II. COEFFIECIENT ESTIMATES FOR THE OMNIBUS ANALYSIS OF THE HTTPS/HTTP MANIPULATIONS EXAMINING $\log(AUC)$

	Est. (β)	**Std. Err**	χ^2	**df**	**P**$(>\chi^2)$
$(Intercept)$	1.40	0.14	108.16	1	< 0.0001
http	-0.06	0.14	0.18	1	0.67
Familiarity	0.07	0.28	0.06	1	0.81
Knowledge	0.77	0.30	6.79	1	0.009
Correct	0.03	0.14	0.04	1	0.84

B. No spoof/spoof Manipulation

Decisions regarding secure logins are dependent not only on security indicators, but also on the possibility that the website is spoofed with an incorrect domain name. Unlike the previous manipulation, all six of these websites were https but evenly divided among Extended Validation (EV), Full Encryption (FP), and Partial Encryption (PE). Three of the websites with different levels of encryption contained a correct domain name (no spoof) and three contained an incorrect domain name (spoof). The analyses were designed to test how familiarity, security knowledge, and encryption level (EV, FE, PE) interacted with the participants responses.

1) Likelihood to Login: The first set of analyses involved participants probability to login. Once again there was a bias to login. The overall likelihood to login to a website was found to be $(\mu = 0.75, SE = 0.03)$. Participants were less likely to log into spoofed websites $(\mu = 0.65, SE = 0.04)$ than not spoofed website $(\mu = 0.85, SE = 0.03)$, but it is important to remember that all logins to spoofed websites were incorrect.

Familiarity with websites again influenced likelihood to login. Participants were more likely to login to more familiar websites, but this response interacted with the website manipulation. For no spoof websites, the likelihood of logging in increased with the familiarity of the websites $(\beta = 1.66, SE = 0.52)$ whereas for spoof websites the likelihood of logging-in was not affected by familiarity $(\beta = -0.42, SE = 0.27)$, (Figure 6). The results for the no spoof websites are very straightforward, but based on previous research [4] we expected that the likelihood to login on spoof websites would also increase with familiarity. One reason that this finding did not occur is that the habit strength to login to familiar websites was offset by detecting some of the spoofed domain names.

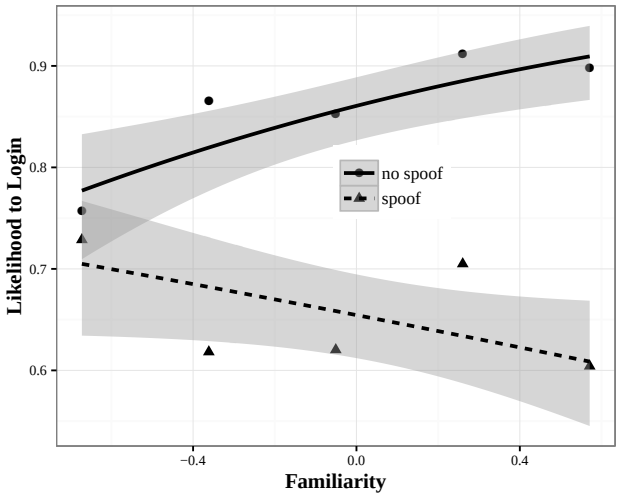


Fig. 6. Effect of manipulation and familiarity on participants' likelihood to login. Familiarity increases participants likelihood to login to "no spoof," but has no statistical effect on their likelihood to login to "spoof" sites.

The availability of encryption information was also an important factor for determining when participants would login. When a lock was present on a website (EV & FE), participants were more likely to login than when a lock was absent. There was no difference between the EV and FE conditions

($\beta_{\text{EV}-\text{FE}} = -0.47, SE = 0.39, z = -1.22, p = 0.44$), but logins to both the EV and FE conditions were more likely than logins to the PE condition ($\beta_{\text{EV}-\text{PE}} = 0.76, SE = 0.26, z = 2.93, p = 0.001$), ($\beta_{\text{FE}-\text{PE}} = 1.23, SE0.37, z = 3.37, p = 0.002$). This encryption information was also found to interact with both manipulation and familiarity ($\chi^2(2) = 8.50, p = 0.01$).

