Comparative Eye Tracking of Experts and Novices in Web Single Sign-on
(full version)

Majid Arianezhad[1] L. Jean Camp[2] Timothy Kelley[2]
Douglas Stebila[3]

[1] *School of Engineering Science, Simon Fraser University, Burnaby, B.C., Canada*
arianezhad@sfu.ca
[2] *Indiana University Bloomington, 107 S. Indiana Ave., Bloomington, IN, 47405, USA*
ljcamp@indiana.edu, kelleyt@indiana.edu
[3] *Science and Engineering Faculty, Queensland University of Technology, Brisbane, Queensland, Australia*
stebila@qut.edu.au

December 13, 2012

Abstract

Security indicators in web browsers alert users to the presence of a secure connection between their computer and a web server; many studies have shown that such indicators are largely ignored by users in general. In other areas of computer security, research has shown that technical expertise can decrease user susceptibility to attacks.

In this work, we examine whether computer or security expertise affects use of web browser security indicators. Our study takes place in the context of web-based single sign-on, in which a user can use credentials from a single identity provider to login to many relying websites; single sign-on is a more complex, and hence more difficult, security task for users. In our study, we used eye trackers and surveyed participants to examine the cues individuals use and those they report using, respectively.

Our results show that users with security expertise are more likely to self-report looking at security indicators, and eye-tracking data shows they have longer gaze duration at security indicators than those without security expertise. However, computer expertise alone is not correlated with recorded use of security indicators. In survey questions, neither experts nor novices demonstrate a good understanding of the security consequences of web-based single sign-on.

Keywords: HTTPS; security indicators; single sign-on; web browsers; usability; eye-tracking; experts

1 Introduction

Web browsers employ certain security indicators—such as the presence of the lock icon, the use of "https" and domain name in the location bar, or certificate information—to help users make decisions regarding potential online threats, particularly related to the security of the web communications transmitted over the Secure Sockets Layer (SSL) / Transport Layer Security (TLS) protocol. Despite having been present in browsers for more than 15 years, many studies have demonstrated, using both self-reported

usage and eye-tracking data, that security indicators are largely ineffective at communicating security information to users.

In other areas of security research, such as phishing, technical expertise has been shown to be a mitigating factor in user susceptibility to online attacks.

Recently, the nature of user authentication on the web has changed. While user authentication was originally site-centric—users had different usernames and passwords for each web site—the use of *single sign-on authentication* has allowed users to use their credentials from a single *identity provider* to log in to multiple sites, which are called *relying parties*. Single sign-on can provide several benefits to users: most notably, they do not need to remember as many username/password combinations, and they do not need to register for new accounts at each site. On the other hand, there is a single point of failure (the identity provider), and users may have less control over their personal information.

Single sign-on systems are of growing importance within organizations and corporations. On the public Internet, several end-user single sign-on systems are currently available, both proprietary, such as ones provided by social networking sites like Facebook and Google, and open, such as the distributed standard OpenID. As of October 2012, one industry study reported that 54% of logins using social networking single-sign on used Facebook's system [Ols12].

Web-based single sign-on typically involves authentication via redirection from the relying party to the identity provider; the user authenticates to the identity provider, and then the browser is redirected back to the relying party with authentication tokens which the relying party can use in a back-channel to obtain the user's profile information from the identity provider. Because of the redirection, information flow is much less clear than the traditional login process and may place a substantial cognitive burden on users.

In this work, we explore two related themes. First, we examine whether users with higher technical expertise make better use of security indicators in web browsers. Second, we examine to what extent users employing single sign-on make use of security indicators in web browsers and their degree of understanding of the flow of information in single sign-on. Our study employs eye-trackers to obtain data on actual user behaviour, using both Facebook and OpenID as single sign-on identity providers.

Our goal is to provide answers to the following questions:

- Do users look for security indicators when using single sign-on in web browsers?
- Does the behaviour of users with respect to security indicators differ between novices and those with computer or security expertise?
- To what extent do users understand the flow of information and risks involved in single sign-on? Do novices and experts have different understandings?

Approach. Our study involved 19 participants who completed a variety of online tasks involving both Facebook and OpenID for single sign-on and then filled out a survey. The surveys were used to compare reported behaviour to observed behaviour. Because eye-tracking is time-intensive, data-intensive, and often perceived as invasive, relatively small sample sizes are common.

While completing online tasks, participants' gazes were recorded using eye-tracking equipment. The online tasks included a variety of social networking tasks, such as rating an item on a movie website, sharing an item onto a social networking profile, and using a social networking account to login to other websites. Participants were asked to use their own Facebook account, but were provided with an alternative account upon request; for tasks involving OpenID, participants used a provided account.

The survey had three sections of questions: demographics, technological expertise, and single sign-on. Using answers from the technological experience section, we classified participants as either (a) novice, (b) computer experts, or (c) computer and security experts.

Results. After classifying users' expertise, we examined a variety of user behaviours and responses within the context of expertise. Here are some of the results of our analysis:

- Security experts have higher self-reported use of security indicators than non-security experts, and this is confirmed with eye-tracking data, both in terms of gaze duration and number of fixations at security indicators.
- Users with only computer expertise, not security expertise, have no more frequent self-reported or actual use of security indicators than novices.
- In general, users have a poor understanding of the flow of information during single sign-on. They do not understand the flow of credentials and profile information between the browser, the identity provider, and the relying party. They cannot correctly say whether relying parties learn the password for their account at the identity provider; computer experts are somewhat better than computer novices at this, though surprisingly we cannot say the same for security experts.
- Users do not always realize that they are using single sign-on, especially when doing so within the context of a single organization whose services are distributed across multiple internal web servers.
- Users *do* understand that, after logging in to a relying party via an identity provider, they need to logout of the identity provider when terminating their session at a public computer.

Outline. Section 2 reviews background on single sign-on, security indicators, and expertise in security usability. In Section 3, we present our detailed methodology. Analysis and discussion is presented in Section 4. Additional discussion, including study limitations, appears in Section 5, and Section 6 concludes. The study tasks, survey, and statistical analysis methodology appear in the appendices.

2 Background

2.1 Single sign-on

Single sign-on (SSO) protocols allow a user with an account at an *identity provider* to identify herself to a third-party service, called a *relying party*. Single sign-on can be used within a single organisation or across multiple organisations. Pashalidis and Mitchell [PM03] classified single sign-on systems into two classes: *pseudo-SSO systems*, in which a user authenticates to a single identity provider and the identity provider internally manages multiple credentials for that user to authenticate to relying parties; and *true SSO systems*, in which a user authenticates to the single identity provider but the identity provider only provides authentication assertions to relying parties. They also divide SSO architectures based on whether the identity provider component is *local* or a third party (called *proxy-based*). For example, the Kerberos protocol can be viewed as proxy-based true SSO system, whereas client-side public key certificates can be seen as a local true SSO system. We focus on proxy-based true SSO systems.

Only recently has single sign-on seen widespread implementation on the public Internet. OpenID [Ope10] is a standard for federated authentication in which anyone can setup an identity provider and anyone can be a relying party, with no formal relationships required between relying parties and identity providers. Several commercial OpenID providers exist, and many webmail services act as OpenID providers, but at present relatively few relying parties exist.

Closely related to single sign-on is the notion of delegated authorisation, such as in the OAuth protocol [RFC10], where a user can delegate authority to a third party to access a particular resource on a server. For example, in August 2009, the popular microblogging site Twitter started requiring OAuth for all delegated authorisation.

In December 2008 the social networking site Facebook started offering a feature called "Facebook Connect" in which third party websites can allow users to login using their Facebook credentials rather than having to register for a separate account; this proprietary single sign-on service is built in part on the OAuth protocol.

For OpenID, Facebook Connect, and OAuth, single sign-on works via a sequence of redirects between webpages:

1. The user is on the website of a relying party, such as the movie review site Rotten Tomatoes.
2. The user clicks the "Login with Facebook" button on Rotten Tomatoes.
3. The user is redirected from Rotten Tomatoes to a Facebook login screen.
4. The user enters their Facebook username and password on the Facebook login screen and clicks "Submit".
5. Facebook verifies the credentials and asks the user to authorise the release of certain profile information.
6. The user consents to the release of profile information and then is redirected back to Rotten Tomatoes. The redirect includes cryptographic tokens that Rotten Tomatoes uses to subsequently request profile information from Facebook for that user.

Sun et al. [SPM+11] performed the first usability study of single sign-on protocols on the web. Participants using OpenID performed single sign-on related tasks using the existing browser interface and a proposed browser interface. They also surveyed attitudes towards single sign-on and comprehension of the risks and functionality of single sign-on.

2.2 Security indicators in web browsers

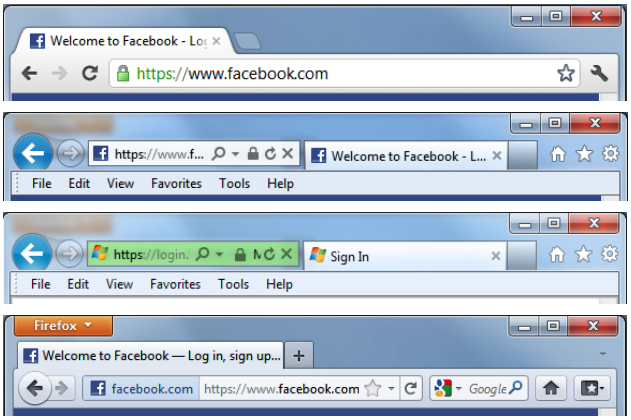
The Secure Sockets Layer (SSL) / Transport Layer Security (TLS) protocol provides encryption and authentication of communication on the Internet. The combination of SSL/TLS with web content delivered over the Hypertext Transport Protocol (HTTP) is jointly referred to as HTTPS. Authentication is performed using public key certificates a certificate authority (CA) who has verified that a given public key belongs to the legitimate owner of the given domain name and, in the case of extended validation certificates, that the key belongs to that real-world entity or business. Multiple CAs exist, and today's popular web browsers typically trust upwards of 650 CAs [Ele10].

