SECURITY MESSAGES

OR:

HOW I LEARNED TO STOP DISREGARDING AND HEED THE WARNING

by

David William Eargle

Bachelor of Science in Information Systems, Brigham Young University, 2013

Master of Information Systems Management, Brigham Young University, 2013

Submitted to the Graduate Faculty of

Katz Graduate School of Business in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

KATZ GRADUATE SCHOOL OF BUSINESS

This dissertation was presented

by

David Eargle

It was defended on

April 12, 2017

and approved by

Laurie Kirsch, PhD, Professor

Narayan Ramasubbu, PhD, Associate Professor

Scott Fraundorf, PhD, Assistant Professor, Department of Psychology

Anthony Vance, PhD, Associate Professor, Brigham Young University

Dissertation Advisor: Dennis Galletta, PhD, Professor

Security Messages

or:

How I Learned to Stop Disregarding and Heed the Warning

David Eargle, PhD

University of Pittsburgh, 2017

Attacks on information security continue to be reported in the media, and result in large losses for organizations. While some attacks are the result of sophisticated threats, others can be traced to failures by organizational insiders to observe basic security policies such as using caution when opening unsolicited email attachments. Faced with the challenges and time demands of everyday stressors, security policy compliance can be costly for individuals; security actions require time and distract attention from other primary tasks. This costliness can lead individuals to ignore prompts to perform security updates, scan their computers for threats, or reboot their computers to apply security updates.

This dissertation contains three studies that address the following overarching research question: How can end-user adherence to security messages be better understood and improved, and how can theory inform security-message design? First, two complementary studies are presented that examine the integration of media naturalness theory into a security message context using field study and fMRI designs. Study 1, the field study, unobtrusively captures objective measures of attention from Amazon Mechanical Turk users (N=510) as they perform a between-subjects deception protocol. Study 2, the fMRI study, examines neural activations from a within-subjects participant design (N=23) in response to different security message designs with integrated emotive human facial expressions. Data from studies 1 and 2 show that warnings with integrated facial expressions of threat (fear, disgust) generally elicited greater adherence rates and higher evidence of cognition and elaboration than did warnings with integrated neutral facial expressions or than did warnings with no integrated facial expressions, supporting our hypotheses. Study 3 explores the pattern of risk taking and analysis that users engage in when interacting with interruptive security messages. The corroboration of multiple behavioral dependent variables suggests that users predominantly use a bimodal risk tradeoff paradigm when interacting with interruptive security messages. All three studies address the overarching research question of understanding and improving end user adherence to security messages.

Keywords: security messages, threat attention, media naturalness theory, NeuroIS, risk tradeoff, heuristic-systematic model

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Attacks on information security continue to be reported in the media, and result in large losses for organizations. For example, Target recently reported fourth quarter 2013 profits to fall by $441 million, or 46% from the same period the previous year (Harris 2014). While some attacks are the result of sophisticated threats, others can be traced to failures by organizational insiders (Willison and Warkentin 2013) to observe basic security policies such as using caution when opening unsolicited email attachments (see the RSA hack, Schneier 2011). Individuals may ignore security warnings and fail to perform security updates, scan their computers for threats, or reboot their computers to apply security updates. Understanding why individuals ignore security warnings, and how to decrease rates of users disregarding warnings, is of paramount importance.

One view of why warnings are disregarded places the blame squarely on the warning's design for users' bad security behaviors. In this paradigm, security policy compliance can be costly for individuals, who are faced with the challenges and time demands of everyday stressors. Security actions require time and distract attention from other primary tasks (Adams and Sasse 1999; West 2008). The onus is on the design of security warnings to first capture, and then hold, attention. That goal met, the warnings then need to effectively educate users about both the threat at hand and about the available responses to the threat, with the aim of motivating an informed secure choice (Wogalter 2006a). Optionally, warnings can be designed that guide users towards choices that the designers consider to be safest (termed "opinionated design", Felt et al. 2015). Unfortunately, in the eyes of many, warnings often fail at the basic requirement of garnering attention, despite much research exploring and attempting to mitigate poor attention (Anderson et al. 2016b).

Contrasting sharply with the soft-paternalism approach of blaming the design and guiding the user is the following view: users are purely self-interested actors who have little to personally gain from complying with security policies, and the way to boost secure behavior is through punishment and sanctions. In this view, users conscientiously make risk tradeoff decisions when they encounter security warnings, and they get away with as much as they can without getting caught. Research that takes this view applies theories from criminology towards tipping the scales of the risk tradeoff decision (e.g., D'Arcy et al. 2009; Siponen and Vance 2010).

Both views of user interactions with security decisions can fit within a framework developed by Wogalter called the communication-human information processing model (C-HIP; 2006a). This framework outlines a process model with stages through which individuals must pass before a warning communication can effectively elicit a desired behavior. In this framework, the first and most basic requirement is attention – a warning communication must capture and hold an individual's attention. Following this, an individual must comprehend the warning. Then, attitudes and beliefs must be in alignment with the individual carrying out the desired behavior. Lastly, an individual must be motivated to align with the desired behavior. In C-HIP, if any of these "gates" are not successfully passed, the warning communication has failed.

While C-HIP can apply to warnings from contexts as diverse as road signs or a recorded message played at an airport, it also can be applied to an interruptive security message context. Security messages aim to draw attention away from a primary task towards themselves, educate a user about a threat, and motivate them to behave securely – or at least to make an informed decision. If we focus on the view that users pay little attention to security warnings, regardless of the content of a security warning, then we focus our efforts on improving warning designs so that they have better likelihood of passing the first attention gate in C-HIP. If, on the other hand, we take the view that users make conscientious security decisions each time, then we assume that the first attention C-HIP gate is being sufficiently met, and we should instead focus our efforts on designing security warnings to the end that they better motivate secure behavior (i.e., with sanctions, rewards, etc.). Addressing this problem domain, this dissertation includes three studies that fall under the following overarching research question:

RQ: How can end-user adherence to security messages be improved, and how can theory inform security-message design?

The first two studies in this dissertation are complementary. Using field study and fMRI designs, they focus on the design of interruptive security messages, using media naturalness theory (Kock 2009) to design security warnings with full images of human facial expressions in an attempt to overcome attention hurdles. They are anchored in the research view that "users are not the enemy" and that better warning design is the best path to take towards improving secure behaviors. Study 3 is then described, which explores the pattern of risk-taking that users engage in when interacting with interruptive security messages. The study complements the first two by investigating the extent to which it is correct to view users as volitional policy violators as opposed to being well-meaning but inattentive. The three studies as a whole will allow us to know which stage of the C-HIP model is most salient on which to focus research and design efforts, and under which conditions one may become more salient than another. A conclusion follows the three studies, highlighting what can be learned from the three studies as a whole.

2.0 STUDIES 1 AND 2 – INTEGRATING FACIAL CUES OF THREAT INTO SECURITY WARNINGS

2.1 MOTIVATION

A pressing reason for poor security behavior is users failing to respond appropriately to security messages (Anderson et al. 2016b). Security messages seem to be failing to capture users' attention. Without the basic requirement of user attention, a security message's components have little to no effect towards effectively communicating about a threat and assisting users in making an informed choice (Wogalter 2006a). Unfortunately, failing an understanding of the seriousness of security threats, users' default choice is to dismiss the warnings (known as the "dancing pigs" problem, see Schneier 2004). This makes sense when considering that security warnings are often interruptive to some primary task.

In a bid to draw attention, security message designs commonly use threat cues such as yellow triangles, ominous exclamation marks, or cartoonish faces; yet the warnings still have troubling levels of non-adherence. Perhaps this poor security behavior is because the threat cues are too abstract (Felt et al. 2015), or perhaps because users become habituated to them and fail to give the warnings conscious attention after repeated exposures (Anderson et al. 2015).

One theoretical approach to boosting engagement with security warnings is offered by media naturalness theory, which predicts that the more closely a systems interface maps to natural human communication patterns, the more engaging it will be for a user (Kock 2009; Riedl et al. 2014). Face-to-face communication between humans rates especially high on naturalness. Humans are thought to have adapted to this form of communication over centuries of evolution, to the point that human facial expressions can carry vivid environmental cues for an observer. This study focuses on facial expressions of threat, including fear and disgust, which are potent cues of danger in the immediate environment. Fearful facial expressions indicate threat of physical attack (Gray 1987), and disgust facial expressions indicate contamination in the environment (Rozin and Fallon 1987). Furthermore, because of their deeper evolutionary ties, warnings with facial threat cues may be more resilient to habituation over repeated exposures.

This work uses NeuroIS methods to examine user interactions with security warnings. NeuroIS methods are apt for use in a security context because reactions such as fear and threat processing, which are important for security contexts, are challenging to measure. For example, they may be too subtle to rise to a level of consciousness for users to be able to accurately self-report them (Anderson et al. 2016b; Dimoka et al. 2011). Two studies were used to test several hypotheses. The first uses mouse cursor tracking in a field study protocol, and the second uses a functional magnetic resonance imaging (fMRI) protocol. Behavioral and perceptual measures complement the neural measures for both studies.

This work informs the design of security messages in practice. It also further extends media naturalness theory into the domain of IS research. While other IS research has considered the impact of photo-realistic faces vs avatars on trust in an ecommerce setting (Riedl et al. 2014), to our knowledge no research has considered the impact of emotive facial expressions in an IS context, let alone facial expressions of threat in an IS security context.

2.2 LITERATURE REVIEW

2.2.1 Attention to security messages (or the lack thereof)

A major contributor to security message failure is a simple lack of attention. We define security-message attention as does the computer-human information processing framework (C-HIP, Wogalter 2006a), where attention to a warning message is described as the first behavioral gate a user must pass on the way towards adhering to a warning message. If the basic requirement of attention to the warning communication is met, then other gates must be passed on the way towards effective threat communication and motivation by the warning. The gates following attention include comprehending the message and the communicated threat, as well as being motivated to comply with the message (see Figure C-1).

However, in a review of HCI literature, a failure of attention was implicated in poor user reactions towards security messages in 22 out of 29 papers (see Anderson et al. 2016b). Some laboratory experiments have pointed to the role of habituation in users' failure to heed warnings and security indicators (Dhamija et al. 2006; Good et al. 2005; Schechter et al. 2007; Sharek et al. 2008; Wu et al. 2006). Egelman et al. (2008) found a significant correlation between recognition and disregard of security warnings. Sunshine et al. (2009) observed that participants remembered their responses to previous security warnings and applied them to other websites even if the level of risk had changed. Felt et al. (2012) found that 42% of participants were not aware of having interacted with security permission dialogs before installing an Android app on their devices. Similarly, some participants in Sotirakopoulos et al. (2011) study clicked through security warnings during a task, and later reported that they had not seen any security warnings (see also user account control prompts in Motiee et al. 2010).

These laboratory study results reflect those in the field. Akhawe and Felt (2013) found that in approximately 50% of the most common type of secure sockets layer (SSL) web browser warnings in Google Chrome, users decided to click through in 1.7 seconds or less, a finding that "is consistent with the theory of warning fatigue" (Akhawe and Felt 2013, p. 14). Felt et al. (2014) found that warning design explained between one-third and one-half of the difference between Chrome and Firefox SSL warnings. Bravo-Lillo et al. (2013) conducted a large field experiment using Amazon Mechanical Turk in which users were rapidly exposed to Windows operating system software installation security confirmation dialog messages (i.e., the "Are you sure you want to run this software?" prompt). After a period of 2.5 minutes and a median of 54 exposures to the dialog message, only 14% of the participants recognized a change in the content of the confirmation dialog in their control (status quo) condition.

Researchers from industry and academia have tested various interface designs, iterating towards solving the problem of low attention to security messages. Felt et al. (2014) and Felt et al. (2015) integrated various visual threat icons, including padlocks, police officers, and cartoon criminals. Anderson et al. (2016c) tested a battery of different visual designs for warning messages, and from them created a set of designs most resilient to habituation of attention. However, the problem of inattention persists, and a fresh approach is called for.

2.2.2 Media naturalness theory

Theories from evolutionary psychology can provide fresh insights for IS research on security warnings (Kock 2009). Evolutionary psychology is based on Darwin's theory of evolution (Darwin 1859) which describes how humans have evolved to have behavioral modules through natural selection which increase the overall fitness of the race, which indicates its ability to survive.

Considering humans' communication abilities, for most of the history of mankind the predominant form of communication has been face-to-face (see Kock 2009). Therefore, humans should excel at this form of communication, having evolved to do so. Media naturalness theory (Kock 2004; Kock 2009) is based on this evolutionary communication argument. It proposes that the more closely a communication medium aligns with traditional face-to-face human communication patterns, the more likely it will invoke a humanoid's innate evolved communication module. The theory defines natural communication by observing five characteristics of face-to-face communication: "(1) a high degree of colocation, which would allow the individuals engaged in a communication interaction to see and hear each other; (2) a high degree of synchronicity, which would allow the individuals to quickly exchange communicative stimuli; (3) the ability to convey and observe facial expressions; (4) the ability to convey and observe body language; and (5) the ability to convey and listen to speech." (Kock 2009, p. 407). It measures communication by how "natural" it is according to these characteristics and makes predictions on three dependent variables: cognitive effort, communication ambiguity, and physiological arousal.

Media naturalness theory was developed to explain theoretical gaps of its non-evolutionary cousin, media richness theory (Daft and Lengel 1986; Daft et al. 1987). There are two main differences between the theories: first, a difference in baseline comparison, and second, a difference in predicted outcomes. Media naturalness theory anchors itself in the ideal of face-to-face communication, whereas media richness theory has no such anchor and whose baseline is a simple absence of richness. Media naturalness theory's dependent variables of cognitive effort, communication ambiguity and physiological arousal are notably different from media richness' less cognitively-oriented predictions of media choice and task outcome. One of the most important results of these differences is that media naturalness theory would predict that if a communication medium is less *or more* rich than face-to-face, outcomes will suffer, whereas media richness would always predict a positive outcome for increased richness.

2.2.3 Reactions to threat cues

The human response to emotional stimuli discriminates between specific emotions of the stimuli, not simply their valence. For example, fearful and sad stimuli are reacted to differently, despite both fear and sadness having negative valence (e.g., Öhman et al. 2001; van Hooff et al. 2013). Emotional images triggering threat processing may include images of sharks, spiders, and snakes (e.g., Kock et al. 2008; Ohman and Soares 1994). Human reactions to threatening stimuli are thought to be an evolutionary module. Supporting the idea of responses to threat cues being an evolutionary module is that threat cues are processed automatically. Physiological markers suggest evidence of threat processing when individuals observe threatening stimuli even when the stimuli are masked – that is, shown for so short a time that individuals cannot report having seen them (e.g., fMRI activations in Nomura et al. 2004; skin conductance response [SCR] in Ohman and Soares 1994).

Notably, images of facial expressions of threat have also been used to prompt threat-processing. The essential elements of a facial expression are the eyebrows, mouth, and eyes (Ekman and Friesen 2003), and Lundqvist et al. (2004) ranks the importance of those elements in that order. In Öhman et al. (2001), participants were able to identify the presence of a face of threat (an angry facial expression) from a crowd of faces more quickly and accurately than of faces displaying sadness or happiness. As for neural correlates of processing facial cues of threat, fMRI studies most often associate the processing of images of expressions of fear and disgust with activations in the amygdala and insula respectively (see Phillips et al. 2004). However, under conditions of partial inattention, disgust images have also been associated with amygdala activation (Anderson et al. 2003).

We hasten to emphasize that processing threat signals does not necessitate experiencing fear. Threat processing would only lead to the emotion of fear if (1) a legitimate threat to one's safety is perceived and if (2) the threat is deemed immitigable. See the "psychological level" column in Figure 1 for an illustration of the process leading up to the emotion of fear, where fear is represented in the figure by the autonomic response stage late in the psychophysiological level. Therefore, potent threat signals integrated into a security message context should prompt low-level neural emotional threat information processing, i.e. threat attention. However, the threat signals will not necessarily cause palatable fear in computer users because of low perceived relevance of the threat to one's physical well-being. Fear is typically reserved for perceived threats to one's physical safety, whereas security warnings relate to threats to data or information. (Johnston et al. 2015). To the degree that this low likelihood of fear is true, there should be lower danger of user emotional burnout from feeling fear every time they are exposed to an evolutionary-threat-cue-integrated security message. Users are also not likely to enter fight-or-flight state (Cannon 1932) and make automatic, perhaps unwise responses to the security messages. Furthermore, the increased threat attention should not preclude future reasoned judgement and action by users. We note that we chose to use a human facial expression over "scary" images such as ones of snakes, spiders, and monsters, because frightening images aim to evoke fear and attendant fight-or-flight responses, while our goal is to only trigger early-stage threat attention while explicitly *avoiding* automatic, mindless fight-or-flight reactions.