As shown in Figure 7, the presence of EV improves participants' ability to discriminate between "spoof" and "no spoof" sites. Not only does familiarity increase the likelihood to login to non-spoofed sites when an EV is present ($\beta_{\text{Familiarity}_{\text{EV}_{\text{no spoof}}}} = 2.17, SE = 0.61$), but it also reduces the chances they will login to a spoof site displaying an EV ($\beta_{\text{Familiarity}_{\text{EV}_{\text{spoof}}}} = -1.22, SE = 0.47$), ($\beta_{\text{Familiarity}_{\text{EV}_{\text{no spoof}-\text{spoof}}}} = 3.38, SE = 0.78, z = 4.32, p < 0.0001$). Aside from the encryption level associated with the EV indicator, neither of the other two encryption levels were affected by familiarity. It may seem surprising that this interaction was attributable only to EV and not to FE encryption level as well, since they both include locks, but we suspect that it is the highlighting of the domain name as part of the EV indicator that increases the likelihood that participants will notice the spoofed domain name and then respond accordingly.

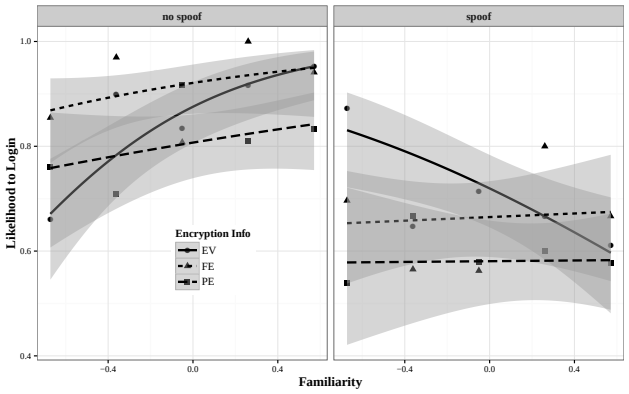


Fig. 7. Effect of manipulation and familiarity on participants' likelihood to login. Familiarity increases participants likelihood to login to "no spoof," but has no statistical effect on their likelihood to login to "spoof" sites.

Security domain knowledge also increased participants' ability to discriminate between "spoof" and "no spoof" sites ($\chi^2(1) = 5.48, p = 0.02$). As Figure 8 shows, participants with higher security domain knowledge were better able to identify malicious and secure sites. Not only does higher domain knowledge increase participants' likelihood to login to "no spoof" sites ($\beta = 1.27, SE = 0.52$), it also reduces their likelihood to login to "spoof" sites ($\beta = -1.55, SE = 0.32$).

2) Area Under the Curve: The main goal of this analysis was to assess decisionshow much of the website was considered before making a decision. The omnibus analysis revealed that none of the independent variables (manipulation, encryption information, familiarity, security knowledge, accuracy) showed significant differences with regard to the dependent measure, $\log(AUC)$. Instead, it was the interactions that were significant and most informative.

The first significant interaction involved both accuracy and the no spoof/spoof manipulation ($\chi^2(1) = 11.37, p = 0.001$). A post-hoc THSD test revealed no difference between incorrect and correct responses in the no spoof condition ($\beta_{\text{no spoof}_{\text{incorrect}-\text{correct}}} = -0.50, SE = 0.30, df = 945.7, t = -1.66, p = 0.35$), but did reveal higher $\log(AUC)$ for incorrect responses in the spoof condition ($\beta_{\text{no spoof}_{\text{incorrect}-\text{correct}}} = 0.79, SE = 0.13, df = 983.00, t = 6.02, p < 0.0001$), (Figure 9).

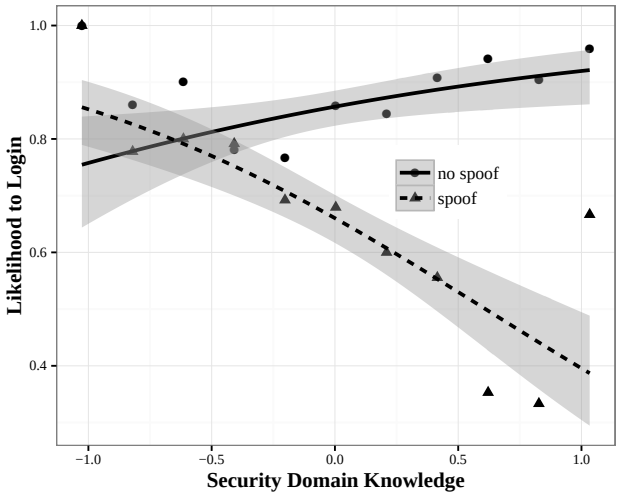


Fig. 8. Effect of manipulation and security domain knowledge on participants' likelihood to login. Increased domain knowledge leads to increased likelihood to login to "no spoof" sites and a reduced likelihood to login to "spoof" sites.