Web browsers use several user interface elements, called *security indicators*, help users judge the security of a connection. Typically, these include the display of the protocol name (https) and domain name in the location bar, a lock icon, and additional colouring or elements for extended validation certificates. These security indicators are displayed within the browser *chrome*, the portions of the window controlled by the browser, as opposed to the *content* portion of the window that displays the HTML page. With several different major web browsers, different computing platforms, and frequent releases of new versions, the placement and semantics of security indicators in web browsers is inconsistent. Notably, Mozilla Firefox versions 4–13 did not display a lock icon to indicate the use of HTTPS, but it returned in version 14, after completion of our study. The indicators for each browser in our study are shown in Figure 1.

Research in the usability of security in web browsers traces its origins to the work of Friedman et al. [FHH+02] who conducted in-depth interviews to understand how users evaluate security of web sites. They collected types of evidence that users employed to decide if a website was secure: these included the above security indicators provided by the web browser chrome, as well as non-chrome indicators such as the type of information requested, the type of the site, the quality of the site, and statements within the page about security.

Several subsequent works have investigated the extent to which users and websites employ these and other security indicators. Whalen and Inkpen [WI05] used eye-tracking equipment and interviews to analyse how users interact with security indicators: most looked at the lock icon, though few made use of its interactive capabilities to display certificate information; less than half looked for the the use of HTTPS in the location bar. Notably, no participants gazed at security indicators prior to being "primed" for security. Stebila [Ste10] observed that popular websites do not consistently cause appropriate

Figure 1: Web browser security indicators. Google Chrome 17.0, Microsoft Internet Explorer 9.0.5, Microsoft Internet Explorer 9.0.5 with an extended validation certificate, and Mozilla Firefox 10.0.2



security indicators to be displayed; for example, of 125 popular websites, 19 had login forms delivered via an HTTP page but submitted via HTTPS, so no security indicators would be visible when entering passwords even though data submission would be secure; notably, this includes the world's second most popular website (according to Alexa Top Sites), Facebook, which is one of the single sign-on identity providers used in this study.

Certificate authorities, in conjunction with browser manufacturers introduced *extended validation (EV) certificates* in 2007; CAs would perform more extensive identity validation checks on parties (in exchange for more money), and browser manufacturers would introduce user interface elements, such as colouring the location bar green, to convey the purportedly greater trustworthiness of sites with EV certificates. Sobey et al. [SBvOP08] analysed the relative effectiveness of the indicators of Mozilla Firefox 3 and their own modification; users did not generally notice the EV indicators in the standard Firefox 3, but their own modification was more successful. However, no major browser currently employs an interface similar to their modification.

Schechter et al. [SDOF07] observed that users continue to login to websites when security indicators have been removed, even when security warnings are presented. Sunshine et al. [SEA+09] tested the effectiveness of various SSL security warnings; some designs were more effective than others, but in all cases a large proportion of users clicked through warnings.

Several works [Pat07, SHB10, SHB11] have raised questions about the extent to which this insecure behaviour can be explained by the artificial study environment. Complicating factors may include: participants using artificial credentials may feel less motivation to protect them; participants being "task focused"; and participants trusting that performing these operations in a study at a university means there is no risk.

2.3 Experts versus non-experts

Early research by Friedman et al. [FHH+02] observed that users from a high-technology neighbourhood were better able to describe the security indicators associated with an encrypted channel compared to users from a less technical neighbourhood. Sobey et al. [SBvOP08] found that expert users were better able to identify extended validation certificate security indicators in web browsers. Sunshine et al. [SEA+09] in their research on the effectiveness of SSL warnings briefly consider whether technical expertise influences ability to identify warnings; they observed that experts made slightly better decisions

than non-experts in some specific situations.

A real-life phishing attack performed by Jagatic et al. [JJJM07] on students at Indiana University found that students majoring in technical fields were roughly half as likely to fall for spear phishing emails as students in non-technical fields. Wright and Marett [WM10] confirmed that individuals with high self-reported computer self-efficacy or web experience, or participants who had high scores on a security awareness evaluation, were less susceptible to phishing attacks.

Asgharpour et al. [ALC07] used a card-sorting experiment to understand and characterize mental models for security risks of self-identified security experts and non-experts. Experts were much more likely to adopt a medical mental model (e.g., 'potential infection risks are ubiquitous') while nontechnical users perceived online risk as more like property crime (e.g., 'must avoid bad web neighbourhoods')

3 Methodology

In this study, we observed the behaviour of study participants while performing certain social networking tasks; our observation equipment included eye-tracking devices to record where the participant's gaze was during the tasks. After the online tasks, participants completed a survey.

Participants were recruited via email and personal contacts; we aimed to recruit approximately 50% of participants as people we believed might end up classified as being security or computer experts and 50% as novices. Participants received a $15 gift card for participating, and could withdraw from the study at any time while still receiving the full value gift card, although no participants did. The study was conducted in a small computer lab at an off-campus university building. Descriptions of the study indicated to participants that we wanted to observe their of social media; we omitted any references to security in the study description or instructions. The study was approved by the Human Ethics committee of the Queensland University of Technology and by the Institutional Review Board (IRB) of Indiana University.

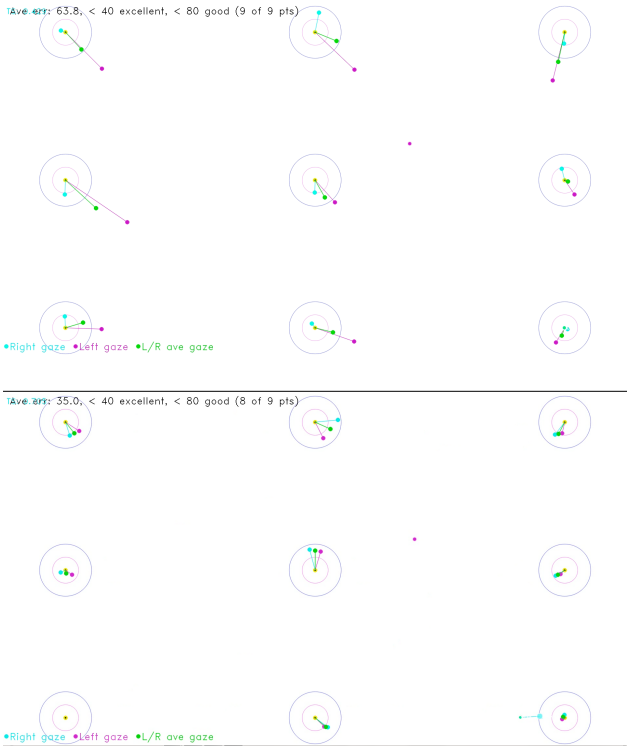
3.1 Eye-tracker calibration

After signing the study consent sheet, participants were seated at a desktop PC running Microsoft Windows 7, with a widescreen 19" monitor (a trial version of this study run earlier noted that some eye-tracking systems perform poorly on small monitors). The PC was equipped with the Mirametrix S2 Eye Tracker, placed just below the monitor. This eye-tracking device has a data rate of 60 Hz with binocular tracking. The accuracy range of the device is 0.5 to 1 degree and the drift range is less than 0.3 degrees.

The device manufacturer's 9-point calibration routine was run. Accuracy varied by participant: participants without astigmatisms had average error of 40 to 50 pixels. Relying solely on manufacturer calibration had two drawbacks. First, the distance between some security indicators was less than 50 pixels, so it would be difficult to distinguish gazes at nearby indicators. Second, reported error for the device was averaged over the 9 calibration points, but subjects had differing inaccuracies: some users had small errors for points close to the centre of the screen but large errors for points near the edge of the screen, and vice versa. See for example calibration errors for two different users in Figure 2.

As a result, we designed a secondary calibration phase in which we showed users additional calibration points which corresponded to points of interest for our study, for example the point at which the lock icon would appear when logging at a certain stage. We identified these points for each of the browsers in our study and prepared calibration videos. Participants were given a choice of web browser: Google Chrome 17.0, Microsoft Internet Explorer 9.0.5, or Mozilla Firefox 10.0.2 (the most recent versions of the browsers at the time of the study). We then showed users the secondary calibration video

Figure 2: Eye-tracking calibration results for (a) a user with poor accuracy near the corners of the screen and (b) a user with poor accuracy near the centre of the screen



and directed users to gaze at the points in our calibration videos. This secondary calibration was done twice: before Facebook tasks and before OpenID tasks.

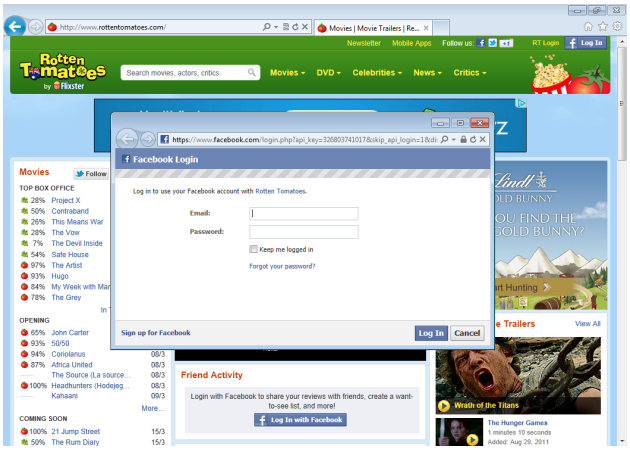
This secondary calibration phase allowed us to identify what the eye-tracker recorded for points of interest of our study, and compare those recorded points with points of gaze during the online tasks to determine whether participants gazed at our points of interest. Since the device's precision error was substantially lower than its accuracy error, this allowed us to obtain higher accuracy. We deemed a participant to have gazed at a security indicator while it was on screen if there was a fixation whose recorded distance to an indicator was within the average, plus two standard deviations, of the distances recorded for that point of interest during secondary calibration; two standard deviations ensured that all gazes recorded during secondary calibration would be accepted as valid, whereas one standard deviation would have missed some of the calibration gazes.