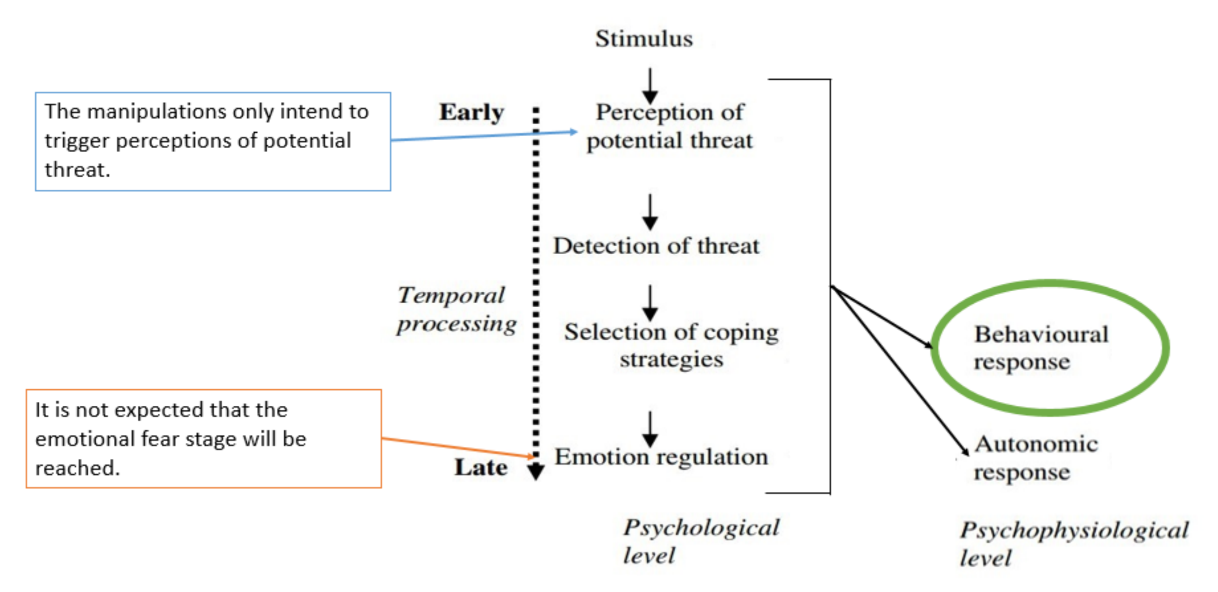

Stimulus
The manipulations only intend to trigger perceptions of potential threat.
Early
Perception of potential threat
Detection of threat
Temporal processing
Selection of coping strategies
Behavioural response
It is not expected that the emotional fear stage will be reached.
Late
Emotion regulation
Autonomic response
Psychological level
Psychophysiological level

Figure 1. A cognitive neurobiological information-processing model of fear and anxiety. Adapted from Hofmann et al. (2012)

While previous research has investigated the effects of integrating faces with eyes into browser security warnings (Felt et al. 2015), the set of studies presented in this dissertation are the first to integrate a full human facial expression. Full human facial expressions should be much more potent than the tiny faces with eyes used in the aforementioned study. Thus, these studies stand to make a novel contribution to both research and practice.

2.3 HYPOTHESES

Our first hypothesis tests for the effect of including novel stimuli into warning messages. Amer and Maris (2007) describe how any dramatic change to a warning design is likely to refresh attention and processing of the image. This can be explained by habituation theory (Rankin et al. 2009). Habituation theory explains that repeated viewing of a stimulus will lead to increasingly attenuated attention, while a change in the stimulus can lead to sensitization and renewed attention.

Given that study participants will begin our studies already in a potential state of warning fatigue (Akhawe and Felt 2013) from already having been exposed to security warnings in their typical computer work, seeing a security warning with an unfamiliar element such as an integrated facial expression of any variety should lead to sensitization and renewed attention.

H1: Warnings with any integrated facial expression will elicit greater levels of attention and elaboration than will security messages without any integrated face.

Our second hypothesis considers the difference between abstract threat cues and natural ones such as threatening facial cues. Security messages commonly contain symbols and cues of threat, including red colors, stop signs, and bolded words such as "warning!" punctuated by exclamation marks. These are intended to boost attention and threat processing. However, media naturalness theory (Kock 2009) would suggest that more natural communication stimuli, such as facial cues, will more effectively prompt threat attention than will abstract cues. This is predicted because one of the earliest forms of communication that humans are thought to have developed is face-to-face communication.

H2: Warnings with threat facial cues will elicit greater levels of attention and elaboration than will security messages with other types of facial cues (e.g., neutral expressions).

The third hypothesis contrasts fear and disgust facial threat cues to determine which of these more-natural stimuli fits best in a security-message context. Anderson et al. (2003) compared the effects of fearful and disgusted facial cues on brain activations. In the study, fearful facial expressions were associated with equivalent levels of amygdala response under conditions of attention and inattention. This suggests that the effect of observing fearful facial expressions is independent of conscious visual attention. However, facial cues of disgust were dependent on attention. Under conditions of inattention, disgust facial expressions were associated with *greater* amygdala activations compared to conditions of attention. Because a face in a security message will likely not be the most prominent component of the message (any of the other message components could also draw visual attention), we predict that integrated facial expressions will not be exclusively attended to. Therefore, they should trigger threat attention patterns, including amygdala activations, similar to the ones seen in the right amygdala for unattended-to stimuli in Anderson et al. (2003).

Furthermore, we predict differences between the two based on the mechanisms behind experiencing the associated emotions. Fear is thought to lead to immediate threat avoidance, but only when the threat is imminent. Experiencing disgust, conversely, leads to threat avoidance regardless of the immediacy of the threat (see Morales et al. 2012 for a review). While we do not expect that participants who observe facial expressions of these emotions will lead to them experiencing the same emotion in our security message context, it is possible that there are subtle differences in the low-level and early threat processing that occurs when exploring potential threats (see Hofmann et al. 2012 for a psychological process model of threat processing). Information security threats often report threats that will not be manifest until some future time, not necessarily immediately. For example, it is not uncommon for ransomware to take some time after the initial point of infection before calling home to its command-and-control server and launching the attack, leading to a considerable delay after a poor security choice before a user is aware that something is wrong (Mourad 2015). In other cases, organizations have had their systems infiltrated for months or years by advanced persistent threats before becoming aware of them (e.g., Tom 2016). Given these kinds of malware behaviors, it is understandable that computer users may perceive that a warned-of threat is not immediate, and therefore may instead freeze if they experience fear, or discount the threat based on its perceived temporal distance (Malhotra et al. 2002). Disgust effects should still prompt attention despite any perceived temporal distance of the information security threat.

> *H3: Disgust facial expressions integrated into security messages will elicit greater levels of attention and elaboration than will fearful facial expressions.*

The face cues should still be effective as long as the faces maintain distinguishable elements such as eyes, eyebrows, nose, and mouth. One method of altering a photograph is to 'posterize' it, or to bin colors using an adaptive threshold to make it more amenable for printing. This has the effect of making a picture seem more cartoonish, or hand-drawn. This adaptability of the faces' stylistic appearance will help warning interface designers have more freedom in their artistic expression of the faces, so as to fit company style guidelines. We summarize this practice-motivated hypothesis as follows:

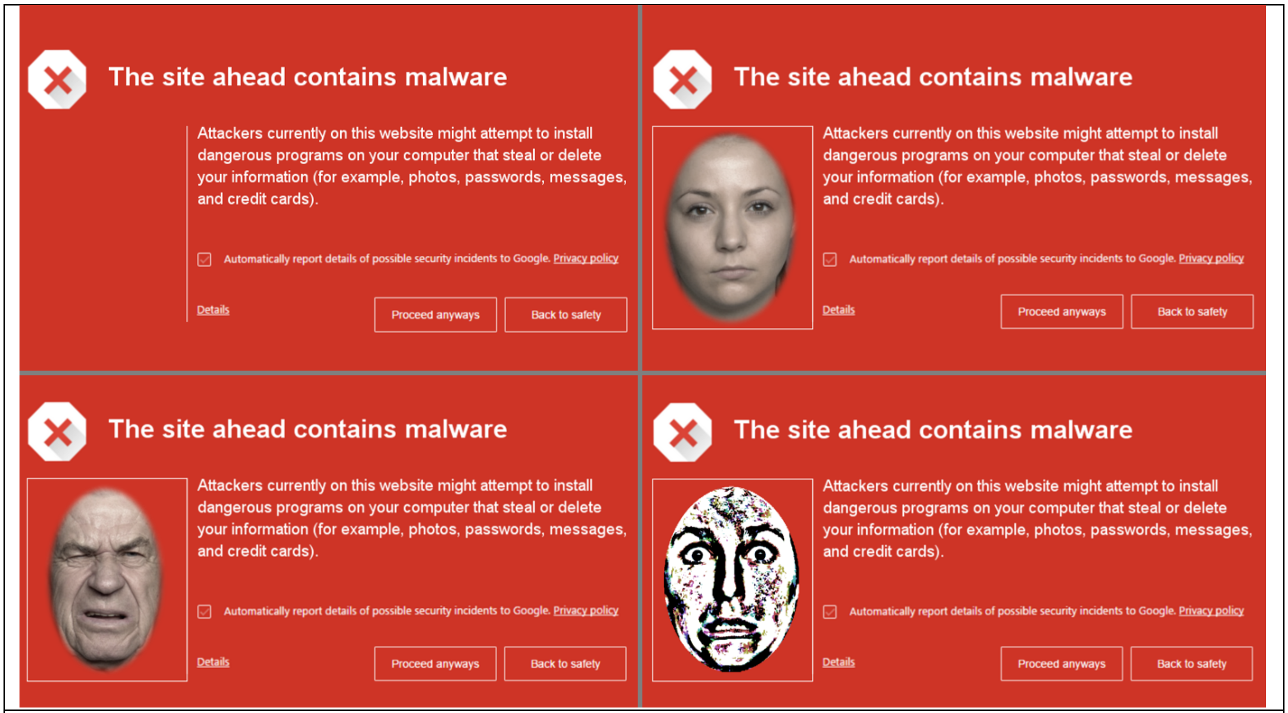
> *H4: The application of a visual artistic image filter to a facial expression will lead to negligible drops in levels of associated attention and elaboration, so long as key facial elements for threat (e.g., eyebrows, mouth, eyes: Öhman et al. 2001) remain discernible in the facial expression.*

We also consider how resilient to habituation the different threat cues will be. Users' attention has been shown to attenuate rapidly when new visual stimuli are integrated into security warning designs (Anderson et al. 2016b). However, facial signals are thought to have deep evolutionary ties. These deep ties may be more likely to consistently activate low-level neural emotional threat information processing (e.g., amygdala activation). While Breiter et al. (1996) shows that reactions to fearful facial expressions do decrease with over repeated impressions, there still remained significant amygdalar response even after repeated exposures. In our context, we explain this greater predicted resilience to threat facial cues as an innate response to natural stimuli. Abstract threat cues such as triangles and exclamation points are less likely to trigger innate responses, given that they do not have linkages to evolutionary modules as facial expressions do. Furthermore, facial signals in an IS context can be portrayed by different actors for each impression. Differing faces should be noticed by users, considering that humans are generally highly skilled at discriminating between faces, even more so than compared to discriminating among other classes of similar objects (such as cars, animals, etc.) (Gauthier et al. 2000). This polymorphism of the precise appearance of the facial signal could help to combat habituation compared to static warning symbols (Anderson et al. 2016c). Therefore, for several of our hypotheses, we propose complimentary habituation counterpart sub-hypotheses.

> *H1–4*: Security messages with integrated facial signals of threat will be more resilient to habituation over repeated exposures than will security messages with more abstract threat cues.*

2.4 INSTRUMENTS

Studies 1 and 2 used instruments from a common set of security warnings with integrated facial expressions that we developed. To make these instruments, we started with a set of 120 facial expressions randomly extracted from an emotion-validated bank of color images of actors faces displaying different emotions (Ebner et al. 2010). As is commonly done in neuroscience protocols using facial stimuli, we took an oval crop of the actor's face, with the hair line and the chin as the upper and lower vertical limits, and up to but excluding the ears as the horizontal limits (e.g., Anderson et al. 2003). Then, we created two new sets from these cropped filters: one which desaturated the colors down to 47%, and another which "posterized" the image using an adaptive spatial threshold. We then integrated these facial expressions into a modified Chrome malware warning, from build 51.0.2704.63 m (original displayed in Figure C-2). The location of the heading text, body text, and buttons was kept consistent with the base Chrome warning in order to control for warning component novelty. We opted to retain the whole text of the Chrome malware source image to increase external validity and warning realism. We shifted the body text a small amount towards the right to make room for our face image on the left, in keeping with past Chrome warning designs that have had images on the left of the body text (Felt et al. 2014). Our control (blank) security warning alongside a sample of facial-expression-integrated warnings is displayed in Figure 2.



The site ahead contains malware

Attackers currently on this website might attempt to install dangerous programs on your computer that steal or delete your information (for example, photos, passwords, messages, and credit cards).

☑ Automatically report details of possible security incidents to Google. Privacy policy

The site ahead contains malware

Attackers currently on this website might attempt to install dangerous programs on your computer that steal or delete your information (for example, photos, passwords, messages, and credit cards).

☑ Automatically report details of possible security incidents to Google. Privacy policy

The site ahead contains malware

Attackers currently on this website might attempt to install dangerous programs on your computer that steal or delete your information (for example, photos, passwords, messages, and credit cards).

☑ Automatically report details of possible security incidents to Google. Privacy policy

The site ahead contains malware

Attackers currently on this website might attempt to install dangerous programs on your computer that steal or delete your information (for example, photos, passwords, messages, and credit cards).

☑ Automatically report details of possible security incidents to Google. Privacy policy

Figure 2. From left to right and top to bottom, (1) our blank, adapted Chrome malware warning, followed by security warnings integrated with (2) a desaturated neutral expression, (3) a disgusted desaturated expression, and (4) a posterized fearful expression.

2.5 STUDY 1 – FIELD STUDY

We used a field study with a between-subjects repeated-measures design. We recruited 550 participants from the United States using Amazon's Mechanical Turk platform. Of the 510 participants whose data were used in the analysis, there were 314 males, 196 females, and ages ranged from 19–69 ($\bar{x}$ ~= 32). Data from Mechanical Turk has been found to be as reliable as data from other U.S. survey panels (Steelman et al. 2014), and more importantly, they are likely to be using their own computers, raising their sense of perceived risk (c.f. Boss et al. 2015; Vance et al. 2014).

Participants were directed to a server under our control running our experiment codebase, built on the psiTurk framework (McDonnell et al. 2012), where they were randomly assigned to one of seven treatment groups from the following design: face emotive expression (fear, disgust, neutral) $fully\ crossed\ with$ image filter (desaturated, posterized) $+$ control (no face). This design is graphically displayed in Figure 3. For each warning impression excepting for participants in control condition, participants saw a security warning with an integrated facial expression that was randomly selected from the associated set of actor images for the assigned treatment group.

	Emotion				
Filter	Fear + Realistic	Disgust + Realistic	Neutral + Realistic	+	Control (no face)
	Fear + Posterized	Disgust + Posterized	Neutral + Posterized		

Figure 3. Experimental design for Study 1

We used an IRB-approved deception protocol. The pretense was that participants were performing an image classification task, when in reality, we were interested in users' behaviors when they were presented with interruptive security messages. The image classification ruse and the security warning presentation are described below.

Participants performed a modified version of the image classification task described in Vance et al. (2014). In our task, participants were told that they would classify a series of images in order to help test a computer classification algorithm. It was explained that a series of live, external websites would be loaded into a frame in the center of the webpage (i.e., an `iframe` HTML element). For each page load, participants were asked to classify whether the image was a photograph of Batman or an artist's rendering. On top of a $1.25 base payment, participants were offered an additional $1.25 performance-based bonus payment. Each incorrect classification resulted in a "penalty" decrease in their bonus payment. Furthermore, to encourage attention to the task, we warned that too many incorrect responses would result in their work being rejected with forfeiture of any payment. Participants' current bonus status was depicted with an animated and labeled bar beneath the central `iframe`. Participants were encouraged to move quickly, limiting them to a maximum of 10 seconds for each classification. A timeout resulted in the classification being marked as "incorrect." After a 4-image practice round, participants classified 75 images. See Figure 4 for an example screenshot of the image classification protocol.

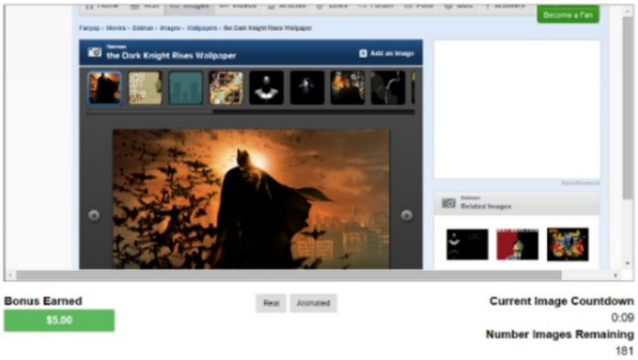
Five times during the main task, the page load within the central window was interrupted with a browser security warning drawn from the appropriate set for the participant's treatment group. The warning, based on Google Chrome malware warning build 51.0.2704.63 m, signaled that continuing to load the page would result in the visitor's computer becoming infected with malicious software ("malware"). The warning had a button allowing the user to proceed past the warning to the website (see Figure 5). Participants were required to use the Google Chrome browser to perform the experiment so that our warning design would seem more natural. We reason that users' familiarity (or lack of familiarity) with the Google Chrome browser and with its typical malware security warnings would likely be randomly distributed across treatment groups.

If, while a security warning was shown, participants made a guess about whether the image on the unseen screen was real or animated, they risked being marked wrong. Because each incorrect classification decreased the bonus earned and increased the likelihood of a participant's work being rejected, participants were financially motivated to ignore the warning.

To treat all participants fairly, at the end of the task, we presented a message saying "some errors were detected during the experiment," and we increased the bonus-earned amount to reflect the amount that would be earned by the lowest penalty group. We also credited back any markdowns that participants endured because of timeouts or incorrect classifications that occurred while a security warning was displayed.

In a post-task survey, participants were asked various questions, including items about their information security concerns and perceptions, whether they noticed the security warnings and the integrated facial expressions, and whether the warnings appeared realistic. A debriefing followed.

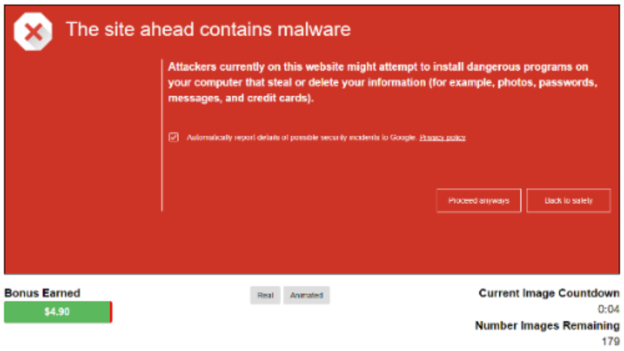
Despite the seeming complexity of the protocol, data from this study, presented in the following sections, indicated that users did notice the embedded warnings, and that they perceived that the warnings were both real and concerning (c.f. Vance et al. 2014).