Furthermore, correct responses in the spoof condition showed a lower $\log(AUC)$ than correct responses in the no spoof condition ($\beta_{\text{Correct}_{\text{no spoof}-\text{spoof}}} = 0.67, SE = 0.12, df = 951.84, t = 5.58, p < 0.0001$). There was no difference between incorrect responses between the no spoof and spoof conditions ($\beta_{\text{Incorrect}_{\text{no spoof}-\text{spoof}}} = -0.61, SE = 0.30, df = 937.46, t = -2.04, p = 0.18$). Presumably, this is due to the fact that when a spoof URL is discovered, participants can immediately avoid logging-in, but, when the URL has not been manipulated, it is necessary to continue searching the site for additional information which for many participants was indeterminate.

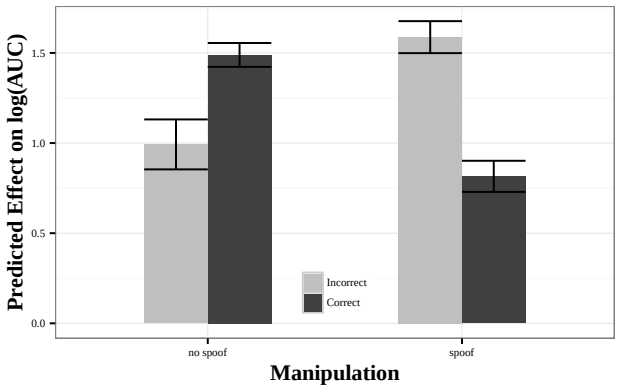


Fig. 9. Effects of manipulation (no spoof/spoof) on participants' $\log(AUC)$ based on the accuracy of their response.

The results in the spoof condition were further affected by the encryption information. A three-way interaction between encryption level, no-spoof/spoof manipulation, and accuracy ($\chi^2(2) = 9.39, p = 0.01$) revealed that the $\log(AUC)$ was greater for incorrect than correct responses in the spoof condition. A post-hoc THSD test revealed lower $\log(AUC)$ in the EV condition than the FE condition when incorrectly logging-in to spoofed websites ($\beta_{\text{Incorrect}_{\text{EV}-\text{FE}}} = 0.62, SE = 0.17, df = 860.98, t = -3.74, p = 0.01$). Moreover, $\log(AUC)$ was greater in both FE ($\beta_{\text{FE}_{\text{Incorrect}-\text{Correct}}} = 1.25, SE = 0.22, df = 952.98, t = 5.62, p < 0.0001$) and EV ($\beta_{\text{FE}_{\text{Incorrect}-\text{Correct}}} = 0.85, SE = 0.22, df = 954.41, t = 3.85, p = 0.01$) conditions. By contrast, there were no significant effects for the no-spoof manipulation, nor were there differences between incorrect and correct responses in the PE condition. It is also noteworthy that the $\log(AUC)$ did not differ between correct and incorrect responses for the PE encryption condition suggesting that participants were confused about this indicator, and thus their search behavior was less likely to terminate even when they detected a spoofed domain name.

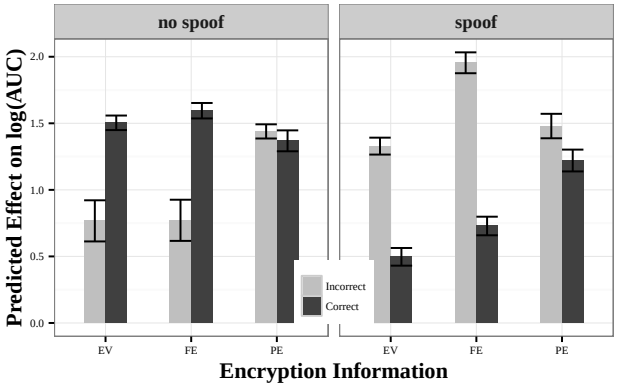


Fig. 10. Effects of encryption information on participants' $\log(AUC)$ based on the manipulation and the accuracy of their response.

V. Discussion

These results are noteworthy for at least four reasons: (1) Survey data are not a good predictor of a users' risky decision making on the web; (2) spoof websites present a greater risk to the all users than unencrypted websites; (3) semi-naturalistic web behaviors can be studied experimentally; and (4) mouse tracking can provide important insights into the cognitive processes underlying web-based decisions. Each of these conclusions will be elaborated below.