3.2 Online tasks

After calibration, participants were instructed to begin the online tasks. Prior to the participant's arrival, we randomly decided whether the participant would be assigned to complete tasks involving Facebook first or tasks involving OpenID first. In this explanation, we will proceed for a participant who was assigned Facebook tasks first; the Facebook and OpenID descriptions below would be swapped for participants assigned to completed the OpenID tasks first.

Facebook. The participant was given the list of Facebook tasks as in the appendix. In summary: in task F1, the participant was asked to navigate to the movie rating site Rotten Tomatoes, login with their

Figure 3: Login screen for Facebook single sign-on from Rotten Tomatoes in task F1 in Microsoft Internet Explorer 9.0.5.



Facebook account, and rate a movie; in task F2, they were asked to post a story about a movie from Rotten Tomatoes to their Facebook profile and then log out "as if you were walking away from a public computer"; in task F3, they were asked to visit a blog on LiveJournal and post a comment on a story using their Facebook account, then log out; in task F4, they were asked to share something from Amazon on to their Facebook profile; finally in task F5 they were asked to log out, go back to Facebook, log in, and then log out one last time.

Each Facebook logins resulted in a pop-up window being displayed in the centre of the screen as shown in Figure 3. Participants were asked to use their own Facebook account if they had one, however they were instructed that they could request alternative credentials to use instead of their own. After logging in to Facebook, participants were asked to grant the relying party access to certain personal information. In the statistical analysis, we analyzed the login portion and the personal information grant portion of the tasks separately.

We did not alter the behaviour or code of any of the websites used in this portion of the study. Although the front page of facebook.com is not delivered over HTTP, HTTPS is used to display the login page when Facebook single sign-on is accessed via a secondary website, so HTTPS security indicators were present during Facebook single sign-on login; no extended validation indicator was displayed as Facebook does not have an EV certificate. After logging in to Facebook, an additional page was displayed asking the user to share profile information with the relying party, security indicators may or may not be present on this screen depending on whether users have enabled Facebook's "Secure browsing" setting.

OpenID. The participant was next given the list of OpenID tasks as in the appendix. In contrast with the Facebook tasks, participants were not asked to use their own credentials. Instead, they were given credentials (an identity URL and password) for the OpenID provider we set up for this study.[1] In task O1, the participant was asked to visit a blog on LiveJournal and post a comment on a story using the provided OpenID account; in task O2 they were asked to visit a blog on BlogSpot and post a comment on a story, again using the provided OpenID account. Our OpenID provider operated entirely over HTTP, so no security indicators were ever present during interaction with our provider.

[1] http://barnraiser.org/prairie

3.3 Survey

After completing the above online tasks, participants were given a 39-question survey to complete online. There were three components to the survey: (1) participant general demographic information, (2) information to assess the participant's computer and security expertise, and (3) information pertaining to their understanding of single sign-on and their behaviour in the online tasks. Some of our questions were based on questions in existing survey instruments: section 2 included questions on technology expertise from Egelman [Ege09] and Sotirakopoulos [SHB11]; section 3 included questions on single sign-on comprehension from Sun et al. [SPM+11].

Some of the questions in our survey tried to identify which security indicators users use when signing in to websites. Where possible, we designed the sequence of questions in our survey to avoid priming participant responses: for example, in question 32 we asked the free-form question "How do you decide if it is safe to enter your username and password on a particular website?", but not until question 38, several screens later, did we explicitly list various security indicators and ask users to indicate which ones they used.

Upon completion of the online tasks and survey, participants were given a debriefing sheet with tips on using social networking sites, specifically Facebook, more securely. For participants that used their own Facebook account during the study, we offered to help them remove artifacts of the study from their account, including posts added to their wall/timeline and apps/websites linked to their account.

3.4 Classifying expertise

We used survey answers to classify participants on two dimensions: computer expertise and security expertise.

Computer expertise In the city and country in which we conducted our study, most people are indeed highly proficient at using computers. For example, no participants in our study answered below 3 on our 5-point Likert scale question (#9) where 1 was "I often ask others for help with the computer" and 5 was "Others often ask my for help with the computer". Thus, our rating of computer expertise was relative within this context. In particular, participants were assigned points for computer expertise as follows:

- 0.5: "Yes" to #8 "Do you use a computer daily for work?"
- 0.2–1.0: Answer to #9 "Rate yourself on this scale: 1—I often ask others for help with the computer . . . 5—Others often ask me for help with the computer"
- 1.0: "Yes" to #12 "Do you have a degree in an IT-related field?"
- 0.5 each: "Yes" to #13 "Have you ever. . . designed a website . . . created a database . . . written a computer program?"

The maximum possible score was 4.0. Participants with scores ≥ 2.5 were classified as computer experts.

Security expertise We used answers from the following questions to assign points for security expertise as follows:

- 0.5 each: "Yes" to #13 "Have you ever. . . used SSH . . . configured a firewall?"
- 1.0: "Yes" to #20 "Have you ever taken or taught a course on computer security?"
- 1.0: "Yes" to #21 "Have you attended a computer security conference in the past year?"
- 1.0: "Yes" to #22 "Is computer security one of your primary job responsibilities?"
- 0.5: "Yes" to #24 "Do you have an up-to-date virus scanner on your computer?"

The maximum possible score was 4.5. Participants with scores ≥ 2.5 were classified as security experts.

While the survey included several security-related free-form questions (#18 "If you know, please describe what a 'security certificate' is in the context of the Internet.", #19 "If you know, please describe

what is meant by 'phishing'."), we explicitly did not use answers to these free-form questions in deciding security expertise. Instead, we used answers to the free-form questions to cross-check validity of the security expertise score above. Points for the free-form answers were as follows, up to 1 point for each question:

- #18 "If you know, please describe what a 'security certificate' is in the context of the Internet."
  - 0.5: Mentioned SSL or HTTPS.
  - 0.5: Mentioned use to secure communication or demonstrate trust of a website.
  - 0.5: Mentioned ownership of a public key.
- #19 "If you know, please describe what is meant by 'phishing'."
  - 0.5: Mentioned stealing user information.
  - 0.5: Mentioned fake email or fake website.

3.5 Eye-tracking data

During our analysis, we found that using eye-tracking data to answer the question "did the user look at this point?" is somewhat difficult. Users have a lot of eye movement during web browsing tasks, and may fixate near a point for just a fraction of a second; how long does the gaze need to be in order to "count" as having looked at that point? We will consider both the number of fixations over a security indicator and the duration of gazes at security indicators.

3.6 Statistical analysis

We analyzed the eye-tracking data using Bayesian two-way analysis of variance and cross-validated our results using standard null-hypothesis testing. We examined mean gaze duration per fixation, mean number of fixations, and mean total gaze duration per task. The full methodology is in Appendix C; our source code is available online.[2]

4 Results and Discussion

Our study had 19 participants overall but our eye-tracking equipment failed to record data for 1 of them.

During the online tasks, two participants requested to use alternative Facebook credentials rather than their own.

Note that based on survey question #33, few participants found the online tasks difficult, with no more than 3–4 participants (out of 19) rating any task "hard" or "very hard".

4.1 Participant demographics

Our participant pool consisted of 3 females and 16 males, with an average age of 26 and an age range of 18–39. Although our participants' gender and age distributions do not match that of the general population, several previous studies on Internet security suggest that gender and age do not affect participant security behaviour [DTH06][SHB11, §4.3, §5.2].

In terms of education, 1 participant had completed at most high school, 8 had studied some of or completed an undergraduate degree, and 10 had some postgraduate education. For those with some university education, 9 responded that they studied in a subject area related to information technology, 4 were in a subject area not related to IT, and 5 gave no answer. Only 5 of our 19 participants had English

[2] http://eprints.qut.edu.au/55714

as a first language, but no participant appeared to have any trouble understanding instructions during the experiment.

All participants indicated that they had a Facebook account, and 17 had used their Facebook account at least once in the past month. Four participants with Facebook accounts reported having previously used their Facebook account to sign in to another website. No participant indicated that they had an OpenID account. However, given that many major webmail services are also OpenID providers, it is likely that many of the participants did indeed have an OpenID account but did not realize it.

In terms of web browser usage, during the study 9 participants chose to use Google Chrome, 2 chose to use Microsoft Internet Explorer, and 8 chose to use Mozilla Firefox. All but one user reported using one of the available browsers as their primary web browser.

4.2 Classifying expertise

Based on the methodology of Section 3.4, we classified participants on their expertise.

Per-participant results are displayed in the following sparklines; values for each participant are aligned between the two sparklines. The grey box indicates the range of values that were considered as "novice", namely scores less than 2.5.

Security expertise: 

Computer expertise: 

- *novices*: 9 participants had computer and security expertise scores < 2.5
- *computer experts*: 4 participants had computer expertise scores ≥ 2.5
- *security and computer experts*: 6 participants had computer and security expertise scores ≥ 2.5

No participants had high security expertise but low computer expertise. As a result, from here on we use "*security expert*" to mean "security and computer expert"; "*non-security experts*" includes both novices and computer experts who were not security experts.