Bonus Earned
$5.00

Current Image Countdown
0:09
Number Images Remaining
181

Figure 4. Example of image classification task demonstrating loaded page window and task control panel (adapted from Vance et al. 2014).



The site ahead contains malware

Attackers currently on this website might attempt to install dangerous programs on your computer that steal or delete your information (for example, photos, passwords, messages, and credit cards).

☑ Automatically report details of possible security incidents to Google. Privacy policy

Bonus Earned
$4.90

Current Image Countdown
0:04
Number Images Remaining
179

Figure 5. The security warning as it appeared to participants. Based on the Google Chrome malware warning, from build version 51.0.2704.63 m

2.5.1 Metrics

We consider various markers of security behaviors and cognition. From one view, the end goal of presenting a warning message in any context is for the message recipient to adhere to the warning (Wogalter 2006b). An intermediary aim is for the message recipient to carefully consider the warning and make an informed choice as to the risks involved, whether or not that choice involves distancing or approaching the warned-of threat (Wogalter 2006a). We test for both. First, we test for differences in actual security message adherence (choosing to load the site despite the warning) among treatment groups – i.e., did participant choose to continue to load the site despite being warned that it contained malware.

Second, we test for differences in reaction time among treatment groups. Reaction time, a form of mental chronology, is a commonly-used metric for cognitive effort (Jensen 2006). We measure reaction time from when a security warning first appears to when a security choice is made (i.e., when the warning disappears).

Third, we test for differences in cognitive engagement by examining two mouse-cursor movement statistics: (1) click latency, and (2) time idle. Both of these measures are markers of cognitive processes such as elaboration and uncertainty (Hibbeln et al. 2016; Jenkins et al. 2016). Click latency was considered for the first button click that occurred for a given warning impression. Time idle was measured in milliseconds (ms), and was accrued whenever the mouse remained unmoved for greater than 200 ms for a given warning impression. For click latency, higher latency is associated with greater cognitive processing and uncertainty. Greater time idle is likewise positively associated with elaboration and uncertainty.

2.5.2 Analysis and Results

We next report the analysis and results for tests of the main effects of face emotion, and also for face filter. Given our sample size of 510 participants, at most 4 treatment groups (4 groups for the face emotion main effect, and 3 groups for the filter main effect), and a high expected correlation among the 5 repeated measures (.95), G*Power 3.1.9.2 reports that we have sufficient statistical power to detect, at least, a medium-small effect size (*f* <= ~0.18).

Although we had data for five security warning impressions per participant, a post-hoc visual analysis of Loess curves for temporal variables such as reaction time suggested that an inflection point existed around the fourth warning impression (e.g., see Figure 6). An examination of participants' free-responses from the survey data suggested that after about four impressions, participants began to suspect the deception in the study. Using this to explain the inflection point, we therefore consider no more than four warning impressions in our analyses.

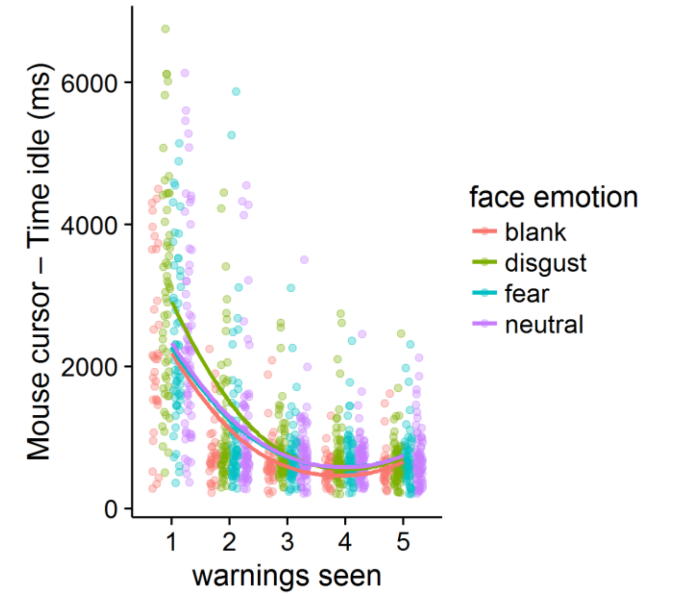


Figure 6. A demonstration of the inflection point that occurred around the fourth warning impression for each participant, which led us to only consider, at most, the first four warning impressions for each participant.

All continuous dependent variables (e.g., reaction times, mouse cursor click latency, and mouse cursor time idle) in models testing effects across time were natural-log-transformed to remedy non-normality of residuals and heteroskedasticity. For these tests, the main effect of number of warnings seen was always significantly negative. However, this effect on its own is not of interest to our research question, so we do not report the effect of time except when it is included in an interaction with treatment group levels.

To determine whether to include covariates in our analyses, we tested for whether several items were predicted by the emotion or filter treatment groups. The potential covariates that we tested were participant age, gender, preferred operating system, preferred browser, task performance accuracy, and whether English was their first language. We ran separate ANOVAs for each potential covariate with Type 2 errors on linear models, each including the emotion effect, the filter effect, and their interaction as independent variables. None of the omnibus *F* tests from any of these tests were significant at an alpha level of .05, so no covariates were included in any analyses. However, a few of the covariates showed overall significance at an alpha level of .10. Filter treatment group was predictive of English-as-a-first-language (p = .080). Also, emotion treatment group was predictive of preferred operating system (OS) (p = .066), and of task performance accuracy (p = .082). Given this, we describe how the pattern of differences for each test is impacted when the associated covariates are controlled for.

Because our task employed repeated measures on a binary dependent variable (whether or not the warning was dismissed), as a preliminary analysis we explored the ratio of ignored warnings for each participant and for each treatment group. This exploration is included in Appendix B, and it shows that the majority of participants (80%) were perfectly consistent across warning exposures in whether they ignored or heeded the warning. The likelihood of perfect consistency was not dependent on the assigned emotion or filter treatment group. Because this finding was not dependent on treatment group, a discussion of it will be postponed until section "4.0 – Final Discussion and Conclusions".

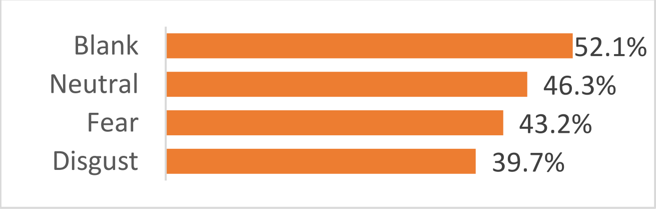
Because all comparisons tested a priori hypotheses, no corrections for multiple comparisons were made. One-tailed p-values are reported where appropriate.

2.5.2.1 Face emotion. We first tested for the impact of the emotion of the integrated facial expression. Each test averages across the different filter factor levels, and compares the responses to the control warning, which had no face.

Warning adherence rates. We tested for differences on proportions of warnings ignored for first exposures only between treatment groups. An overall difference among proportions was found, $\chi^2(df = 4) = 9.908, p = .042$. By rank order, blank warnings were the most likely to be ignored, followed by neutral, fear, and then disgust warnings (see Table 1). However, pairwise contrasts from a logistic regression model predicting whether the first warning was ignored by treatment group only found significant differences between blank and disgust warnings (*one-tailed p*=.043). This pattern of differences holds when controlling for task performance accuracy and for preferred OS, with the exception that the difference between blank and disgust warnings falls into marginal significance (one-tailed *p* = .064).

Table 1. Proportion that ignored first warning by emotion treatment

Condition	***n***	**Proportion that ignored first warning**
Blank	63	52.1%
Neutral	130	46.3%
Fear	134	43.2%
Disgust	122	39.7%



Reaction Time. We tested for differences on reaction times among emotion treatment groups, averaging across filter conditions using a linear mixed model with logit link using the first four warning impressions per participant. Emotion condition and impression order were specified as fixed effects, and a random effect was included for each participant to account for repeated measures. An omnibus ANOVA with Type 2 sum of squares[1] found differences in reaction times among emotion condition groups ($\chi^2(df = 3) = 9.892, p = .020$), the number of warnings seen ($\chi^2(df = 1) = 1667.43, p < .001$), and the interaction between the two ($\chi^2(df = 3) = 7.95, p = .047$). Follow-up contrasts for the main effect of emotion averaged across number of warnings seen found that participants in disgust, fear, and neutral conditions had longer reaction times than did the blank condition (one-tailed *p* = .005, *p* = .002, and *p* = .038 respectively), but the test also found that warnings with faces did not have statistically significant differences from one another for reaction times. This pattern of differences holds when controlling for task performance accuracy and for preferred OS.

[1] It is worth noting that when we performed the omnibus ANOVA with Type 3 sum of squares, the main effect of emotion treatment group was washed out by the interaction effect. This is understandable, given that Type 3 sum of squares tests for main effects in the presence of interaction effects, while Type 2 removes the interaction effects. We opt for Type 2 given that we have separate hypotheses for the main effects and for the slopes.

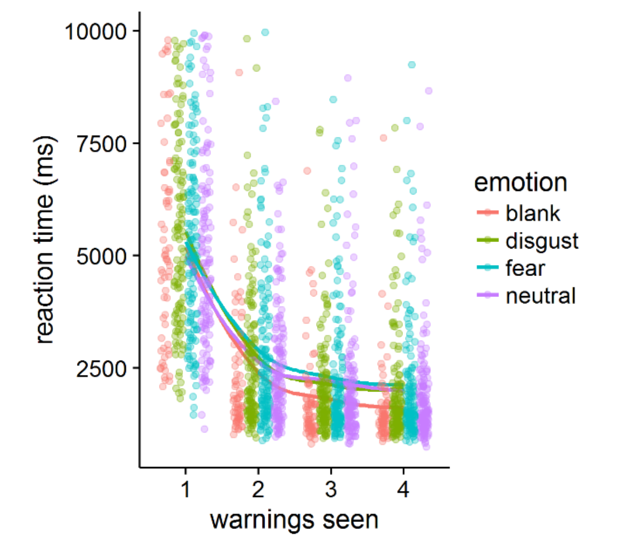


Figure 7. Loess curve plotting reaction times across warning impressions by emotion treatment group. Data points are horizontally dodged to aid in density visualization.

As for differences in slopes over time among emotion treatment groups, blank had a steeper downward slope than did fear or neutral conditions (one-tailed *p* = .030 and *p*=.006 respectively), but was not significantly different from disgust's slope. Disgust had a steeper downwards slope than did neutral, (one-tailed *p*=.028), but fear was not significantly different from neutral. Slopes for disgust and fear were not statistically significantly different from one another. (See Figure 7).

Mouse-Cursor: Click latency. An analysis of mouse cursor click latency was performed for only first warning impressions, and for only those who chose to ignore the warning (see Table 2 for resultant *n* distributions). Significant differences were found among treatment groups, $F(3,206) = 4.712, p = .003$. Pairwise contrasts of parameter estimates from a linear model found that disgust and fear had longer click latencies than did blank warnings (both one-tailed p's < .001), and that neutral warnings had longer click latencies than blank (one-tailed p = .041). Click latencies for disgust and fear did not differ from one another (p=.934), but disgust and fear were each longer than neutral (one-tailed *p*=.033 and *p*=.036 respectively). When preferred operating system and task performance accuracy are controlled for, the partial effect of emotion treatment group becomes much smaller; only disgust and fear are significantly different from blank (one-tailed p = .013 and p = .009 respectively), and no other statistically significant differences are seen. Interestingly, users who prefer the Mac OS have shorter clicking durations than do users who prefer the Windows operating system ($p < .0001$).

Table 2. Sample size distributions for the mouse cursor tests.

	Mouse cursor test *n*	
Emotion condition	**Click latency (first impression only)**	**Time idle (first four impressions)**
Blank	34	335
Neutral	66	698
Fear	53	666
Disgust	57	694

* Only the impressions where participants chose to ignore the warning were selected. DV values of 0 were removed because of incompatibility with the log transformation.

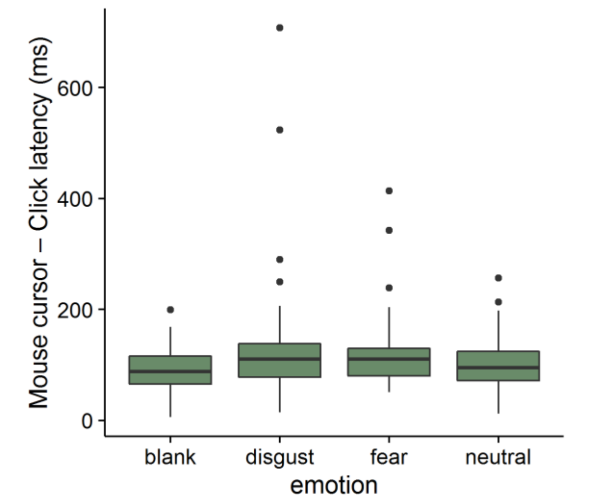


Figure 8. Click latency among emotion treatment groups for only the first warning impression for those who ignored the warning.

Mouse-Cursor: Time Idle. For the test of the impact of emotion condition on mouse cursor idle time, we report the results from a test of the first four impressions. Significant differences were found among emotion treatment groups, but the interaction effect of emotion condition and warning impression order was not significant ($\chi^2(df = 3) = 10.243,\ p = .017$ and $\chi^2(df = 3) = 4.105, p = .250$ respectively, see Table 2 for *n*). Pairwise contrasts of emotion condition parameter estimates from a linear mixed model accounting for repeated measures were examined. Blank warnings had significantly less idle times averaged across warning exposure order than disgust, fear, and neutral warnings (one-tailed p < .001, p=.013, *p*=.018 respectively). Disgust warnings had longer idle times than did fear or neutral warnings, (one-tailed p = .042 and p= .025 respectively), but no differences were observed between fear and neutral warnings (one-tailed p=.424). See Figure 9. This pattern of differences holds when controlling for task performance accuracy and for preferred OS.

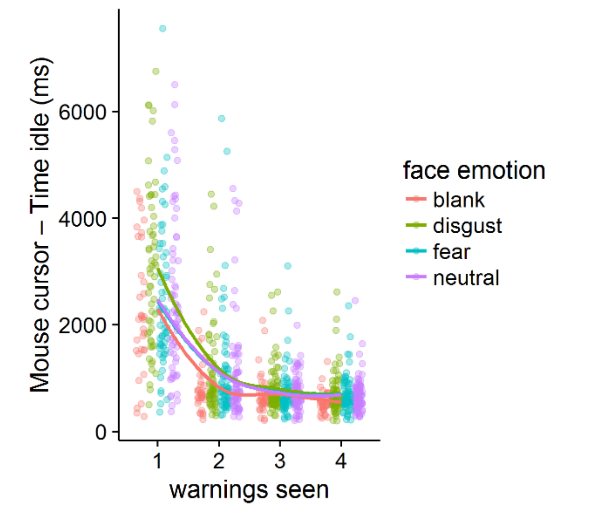


Figure 9. Mouse cursor idle times over time for each emotion condition.

Survey. All participants reported that they noticed the warnings. We tested for differences in self-reported levels of warning realism and concern felt over the warning among emotion treatment groups. On a scale of 1 to 10, blank warnings were reported to seem more realistic than any of the warnings with integrated facial expressions (all p's < .0001), but there were no differences observed among the different kinds of faces. No differences were observed among reported levels of concern felt when exposed to a security warning among any of the emotion treatment groups, including the blank group. See Figure 10. This pattern of differences holds when controlling for task performance accuracy and for preferred OS.

No differences were observed among emotion treatment groups for self-reported measures of perceived risk, threat susceptibility, threat severity, or fear, whether or not task performance accuracy and preferred OS were controlled for (all contrast families Bonferroni-adjusted). Raw (unadjusted) means among emotion treatment groups are displayed in Figure 11.

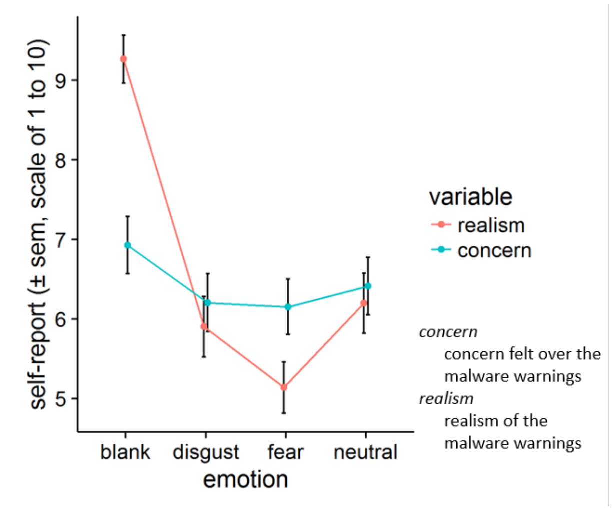


Figure 10. Differences in self-reported levels of warning realism and concern felt over the warning among emotion treatment groups. Note: practical effects may be smaller than they appear – y-axis is scaled to the data range for this and several future graphs.

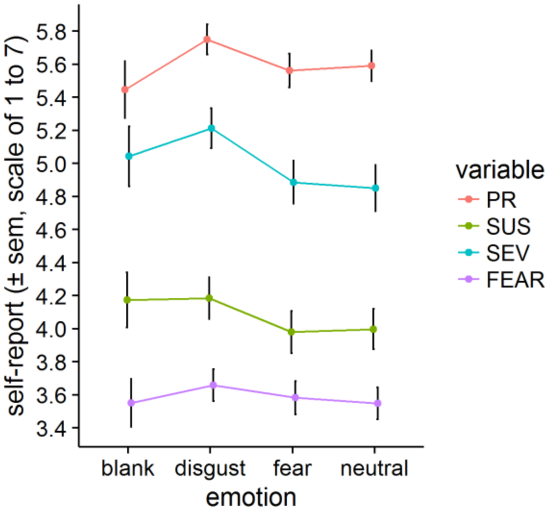


Figure 11. Unadjusted means among emotion factor levels for four variables: PR (perceived risk), SUS (threat susceptibility), SEV (threat severity), and FEAR (fear of threat).