A substantial number of studies have relied on surveys to inform us about users behaviors on the web. An implicit assumption in these usability and security risk surveys is that users knowledge will determine their likelihood of making risky decisions on the web, such as whether or not to login to a specific site. For many reasons, these survey results have been suspect since even knowledgeable users often do not devote sufficient time and attention to the information provided on a browser chrome to decide whether or not to login.

The findings from the current study confirm this intuition since even users with considerable security knowledge made a significant number of errors in judging whether or not it was secure to login to a specific website. It is conceivable that at least some of these errors were prompted by the time pressure included in our design, but this was done intentionally to simulate fairly typical situations for users. In the future, it will be important to compare these results with other conditions which are designed to reduce stress and increase accuracy so that the effects of a stressful condition can be assessed by more directly.

Although there was a bias to login, it was modulated by both the familiarity of the websites as well as the security knowledge of the participants. Although familiarity contributed to this bias, security knowledge interacted with whether or not the protocol was http or https. Participants with high security knowledge were more likely to login to secure sites, but it did not decrease the likelihood to login to insecure sites. Still, these participants were more successful than less knowledgeable participants in deciding whether or not to login. Interestingly, this greater technical knowledge was associated with larger mouse trajectories suggesting that they reviewed more information at the website.

The second conclusion from this research should not be surprising to most researchers. A spoof website is much more deceptive than an http website, because many of the security indicators may still be present but the website is nevertheless insecure. For https/http websites, decision making performance was primarily a function of whether or not a lock was present. The likelihood to login was higher for the https sites, especially because a lock was also visible on two-thirds of these trials. When a lock was not visible, users were much less likely to login to the website. For no-spoof/spoof websites, logins were again more likely with a lock in the browser chrome than without a lock.

Taken together, these last two findings challenge the view that security indicators are ignored. The problem for users is that different websites include different levels of encryption and these inconsistencies can wreck havoc on trying to ensure secure behaviors. In particular, websites containing only partial encryption (PE) result in a good deal of confusion, and this was reflected in the current study by participants' mouse trajectories showing no systematic relation with their accuracy in the spoof or no spoof condition. By contrast, correct responses in the spoof condition when EV or FE was present was associated with smaller mouse trajectories for correct as opposed to incorrect responses. These decreased log(AUCs) were attributable to participants detecting a spoofed domain name and then terminating their search for additional information.

In the no-spoof condition, participants faced the prospect of conducting an exhaustive search for security information, and thus they were more accurate in this condition when their mouse trajectories were larger. Interestingly, the EV indicator improved responding correctly to both spoof and no-spoof conditions, because it increased participants attention to the domain name that appeared highlighted in green in the browser chrome. As a consequence, particpants were more sensitive to whether this domain name was accurate or spoofed.

Logically, we had predicted that detecting a spoofed website would have increased with the familiarity of the website, and this was in part correct. Logging in to a no-spoof website

increased with familiarity, and while logging in to a spoof website did not decrease with familiarity, at least it did not increase. As previously discussed, we suspect that highlighting the domain name with the EV indicator increased the likelihood of detecting an incorrect domain name and thus offset the familiarity bias to login.

Critically, performance in the spoof condition was modulated by the presence or absence of a lock, but the direction of effects was reversed and users were more likely to err when a lock was present. One reason for poorer decision making with the lock present than absent is that the combination of a familiar website and the presence of a lock provided users with false assurance that the website was secure and they were therefore less likely to explicitly review the domain name. Some preliminary evidence from an eye tracking study supports this interpretation: users were less likely to check the domain name when the website was familiar and a lock was present in the browser chrome. This result underscores how habitual behaviors can result in greater risk for users on the web. Knowledge facilitates responding

One of the most intriguing findings was that participants with low security knowledge often performed no differently than high security knowledge participants. If we relied exclusively on the their responses to login or go back, then the reason for why there was no difference between the two groups would have remained a mystery. Instead, neither low- nor high-knowledge participants were simply guessing because their AUCs were higher when they were incorrect than correct. These higher AUC scores suggest that participants were uncertain about their decisions because they possessed only partial knowledge or observed conflicting information about the security of the websites. For these participants, the mouse trajectories revealed much higher area under the curve scores.

The third conclusion concerns the paradigm, itself, which was innovative and designed to simulate for users a more naturalistic situation than most previous experimental studies. The websites visited by participants were extremely realistic since they were screen-captured images, and participants were instructed to login or sign-in as they would when normally visiting these sites. Critically, performance varied as a function of the familiarity of the website which would be expected if the stimuli were viewed as realistic.