To assess the validity of our security expertise questions, we included some free-form questions (#18, #19) in our survey, and scored participants on their answers to those questions as described in Section 3.4. We then compared the score on free-form answers to the classification based on the non-free-form answers. The mean score on free-form answers by the 6 security experts was 1.0, while the mean score by the 13 security non-experts was 0.423. The difference in means was statistically significant (Mann-Whitney $U = 62, n_1 = 6, n_2 = 13, p = 0.0398$).[3] We argue this provides evidence for the validity of our security expertise classification.

4.3 Use of security indicators

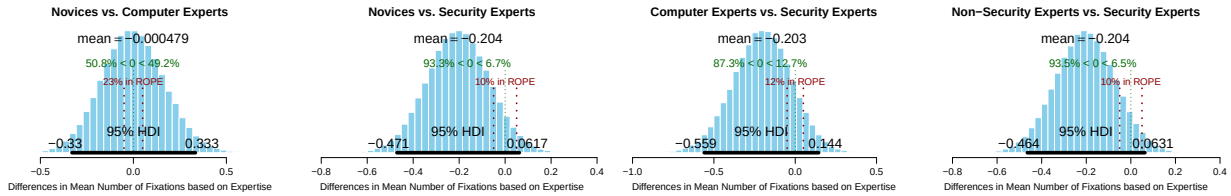
Whalen and Inkpen [WI05] were the first to analyze the use of security indicators use eye trackers. For a variety of security indicators, they compared self-reported use for each indicator with verification of use of that indicator via eye-tracking data. For example, in their study, 7 (out of 16) participants self-reported looking for https, and those 7 participants were verified as indeed having looked for https via eye-tracking; similarly, 12 participants self-reported looking for the lock or key icon, but only 11 actually did.

[3]The Mann-Whitney U test is a standard statistical tool that measures whether one set of independent observations tends to have larger values than another; in other words, whether the difference between two means is statistically significant. The test is *non-parametric*, meaning it makes no assumptions on the underlying distributions. Values of $p < 0.05$ indicate significance. Large values of p indicate that the data obtained does not demonstrate a statistically significant difference, though that alone does not prove the null hypothesis that the underlying behaviours are identical.

Table 1: Use of security indicators. Average total gaze duration (seconds) and average number of fixations on security indicators by task and classification for login and personal information grant dialog boxes.

Task	**Security experts** /6		**Computer experts** /3		**Novices** /9	
	dur.	# fix.	dur.	# fix.	dur.	# fix.
F1: Rotten Tomatoes / Facebook						
Login	1.788	2.667	0.679	1.000	0.794	1.556
Personal info	0.282	0.500	0.170	0.333	0.064	0.111
F3: LiveJournal / Facebook						
Login	0.110	0.167	0	0	0	0
Personal info	0.058	0.167	0.016	0.333	0	0
F4: Amazon / Facebook						
Login	0.036	0.167	0.186	0.333	0.274	0.556
Personal info	1.188	2.167	0.701	1.667	0.617	1.444
O1: LiveJournal / OpenID						
Login	0	0	0.455	1.000	0.142	0.444
Personal info	0.225	0.333	0.099	0.333	0.194	0.444
O2: BlogSpot / OpenID						
Login	0.126	0.333	0	0	0.049	0.111
Personal info	0.479	0.833	0	0	0.029	0.111

Figure 4: Differences in mean number of fixations based on expertise.



4.3.1 Eye-tracking evidence

As reported in Table 1, the majority of users in all expertise classifications did have a gaze point near the https or domain name security indicators. We now explore in detail the extent to which task and expertise affected number and duration of gazes.

Number of fixations *Expertise effects.* When we examine the overall mean number of fixations between different expertise groups, we find that, while there appear to be differences between the groups in terms of number of fixations—with security experts having a higher mean number of fixations—those differences fall within our uncertainty measures and cannot be considered credibly different (Figure 4).

Task effects. However, when we examine the data by task, a different story emerges. When we consider all of the Facebook tasks and compare them with the OpenID tasks we observe that while a mean difference of 0.0 is to the left of the distribution, it is fully within our 95% highest density interval, meaning that we cannot credibly conclude that the mean number of fixations differs in terms of task groups (Figure 5).

However, we note that Task F1 login and Task F4 personal information grant have noticeable and credible differences in terms of the mean number of fixations they receive. Furthermore, when we remove those tasks from consideration, any differences between Facebook tasks and OpenID tasks disappear (Figure 5).

Cross validation. These results correspond to our standard two factor ANOVA analysis looking at

Figure 5: Differences in mean number of fixations based on task. In the charts, Task 1 refers to the login portion of the Facebook F1 task on Rotten Tomatoes, and Task 6 refers to personal information grant portion of the Facebook F4 task on Amazon.

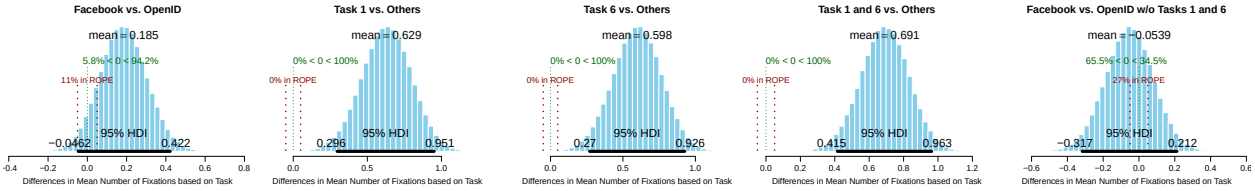


Table 2: Use of security indicators

	Df	**Sum Sq**	**Mean Sq**	***F* value**	Pr(> *F*)
A	2	5.10	2.548	1.647	0.196
B	9	70.96	7.885	5.095	5.85e-06
A:B	18	17.65	0.980	0.634	0.868
Residuals	140	216.67	1.548		

Results of ANOVA measuring Expertise and Task on mean number of fixations.

expertise (factor A) and task (factor B). Only factor B was found to be statistically significant ($p =$ 5.85e-06). Neither factor A nor the interaction were found to be statistically significant (Table 2). When we use Tukey's Honestly Significant Difference (Tukey HSD) to investigate the results, we find that only tasks F1 and F4 have any significant differences compared to the other tasks, supporting our initial Bayesian analysis.

These results demonstrate that users checked security indicators more during their initial login to Facebook, as well as when being asked to confirm sharing personal data on Amazon, but that overall number of gazes did not change for the OpenID login tasks[4].

Gaze duration The number of fixations gives us a picture of what tasks subjects consider important, but it does not give us the full picture. We also want to know the length of consideration each subject gives to each fixation and the total time they spend gazing at security indicators. We analyzed mean gaze duration per fixation in two ways: First we look at a more fine grained model of expertise while only considering if the task is from Facebook or OpenID, then we look at a fine grained model of individual tasks, but treat subjects as either having security expertise (security experts) or not (non-security experts).

Expertise effects. Looking at our fine grained model of expertise, we find that security experts, on average, gaze longer than novices[5]. However, our results suggest a bit of uncertainty: the mean difference of 0.0 falls outside our 95% HDI, but the 95% HDI includes part of the ROPE (Figure 6). A similar situation arises when we compare security experts with non-security experts. No mean difference lies outside our 95% HDI, but part of the ROPE is contained within the 95% HDI. We find no credible difference between security experts and computer experts, nor between novices and computer experts.

Cross validation. When we cross-validate we find that the results of our Bayesian analysis are confirmed using a two factor ANOVA considering expertise (factor A) and task type (factor B). We find that expertise is an important factor (Table 3). However, when we use Tukey's HSD to examine the

[4]Recall that OpenID login in our study occurred over HTTP, so no SSL security indicators were present. However, we still analyzed whether participants gazed at where those indicators would have been.

[5]Recall that "novices" excludes computer experts.

Figure 6: Differences in mean log(duration) based on expertise.

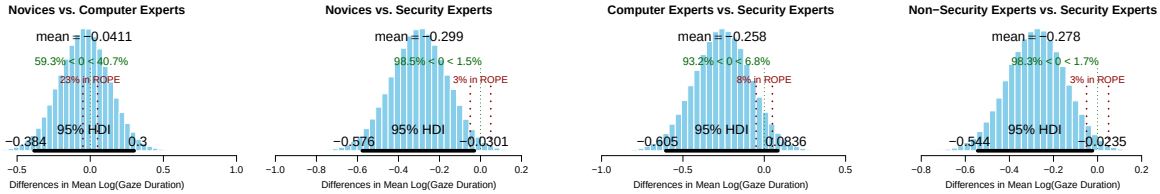


Table 3: Two-way ANOVA Analysis of Mean log(Gaze duration/fixation)

	Df	**Sum Sq**	**Mean Sq**	***F* value**	$Pr(>F)$
A	2	2.049	1.0245	3.152	0.0472
B	1	0.352	0.3521	1.083	0.3005
A:B	2	0.429	0.2145	0.660	0.5191
Residuals	96	31.200	0.3250		

Results of ANOVA measuring Expertise and Task on mean duration of fixations, given fine grained expertise.

paired comparisons, we find that only the difference between security experts and novices is significant ($p = 0.061$), roughly corresponding to our Bayesian results.

When we compare security experts to those without security expertise (non-security experts)[6], the cross validation is stronger than our Bayesian results. The log transformation of the data gives us approximately normal data. This allowed us to use a Welch two sample t-test to compare the results due to differences in group variances. We found that differences between the groups were statistically significant ($\mu = -0.2858$, 95% CI $= -0.50284292, -0.06874345$, $t(99.869) = -2.6124$, $p = 0.01038$), confirming that security experts gaze longer than non-security experts.