2.5.2.2 Image filter. We next tested for the impact of image filter. In these analyses and figures "desat" refers to "desaturated", or in other words, to the photo-realistic face images with toneddown coloration. “Posterized” refers to the cartoonized versions of the images. Each test averages across the different emotion conditions for each filter, and compares the responses to the control warning, which had no face. Therefore, the three levels for the filter tests were “desat”, “posterized”, and “blank”.

Warning adherence rates. A linear mixed model with empirical logit link with a fixed effect for filter levels, order of warning seen, an interaction between the two, plus a random effect for each participant, was fitted to the data. A marginal difference was found among the slopes of filter levels on whether the warning was ignored, $\chi^2(df = 2) = 4.689, p = .096$. Pairwise contrasts reveal that participants who saw either desaturated or posterized images became more likely to ignore the warning over time, approaching the flat-line probability of ignoring that participants in the blank condition showed (slope of desat vs. blank *p* =.016 and posterized vs. blank *p* = .007). No differences in slopes were observed between desat and posterized groups. See Figure 12. No differences among filter treatment groups for marginal means were observed. The same pattern of differences is observed when controlling for English-as-first-language.

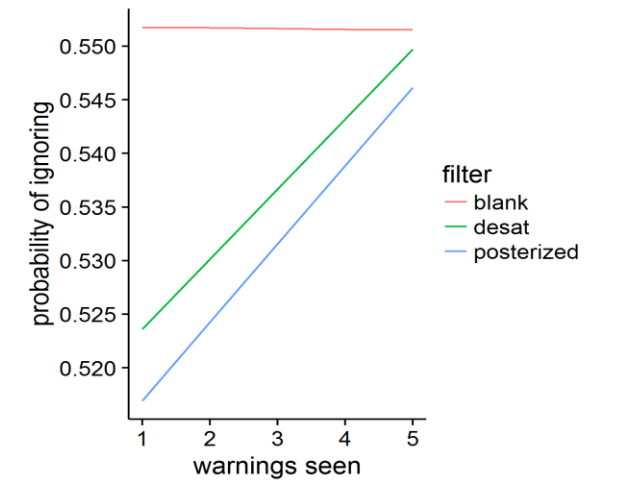


Figure 12. Plot of parameter estimates from linear mixed model with logit link, with dv converted to probabilities.

Reaction Time. A linear mixed model testing response times among filter factor levels found marginal significance for the interaction effect of filter and number of warnings seen ($\chi^2(df = 2) = 5.880, p = .053$), as well as significance for the main effect of filter ($\chi^2(df = 2) = 8.017, p = .018$). Pairwise comparisons of the trend over time among filter levels found that posterized warnings had less steep downward slopes than blank warnings (*p* = .018), but no significant differences were found between posterized warnings and desaturated warnings (*p* = .198). No differences in slopes were observed between blank and desaturated warnings (*p* = .148). As for the main effect of filter averaged across time, blank warnings had quicker reaction times than either desat or posterized warnings (*p* = .006 and *p* = .018 respectively), but no differences were observed between desat and posterized warnings (*p* = .567). The same pattern of differences was observed for main effects and for slopes when English-as-first-language was controlled for.

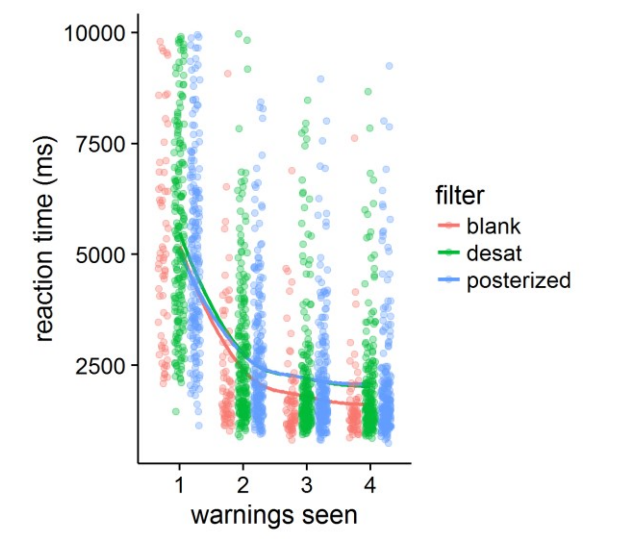


Figure 13. Loess curve of reaction times among filter factor levels across time. Data points are dodged horizontally to aid in density visualization.

Mouse-Cursor: Click latency. The effect of filter treatment on click latency across time was examined. Significant differences were found among the main effect of different filter levels,

$\chi^2(df = 2) = 7.400, p = .025$. Posterized warnings had higher click latency than did either blank or desaturated ones (p=.023 and p=.032 respectively), but no difference was observed between blank and desaturated warnings (p=.453). See Figure 14. The same pattern of differences was observed when controlling for English-as-first-language.

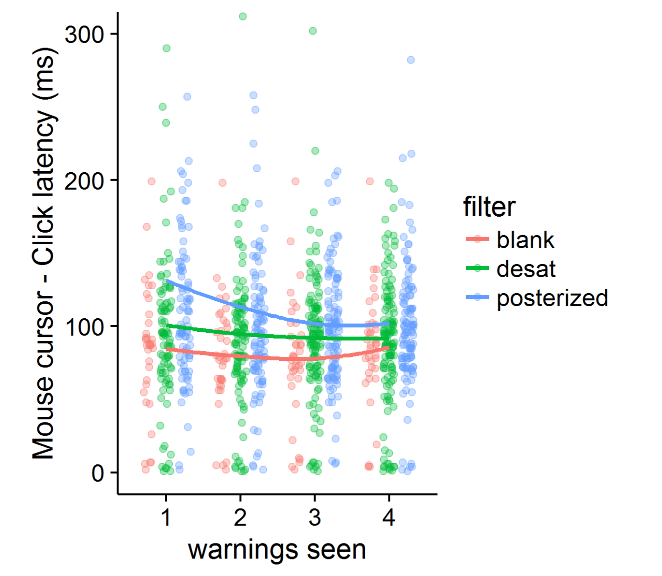


Figure 14. Loess curve of click latency over time among filter group levels.

Mouse-Cursor: Time Idle. The effect of filter treatment on mouse cursor time idle across time was examined. Significant differences were found among the main effect of different filter levels, $\chi^2(df = 2) = 10.100, p = .006$. Blank warnings had less time idle than either posterized or desaturated warnings (p=.003 and p=.022 respectively), but no differences were observed between desaturated and posterized warnings (p=.312). See Figure 15. The same pattern of differences was observed when controlling for English-as-first-language.

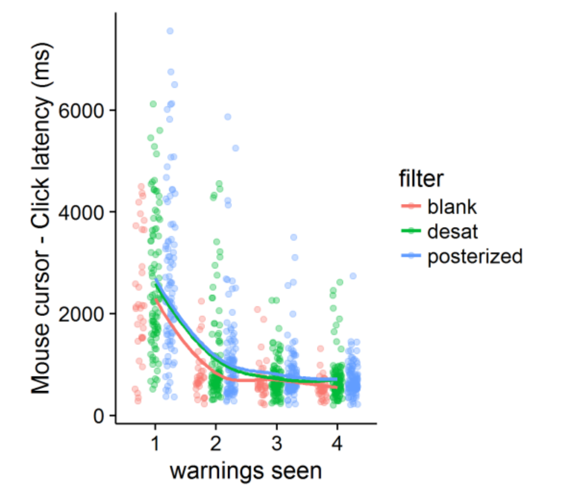


Figure 15. Loess curve of mouse cursor time idle over time among filter treatment groups.

Survey. While no differences were found between desaturated and posterized filter groups on survey responses for reported warning concern or realism, differences were found on aggregated protection motivation theory items (Johnston and Warkentin 2010) and fear items (Osman et al. 1994). Specifically, participants who saw desaturated face warnings reported higher perceived *information security risk* than did participants in either the posterized or in the blank treatment groups ($p = .037$ and $p = .062$ respectively). Also, participants who saw desaturated warnings reported higher *threat severity* than did participants in the posterized treatment group ($p = .022$). No differences were observed among filter treatment groups on reported threat susceptibility or information security-related fear. See Figure 16 and Table 3. The same pattern of differences appears when controlling for English-as-first-language. The differences among filter levels for perceived risk fall out of significance when applying a Bonferroni correction, whether English-as-first-language is controlled for or not.

Table 3. Summaries for perceived risk survey items among filter treatment groups

	Aggregated survey item means (std. dev.)			
Filter Condition	**Perceived Risk**	**Threat Severity**	**Threat Susceptibility**	**Fear of Malware**
Blank	5.44 (1.45)	5.04 (1.53)	4.17 (1.41)	3.55 (1.22)
Desaturated	5.75 (1.02)	5.15 (1.54)	4.05 (1.52)	3.65 (1.13)
Filter	5.51 (1.27)	4.81 (1.63)	4.05 (1.51)	3.54 (1.23)
Grand Mean	5.61 (1.20)	4.99 (1.58)	4.07 (1.50)	3.59 (1.19)

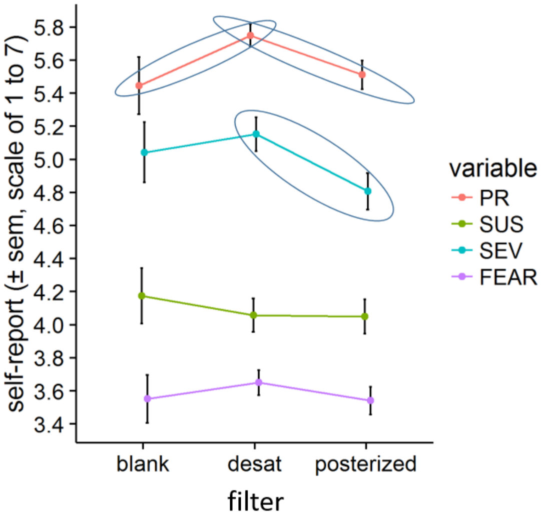


Figure 16. Self-reported risk perception items. PR = perceived risk of malware, SUS = susceptibility to malware, SEV = malware severity, FEAR = fear about getting malware. Ovals highlight statistically significant differences (*p's* < .10). Scores are aggregates of items from Johnston & Warkentin (2010) and Osman et al. (1994).

2.5.3 Discussion

The field study's design and various measures allow us to answer several questions. First, they allow us to examine the overall and differential impact of the kind of facial expression on markers of elaboration and uncertainty. Second, they allow us to compare the effects of different stylizations of the facial expressions. Third, we can investigate the endurance of these effects over repeated exposures. A summary of the findings for each hypothesis across each dependent measure is displayed in Table 4.

Table 4. Summary of hypothesis tests for Study 1 for each dependent variable

Hypothesis	Contrast		Analysis type	Supported*	Notes
H1 – Overall effect of face	Blank	< neutral	Warning adherence	N	Δ11.8% in the predicted direction, but insufficient statistical power to detect significance (*p*=.418).
			RT	N	4-warning marginal mean response ratio = 0.915, *p*=.286.
			M: Click latency	Y	1st only. Marginally supported, response ratio = -0.805, one-tailed *p* = .041.
			M: Time Idle	Y	4-warning marginal mean response ratio = .859, one-tailed *p* = .018
		< fear	Warning adherence	N	Δ14.9% in the predicted direction, but insufficient statistical power to detect significance (one-tailed *p*=.107).
			RT	Y	4-warning marginal mean response ratio = 0.862, one-tailed *p*=.009.
			M: Click latency	Y	1st only. Supported, response ratio = 0.664, one-tailed *p* < .001
			M: Time Idle	Y	4-warning marginal mean response ratio = 0.849, one-tailed *p*=.013.
		< disgust	Warning adherence	Y	Δ18.4% in the predicted direction, one-tailed *p* = .043.
			RT	Y	4-warning marginal mean response ratio = 0.876, *p*=.022.
			M: Click latency	Y	1st only. Supported, marginal mean = .657, one-tailed *p* < .001
			M: Time Idle	Y	4-warning marginal mean response ratio = 0.763, *p* < .001.
H2 – Overall effect of threat face	Neutral	< fear	Warning adherence	N	Δ3.3% in the predicted direction, but insufficient statistical power to detect significance (*p*=.593).
			RT	N	4-warning marginal mean response ratio = 1.060, *p*=.478.
			M: Click latency	Y	1st only. Marginally supported, response ratio = 1.213, one-tailed *p* = .036
			M: Time Idle	N	4-warning marginal mean response ratio = 1.011, *p* = .848.
		< disgust	Warning adherence	N	Δ6.6% in the predicted direction, but insufficient statistical power to detect significance (*p*=.259).
			RT	N	4-warning marginal mean response ratio = 1.044, *p*=.714.
			M: Click latency	Y	1st only. Marginally supported, response ratio = 1.224, one-tailed *p* = .033
			M: Time Idle	Y	4-warning marginal mean response ratio = 1.125, one-tailed *p* = .025.
H3 – disgust vs fear face	Disgust	> fear	Warning adherence	N	Δ3.5% in the predicted direction, but insufficient statistical power to detect significance (*p*=.553).
			RT	N	4-warning marginal mean response ratio = 0.985, *p*=.983.
			M: Click latency	N	1st only. Not supported, response ratio = 1.009, *p* = .934
			M: Time Idle	Y	4-warning marginal mean response ratio = 1.112, one-tailed *p* = .042.
H4 – Filters	Realistic	= Posterized	Warning adherence	Y	No difference, *p* = .690.
			RT	Y	4-warning marginal mean response ratio = 1.020, *p*=.577.
			M: Click latency	Y	Supported, posterized not less than realistic. (In fact, they were *greater than* realistic.) Encouraging for practitioners. 4-warning marginal mean response ratio = 0.772, *p* = .032.
			M: Time Idle	Y	4-warning marginal mean response ratio = 0.948, *p*=.312.

*For “supported” column, “Y”= Supported, N=”Not supported”.
“M:” indicates “mouse” statistic.
“RT” = reaction time.

Table 5. Summary of hypothesis testing across Study 1

Hypothesis	Contrast	# Supported	
		Y	N
H1 – Overall effect of face	Blank < [neutral, fear, disgust]	9	3
H2 – Overall effect of threat face	Neutral < [fear, disgust]	3	5
H3 – disgust vs fear face	Disgust > fear	1	4
H4 – Filters	Realistic = Posterized	4	0

2.5.3.1 Face emotion

A comparison of warnings with faces against the standard warning with no integrated facial expression generally showed that *some* face, represented by the warnings with integrated neutral expressions, was better than *no* face. We would expect such findings if we had integrated any novel and noteworthy stimulus into the warnings – novelty garners visual attention (Amer and Maris 2007). This was seen in warning adherence rates, where first-impression blank warnings had nearly 59% ignore rates, whereas the next-highest rank of ignoring was the neutral face treatment, with about 46% first-impressions ignored. Click latency also found that warnings with neutral faces showed marginally greater evidence of uncertainty and elaboration than did blank warnings. Mouse movement idle time was also lowest for blank warnings than for any of the warning variations with faces. However, reaction times to neutral warnings averaged across exposures were no different from reactions to blank warnings – although blank warnings had steeper downward slopes for reaction times than did warnings with neutral expressions. Thus, H1 is supported, with significant differences being found in 75% of the tests.

The more important question of interest to our study is whether the facial expressions of threat elicited greater threat attention and more secure behavior than did neutral expressions, compared to blank warnings. Several dependent measures suggest that this holds true. Warning ignore rates were higher for both fear (43.2%) and disgust (39.7%) expression integrated warnings than for neutral-expression-integrated warnings (46.3%). For reaction times, participants in the disgust and fear groups had longer reaction times averaged across time than did participants in the blank treatment group, while participants in the neutral expression group did *not* have longer reaction times compared to the control group. For click latency, participants in the disgust and fear treatment groups had marginally longer latencies than did users in the neutral group. And for idle mouse cursor times, participants in the disgust group had greater idle times than did neutral participants. However, no difference was found on idle times between the fear and neutral groups. These findings across multiple measures give support to the notion that not just any facial expression can be integrated into a warning – threat faces appear to elicit greater threat attention and secure behavior in a security message context than do neutral faces. In summary, H2 finds some support, with 37.5% of the tests showing statistically significant differences.

We also can test for differences in performance between kinds of threat faces integrated into security messages. The findings here are not as clear, although they tend towards the conclusion that there is little to no difference between disgust and fear facial expressions. While participants in the disgust group had 3% lower warning ignore rates than did participants in the fear group, and while they had marginally greater mouse cursor idle times than did participants in the fear group, no differences were observed between fear and disgust treatment groups on reaction times or click latency. In summary, there is the possibility that disgust warnings elicit better security behavior than do fear warnings in a security message context, but the effect is small and was not detected by this study. In summary, H3 is supported by only 25% of the tests.

We also hypothesized differences in the endurability of the facial expression treatments over time. However, in hindsight, performing an experiment may have precluded the ability to test this in an ideal context. While our warning impressions were separated by at least 10 regular image classifications, the warnings still would have appeared at a greater frequency than would be typical in everyday computing, and habituation rates to repeated stimuli depend heavily on the interval between exposures (Rankin et al. 2009, characteristic #4). As for our measures, warning adherence rates did not differ largely from what was reported for first-impressions only when we included more than one warning impression per participant, largely because the correlation between repeated measures for each participant was very high. However, other non-binary measures had more variance within-subject. Reaction times for fear and neutral emotion treatment groups were less subject to attenuation than were reaction times to blank warnings. While the slope for disgust warnings did not differ from that of blank warnings, it was trending towards being less steep (see Figure 7). Differences among emotion treatment groups on click latencies fell out of significance when multiple impressions per participant were considered. Similarly, differences in slopes of mouse cursor idling time were not found among emotion treatment groups. This suggests that once participants made a decision *in the context of the image classification task*, that making that decision again did not impact their click latencies or idle time. This conflicts with the findings in Authors (2016) in which differences in slope were observed between treatment groups over repeated security decisions within one study, although in that study, participants may have perceived greater differences between their security decisions since each one was hosted on a different website, and thus participants had more information on which to base security decisions. Thus, the supplemental hypotheses about differences in slope are not supported with data from our protocol. We call for future research to find more efficient ways to test these hypotheses.