Unlike the real site, however, no personal information was submitted, and thus there was no risk to the security of the participants. Although this protection from risk may have allowed for a more cavalier attitude, we believe that the inclusion of a financial incentive and a penalty for wrong responses raised the stakes and motivated participants to make these decisions as carefully as they could even though the time pressure likely motivated them to respond more quickly and take more chances.

Finally, this study is perhaps the first to apply mouse tracking as a cognitive measure to usability and security studies. In the psychological literature, this approach is becoming increasingly common as it represents one of the most direct methods for studying the dynamics of perception, categorization, and decision making. The current application represents a somewhat novel approach because unlike the standard psychological paradigm, the two alternative choices are not located in fixed and symmetrical positions.

Instead, this new application demonstrates that it is possible to normalize the locations between the login and back button using geometric transformations. It is our impression that the ability to conduct security-related experiments with hundreds of participants on crowd sourcing websites and to also be able to capture mouse tracking data bode well for the future of experimentally studying risky decision making on the web.

A. Limitations

This experiment was designed to study the effects of security indicators on participants' decision making and their willingness to login in a realistic scenario, but there are several limitations that could affect the results. First, participants were asked to interact with image-mapped versions of different websites. These image maps provided only the necessary functionality for participants to proceed through the experiment by clicking account links, then either the back button or the login button on each website. This limited participants' ability to interact with the website to collect additional clues about potential risks.

Second, Firefox was the only browser that was permitted for this experiment. Even though a majority of users reported Firefox was their primary browser, this may have affected participants that were more familiar with other browsers, but chose to use Firefox to complete the tasks.

Additionally, because the experiment was conducted on Amazon's Mechanical Turk, participant input devices were unknown and could have added additional variability to results. One subject was manually identified as using a touch pad, or keyboard shortcuts, due to the structure of their mouse tracking data, but participants using a mouse ball cannot currently be differentiated from participants using a mouse.

Further validation that inability to counterbalance location of login and back button did not confound the results must be done. For the most part, this confound is not a concern because this study did not compare responses to login and back but rather pooled responses (login or back) to different websites.

VI. Conclusion

This study introduced a novel Two-Alternative Forced Choice experimental paradigm that can be used with mouse tracking to study the effects of security indicators on participants decision making when identifying potentially risky websites. Participants likelihood to login was examined, and mouse tracking was used to gain additional insights into participants decision processes guiding the responses. Results show that participants use their experience and knowledge when determining when to login to a given website. Participants security domain knowledge, familiarity with websites, as well the presence of security indicators all affect their willingness to login.

Acknowledgment

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number

W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. The authors would also like to acknowledge the following people and organizations for their assistance: NSWC Crane, L. Jean Camp, Prashanth Rajivan, Rachel Huss, and Tom Denning.

REFERENCES

[1] M. R. Meiss, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani, "Ranking web sites with real user traffic," *Proceedings of the international conference on Web search and web data mining - WSDM '08*, p. 65, 2008.

[2] V. Garg and J. Camp, "End User Perception of Online Risk under Uncertainty," in *2012 45th Hawaii International Conference on System Sciences*. IEEE, jan 2012, pp. 3278–3287.

[3] D. Stebila, "Reinforcing bad behaviour," in *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction - OZCHI '10*. New York, New York, USA: ACM Press, 2010, p. 248.

[4] A. Hazim, A. P. Felt, R. W. Reeder, and S. Consolvo, "Your Reputation Precedes You: History, Reputation, and the Chrome Malware Warning," in *Symposium on Usable Privacy and Security (SOUPS)*, 2014.

[5] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer, "The emperor's new security indicators an evaluation of website authentication and the effect of role playing on usability studies," in *Proceedings - IEEE Symposium on Security and Privacy*, vol. 0. Oakland/Berkley, CA, USA: IEEE, may 2007, pp. 51–65.

[6] M. Alsharnouby, F. Alaca, and S. Chiasson, "Why phishing still works: User strategies for combating phishing attacks," *International Journal of Human-Computer Studies*, vol. 82, pp. 69–82, 2015.

[7] T. Kelley and B. I. Bertenthal, "Tracking Risky Behavior On The Web: Distinguishing Between What Users Say' And Do'," in *Ninth International Symposium on Human Aspects of Information Security & Assurance (HAISA 2015)*, S. Furnell and N. Clarke, Eds., no. HAISA. Lesvos, Greece: CSCAN Open Access Repository, 2015, pp. 204–214.