Task effects. When we consider gaze durations per fixation based on specific expertise, and analyze the general task type differences we find there is no credible difference between Facebook and OpenID in terms of mean log(duration), meaning that subjects gaze for roughly the same amount of time during Facebook tasks and OpenID tasks (Figure 7).

However, when we consider our fine grained task model to see if any tasks receive more time per fixation based on expertise, we find that participants of all expertise levels gazed somewhat longer at

[6]Recall that "non-security experts" includes both novices and computer experts who are not security experts.

Figure 7: Differences in mean log(duration) based on task type.

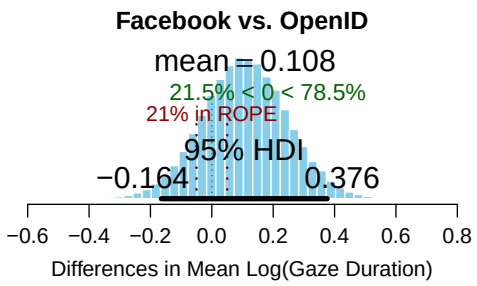


Figure 8: Differences in mean log(duration) based on tasks.

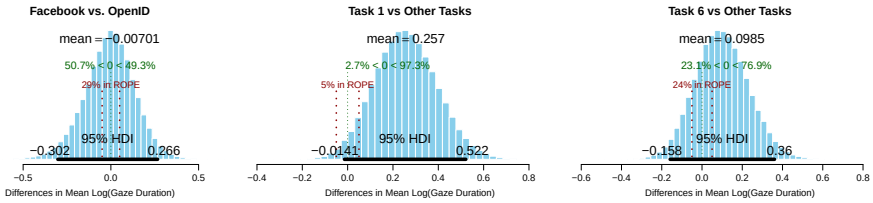


Table 4: Two-way ANOVA analysis of mean total gaze duration per task.

	Df	**Sum Sq**	**Mean Sq**	***F* value**	Pr($>F$)
A (Expertise)	2	2.60	1.3018	2.193	0.116
B (Task)	9	24.49	2.7213	4.584	2.9e-05
A:B	18	7.29	0.4048	0.682	0.824
Residuals	130	77.18	0.5937		

security indicators during the login portion of task F1 and the personal information grant portion of task F4, but the difference in average gaze duration during these tasks compared to other tasks fall within our uncertainty measures and cannot be considered credibly different (Figure 8) .

Cross validation. With a two factor ANOVA test, we did find an statistically significant effect for task on mean per fixation duration. However, in the post-hoc analysis, the only truly significant differences were between task F3 and all other tasks. Looking at the data, this is due to an outlier effect, rather than any true artifact. No other tasks had any significant differences in mean per fixation durations.

Total duration. Having both the number of fixations and mean per duration of fixation, we also wanted to examine differences between expertise (factor A) and tasks (factor B) on the total fixation duration per task, rather than per fixation. Due to difficulty of specifying an accurate sampling distribution, we analyzed total duration using standard NHST techniques.

We find that task has a significant effect on total gaze duration, but there is still too much uncertainty about the effect of expertise to make a strong claim that it has an effect Table 4). When we look at post-hoc Tukey HSD analysis we find that the only two tasks that are significantly different from the mean are tasks F1 and F4, which makes sense, given that they receive more fixations total (Figure 5).

Discussion. It is difficult to directly compare our rates of security indicator usage with those of Whalen and Inkpen [WI05], since they simply reported the number of participants for which they "verified" that the user checked the indicator. In Table 1 we report the proportion of participants that gazed at security indicators, but we see from the average duration of these gazes that the time spent looking at security indicators varies significantly. As a result, we cannot say whether our participants' observed use of security indicators matches or disagrees with that of Whalen and Inkpen. This does highlight the open issue of how to report and compare usage of security indicators from eye-tracking data.

Unlike Whalen and Inkpen, who reported observing no fixations before priming their participants for security, we do observe fixations without any security priming. However, the nature of the fixations focuses primarily on tasks F1 (initial Facebook login) and F4 (share information with Amazon). The first represents the first time a subject logs in, while the second represents interacting with a commercial website. It appears that, aside from these tasks, subjects give only cursory attention to security indicators, supporting Whalen and Inkpen's results that without security priming, subjects will not pay attention to security indicators.

On the other hand, our results seem to suggest that most users, regardless of security experience, are

Table 5: Self-reported use of security indicators.

Indicator	Security experts /6	Computer experts /4	Novices /9
https	6	1	6
lock icon on the page	2	1	5
certificate	4	1	6
website privacy statements	2	3	4
type of website	6	2	6
professional-looking site	2	1	3
lock icon in browser	5	1	3

Other security indicators reported: "brand", "lack of ads", "firewall", "anti-virus safe browsing feature" (×2).

aware of security indicators and consult them in tasks they view as risky. This is an encouraging result, but, our experimental design does not allow us to determine the effects of the security indicators on decision making.

4.3.2 Self-reported use

Next, we compared self-reported use of indicators versus gaze duration. Recall that survey questions #32 and #38 asked subjects to report which security indicators they used to decide if it is safe to enter their username and password; in question #32, it was a free-form question, whereas in the later question #38 subjects were presented with a list of indicators and asked to check the ones that they looked for.

In the free-form question #32, only 4 subjects' responses mentioned one of the three accepted SSL security indicators (https, lock icon in browser, certificate); all 4 of these subjects were classified as security experts. (We emphasize that the classification in Section 3.4 of subjects as security experts did not depend on question #32 or question #38.)

In contrast, when presented with a list of security indicators in question #38, many more users reported using security indicators (Table 5). When prompted, participants self-report much higher use of security indicators.

To analyze this data, we assigned each user a score between 0 and 3, with one point for each of "https", "lock icon in the browser", and "certificate", which are the only security indicators presented in the browser chrome. In particular, we omitted "lock icon in the page": as an element of the web page content, a lock icon on the page is not a trusted user interface element [Ste10]. All but 3 participants reported looking for at least one of these three security indicators. The average score for security experts was 2.5, whereas the average score for security non-experts was approximately half that at 1.38, with the difference in these averages being statistically significant (Mann-Whitney $U = 62, n_1 = 6, n_2 = 13, p = 0.0408$). In contrast, the average score for computer experts was 1.8 compared to 1.67 for computer non-experts; the difference was not statistically significant (Mann-Whitney $U = 50.5, n_1 = 9, n_2 = 10, p = 0.6722$).

We then compared the self-reported use of security indicators with eye-tracking data. Of the 4 users that self-reported use of security indicators in free-form question #32, 2 did gaze for security indicators and 2 did not. To analyze self-reported use of security indicators in prompted question #38, we used the security indicator score computed in the previous paragraph and compared it with security indicator gaze duration during Facebook login in task F1. The correlation was quite low and not statistically significant (Spearman rank correlation coefficient $\rho = 0.188, P = 0.4546$). [7]

[7] Spearman's rank correlation coefficient is a non-parametric measure of dependence between two variables, in particular

Discussion. The only statistically significant relationship we observed regarding self-reported use of security indicators was that, when prompted, self-reported use of security indicators by security experts was substantially higher than security novices. For all other relationships we considered, we observed no statistically significant differences. In particular, we observed no statistically significant difference between the security indicator gaze duration of security experts compared to security novices, or between participants who self-reported using security indicators and their gaze duration.

Compared with previous results by Whalen and Inkpen [WI05], our participants had higher self-reported use of the https and certificate security indicators and somewhat lower self-reported use of the browser lock icon.

4.4 Understanding of single sign-on

To determine participants' understanding of single sign-on, we considered participants' successful completion of logout tasks and their responses to survey questions related to the flow of information in single sign-on.

We asked participants questions about their previous use of single sign-on. Nine of 19 participants had heard of single sign-on, all of whom provided a reasonably correct definition (question #27); 4 were security experts, 2 were computer experts, and 3 were novices. Thirteen of 19 participants reported having "previously experienced using a single username and password to access different systems" (question #28). Of the 7 who responded "No" to that question, we have reason to believe as least 6 of them had in fact used single sign-on systems before, as they were or had been students at a university that the authors know uses single sign-on for a variety of services. This suggests that in today's web-based environment, users and system administrators do not have the same view of what constitutes "different systems".

4.4.1 Logout

We directed the participants to log out several times: during Facebook tasks F2, F3, and F5, the participants were instructed to "Log out of the web browser as if you were walking away from a public computer. (Do not log out of Windows, however.)" Subsequently in task F5, they were asked to "Go back to the Facebook site and log in." and then "Log out of Facebook." All participants completed the last part of task F5—"Log out of Facebook". Participants' behaviour at earlier tasks is more interesting; we focus on the first logout at task F2.

The participant is actually logged in to two websites during task F2: Rotten Tomatoes and Facebook. Rotten Tomatoes appears to make use of Facebook's single sign-on API for logout: users that log out via the link on Rotten Tomatoes are also logged out of Facebook. This is not a required feature of the Facebook single sign-on API, and users do not know a priori if dual logout will occur.

For task F2, we recorded whether users explicitly logged out of Rotten Tomatoes website and whether they explicitly logged of Facebook (Table 6). Overall, 14 of 19 users successfully logged out of either Rotten Tomatoes or Facebook, with 11 actually visiting Facebook to logout or check that they were logged out. There was no significant difference between the behaviour of experts and novices.

We did not specify any logout task for OpenID, although our OpenID provider did have a logout function. Curiously, one participant did return—unprompted—to the URL for the OpenID provider to logout after task O2.