The survey data indicated that participants perceive warnings without integrated facial expressions to be more realistic than ones with integrated facial expressions. This was somewhat expected, given that our facial expression integration did not meet Google's design guidelines. This would be remedied were a warning design to go through an organization's official design review process. However, no differences were found between emotion treatment groups on the concern they felt when they encountered a warning. This is interesting – despite lower realism, concern was unaffected. We concede that concern following a suspected fake warning may not be the same as concern following perceived legitimate warnings. However, our objective measures suggest that users may be more concerned than they realize. This highlights the importance of measuring security behavior such as reactions to security messages using measures that are less subject to biases and that can capture subtle reactions that users may not be aware of (Dimoka et al. 2011).

While the follow-up contrasts on actual adherence choices only found a significant difference between the blank (58.7% ignored) and the disgust (39.7% ignored) warnings, we observe that the differences were trending in the predicted direction, with disgust and fear showing the highest warning adherence rates, followed by neutral, and then blank warnings. If we had a larger sample size, we expect that these differences would become statistically significant. As illustrated in Wogalter's C-HIP model (Wogalter 2006a), warning adherence behavior is a multi-gated process. We only expected our face treatments to increase threat attention to the warnings, and our other measures suggest that this goal was met. After attention is garnered, other elements of the warning and of the user will impact warning comprehension, user security attitudes, and motivation to adhere. Increasing threat attention through incorporating facial cues therefore encourages users to be more thoughtful in their decision, but may have a small effect on the moredistant actual behavioral outcome. We call for more research to address the remaining steps towards improving secure end-user behavior.

2.5.3.2 Face Filter

We now explore the differences on user behaviors between photo-realistic (desaturated) and posterized filtered versions of the face images, averaged across face emotions. For reaction times, desaturated warnings had steeper downward slopes compared to those of posterized warnings. A similar pattern was seen for click latencies – posterized warnings elicited longer click latencies compared to realistic warnings. However, no difference in mouse cursor idling was observed between desaturated and posterized warnings, and no differences were observed in actual security choice either. We had hypothesized that posterized images would perform no differently than photo-realistic images on invoking security attention – or, at worst, that they would only slightly underperform in these areas. Instead, we see that the opposite may be true – posterizing the images *improves* attention.

Further follow-up analyses can compare posterized to photo-realistic images for only threat faces to see if the positive impact holds when neutral faces are excluded, but we expect that this would only amplify the strength of the relationships already found. Furthermore, we reason that posterizing the images had the unexpected impact of making them more interesting to look at than a photo-realistic face. Follow-up work can involve hiring an artist to create hand-drawn versions of the faces that are more equally interesting to look at compared to the photo-realistic ones, and the tests can be rerun. We expect that were this done, the differences between posterized and photo-realistic images would drop into insignificance. From a practitioner's standpoint, this would mean that hand-drawn facial images would have the same positive effect as would photo-realistic ones, as long as the essential facial elements were discernible (eyebrows, mouth, eyes). Thus, while H4 was contraindicated, we find support for the practitioner spirit of H4, which was to ensure that photo-realism is not required to elicit the same threat attentional effects.

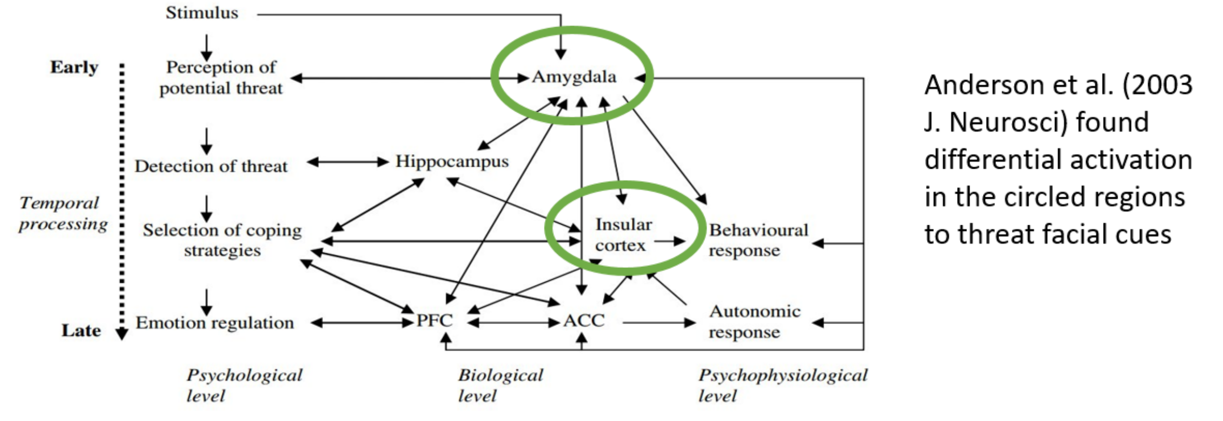
2.5.3.3 Survey perceptions

We also examined differences on self-reported perceptions of security risk and fear between emotion treatment groups and between filter treatment groups. Significant difference in these measures were only found among comparisons for image filter treatment groups. Specifically, photo-realistic images elicited higher reports of perceived risk and malware threat severity than did posterized images. This could be because the photo-realistic threat images connect more deeply with the human psyche, in turn prompting users to elaborate over the warnings they saw more carefully. And as Vance et al. (2014) showed, recent information security events can raise users' security risk perceptions. It is possible that the effects of the treatments are amplified for threat faces compared to neutral faces. However, follow-up analyses, testing for an interaction between the face emotion and the face filter, found no significant differences.

We also note the lower-than-hoped-for overall reported levels of threat susceptibility and malware fear. The grand means for threat susceptibility and fear of malware were only 4.07 (1.5 *SD*) and 3.59 (1.19 *SD*) on 7-point Likert scales. The modest reported levels for threat susceptibility are consistent with other recent literature, which has found that users, by default, do not feel particularly susceptible to or raw emotional fear about information security threats (Boss et al. 2015), given their low relevance to one's physical safety (e.g., Johnston et al. 2015). We call for future research to inform security message design in a way that will increase users' perceptions of threat susceptibility.

2.6 STUDY 2 – FMRI

Study 2's fMRI protocol allowed us to assess whether exposure to security message variations prompted differential *threat processing* attention as opposed to simple visual attention. In Study 2, we test all of the hypotheses except for the ones relating to differences in attenuation over time between design cells, due to limitations in the length of time that we could keep participants in the MRI scanner. Early-stage threat processing is thought to occur in the amygdala region of the brain, with later stages showing activity in the hippocampus and lateral cortex (Hofmann et al. 2012) (see Figure 17).



Anderson et al. (2003 J. Neurosci) found differential activation in the circled regions to threat facial cues

Figure 17. A cognitive neurobiological information-processing model of fear and anxiety. Adapted from Hofmann et al. (2012)

2.6.1 Design

We drew our security warning stimuli from the same set that was used for Study 1. After a 5-participant pilot study, we analyzed scan and behavioral data from 23 participants. In our repeated-measures design, each participant saw the same set of 240 unique warnings with integrated facial expressions plus 20 images with no integrated facial expression in a randomized order which were used as a baseline in the fMRI analysis. The 240 images were drawn from a 3x2 design that fully crossed an emotion factor (fear, disgust, neutral) with an image filter factor (posterized, photo-realistic). Each stimulus was presented for 3 seconds with a .5 second break in between (see Appendix A, Figure A1). Participants were instructed to self-report whether each warning captured their attention (yes/no) using an MRI-compatible button box. Dependent variables included recorded brain activity, binary self-reported attention, and reaction time (see Appendix A).

2.6.2 Analysis and Results

Individual-level regressions predicting activations across the whole brain were first performed. In addition to including parameters for the emotion effect and image filter, these regression models also controlled for face actor age group and gender. The individual-level parameter estimates were then entered into a group-level analysis to obtain the result shown in Figure 18. After a spatial extent threshold of 24 voxels was applied to the group-level analysis to correct for multiple comparisons, only the right amygdala and two clusters within the cerebellum were identified as having significant differences in signal response among factor levels. Only the activations for the right amygdala were extracted – the two cerebellum clusters were not considered in our follow-up contrasts because the scanner field of focus did not capture data for these outlying regions for all participants. Follow-up contrasts for the emotion and factor levels were performed by extracting the individual-level parameter estimates for the right amygdala for each design cell. The fMRI parameter estimates were analyzed using the `lmer` function from the lme4 R package v1.1.12, which allowed us to control for repeated measures using a random intercept for each participant. Reaction time was also analyzed using a linear mixed model with random intercept per subject. Contrasts for emotion factor levels and filter factor levels within the right amygdala were performed. `glmer`, also from the lme4 package, was used to analyze self-reported attention by fitting a general linear model with a logit link and a random intercept for each participant.

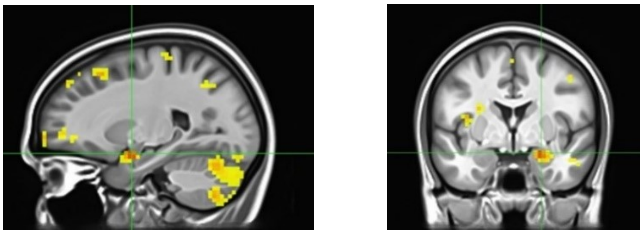


Figure 18. Main effect of emotion in right amygdala, showing, from left to right, coronal and sagittal cuts. Activation threshold set at *p* < .05

Because all comparisons tested a priori hypotheses, no corrections for multiple comparisons were made. One-tailed p-values are reported where appropriate.

2.6.2.1 Face emotion

Right amygdala activations. The right amygdala showed a significant main effect for the emotion factor (see Table 6). Follow-up analyses comparing the extracted parameter estimates among emotion factor levels for each participant for this region indicated that activations for disgust and neutral warnings did not significantly differ from one another. However, disgust and neutral warnings showed higher right amygdala signal responses compared to fear face warnings, (p = .047 and p = 0.027 respectively). Furthermore, only fear significantly differed from activations to warnings with no faces (see Figure 19 and also confidence intervals in Table 7).

Table 6. Main effect of emotion by region of interest

		Coordinates			**Main Effect Stimulus**	
Region	**#Voxels**	**X**	**y**	**Z**	$\chi^2(3)$*	**p**
R. Amygdala	24	-22.0	2.0	-14.9	14.21	.003

Analysis of deviance obtained from a linear mixed model including emotion as a fixed factor and participant_id as a random intercept.

Table 7. Least-squares means and 90% confidence intervals for right amygdala activations among emotion factor levels compared against warnings with no faces (baseline).

Emotion	LS Mean (SE)	Df	90% CI
Disgust	0.088 (0.085)	537	[-0.051,0.228]
Fear	-0.144 (0.085)	537	[-0.284, -0.004]
Neutral	0.106 (0.085)	537	[-0.034,0.245]

Reaction times. An analysis of the reaction time data averaged over repeated exposures showed that participants took more time to respond to warnings with integrated fear and disgust expressions than they did for either neutral or blank images (see Figure 20, all contrast *p*'s < .0001). Neutral images also had longer reaction times than did blank warnings (*p* < .0001). As for trends over time, blank warnings had the most precipitous drop in reaction times compared to any other emotion factor level (all *p*'s < .0001). Reaction times to disgust warnings also had a steeper drop than did fear (one-tailed *p* = .016). Disgust warnings also had steeper drops in reaction times over repeated impressions than did neutral warnings, (one-tailed *p* = .041).

Self-reported attention. The logit regression of the self-reported attention lines up closely with the reaction time data – participants were more likely to say that warnings with fear and disgust expressions captured their attention more than neutral or blank ones, and that neutral warnings captured their attention more than did blank ones (see Figure 21, all contrast *p*'s < .001). There was no significant difference between disgust and fear factor levels for reaction times or for predicted probability intercepts.

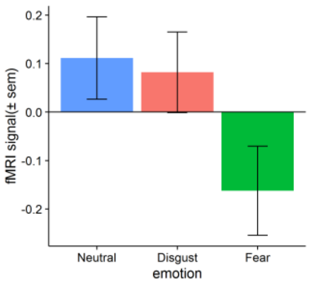


Figure 19. MRI activations for levels of emotion compared to faceless warning stimuli (±sem).

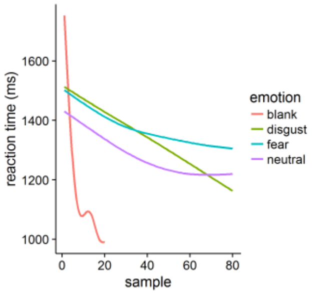


Figure 20. Loess smoothing line of reaction time in milliseconds over repeated exposures by emotion level.

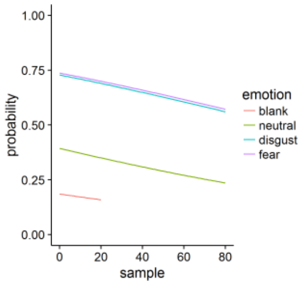


Figure 21. Probabilities of responding that stimulus captured attention over time among stimulus emotion levels.

2.6.2.2 Image Filter

As predicted, a whole-brain analysis did not identify a significant main effect for the filter factor level on the right amygdala. Regardless, we extracted the individual parameter estimates for the region identified by the emotion main effect in order to assess any trends between filter factor levels that might become significant with a more powerful design. Comparisons of filter factor levels averaged across emotion factor levels were largely as expected. There were no significant differences in fMRI signal for right amygdala activation between the two filter levels (p = .467), although signal response to posterized security warnings trended towards lower activation compared to photo-realistic security warnings (see Figure 22 and Table 8). Once again, neither filter factor level differed significantly from 0, meaning that right amygdala activations to these stimuli were not significantly different from blank warnings (the baseline – see Table 8). Reaction times averaged across repeated exposures between the two filter factor levels were not significantly different (see Figure 23). However, participants were more likely to report that posterized faces captured their attention than did photo-realistic ones – Δ log(odds) = 0.358, p < .0001 (see Figure 24).

Table 8. Least-squares means and 95% confidence intervals for right amygdala activations among filter factor levels compared against warnings with no faces (baseline).

Filter	LS Mean (SE)	Df	90% CI
Photo-realistic	-0.016 (0.176)	24.74	[-0.378,0.347]
Desaturated	-0.144 (0.085)	24.83	[-0.440,0.285]

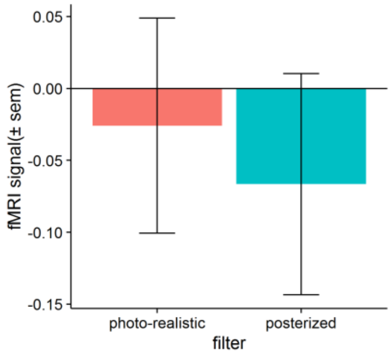


Figure 22. MRI activation betas for levels of filter. No significant difference between levels.

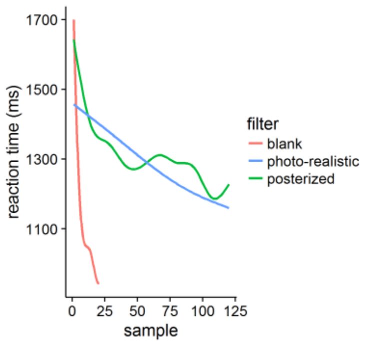


Figure 23. Predicting reaction time over multiple exposures by filter level.

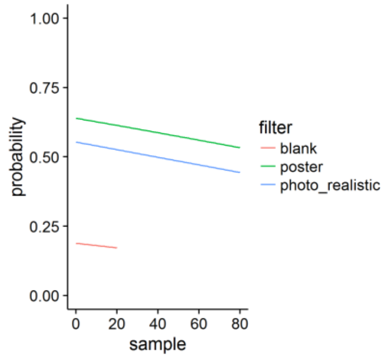


Figure 24. Predicting probability of attentional self-report by filter.

2.6.3 Discussion

We now discuss the results of study 2 from the perspective of each of the hypotheses. A summary of the findings is presented in Table 9.

2.6.3.1 Regions of Interest

Relying on previous literature and on models of threat processing, we expected to find differential activations within two ROIs – the amygdala and the insular cortex (Anderson et al. 2003; Hofmann et al. 2012) (see Figure 17). Our whole-brain analyses for the emotion factor did point to the right amygdala, but not the insular cortex. This may suggest that for small effect sizes such as the ones our protocol was likely to elicit, the amygdala is more sensitive than is the insular cortex. Or, the lack of activation on the insular cortex signal response levels could suggest that our participants never moved beyond the "perception of potential threat" phase to the "detection of threat" phase.

This is desirable – as we stated before, it is not desirable to push users into a state of emotive fear or high stress every time they interact with a security warning. The "perception of potential threat" phase alone should foster threat attention, which should in turn lead to closer engagement with the security warning (see Figure 1).