[8] T. Whalen and K. M. Inkpen, "Gathering evidence: use of visual security cues in web browsers," in *Proceedings of Graphics Interface 2005*, 2005, pp. 137–144.

[9] M. Arianezhad, L. J. Camp, T. Kelley, and D. Stebila, "Comparative eye tracking of experts and novices in web single sign-on," in *Proceedings of the third ACM conference on Data and application security and privacy - CODASPY '13*, no. October. New York, New York, USA: ACM Press, 2013, p. 105.

[10] S. D. Levitt and J. a. List, "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World ?" *The Journal of Economic Perspectives*, vol. 21, no. 2, pp. 153–174, 2007.

[11] A. Petzold, F. Plessow, T. Goschke, and C. Kirschbaum, "Stress reduces use of negative feedback in a feedback-based learning task." *Behavioral neuroscience*, vol. 124, no. 2, pp. 248–255, 2010.

[12] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor, "Crying Wolf: An Empirical Study of SSL Warning Effectiveness," in *Proc. 18th USENIX Security Symposium*, 2009.

[13] J. Sobey, R. Biddle, P. van Oorschot, and A. S. Patrick, "Exploring User Reactions to New Browser Cues for Extended Validation Certificates," in *Proc. 13th European Symposium on Research in Computer Security (ESORICS) 2008*, ser. LNCS, S. Jajodia and J. Lopez, Eds., vol. 5283. Springer, 2008, pp. 411–427.

[14] M. a. Goodale, D. Pelisson, and C. Prablanc, "Large adjustments in visually guided reaching do not depend on vision of the hand or perception of target displacement." *Nature*, vol. 320, pp. 748–750, 1986.

[15] J. B. Freeman, R. Dale, and T. a. Farmer, "Hand in motion reveals mind in motion." *Frontiers in psychology*, vol. 2, no. April, p. 59, jan 2011.

[16] J.-H. Song and K. Nakayama, "Role of focal attention on latencies and trajectories of visually guided manual pointing." *Journal of vision*, vol. 6, no. 2006, pp. 982–995, 2006.

[17] P. Cisek and J. F. Kalaska, "Neural Correlates of Reaching Decisions in Dorsal Premotor Cortex: Specification of Multiple Direction Choices and Final Selection of Action," *Neuron*, vol. 45, no. 5, pp. 801–814, 2005.

[18] C. McKinstry, R. Dale, and M. J. Spivey, "Action dynamics reveal parallel competition in decision making," *Psychological Science*, vol. 19, no. 1, pp. 22–24, 2008.

[19] R. Dale and N. D. Duran, "The Cognitive Dynamics of Negated Sentence Verification," *Cognitive Science*, vol. 35, pp. 983–996, 2011.

[20] N. Ben-Asher and C. Gonzalez, "Effects of cyber security knowledge on attack detection," *Computers in Human Behavior*, vol. 48, pp. 51–61, 2015.

[21] R. Dale, C. Kehoe, and M. J. Spivey, "Graded motor responses in the time course of categorizing atypical exemplars," *Memory & Cognition*, vol. 35, no. 1, pp. 15–28, 2007.

[22] M. J. Spivey and R. Dale, "Continuous Dynamics in Real-Time Cognition," *Current Directions in Psychological Science*, vol. I, no. 5, pp. 207–211, 2006.

[23] G. J. Koop and J. G. Johnson, "The response dynamics of preferential choice," *Cognitive Psychology*, vol. 67, no. 4, pp. 151–185, 2013.

[24] E. Hehman, R. M. Stolier, and J. B. Freeman, "Advanced mouse-tracking analytic techniques for enhancing psychological science," *Group Processes & Intergroup Relations*, pp. 1–18, 2014.

[25] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of Memory and Language*, vol. 68, no. 3, pp. 255–278, 2013.

[26] A. Gelman and Y.-S. Su, *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2015.

[27] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models using lme4," 2014.

[28] J. Fox and S. Weisberg, *An {R} Companion to Applied Regression*, 2nd ed. Thousand Oaks {CA}: Sage, 2011.

[29] R. Lenth, *lsmeans: Least-Squares Means*, 2015.

[30] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014.

[31] W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois, 2015.

[32] H. Wickham, *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[33] ——, "The Split-Apply-Combine Strategy for Data Analysis," *Journal of Statistical Software*, vol. 40, no. 1, pp. 1–29, 2011.

[34] J. W. Tukey, "Comparing individual means in the analysis of variance." *Biometrics*, vol. 5, no. 2, pp. 99–114, 1949.

[35] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press, 2007.