Discussion. The participants seemed to demonstrate conservative logout behaviour, in that when using a single sign-on service such as Facebook on a public computer and when directed to logout upon

measuring how well the relationship can be described using a monotonic function. Values of ρ near ± 1 indicate a high degree of correlation, values of ρ near 0 do not.

Table 6: Participant logout actions on Rotten Tomatoes and Facebook in task F2.

Classification	Logout of			
	RT&FB	RT only	FB only	none
Security experts /6	3	0	1	2
Computer experts /4	3	0	0	1
Novices /9	3	3	1	2

Table 7: Mental models of single sign-on.

Classification	Correct drawing	#34 Does Rotten Tomatoes know your Facebook p.w.?		
		Yes (wrong)	No (correct)	Don't Know
Security experts /6	3	3	3	0
Computer experts /4	0	0	3	1
Novices /9	5	1	4	4

completion of their work, they logged out of both the relying party and the identity provider, and in particular all participants logged out of Facebook at the completion of the study.

4.4.2 Mental model

Several survey questions provide insight into participants' mental models of single sign-on, including the drawing exercise after question #33 and questions about password and profile information.

For the drawing exercise, we used the same methodology as Sun et al. [SPM+11] for assessing the correctness of the mental model expressed in the drawing. As reported in Table 7, security experts did not do significantly better than security non-experts at answering this question.

Question #34 tested participants' understanding of the flow of information during single sign-on: it aksed if they believed that the relying parties, such as Rotten Tomatoes, learned their password for the identity provider Facebook. Table 7 reports the results: security experts did not do better than security non-experts, in fact they did worse.

4.5 Other

In January 2011 Facebook fully deployed its "secure browsing" feature, which allows users to opt-in to having all Facebook pages delivered over HTTPS. Of the 17 participants who used their own Facebook account, we observed that only 3 had activated secure browsing.

5 Additional discussion

In this section we discuss some additional observations.

5.1 Preference for single sign-on

When asked in question #37 if they would use their Facebook or OpenID account to login to third-party websites in the future, only 11% of participants responded "yes"; 42% chose "depends", and 47% chose "no". Contrast this with the results of Sun et al. [SPM+11], where they asked participants whether they would in the future prefer to use single sign-on in the form of OpenID (3%), in the form of Sun et al.'s

identity-enhanced browser (9%), it "depends" on the type of site (36%, of which 30% preferred the ID-enhanced browser and 6% preferred OpenID), or not use single sign-on at all (29%). Our participants were substantially less inclined to use single sign-on than Sun et al.'s participants; since our participants only had the option of using traditional single sign-on as opposed to Sun et al.'s ID-enhanced browser, they did demonstrate a slightly more favourable response than Sun et al.'s participants did to OpenID.

5.2 Nature of task and risk

The Facebook and OpenID tasks involved different levels of risk: in the Facebook tasks, most participants used their own accounts despite having the option to use manufactured accounts, whereas in the OpenID tasks all participants were instructed to use manufactured accounts. However we noticed no significant difference in number of fixations or gaze duration between OpenID and Facebook tasks.

However, we did notice a difference between certain Facebook tasks. As noted in Section 4.3.1, participants of all expertise levels paid more attention to security indicators during their initial Facebook login, and when they granted Amazon, the only e-commerce site in our study, personal information. The difference in security behaviour depending on the nature of the site is interesting and we believe merits further study in future work.

5.3 Study limitations and mitigations

It is well known that there are limitations to the ability of laboratory usability studies to reflect real-world environments [Pat07, SHB10, SHB11]. We consciously made several study design choices aligned with recommendations previous work to try to reduce the impact of the study environment.

Setting. The setting of a study and the demeanour of the person running the study can have an effect on study participants. Individuals participating in a study—particular a security study—at a university can be of the frame of mind that "this is being run at a university, nothing can go wrong". Our ethics restrictions did not permit us to disassociate the study with the university, but we did take some measures to attempt to mitigate these factors. Our study took place in a university building a few blocks from the main campus, in an office tower in the city's central business district. The person running the study was a Master's student, casually dressed in shorts and a t-shirt.

In terms of the electronic "setting" of the study, we tried to match the participants' natural computing environment to some extent. Participants were given a choice of browser. All had previous experience using Facebook, so that single sign-on mechanism was not entirely foreign. One unavoidable unnatural characteristic was the use of eye-tracking equipment and the required calibration stage, though the device itself is relatively unobtrusive, and requires no further user attention once calibration is complete.

Demand characteristics refer to the "tendency for research subjects to guess the reason for a study, and then to attempt to confirm the experimenter's apparent hypothesis" [Pat07]. All materials that our participants saw before and during the online tasks described the study as being interested in 'participants' use of social networking and social media', with no mention of security or privacy. Mentions of security only began in the survey, after completion of the online tasks.

Task focus is a risk in security usability studies: participants in studies are often highly motivated to complete the given tasks. Some previous studies [SDOF07] gave participants tasks to complete and then analyzed whether the participants completed these tasks even when security indicators or site authentication images were removed; participants who so completed the tasks were deemed to have not paid attention to indicators. Patrick [Pat07] criticizes that approach due to task-focused participants being motivated to complete the tasks they have been given. As a result, we did not artificially remove any security indicators during our study, instead relying on eye-tracking data to assess participant attention to security indicators, both on tasks where security indicators were naturally present (single

sign-on with Facebook which uses HTTPS), and naturally absent (single sign-on with our OpenID provider which designed to use only HTTP). Moreover, our participants were promised that they would receive the full value of their compensation regardless of whether they completed the tasks or not. Nonetheless, in the informal discussions we had with participants upon completion of the survey, some participants reported task focus affecting their decisions.

Use of credentials. Schechter et al. [SDOF07] confirmed that study participants who use their own account credentials, rather than provided credentials, behave more securely. As a result, we asked participants to use their own credentials for Facebook tasks; we provided participants with alternative Facebook credentials if asked, which 2 participants did.

6 Conclusions

With ever more websites that users need accounts for, and with the growing popularity of social networking, the use of web-based single sign-on systems is likely to increase. With multiple parties involved—the user's browser, the identity provider, and many relying parties—users may have a hard time understanding what happens with their credentials and personal information, and what conditions should be satisfied for them to believe that a connection is secure or that it is safe to enter their username and password.

We examined users' use of security indicators in web-based single sign-on using Facebook and OpenID by employing eye-tracking equipment and surveyed users on their perception of information flow in single sign-on to determine if users with technical experts behave more securely than novices. Our survey tool for classifying users as computer or security experts adapts existing tools and is cross-validated against other questions in our survey.

We found that users with security expertise did look at web browser security indicators more than those without security expertise; but computer expertise alone was not a predictor. Our participants—security experts and novices alike—in general had very poor understandings of the flow of information and trust in web-based single sign-on.

Future work directly related to the study includes examining the proportion of users that logout without being directed to do so and examining the generalizability of the results to others demographics.

Important future work in this area includes the study of long-term trends. As users continue to use the Internet more and more and as their general computer proficiency advances, do they make better or worse use of security indicators? With the recent popularity of social networking, it seems plausible that web-based single sign-on will become far more prevalent in the coming years, and it will be interesting to see if and how users' understanding of web-based single sign-on improves as frequency of use increases.

7 Acknowledgements

This research was performed while M.A. was a student at the Queensland University of Technology. The authors acknowledge helpful discussions with Tom Busey and Sonia Chiasson and programming assistance from Reza Ahli Araghi.

References

[ALC07] Farzaneh Asgharpour, Debin Liu, and L. Jean Camp. Mental models of computer security risks. In *Proc. Sixth Workshop on the Economics of Information Security (WEIS)*, 2007. URL http://weis2007.econinfosec.org/papers/80.pdf.

[Cra11] Lorrie Faith Cranor, editor. *Proc. 7th Symposium on Usable Privacy and Security (SOUPS) 2011*. ACM, 2011.

[DTH06] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why phishing works. In *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI) 2006*, pp. 581–590. ACM, 2006. DOI:10.1145/1124772.1124861.

[Ege09] Serge Egelman. *Trust me: Design patterns for constructing trustworthy trust indicators*. PhD thesis, Carnegie Mellon University, April 2009. URL http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA502272.

[Ele10] Electronic Frontier Foundation. The EFF SSL Observatory, 2010. URL https://www.eff.org/observatory.

[FHH+02] Batya Friedman, David Hurley, Daniel C. Howe, Edward W. Felten, and Helen Nissenbaum. Users' conceptions of web security: a comparative study. In *Proc. CHI '02 Extended Abstracts on Human Factors in Computing Systems*, pp. 746–747. ACM, 2002. DOI:10.1145/506443.506577.

[JJJM07] Tom N. Jagatic, Nathaniel A. Johnson, Markus Jakobsson, and Filippo Menczer. Social phishing. *Communications of the ACM*, **50**(10):94–100, October 2007. DOI:10.1145/1290958.1290968.

[Kru10] John Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press, 1st edition, 2010. URL http://www.indiana.edu/~kruschke/DoingBayesianDataAnalysis/.

[Ols12] Michael Olson. Janrain social login and social sharing trends across the web for Q3 2012, October 2012. URL http://janrain.com/blog/social-login-and-social-sharing-trends-across-the-web-for-q3-2012/.

[Ope10] OpenID Foundation. Specifications, 2010. URL http://openid.net/developers/specs/.

[Pat07] Andrew Patrick. Commentary on research on new security indicators, March 2007. URL http://www.andrewpatrick.ca/essays/commentary-on-research-on-new-security-indicators.

[PM03] Andreas Pashalidis and Chris J. Mitchell. A taxonomy of single sign-on systems. In Reihaneh Safavi-Naini and Jennifer Seberry, editors, *Proc. 8th Australasian Conference on Information Security and Privacy (ACISP) 2003*, *LNCS*, volume 2727. Springer, 2003. DOI:10.1007/3-540-45067-X_22.