Table 9. Summary of hypothesis tests for Study 2 for each dependent variable

Hypothesis	Contrast		Analysis type	Supported*	Notes
H1 – Overall effect of face	Blank	< neutral	fMRI – right amygdala	N	Greater than baseline but not statistically significantly so. BOLD response = 0.111, 90% CI = [-0.033, 0.255]
			RT	Y	Supported, Δ = 985 ms, t=8.938, p < .0001
			Self-report	Y	Supported. Δ log(odds) = -1.047, p < .0001
		< fear	fMRI – right amygdala	N	Contraindicated. Fear was less than baseline. BOLD response = –0.162, 90% CI = [–0.305, –0.019]
			RT	Y	Supported, Δ = 1098 ms, t = 9.935, p < .0001.
			Self-report	Y	Supported. Δ log(odds) = 2.512, z = 16.261, p < .001
		< disgust	fMRI – right amygdala	N	Greater than baseline but not statistically significant so. BOLD response = 0.082, 90% CI = [-0.060, 0.224]
			RT	Y	Supported, Δ = 1087, t=9.837, *p* < .001
			Self-report	Y	Supported. Δ log(odds) = 2.466, z = 15.917, p < .001.
H2 – Overall effect of threat face	Neutral	< fear	fMRI – right amygdala	N	Contraindicated. Fear less than neutral. Δ BOLD response = –0.273, t=-2.222, p=.027.
			RT	Y	Supported, Δ = 112 ms, t = 6.514, p < .001.
			Self-report	Y	Supported. Δ log(odds) = 1.464, z = 16.486, p < .001.
		< disgust	fMRI – right amygdala	N	No significant difference. Δ BOLD response = -0.029, t=-0.240, p=.811.
			RT	Y	Supported, Δ = 101.59 ms, t=5.828, p < .001
			Self-report	Y	Supported. Δ log(odds) = 1.419, z = 15.839, p < .001.
H3 – disgust vs fear face	Disgust	> fear	fMRI – right amygdala	Y	Supported, predicted direction. Δ BOLD response = 0.244, t=1.994, p=.047.
			RT	N	Not supported. Δ = -10.54 ms, t= -0.572, p=.567.
			Self-report	N	Not supported. No difference in means. Delta log(odds) = -0.046, *z*=-0.498, *p*=.619.
H4 – filters	Realistic	= Posterized	fMRI – right amygdala	Y	Supported, no difference. Δ BOLD response = 0.062, t=0.728, p=.467.
			RT	Y	Supported. No statistically significant difference between reaction times to realistic or posterized stimuli. Δ = 8.276 ms, t=0.606, p=.545.
			Self-report	N	Not supported. Δ log(odds) = 0.3578232, z=5.108. p < .0001.

*For "supported" column, "Y"= Supported, N="Not supported".
"RT" = reaction time.

Table 10. Summary of hypothesis testing for Study 2

Hypothesis	Contrast	Summary	
		YY	N
H1 – Overall effect of face	Blank < [neutral, fear, disgust]	6	3
H2 – Overall effect of threat face	Neutral < [fear, disgust]	4	2
H3 – Disgust vs fear face	Disgust > fear	1	2
H4 – Filters	Realistic = Posterized	2	1

2.6.3.2 Face Emotion

Fear's low activations in Study 2 may be explained by post-study interviews from an fMRI pilot study, which suggested that the fear faces appeared humorous to some subjects. This was an interesting finding because the emotional valence of the photo set we used had been pre-validated (Ebner et al. 2010). Furthering the mystery, standalone fearful facial expressions have been found to consistently elicit right-amygdala activations in other fMRI studies (e.g., Anderson et al. 2003). It is possible that the fearful facial expressions invoked a different response than is typical once they were interjected into our security message instrument. However, after we applied the oval crop and desaturation after the fMRI pilot[2], participants no longer reported that the fearful faces appeared humorous. An alternative explanation for the lower right amygdala activations for fearful facial expressions is that the fearful facial expressions interact with low-level threat concern that participants feel even when exposed to blank information security warnings with their default threat cues (e.g., the color red and the red stop-sign 'X' symbol). It is possible that when a fearful facial expression is interjected into this context, it serves as a social confirmation that the threat is legitimate, calming any uncertainty users may have otherwise felt in the absence of the social cue afforded by observing the reaction of another. Discrediting this proposition, however, is the

[2] Chronologically, the fMRI pilot was performed before the field study, which led us to use oval crop images for the field study as well as for the fMRI study.

findings of Study 1, which showed that, in a realistic field study, security warnings with integrated fearful facial expressions were often just as effective as disgust facial expressions at prompting secure behaviors. Further discrediting this theory is that, if the calming effect were true, we might expect a similar outcome from exposures to security warnings with disgust facial expressions.

The longer reaction times for warnings with integrated facial expressions of threat compared against neutral-expression-integrated warnings and compared against blank warnings suggests that threat-face-integrated warnings elicit greater elaboration over the security warnings, supporting H2. Neutral warnings also had longer reaction times than did blank warnings, supporting H1. However, no differences were found in response times between fearful and disgust factor levels, which suggests that the effect is not discriminatory between threat face type. This weakens H3.

2.6.3.3 Face Filter

We predicted that warnings with integrated photo-realistic faces would show either no difference compared to ones in integrated posterized faces as long as the essential facial elements were discernible (eyebrows, mouth, eyes). Or, that if there were differences, posterization would draw slightly less threat attention than would photo-realism. As with Study 1, the practical intention of this prediction was largely supported. No significant differences were found on right amygdala activation or on response times between photo-realistic and posterized warnings. However, participants were *more* likely to report that posterized images captured their attention than did photo-realistic ones. Also, the difference in right amygdala activation trended towards photo-realistic images eliciting greater threat attention compared to posterized images, although this difference was small. Thus, as in Study 1, we find support for the notion of H4, in that posterizing a facial expression does not substantially negatively impact its performance on prompting threat attention compared to photo-realistic faces in warnings.

2.6.3.4 Limitations and Future Research

The fMRI design has limitations: The intervention inside the fMRI machine lacked external validity because of the absence of an actual threat. Furthermore, each participant saw 240 warning images with integrated facial expressions + 20 warnings without facial expressions over the course of about 20 minutes. While such repetition was necessary in order to simultaneously test emotion and filter factor levels while controlling for actor age and gender, this hampered the generalizability of the protocol. Whereas other studies have studied reactions to images of security messages using an fMRI protocol, those studies varied the base warning template on which their treatments were imposed (i.e., software installation warnings, browser warnings, macro warnings). Contrastingly, in our study, participants saw warning stimuli variations based on only one type of security warning (the Chrome malware warning) for every security warning they were exposed to during our protocol. This high level of visual similarity between all stimuli may have led to high levels of habituation despite the different integrated facial expressions, effectively overwhelming any true differences that differential emotive faces integrated into security messages would invoke (see Rankin et al. 2009, Characteristic #7). Further supporting the notion that the effects may have been overwhelmed is found in comparing the results of Study 2 to the between-subjects limited-warning-exposure design of Study 1.

We also note that, more often than not, the levels for the emotion and filter factors were not found to elicit significantly different activations compared to the baseline. In hindsight, this may be due to the fact that the ratio of baseline warning images to warnings with faces was 1:12. This low relative frequency for the baseline may have led to the "oddball effect" (Squires et al.

1975), which describes how attention is much more likely not when exposed to an infrequent stimulus, not necessarily because of the inherent characteristics of a stimulus, but only because the stimulus's occurrence is rare. We suspect that if our ratio of baseline warning images to warning images with faces were more balanced, then the confidence intervals for the factor levels' least-squared means would have been more likely to differ from zero. This would have led to the fMRI data supporting our first hypothesis – namely, that warnings with faces are more likely to garner attention than are warnings without faces.

One suggestion for future research to increase power without having to expose participants to so many trials is to decrease the repetition time (TR) used during scanning. The TR time is a measure of the interval between successive whole-brain captures that the scanner can take. The lower the TR, the faster the capture. The configuration of the scanner we used could at best lower the TR to 2 seconds. We reasoned that an event such as threat attention to a security message would not last longer than the span of one TR, especially given the finding that many warnings are responded do within 3 seconds or less in the field (Akhawe and Felt 2013). Therefore, we were only able to acquire one sample per warning shown. It is possible, however, to drop the TR to much lower levels. If a TR of 0.5 seconds could be achieved, for example, then at least four samples could be obtained for each warning exposure, quartering the number of warnings that participants would need to see in order to achieve the same statistical power that we had in our protocol, and likely substantially reducing the effects of attenuation and habituation. If the TR were dropped to a sufficiently low number (such as 0.5 seconds with 5 samples per warning impression), and if the baseline stimulus were not a security warning, and if sufficient funding were obtained to run a between-subjects design, then it is possible that each participant would only need to see, at minimum, 2 warnings (which would provide 10 samples – the minimum recommendation requirement per cell for fMRI studies). If only 2 warnings were required, then the doors open wide for new interruptive security message protocols that are significantly more realistic and less susceptible to within-study stimuli attenuation. Thus, power would be substantially improved.

2.7 STUDY 1 AND 2 – GENERAL DISCUSSION

While the discussion subsections specific to studies 1 and 2 examine the particulars of those studies in detail, this discussion section touches on points that overarch the two studies. Table 4 and Table 9 provide hypothesis support summaries for studies 1 and 2 respectively, and Table 11 gives an overview of support for the hypotheses across both studies.

Table 11. Summary of hypothesis testing across Studies 1 and 2

Hypothesis	Contrast	Summary		
			Y	N
H1 – Overall effect of face	Blank < [neutral, fear, disgust]	**Study 1**	9	3
		Study 2	6	3
H2 – Overall effect of threat face	Neutral < [fear, disgust]	**Study 1**	3	5
		Study 2	4	2
H3 – disgust vs fear face	Disgust > fear	**Study 1**	1	4
		Study 2	1	2
H4 – Filters	Realistic = Posterized	**Study 1**	4	0
		Study 2	2	1

Hypothesis 1 found moderately strong support across both studies (75% and 66% of tests supported, respectively). This follows the findings of Amer and Maris (2007) where the introduction of any new stimulus, in this case any facial expression, will renew attention to security messages.

Hypothesis 2 found some support in Study 1 (37.5% of tests supported) and also in Study 2 (66% of tests supported). The lower support in Study 1 is likely due to that study's insufficient power to detect the small effect size. The effect size for Study 2 was larger likely because participants were not as distracted during this protocol. For example, they were not distracted by a bonus loss or by concern for their computer's safety.

Hypothesis 3 did not find strong support in either study (20% and 33% of tests supported respectively). This is surprising, given that differential activations were previously observed between exposures to fear and disgust facial expressions in another study context (Anderson et al. 2003), and because experiencing disgust includes less uncertainty than does feeling fear (Morales et al. 2012). In retrospect, the differential prediction of uncertainty between fear and disgust emotions requires that an individual actually experience that emotion, rather than merely engaging in low-level neural precursors (Hofmann et al. 2012). We may conclude from this that participants did not reach the state of experiencing disgust or fear themselves.

Hypothesis 4 was supported by both studies (100% and 66% tests supported, respectively), in that posterizing a security-message-integrated facial expression did not diminish its likelihood of eliciting secure behaviors such as adherence and attention. In fact, posterizing a facial expression often elicited *higher* levels of attention markers (while no differences were observed on adherence). The solarized effect of the face posterization may have been more visually interesting for participants. We note that even for the posterized warnings, threat faces elicited higher attention marker levels than did neutral faces. We anticipate that more benign cartoonization, such as facial expression line drawings, would still show the patterns of activations that we observed.

Besides the direct contribution provided by the hypothesis tests, these studies also contribute peripherally to the conversation on differences between lab and field studies for human-computer interaction information security contexts. Study 1 used a realistic field context, while Study 2 (fMRI) used a lab context. Lab contexts introduce several challenges to the reliability of security studies because of the security-specific biases that they introduce, such as feeling irrational protection from security threats because of being in a physical lab and because of the proximity to research personnel (Sotirakopoulos et al. 2011). It is interesting that the behavioral dependent variables from Study 2, including reaction times and self-reported attention, still supported our hypotheses, despite these biases (see Table 9). These behavioral measures were also supportive of the same hypotheses in the much more ecologically valid field study setting of Study 1.

2.7.1 Limitations and Future Research

One of the driving limitations of studies 1 and 2 is that their protocols did not afford the opportunity to test the attenuation rates of face-integrated security warnings over repeated exposures. As Rankin et al. (2009) observes, habituation rates to a stimuli are dependent on the length of time between the stimuli presentations. Exposing users to multiple security warnings within a short time period is probably not similar to rates at which users typically see warnings of the same type, let alone of different types. To test the habituation rates ideally, a longitudinal protocol could be employed that would present warnings to users separated by days or weeks. Such a protocol is fraught with difficulty, however, because unless an omniscient tool were installed on users' computers, it would be impossible to know what other kinds of real warnings (warnings not presented by the research software) participants were exposed to during the timeframe. Knowing all of the different types of security messages that users would see would help control for generalized habituation rates, which occurs when being exposed to stimuli from the same class (i.e., security messages from two different software suites with entirely different visual design) impacts a habituation process to security messages *in general*. This variance could be addressed by delegating the issue of generalized security message habituation to random assignment in a mixed between-subjects design with repeated measures, such as was used in the protocol of study 2. One software could be focused on, such as malware warnings presented by one particular browser. Participants could be instructed to use a browser version that had had experimental code injected into it by the experimenters, such as could be done with Chromium. If full control were had over such a browser, then it would be possible to know each *legitimate* security warning that was presented during the study, as well as *feigned* ones. Conversely, the browser could *only* present security messages for *legitimate* threats, varying the presentation of the message between groups. Participants could be instructed to perform some task that would make it more likely that they would encounter legitimate warnings, institutional review boards allowing. Such a design would rate very highly on realism, and internal validity would be fairly well controlled for considering the typical history threat inherent in longitudinal designs.

One might argue that the practical significance of the effect sizes for the different treatments is relatively small – that a difference of a few milliseconds is negligible. The following counterarguments are offered: (1) Any difference in markers of cognition may make a difference in an individual not succumbing to an information security threat and suffering the accompanying stress and other fallout from the incident. (2) Our treatments only aimed to impact the very earliest stage of warning processing according to the Wogalter (2006a) communication-human information processing model (see Figure C-1), and a security warning is a careful orchestration of many components, all of which contribute towards influencing whether a user heeds a security warning. In summary, the incremental value of our security warning treatments is of high value even if it makes a difference for only one user (of course, we expect the number of users benefitting to be much higher than one).

The study design also afforded the opportunity to test for an interaction between the emotion and filter factors. We surmised, post-hoc, that it is possible that the effects of the treatments are amplified for threat faces compared to neutral faces. However, follow-up analyses testing for an interaction between the face emotion and the face filter found no significant differences.

2.8 STUDIES 1 AND 2 – CONCLUSION

The two studies presented in this section tested the integration of human facial expressions into interruptive security messages, with the aim of improving end user security behaviors, including attention to the warnings. They corroborated multiple dependent variables, including self-report, reaction times, mouse-cursor movements, and fMRI data, to test the hypotheses. These were tested in a lab setting and in a field study with participants using their own computers. All dependent variables excepting the fMRI data support the integration of facial expressions of threat into interruptive security messages for improving security behaviors. An improved fMRI protocol is described which would more closely map to typical security message decision contexts. Lastly, tests suggested that making artistic renderings of the faces did not negatively impact the performance of the warnings on eliciting desirable security behavior outcomes. Theory has benefitted from the extension of media naturalness theory into an information security context, and practitioners can benefit from the security warning design guidance that these studies afford. Future research can further test the efficacy of these interventions over repeated exposures in a longitudinal study spanning days or weeks.

3.0 STUDY 3 – APPLYING RISK TRADEOFF PARADIGMS TO EXPLAIN USER INTERACTIONS WITH INTERRUPTIVE SECURITY MESSAGES

3.1 MOTIVATION

While the first two studies have argued that attention to security messages is essential for purposeful adherence by end users, this study highlights that motivation to adhere is also important. Even if attention to a security message is present, attention absent motivation will, in the end, be more likely to result in non-compliance (Wogalter 2006a). Focusing on message motivation assumes that attention to the message is already in place, which is why this study follows the first two.

Information security research has explored why individuals violate security policies and fall victim to attacks. Some studies make an underlying assumption that users make active risk-taking assessments for every security decision, prompted by security messages (Boss et al. 2015; Johnston et al. 2015). A "lazy user" perspective depicts security as an unnecessary burden that should be bypassed if possible. Many studies use deterrence theory, testing the efficacy of using sanctions to influence security-related decision making (e.g., D'Arcy and Herath 2011; Johnston et al. 2015). Another camp takes the position that "users are not the enemy" (Adams and Sasse 1999), eschewing criminology-inspired sanctioning deterrence, and attributing security misbehavior largely to inattention and habituation (Anderson et al. 2016a; Anderson et al. 2016c). In this view, if a security message is ignored, the design of the interface is to blame. We question how these two stances coexist – purposeful risk-taking security decision making does not seem congruous with inattentive dismissal of security messages.

The purpose of this study is to attempt to reconcile the differences between the two camps of research on user interactions with security messages. In our study, we employ a between-subjects repeated-measures field study using Amazon Mechanical Turk with 510 subjects. In our design, we influence the risk-taking tradeoff by varying the value of adhering to security messages. Corroborating several dependent variables, including security choice, reaction times, and mouse-cursor movements measures, we discover an interesting bimodal pattern where attention is not dependent on risk tradeoff levels, but where warning adherence *is* dependent on the risk tradeoff levels. The findings suggest that participants make security decisions ahead of time, with the decisions being dependent on the risk tradeoff values.

3.2 LITERATURE REVIEW AND HYPOTHESES

3.2.1 Security message inattention

As discussed in the literature review for Studies 1 and 2, a major contributor to security message failure is a basic lack of attention to the message (e.g., Anderson et al. 2016b; Bravo-Lillo et al. 2014; Schechter et al. 2007). Drawing on the attention findings presented earlier, one possible pattern is that users rarely engage in risk-taking assessments when interacting with security messages, regardless of varying levels of tradeoff in the risk-taking decision. This would suggest that users are habituated to the messages, and are performing automatic, learned responses when encountering new ones. If this is true, then research should focus mainly on fostering attention to the messages, so as to increase the likelihood that users will engage with the messages and make meaningful choices.

H1: There will be no difference in markers of cognition between varying risk-taking tradeoff levels (i.e., there will be consistently ***low attention*** *and* ***low warning adherence****).*

3.2.2 Risk tradeoffs

Risk has been studied in an information security context typically through the lens of protection motivation theory (Rogers 1983), wherein the constructs of threat severity and threat susceptibility essentially represent the security threat's risk levels (Boss et al. 2015; Johnston et al. 2015; Johnston and Warkentin 2010). Individual differences in risk perceptions have also been used to predict security message disregard (Vance et al. 2014).

In this study, we consider a different facet of information security risk -- the risk tradeoff associated with *adhering* to the security message. Inherent in the idea of risk is that there is something to be gained from taking the risk. In the finance literature, risk tradeoff is quantifiable as the potential return on investment, with willingness to accept the risk being a function of the magnitude of the return (Ghysels et al. 2005). This same concept of risk-taking behavior being positively associated with the potential gains or loss-avoidance involved has also been described in the behavioral economics literature (e.g., Kahneman and Tversky 1979).