[RFC10] The OAuth 1.0 protocol, April 2010. URL http://www.ietf.org/rfc/rfc5280.txt. RFC 5849.

[SBvOP08] Jennifer Sobey, Robert Biddle, Paul van Oorschot, and Andrew S. Patrick. Exploring user reactions to new browser cues for extended validation certificates. In Sushil Jajodia and Javier Lopez, editors, *Proc. 13th European Symposium on Research in Computer Security (ESORICS) 2008*, *LNCS*, volume 5283, pp. 411–427. Springer, 2008. DOI:10.1007/978-3-540-88313-5_27.

[SDOF07] Stuart E. Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The emperor's new security indicators: An evaluation of website authentication and the effect of role playing on usability studies. In *Proc. IEEE Symposium on Security and Privacy (S&P) 2007*, pp. 51–65. IEEE Press, 2007. DOI:10.1109/SP.2007.35. EPRINT http://usablesecurity.org/emperor/.

[SEA+09] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. Crying wolf: An empirical study of SSL warning effectiveness. In *Proc. 18th USENIX Security Symposium*, 2009. URL http://www.usenix.org/events/sec09/tech/full_papers/sunshine.pdf.

[SHB10] Andreas Sotirakopoulos, Kirstie Hawkey, and Konstantin Beznosov. "I did it because I trusted you": Challenges with the study environment biasing participant behaviours. In *SOUPS Usable Security Experiment Reports (USER) Workshop*, 2010. URL http://cups.cs.cmu.edu/soups/2010/user_papers/Sotirakopoulos_environment_biasing_participants_USER2010.pdf.

[SHB11] Andreas Sotirakopoulos, Kirstie Hawkey, and Konstantin Beznosov. On the challenges in usable security lab studies: Lessons learned from replicating a study on SSL warnings. In Cranor [Cra11]. DOI:10.1145/2078827.2078831.

[SPM+11] San-Tsai Sun, Eric Pospisil, Ildar Muslukhov, Nuray Dindar, Kirstie Hawkey, and Konstantin Beznosov. What makes users refuse web single sign-on?: an empirical investigation of OpenID. In Cranor [Cra11], pp. 4:1–4:20. DOI:10.1145/2078827.2078833.

[SR93] A.F.M Smith and G.O. Roberts. Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *J. Royal Statistical Society. Series B (Methodological)*, **55**(1):3–23, 1993.

[Ste10] Douglas Stebila. Reinforcing bad behaviour: the misuse of security indicators on popular websites. In *Proc. 22nd Australasian Conf. on Computer-Human Interaction (OzCHI) 2010*, pp. 248–251. ACM, 2010. DOI:10.1145/1952222.1952275.

[WI05] Tara Whalen and Kori M. Inkpen. Gathering evidence: use of visual security cues in web browsers. In Kori M. Inkpen and Michiel van de Panne, editors, *Proc. Graphics Interface 2005*, *Graphics Interface*, volume 112, pp. 137–144. Canadian Human-Computer Communications Society, 2005. URL http://portal.acm.org/citation.cfm?id=1089532.

[WM10] Ryan T. Wright and Kent Marett. The influence of experiential and dispositional factors in phishing: An empirical investigation of the deceived. *J. Management Info. Sys.*, **27**(1):273–303, July 2010. DOI:10.2753/MIS0742-1222270111.

A Online tasks instructions

[As noted in Section 3.2, participants were randomly given either the Facebook tasks first or the OpenID tasks first.]

A.1 Study Tasks—Facebook

Task F1—Rotten Tomatoes

1. RottenTomatoes.com is a movie information and review website. It allows users to login using their Facebook account. Please use your Facebook account to sign in to the Rotten Tomatoes website.
2. Please pick a movie on Rotten Tomatoes, rate it.

Task F2—Rotten Tomatoes

1. Please post a story about a movie from Rotten Tomatoes on your Facebook profile.
2. Log out of the web browser as if you were walking away from a public computer. (Do not log out of Windows, however.)

Task F3—LiveJournal

1. LiveJournal is a blogging and community site. It allows users to post comments using their Facebook account. Please visit the blog at the address "[omitted from paper]" and post a comment on a story using your Facebook account.
2. Log out of the web browser as if you were walking away from a public computer. (Do not log out of Windows, however.)

Task F4—Amazon

1. Amazon.com is an online shopping website. It allows users to post items from Amazon onto their Facebook profile. Please visit Amazon, find an item, and share it to your Facebook profile.

Task F5—Facebook

1. Log out of the web browser as if you were walking away from a public computer. (Do not log out of Windows, however.)
2. Go back to the Facebook site and log in.
3. Log out of Facebook.

A.2 Study Tasks—OpenID

For these tasks, you will be using a different account to login to websites. Here are the credentials you should use:

- Identity URL: [omitted]
- Password: [omitted]

Task O1—LiveJournal

- LiveJournal is a blogging and community site. It allows users to post comments using their OpenID account. Please visit the blog at the address "[omitted]" and post a comment on a story using your OpenID account.

Task O2—BlogSpot

- BlogSpot is another blogging and community site. It allows users to post comments using their OpenID account. Please visit the blog at the address "[omitted]" and post a comment on a story using your OpenID account.

B Survey

[In the survey questions reproduced below, we use _____ to indicate that the question allowed a free-form answer, ◯ to indicate that a single choice could be made, and □ to indicate that multiple choices could be made.]

B.1 Background

You can skip any questions you prefer not to answer.

1. What is your participant number? _____
2. What is your age? _____
3. What is your gender?
   ◯ Male ◯ Female ◯ Prefer not to say
4. What is the highest level of education you have completed?
   ◯ Some high school
   ◯ High school diploma
   ◯ TAFE diploma[8]
   ◯ Some university education
   ◯ Bachelor's degree
   ◯ Master's degree
   ◯ Doctoral degree
   ◯ Other
5. Are you currently a student?
   ◯ Yes ◯ No
   If yes, what is your year and major? _____
6. Are you currently employed?
   ◯ Yes ◯ No
   If yes, what is your occupation? _____
7. Is English your first language?
   ◯ Yes ◯ No

B.2 Technology Experience and Usage

You can skip any questions you prefer not to answer.

8. Do you use a computer daily for work?
   ◯ Yes ◯ No
9. Rate yourself on this scale:
   ◯ 1—I often ask others for help with the computer
   ◯ 2 ◯ 3 ◯ 4
   ◯ 5—Others often ask me for help with the computer
10. Please specify the brand and model of your mobile phone. _____
11. Do you know any programming languages?
    ◯ Yes ◯ No
    If yes, which programming language(s)? _____

[8] *[In Australia, TAFE stands for Technical and Further Education, and such institutions typically offer vocational tertiary education courses.]*

12. Do you have a degree in an IT-related field (e.g., information technology, computer science, electrical engineering, etc.)?
    ◯ Yes ◯ No
13. Have you ever (select all that apply)
    ☐ Designed a website
    ☐ Registered a domain name
    ☐ Used SSH
    ☐ Configured a firewall
    ☐ Created a database
    ☐ Installed a computer program
    ☐ Written a computer program
    ☐ None of the above
14. Please name a few operating systems that you know, if any: _____
15. How much time do you spend on the Internet per week?
    ◯ 1-5 hours ◯ 6-10 hours ◯ 11-20 hours
    ◯ 21-30 hours ◯ 31+ hours
16. Please name the browser you most frequently use: _____
17. How many non-spam email messages do you receive on average each day?
    ◯ Less than 10 ◯ 10-30 ◯ 30-50
    ◯ 50-100 ◯ More than 100
18. If you know, please describe what a "security certificate" is in the context of the Internet, or write "Don't know"? _____
19. If you know, please describe what is meant by "phishing", or write "Don't know"? _____
20. Have you ever taken or taught a course on computer security?
    ◯ Yes ◯ No
21. Have you attended a computer security conference in the past year?
    ◯ Yes ◯ No
22. Is computer security one of your primary job responsibilities?
    ◯ Yes ◯ No
23. Please check all of the following statements that describe your password habits.
    ☐ I use the same password for every website.
    ☐ I have a few passwords that I use interchangeably.
    ☐ I have one password that I use for important sites and another password I use for less important sites.
    ☐ I use different passwords for each site.
    ☐ I use my web browser's password manager to store my passwords.
    ☐ I write my passwords down on a piece of paper.
    ☐ I use a separate program to store my passwords.
24. Do you have an up-to-date virus scanner on your computer?
    ◯ Yes ◯ No
25. In the past month, which social networking sites have you used (check all that apply)?
    ☐ Facebook ☐ Google+ ☐ Twitter
    ☐ Tumblr ☐ WordPress ☐ MySpace
    Other: _____

B.3 Single Sign-On

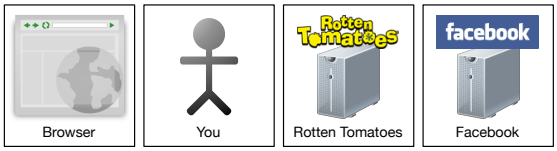
26. Do you have a Facebook account?
    ◯ Yes ◯ No
27. Have you heard of "single sign-on"?
    ◯ Yes ◯ No
    If yes, what do you think "single sign-on" means? _____
28. Have you previously experienced using a single username and password to access different systems?
    ◯ Yes, within a single organization
    ◯ Yes, across multiple organizations

◯ Yes, on the web
◯ No

If you selected any of these, please name the organization(s) and/or website(s). _____

29. Do you have an OpenID account?
    ◯ Yes ◯ No ◯ Don't know
    If yes, what is the name of the OpenID provider? _____
30. Have you ever used a "share" button on a website to share something on your Facebook profile?
    ◯ Yes ◯ No
    If yes, which website(s)? _____
31. Have you ever used your Facebook account to sign in to another website?
    ◯ Yes ◯ No ◯ Don't know
    If yes, what other websites you have signed in to using your Facebook account? _____
32. How do you decide if it is safe to enter your username and password on a particular website? _____
33. Please rate each task in the study based on how difficult it was to complete the task (1=very easy, 2=easy, 3=hard, 4=very hard).
    (a) F1: Rate a movie on Rotten Tomatoes.
    (b) F2: Post a story about a Rotten Tomatoes movie on Facebook.
    (c) F3: Post a comment on LiveJournal using Facebook.
    (d) F4: Share an item from Amazon on Facebook.
    (e) O1: Post a comment on LiveJournal using OpenID.
    (f) O2: Post a comment on a blog at BlogSpot using OpenID.

    Using the provided picture cut-outs, please draw how you think information (your username, password, and profile information) flows from one to the other when you sign on to the Rotten Tomatoes website:



[Cut-outs of the 4 images above were provided to participants to paste onto a page and then draw on.]

34. After completing these tasks, do you think the websites in the study (i.e., Rotten Tomatoes, LiveJournal, Amazon, BlogSpot) know your password from Facebook or OpenID?
    ◯ Yes ◯ No ◯ Don't know
    If yes/no, please explain why: _____
35. If someone were to hack into Rotten Tomatoes, do you think they could use that information to login in to your Facebook account?
    ◯ Yes ◯ No ◯ Don't know
    If yes/no, please explain why: _____
36. If someone were to hack into Facebook, do you think they could use that information to login to your Rotten Tomatoes account?
    ◯ Yes ◯ No ◯ Don't know
    If yes/no, please explain why: _____
37. In the future, if you encounter a website that supports using third-party accounts to login (similar to the websites in the study), will you use your existing account from Facebook or OpenID to login?
    ◯ Yes ◯ No ◯ Depends ◯ Don't know
    If yes/no/depends, please explain why: _____
38. Please indicate which of the following indicators you use to decide if it is safe to enter your username and password on a particular website?
    ☐ https
    ☐ lock icon on the page
    ☐ certificate
    ☐ website privacy statements
    ☐ type of website

☐ professional-looking site
☐ lock icon in the browser
Other (please specify) _____

39. When you went back to Facebook to login in task F5.2, did you actually need to re-login or was your account still logged in?
◯ Yes ◯ No ◯ Don't remember
If yes, was this what you expected? Why or why not? _____

C Bayesian models

We analyzed the recorded eye-tracking data using Bayesian two-way analysis of variance and standard null-hypothesis testing. We examined mean gaze duration per fixation, mean number of fixations, and mean total gaze duration per task. After examining the initial data, we found that a log transform of the mean gaze duration per fixation would transform the data into a normal distribution, rendering it more amenable to standard null-hypothesis testing.

For all of our analysis we at looked two nominal predictors: *expertise* and *task*. While our experiment design is within-subjects, due to the nature of data collection, there are tasks where subjects may not fixate at all, leading to missing data points for groups of subjects. This is fine for number of fixations and total fixation duration per task, where we can record a 0 for number of fixations, or total fixation duration, respectively. However, for mean gaze duration per fixation, it makes analyzing the task factor difficult, as a task with no fixations has an undefinded mean fixation duration. To address this issue, we looked at two situations:

1. 3 levels of expertise: novice, computer expert, and security expert; but 2 levels of task: Facebook or OpenID.
2. 2 levels of expertise: security non-expert, and security expert; but 7 levels of task.

The analysis in condition 1 demonstrated no credible difference between novices and computer experts in terms of mean gaze duration, providing support for the validity of the analysis in condition 2.

For our statistical analysis we used `R` with `JAGS`, through the library `rjags`. We also used the library `coin` to faciliate the use of the Mann-Whitney Rank Sum Test during our cross-validation.

C.1 Bayesian model definitions

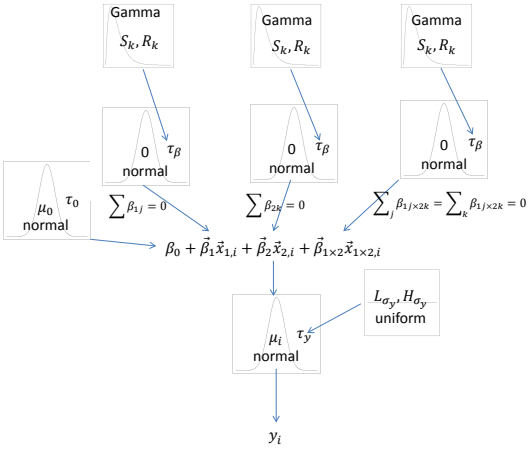
We used two different Bayesian ANOVA models using two nominal predictors of mean log-fixation duration and mean number of fixations. We evaluated the Bayesian model for the conditions in Section C and each of the duration models had the same parameters, but evaluated under different factor conditions. We looked at individual factors as well as interactions between factors [Kru10]: $y = \beta_0 + \vec{\beta}_1\vec{x}_1 + \vec{\beta}_2\vec{x}_2 + \vec{\beta}_{1\times2}\vec{x}_{1\times2}$ where the each dot product of the vectors β_i and x_j sums to 0. This allows us to understand how each factor relates to the baseline β_0.

The hierarchical model used for the mean log duration analysis (Figure 9) has the same structure between conditions 1 and 2, but has different levels j and k, depending on the condition. The model for mean number of fixations has the same overall structure, but takes its sample values from a Poisson distribution (Figure 10) .

We use skeptical priors, informed by the data to reduce the need for long burn-in times. Our priors are based on the means of the data, but use larger standard deviations than what are observed. For example, we calculate the shape and rate parameters by using the observed standard deviation of the entire dataset as the mode and standard deviation of the gamma distribution. This gives a prior distribution that is based on the data, but is broad enough to avoid an overly biased assumption.

We assume all of the data comes from the same overarching distribution, but, as the data is integrated, the posterior will reveal any differences based on the factors being investigated. This is akin to standard NHST in that we can evaluate the differences between the distributions, and if those differences are credibly different than 0.0, we can conclude that, given the data, we can be reasonably certain that the groups differ. In order to get a good estimation of the desired distribution, we sample the model through a Markov Chain Monte Carlo (MCMC) process.

Figure 9: Hierarchical Bayesian Model for mean log(duration) of fixations. μ_0 and τ_0 represent the baseline mean and precision of the normal distribution representing the log transformed duration data as a whole. S_k and R_k are the shape and rate parameters of the gamma distributions of σ_β, which are used to calculate τ_β. τ_β is the precision $(1/\sigma_\beta^2)$ of the normal distributions used to model the deflections β_1 and β_2. β_{1j} represents factor 1 (experience), across levels j. β_{2k} represents factor 2 (task), across levels k. L_{σ_y} and H_{σ_y} represent the low and high values for the uniform distribution describing τ_y. τ_y is the precision of the normal distribution with mean μ_i. This final normal is generated for each data point according to the equation given in the figure. The sample data point y_i is taken from this final distribution. This illustration shows the prior distributions, which are adjusted by the data before the final samples are made.



C.2 MCMC features

The goal of the hierarchical model is to describe a Markov Chain Monte Carlo process to recover the correct distributions of the factors under consideration by sampling from the given state space. When a long enough chain of samples is considered, it can recover the features of a given equilibrium distribution [SR93]. However, given the nature of the sampling, there can be correlations between samples at time t and $t-1$. We also need to be certain that the MCMC process is sampling from the correct distribution.

In order to address these issues we use multiple sample chains, each with a period of time to adapt and burn-in the MCMC sample. Because our prior assumptions are initialized in an informed way, our burn-in and adaption phases need not be too long, but we use 12000 steps to adapt each MCMC sample chain, and 15000 steps to burn the chain in to the correct distribution [SR93].

Once the MCMC is in place to sample the distribution, we save a total of 250000 samples. However, we only take 1 out of every 75 samples to reduce autocorrelation. Thus, each one of our 25 chains has a length of $7.5e05$[9]. This lets us perform convergence analysis to ensure that the chains in fact, converge to an equilibrium distribution [SR93]. After these assurances that our Bayesian analysis is robust, we also cross-validated our results with standard null-hypothesis testing.

C.3 Standard null-hypothesis testing

Since Bayesian analysis is not yet standard, we compared the Bayesian results with standard null-hypothesis testing using a two-way analysis of variance with interactions between expertise and task. We corrected for the unbalanced nature of the results using Tukey's Honest Significant Difference. We performed analogous ANOVA comparisons on all of the Bayesian analysis to test agreement.

[9] $\frac{7.5e05 \text{ samples}}{\text{chain}} \times 25 \text{ chains} = 1.875e7 \text{ total samples} \times \frac{1}{75} \text{ samples} = 250000 \text{ kept samples}$

Figure 10: Hierarchical Bayesian Model for mean number of fixations. μ_0 and τ_0 represent the baseline mean and precision of the normal distribution representing the mean number of fixations data as a whole. S_k and R_k are the shape and rate parameters of the gamma distributions of σ_β, which are used to calculate τ_β. τ_β is the precision $(1/\sigma_\beta^2)$ of the normal distributions used to model the deflections β_1 and β_2. β_{1j} represents factor 1 (experience), across levels j. β_{2k} represents factor 2 (task), across levels k. This final Poisson distribution is generated for each data point according to the equation given in the figure. The sample data point y_i is taken from this final distribution. This illustration shows the prior distributions, which are adjusted by the data before the final samples are made.