Risk-tradeoff applies to the context of information security messages in that one risks a security threat in exchange for some benefit. Guo et al. (2011) captures the motivation to intentionally violate organizational information security policies with their "relative advantage for job performance [from violating a policy]" measure. Interruptive security messages often block or hinder users from completing their primary tasks (Jenkins et al. 2016). Observance of the security policy adds stress and requires more effort to complete the primary task. Failing to complete the task or taking longer to complete it may lead to poor employee performance evaluations. Users, perceiving this, may be motivated to disregard security policy and ignore security messages (Lowry and Moody 2015). To capture these tradeoffs, we will vary the "penalties" associated with *heeding* the message, while holding constant security threat severity and susceptibility.

We use the risk-taking paradigm to propose the first component of an alternative to H1, wherein users nearly *always* engage in risk-taking assessments when encountering security messages, with the likelihood of security message adherence depends on the tradeoff weights. This view assumes that attention is sufficiently present to prompt risk-tradeoff appraisals, and supports studying the impact of levels of perceived risk on users' security message risk-taking assessments. The existence of the tradeoff assessments can be discerned if lower rates of adherence are present as the tradeoff scale is increasingly tipped towards heeding a message being the more penalizing choice (i.e., the choice with greater tradeoff), holding risk constant.

> *H2a: The relationship between the **balance of risk-taking** and **adherence** to interruptive security warnings will be **strictly monotonic**: e.g., security message heedance will always decrease as heedance increasingly becomes the more penalizing choice.*

3.2.3 Cognitive elaboration as a function of risk-tradeoff balance

We now develop a second component to the alternative to H1. Whereas H2a focused on patterns of adherence, this component of H2 pertains to the degree of attention to (elaboration over) an interruptive security warning, dependent on the balance of the between risks and benefits. This hypothesis component draws from principles of the heuristic-systematic model of information processing (HSM, Chen and Chaiken 1999) to predict whether a user will cognitively engage with a security warning. HSM, a theory of persuasion, finds early expression in the script concept (Abelson 1981). The script concept asserts that an individual will follow a "script" and grant small requests without cognitive elaboration, as long as a reason is given. Individuals will be likely to perform this script unless (1) the script is broken by not providing a reason or if (2) the request is large, in which case they will elaborate over the request and the reason before deciding whether to accept it.

We predict that the perceived risks involved will impact whether or not a user elaborates over a security-message decision. To our knowledge, while HSM has been evaluated in a risk judgement paradigm (Trumbo 2002), the impact of the *balance* of the risk tradeoff has not been examined in an HSM frame. We will manipulate the risks involved for *heeding* the warning. If the script theory concept or HSM elaboration prediction holds, we expect to see a bimodal distribution of behavior across risk levels, where after a certain threshold of risk tradeoff for *adhering* to the message is surpassed, elaboration will be much less likely. From the habituation-theory lens, the scripted behavior would be to rely on memory, which would result in little scrutiny of the at-hand security message (c.f. Böhme and Köpsell 2010; Sunshine et al. 2009). The tradeoff behavior will involve whatever task was interrupted by the security message. If adhering to the message will not adversely impact the interrupted task, then the risk-benefit balance will be more balanced, and elaboration over the decision should be more likely to occur. In summary, we posit that if users perceive that the benefits of heeding a warning are close to the accompanying losses (e.g., time lost or inability to complete an objective), then they will more carefully consider the risk tradeoff before making a decision. However, if the tradeoff choice is clear, then users will be less likely to engage in elaboration, and instead their behavior will more closely follow patterns of lower attention and automatic choices.

> *H2b: The pattern of* ***attention*** *to interruptive security warnings will be* ***bimodal****: the risk decision will either be elaborated over or not, depending on how (un)balanced the risk-benefit balance is.*

3.3 RESEARCH DESIGN

We used the same deception protocol that was described in Study 2. On top of a $1.25 base payment, participants were offered an additional $1.25 performance-based bonus payment. Each incorrect classification results in a "penalty" decrease in their bonus payment, with the penalty amount depending on a participant's randomly assigned treatment group. Four penalty level treatment groups were used: 1, 5, 10, and 25 cents. We chose these increments because they mapped naturally to U.S. coinage so that the penalty would have high salience (i.e., high tangibility or ability to visualize, compared to $1.57 vs $1.92 vs. $3.24, which would take greater visualizing effort for users to compare). Dependent measures tested were the same ones as collected for Study 2: warning choice, reaction time, and mouse-cursor movement measures.

3.4 ANALYSIS AND RESULTS

As in Study 1, with our sample size of 510 participants, G*Power 3.1.9.2 reports that with four treatment groups (one for each amount level) and a high expected correlation among the repeated measures (.95), we have sufficient statistical power to detect, at least, a medium-small effect size. Following the findings from Study 1 about participants suspecting deception after the fourth warning, we only included up to four warning impressions for each participant in any analysis. All continuous dependent variables (e.g., reaction times, mouse cursor click latency, and mouse cursor time idle) in models testing effects across time were natural-log-transformed to remedy non-normality of residuals and heteroskedasticity.

To determine whether to include covariates in our analyses, we tested whether several items were predicted by the emotion or filter treatment. The potential covariates that we tested were participant age, gender, preferred operating system, preferred browser, task performance accuracy, and whether English was their first language. We ran separate ANOVAs for each potential covariate with Type 2 errors on linear models, each including the emotion effect, the filter effect, and their interaction as independent variables. None of the omnibus *F* tests from any of these tests were significant at an alpha level of .05, so no covariates were included in any analyses. None of the omnibus F tests were statistically significant even when the alpha level was relaxed to .10.

The exploration of the pattern of ratio of warnings ignored over time for each participant presented in Appendix B is also relevant to this study. Again, the exploration shows that the majority of participants (80%) were perfectly consistent across warning exposures in whether they ignored or heeded the warning. As in study 1, the likelihood of perfect consistency was not dependent on the assigned treatment group (in this case, the assigned penalty level). Because perfect behavioral consistency was not dependent on penalty treatment group, a discussion of it will be postponed until section “4.0 – Final Discussion and Conclusions”.

3.4.1 Warning adherence rates

To test for differences in adherence rates (whether a participant ignored a warning), we performed an empirical logit analysis (Barr 2008). We specified a fixed effect for treatment group, a fixed effect for the number of warnings seen, an interact effect between these two fixed effects, and a random intercept for each participant. The interaction effect between number of warnings seen and penalty group (the slope) was not statistically significant, $Wald\ \chi^2(3) = 7.537, p =$

.057. An ANOVA found significant differences among treatment groups on whether the warning was ignored, $Wald\ \chi^2(3) = 23.308, p < .001$. Averaged across warning exposures, participants in the 1-cent penalty treatment group were not significantly more likely to ignore warnings than were participants in the 5-cent penalty group ($\Delta probability = 10\%, p = .128$). But, participants in the 1-cent treatment group were 14% less likely to ignore warnings than were participants in the 10-cent and the 25-cent treatment groups (p's < .0001). Participants in the 5-cent treatment group were 12% less likely to warnings than were participants in either the 10-cent or 25-cent treatment groups (p = .011 and p=.014 respectively). Participants in the 10-cent treatment group were just as likely to ignore warnings than were participants in the 25-cent treatment group (p=.958) 50% 5-cent penalty treatment group were 32% less likely to ignore the warning than were participants in the 10-cent penalty treatment group (p= .001), and 39% less likely to ignore the warning than participants in the 25-cent penalty treatment group ($p <$ .001) (see Figure 25 and Table 12).

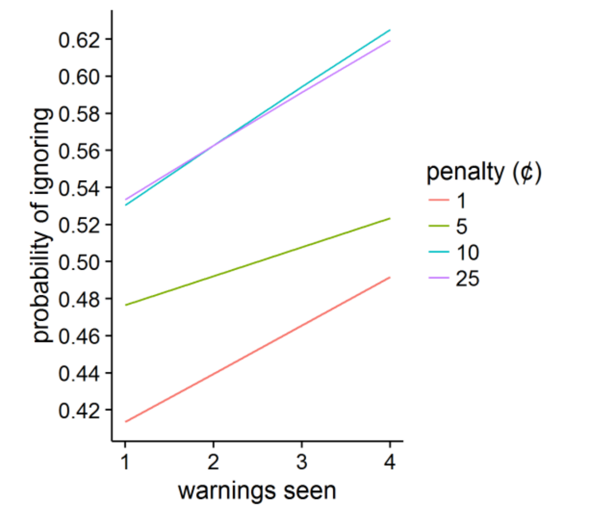


Figure 25. Empirical logit comparing probabilities of ignoring warnings across exposures for each penalty treatment group.

Table 12. Log odds and pairwise comparisons (alpha < .05) of empirical logit analysis predicting warning adherence

Penalty treatment group (cents)	log odds	SE	df	Pairwise comparison group
1	-0.188	0.0888	497.47	1
5	0.002	0.0872	499.66	1
10	0.321	0.090	497.04	2
25	0.314	0.091	497.02	2

3.4.2 Reaction time.

We tested for the impact of treatment group on a log transformation of reaction time using a linear mixed model with random intercept for each participant, along with fixed effects for treatment condition, number of warnings seen, plus an interaction between the fixed effects. However, neither the omnibus test for the interaction effect nor for the main effect found significant differences ($\chi^2(3) = 3.324, p = .334$ and $\chi^2(3) = 4.635, p = .201$ respectively) (see Figure 26).

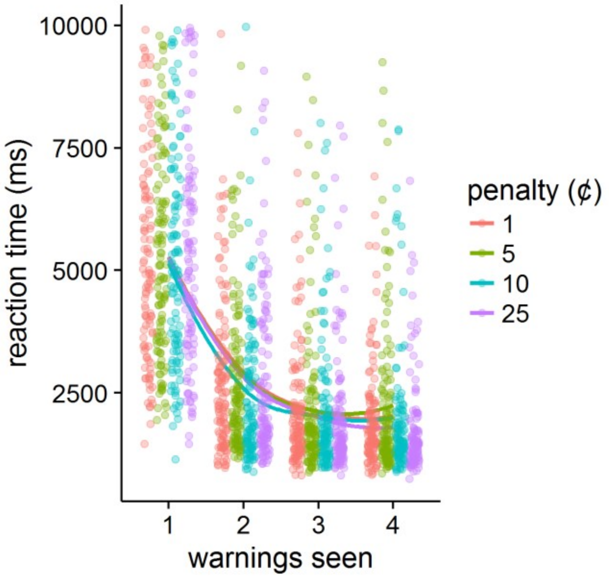


Figure 26. Untransformed reaction time (in milliseconds) plotted across time for each penalty treatment group.

3.4.3 Mouse cursor measures

While we had many mouse cursor measures available to test, we only present the two that were analyzed in Study 1, for the sake of symmetry between the two studies. We explored other mouse cursor measures not reported below and found similar result patterns for the penalty treatment group factor, for first-impression-only and for first-four impression analyses.

3.4.3.1 Mouse-cursor: Click latency

We first performed analyses on the log-transformed transformation of the click latency measure. For our first analysis of this dependent variable, we only considered the mean difference between treatment groups for the first warning exposure for each participant. However, no significant differences were found among penalty treatment group levels in the omnibus test, $F(3{,}205) = 0.441, p = .724$. We also tested a model that included the first four warnings for each participant. This model had an interaction effect for the number of warnings seen with the penalty treatment group, a fixed effect for the penalty treatment group, as well as a random intercept for each participant. However, neither the interaction effect nor the main effect was significant in this model, either ($\chi^2(3) = 2.638, p = .451$ and $\chi^2(3) = 3.541, p = .315$ respectively) (see Figure 27 and Figure 28).

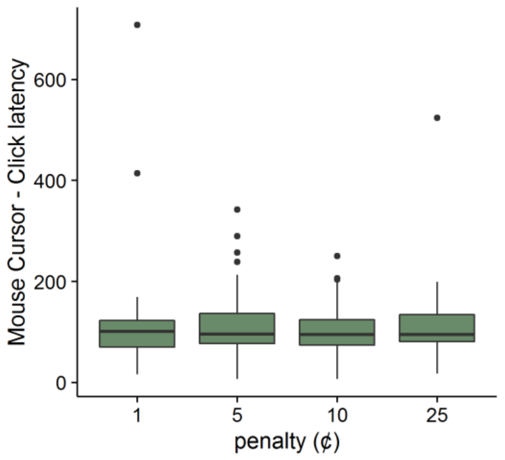


Figure 27. Click mean latency for first warning impressions only.

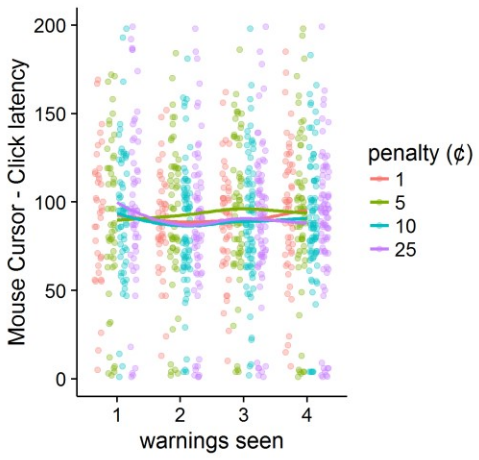


Figure 28. Click mean latencies for each penalty treatment group for the first four warning impressions.

3.4.3.2 Mouse-Cursor: Time Idle

We followed the same approach for the mouse-cursor time idle measure as we did for the click latency measure. We first performed analyses on the log-transformed transformation of the click latency measure. For our first analysis of this dependent variable, we only considered the mean difference between treatment groups for the first warning exposure for each participant. However, no significant differences were found among penalty treatment group levels in the omnibus test, $F(3{,}205) = 0.536, p = .658$. We also tested a model that included the first four warnings for each participant. This model had an interaction effect for the number of warnings seen with the penalty treatment group, a fixed effect for the penalty treatment group, as well as a random intercept for each participant. However, neither the interaction effect nor the main effect was significant in this model, either ($\chi^2(3) = 1.724, p = .632$ and $\chi^2(3) = 1.696, p = .638$ respectively) (see Figure 29 and Figure 30).

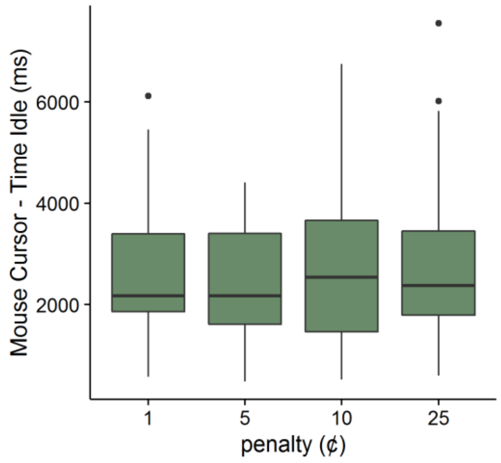


Figure 29. Time idle for first warning impressions only.

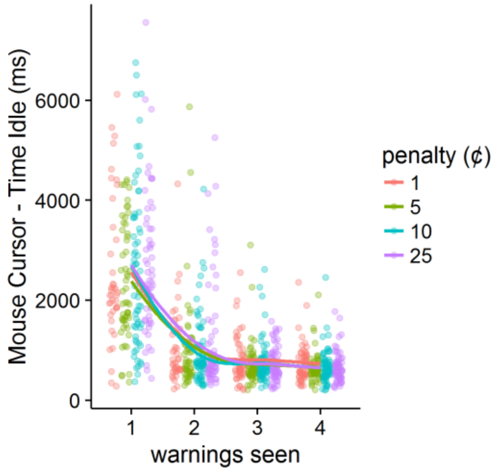


Figure 30. Time idle for each penalty treatment group across for the first four warning impressions.

3.4.4 Survey results

When participants were asked whether they noticed a bonus penalty for each incorrect response, 500 reported "yes", while 10 reported "no". Conflictingly, 8 of the 10 participants who reported not noticing the bonus penalty also reported having at least some concern about the bonus status bar. Despite these oddities, all participants were retained in the analysis to provide "a more robust testing of the hypotheses" (Straub et al. 2004, p. 408).

ANOVAs were performed on three survey items to test for differences among groups. The items measured warning concern, penalty concern, and concern over the bonus status. Warning concern and penalty concern were measured on a scale of 1 ("not at all concerned") to 11 ("extremely concerned"). Bonus status was on a scale of 1 ("Never") to 5 ("Very often"). Items were measured with the following questions: For *warning concern*, "how concerned did the warning make you feel?"; for *penalty concern*, "How concerned did the bonus penalty make you feel?"; for *bonus status*: "How often did you check the bonus status bar?"

No differences were observed among responses to warning concern or bonus status concern ($F(3{,}506) = 2.187, p = .089$ and $F(3{,}506) = 1.71, p = .164$ respectively). As would be expected, differences were observed among concern over the penalty ($F(3{,}506) = 11.82, p < .0001$. Participants in the 1-cent penalty group reported lower concern over the penalty than did participants in any other group (all p's < .0001, Bonferroni-adjusted). No other differences among treatment groups on penalty concern were observed, although an expected upwards trend is observed (see Figure 31 and Figure 32).

We also tested for differences among reports of perceived malware risk, malware threat severity, threat susceptibility, and malware fear. Differences were observed among penalty treatment groups for reports of perceived threat severity ($F(3{,}506) = 3.322, p = .020$). Participants in the 1-cent penalty group reported marginally higher perceived threat severity than did participants in the 10-cent penalty group (*p* = .068, Bonferroni-adjusted). Participants in the 5-cent penalty group also reported higher perceived threat severity than did participants in the 10-cent penalty group (*p* = .033, Bonferroni-adjusted). No differences were observed among penalty treatment groups for perceived risk ($F(3{,}506) = 0.800, p = .495$), threat susceptibility ($F(3{,}506) = 1.549, p = .201$), or fear of security threats ($F(3{,}506) = 0.878, p = .452$) (see Figure 33).

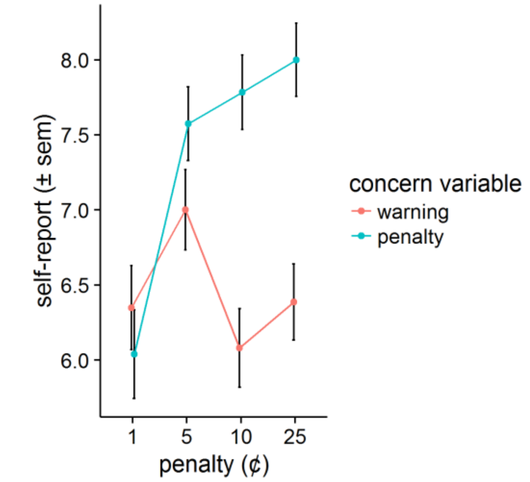


Figure 31. Self-reported values for two "concern" items, scale of 1 ("not at all concerned") to 11 ("extremely concerned").

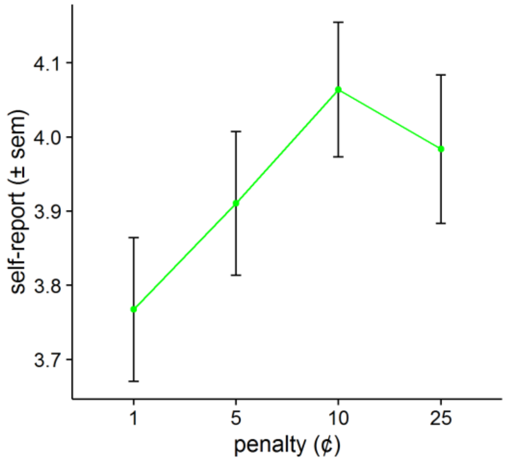


Figure 32. Self-report for "How often did you check the bonus status bar?", scale of 1 ("Never") to 5 ("Very often").

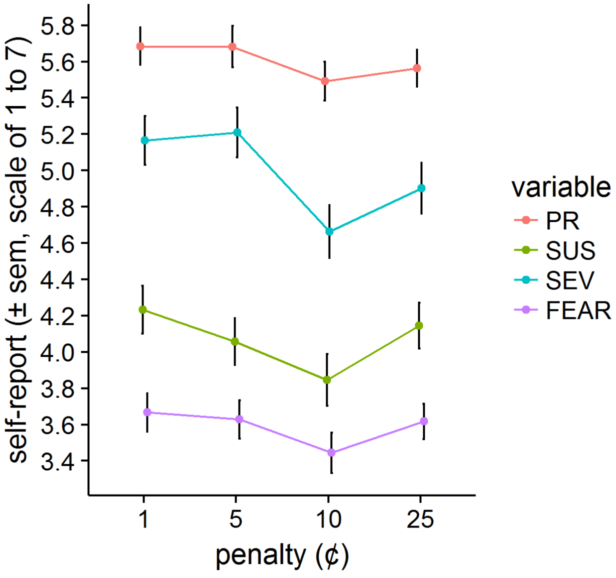


Figure 33. Unadjusted means for four self-reported measures for each penalty treatment group. The variables are perceived risk (PR), malware threat susceptibility (SUS), malware threat severity (SEV), and fear of malware (FEAR).

3.5 STUDY 3 – DISCUSSION

We tested for differences on various outcomes: (1) actual adherence rates, (2) reaction times, and two mouse cursor movement statistics: (3) click latencies and (4) time idle. Our hypotheses were informed by competing bodies of literature. The view that users are desensitized to warnings and that they do not consider underlying risks informed H1, while H2a and H2b were based on the assumption that users engage in meaningful threat assessments for each security decision.

In our results for actual adherence rates, participants who were only penalized 1 or 5 cents per incorrect answer were much less likely to ignore the warning than were participants were penalized either 10 or 25 cents per incorrect response. This was expected – a penalty of 10 cents and 25 cents represented losses of 8% and 20% of the available $1.25 bonus that participants stood to earn, respectively. Using a risk tradeoff paradigm, participants appeared to be more likely to trade anywhere 4% or less of their bonus to avoid a security risk than they were to trade anywhere above 12.5%. One interesting observation is that there were no observed differences in adherence rates between the 10-cent and 25-cent treatment groups, or between the 1-cent and 5-cent treatment groups. This suggests that the risk-analysis tradeoff in which individuals engage is not necessarily strictly monotonic, but rather, that it is modal. This conflicts with a strict interpretation of H2a, which predicted that adherence rates would always change with a change in penalty levels. However, one alternative interpretation of our findings is that the change in adherence rates reached nearly asymptotic levels along a very steep curve at the level of the 10-cent penalty group. Supporting this interpretation, differences, albeit statistically insignificant ones, were observed between the 1-cent and 5-cent penalty groups (see Figure 25). With this interpretation and with the asymptote assumption, H2a is supported.

In this study, an 8% penalty – a mere increase of 4 percentage points over the next-lowest treatment group – was sufficient to substantially boost the rates of security warning disregard (the 10-cent treatment group was 32% more likely to disregard the warning than was the 5-cent group, see Figure 25). These results may generalize to the workplace: individuals at work may engage in these risk-taking tradeoffs when they are interrupted by security messages. Time lost through adhering to the security warning and finding a workaround may result in negative outcomes such as missing a work deadline. Depending on the weight of these negative outcomes compared to the perceived benefits of avoiding the security threat, similar warning adherence patterns as seen in our study may be observed in the workplace.

Compared to the adherence findings, our outcome measures indicative of attention and cognitive processing – namely, reaction times and mouse cursor click latency and time idle – showed different result patterns. No significant differences among penalty treatment groups were observed for *any* of the cognitive processing dependent variables. Unlike the findings for adherence rates, this lends support to the H1 – that participants do not attend to and elaborate over the warning regardless of the risk tradeoff values. However, when we consider the adherence rates findings vis-à-vis the attention findings, we make a striking observation. Users made security choices following a risk-tradeoff paradigm, but required no differential time to analyze those tradeoffs in the moment. They must have made the decision *ahead of time*. Put another way, users appeared to already have values in mind for their system's security and for continuing the task at hand. They did not require time to analyze, value, and weigh the tradeoff decisions in the moment. Thus H1 finds support in that participants did not differentially make risk-tradeoff assessments when in the moment of being exposed to the warning (rejecting H2b), but H2a finds support in that users acted differentially between treatment groups in their adherence, apparently following pre-determined mental models of information security risk values. Hence, users do *not* indiscriminately disregard all security warnings (supporting H2a over H1), but they also *do not* attend closely give the decision differential attention in the moment despite varying risk tradeoff levels (supporting H1 over H2b). This leaves us with an unexpected takeaway for our comparison of the two conflicting literature streams that informed H1 and H2a-b. The view that users are habituated to messages, and that their attention to them does not depend on the content or context of the warning, finds support from the findings that backed H1. Meanwhile, simultaneously, the view that users' behavior is indeed impacted by risk tradeoffs was supported by the findings that backed H2a. The main issue is that while attention may be low, discriminatory security choices are *still* made, albeit in the form of pre-made decisions, and not *in-situ* ones.

The survey results showed that participants in the 1-cent penalty group had only moderate concern about the penalty, while participants in any of the higher penalty treatment groups had greatly amplified concern compared to the 1-cent group. Concern with the penalty was not correlated with self-reported concern for the security warning, although we know that participants in the higher penalty groups were more likely to ignore the warning. It is surprising that penalty concern and warning concern did not correlate – we would have thought that concern over the penalty should have been closely related to concern over the warning, because the appearance of the warning threatened a participant's reward. It is possible that participants interpreted the question about concern for the warning to be asking about how worried they were about actually receiving malware on their machine. With this interpretation, the penalty level and penalty concern would not correlate with warning concern. No differences were observed among penalty treatment groups on warning concern – all participants reported moderate levels of concern for the warning, reporting a concern level of about 6.5 out of 11. This warning concern level appears to have been high enough for participants to take the warning seriously, as evidenced by about 50% of participants choosing to heed the warning (with the actual heedance rate depending on penalty treatment group).

Survey responses to measures of perceived risk, threat susceptibility, threat severity, and fear of threat showed an interesting inflection point at the 10-cent penalty group mark. While many of the variables do not show significant differences among penalty group levels, the inflection point is apparent from a visual inspection (see Figure 33). We are unsure why it would be that 10-cents is a low mark. It is possible that participants responded differently to denominations ending in a 0 compared to ones ending in a 5, although we know of no theory to support this argument. Or perhaps participants had a visualization of the actual size of the coin, and measured the threat this way. A dime is the smallest coin, so it may have elicited the lowest threat and fear appraisals. Future research with larger budgets can control for penalty denomination size by using different penalty level denominations that come from bills (e.g., $1, $5, $10, etc.). However, using such high denominations may be unethical – payments to Amazon Mechanical Turk workers tend to be lower than the U.S. federal minimum wage (Hitlin 2016), so a task with a $10+ potential reward would be overwhelming. By comparison, our task offered up to $2.50, an equivalent about $12 per hour. Even with this amount, we received complaints that the amount was so high that some participants felt great emotional distress each time they lost some of their bonus, or when they felt that they had to return the task to avoid risk of being rejected. Said one worker, because she had to return our $2.50 task, she no longer would be able to afford a meal of rice and beans for her children. Perhaps a task could be used where the bonus-earned displayed on the screen would not directly map to what participants would receive in actual payment at the end of the task. However, this would hinder the generalizability of the risk tradeoff amounts that participants used.

The slopes for warning adherence showed an upwards trend; over repeated warning exposures, users became more likely to ignore the warning. This may be evidence of a learning effect. By the second warning exposure, users may have decided that the warning was “crying wolf” (Sunshine et al. 2009); nothing bad appeared to happen after the first warning, so confidence in future warnings may have dropped. A loss of trust such as this would be unfortunate. While older strands of malware were noisy and had graffiti or destruction as their primary aim(Ducklin 2016), modern malware may install itself silently and enslave the computer into a botnet or steal login passwords, unbeknownst to the user (Goodin 2017). Or only make itself known after encrypting important files on the device (ransomware) (Krebs 2016). Because of such threats, subsequent malware warnings should be taken just as seriously as the first ones. Research is needed for best ways to educate users about such threats, and for how to avoid the loss in credibility that appears to be such a common issue.

A greater number of treatment groups would be necessary to determine the number of modes for interruptive security warning risk-taking decisions. For situations where the tradeoff quantification is less immediately quantifiable than our money-penalty operationalization, a model would be useful to describe what perceptual factors best predict the tradeoff values that participants use when they engage in evaluation of the security message risk-taking tradeoff. Such a model could build on the information security policy violation intention models already in existence (e.g., D'Arcy and Herath 2011). Organizations can modify their incentive structures to *decrease* the tradeoff amount that organizational insiders discern when considering whether to adhere to a warning, perhaps through threat of sanctions for non-security-message adherence (D'Arcy and Herath 2011; Johnston et al. 2015), or through rewards for good security hygiene. Security message design can also aim to boost perceptions of threat severity and susceptibility, which may also tip the risk-tradeoff decision further.

3.6 STUDY 3 – CONCLUSION

This study has investigated a gap in information security literature between assumptions of high and low user attention and adherence to interruptive security messages. Using an interruptive security message context, the corroboration of multiple dependent variables from a field study supported the existence of users behaving under a bimodal risk tradeoff paradigm, where security message adherence was dependent on the risk tradeoff balance between the perceived information threat and the losses involved in not being able to perform the interrupted task. Users who were penalized 1 or 5 cents from a bonus payment for heeding an interruptive security warning were more likely to ignore that warning than were participants who were penalized 10 or 25 cents. While, simultaneously, users did not show differential levels of elaboration over security warnings regardless of how much money they lost from their available bonus. This lack of differential elaboration was evident in reaction time data as well as in data from two mouse cursor measures – click latency and movement idle time. A likely explanation for the findings is that users rely on predetermined mental models of information security valuation and risk. Future research should be performed to further investigate users' risk perceptions when interacting with interruptive security messages, including how to manipulate these perceptions, and how deterrence approaches such as sanctioning apply to the context of interruptive security messages.

4.0 FINAL DISCUSSION AND CONCLUSIONS

We now reflect on what has been learned across all three studies presented in this dissertation. The overarching theme was to both better understand and also to improve outcomes of user interactions with interruptive security messages. The first two studies sought to influence users' elaboration over and attention to interruptive security messages through manipulating the design of warnings. To this end, media naturalness theory informed the integration of human facial expressions of threat into internet browser malware warnings. A field study and an fMRI study were both performed in a bid to corroborate the findings and gain both internal and external validity. The field study showed that integrating facial expressions of threat may indeed improve user attention to interruptive security messages, despite small practical effect sizes. Applying an image filter to the facial expressions that "posterized" or "cartoonized" the expression did not decrease the potency of those messages for garnering user elaboration compared to warnings with photo-realistic (unfiltered) integrated facial expressions. The survey revealed generally low perceived threat susceptibility to and fear caused by security messages across all emotion and filter treatment groups, while perceived threat severity and general perceived risk were reported to be generally high. The low perceived susceptibility and fear suggest a possible troubling invincibility complex under which users may be operating. Also, while self-reported warning realism was lower for face-integrated warnings, participants reported equal concern over the warnings regardless of whether they had integrated facial expressions. These last survey findings, coupled with objective behavioral measures, suggest that users considered the warnings to be *real enough* to warrant increased attention if they had integrated facial cues of threat.

The third study investigated contrasting stances of information security research, with one holding that users respond automatically and consistently poorly to security warnings regardless of the balance of the underlying risks. The other stance assumes that users always engage in risk-taking assessments, and that the balance of the risk tradeoff has a strong and conscious impact on users' security decisions. To investigate this, the third study held the warning design constant, and instead varied risk tradeoff values for the interruptive security warning decisions. This was done by manipulating the monetary penalty tradeoff for users to stay safe and heed the warning. This study showed interesting results – while the amount of the monetary tradeoff had a strong impact on user's security decisions, no evidence of differential elaboration over the interruptive security warning risk-taking moment was observed. These findings suggest that while users may be responding automatically to warnings without updating mental models for each exposure, they are *not* consistently choosing to disregard – instead, the value of disregarding an interruptive security warning (one of the risk tradeoff values) is predictive of actual warning disregard. Study 3 calls for more research into valuating the mental models that users employ when interacting with interruptive security warnings specifically, and for ways to predict and influence the tipping point for when users will begin to heed or disregard an interruptive warning.

Studies 1 and 3 demonstrated measuring actual user security behaviors via a deception protocol. Both binary decisions (heed/disregard) as well as unobtrusive and objective mouse cursor measures of elaboration and attention were examined. These studies answer the call of Crossler et al. (2013) to capture objective measures of information security behaviors. They also highlight the challenge of measuring habituation to the warnings over a short timeframe, given rapid habituation rates when intervals between stimuli are short (Rankin et al. 2009). Study 2 followed the example of other recent NeuroIS security studies in that it corroborates data collected via high-resolution techniques such as fMRI alongside externally valid field study findings (e.g., Anderson et al. 2016c). The three studies in this dissertation seek to be an example for other behavioral information security studies in approaching information security research questions via mixed methods, including NeuroIS.

One might propose an alternative interpretation of the hypotheses in studies 1 and 3 related to response times. Whereas we proposed that longer response times are evidence of greater elaboration and attention to the security warning, an alternative interpretation could be that quicker response times are evidence of greater problem-solving ease. With this alternative interpretation, quicker response times would be more desirable than longer ones. We would counter this interpretation by referencing Akhawe and Felt (2013), who found that, in their field study, "47% of users who clicked through the warning made the decision within 1.5s, whereas 47% of users who left the page did so within 3.5s. We interpret this to mean that users who click through the warning often do so after less consideration." Rewording this, the status quo is that users who ignore warnings do so quickly, with hardly enough time for any meaningful "problem solving" to be engaged with. Reading and pondering the warning takes a minimum threshold of time, and given the short timespan ranges within which our data falls (<5 seconds), we argue that our hypotheses' assumption of longer reaction times being indicative of greater elaboration is appropriate.

We note an ethics concern that arose during our data collection. When designing our protocol, we were concerned that some participants would game the task by starting the main task, and then switching to a different task while allowing our task to run in the background. With this method, each image would time out after 10 seconds, and they would end up with an accuracy of 0%. This would lead them to lose all of their bonus payment. But, they would still earn their base payment, ostensibly while simultaneously working on another task. To guard against this kind of gaming, we warned participants that if their performance accuracy was too slow, we would reject their submission. Participants later complained to us that the threat of rejection was too harsh – they became worried when our security warnings appeared, because they had to choose between their computer's security and a risk of rejection. A rejection means more than a loss of payment to an Amazon Mechanical Turk worker – it is also a bad mark on their overall approval rating. Too many bad marks can mean that they become disqualified from doing higher-paying tasks. So, some participants felt that we were asking them to not just risk a few cents per warning, but rather, that we were asking them to risk their livelihoods. We reported this adverse unexpected situation to our institutional review board. For future data collection using this protocol, we will not threaten to reject tasks with accuracy levels that are too low. We feel that the per-warning money risk is sufficient to simulate a risk tradeoff scenario. It is not every day that a user has to choose between complying with a security policy and losing their job (although such a draconian scenario is not outside the realm of possibility!). In all, we noticed that 63 participants quit the task before finishing, compared to 555 participants who completed the task. A substantial number of these dropouts could have been participants not wanting to risk their approval rating *or* the security of their system. We do not have dropout rates to compare ours against, but we imagine that our dropout rate would fall on the high side. Instances of dropping out appear to be randomly distributed across treatments. Any participant who dropped out who contacted us with a concern was debriefed and paid in full for their time worked. We will monitor dropout rates for future tasks to discern if users are feeling similar unnecessary pressure.

Because studies 1 and 3 used a repeated-measures design, we were able to explore the degree of users' consistency in how they respond to security warnings. Would users always do the same thing regardless of how many warnings they saw, or would they vary their responses? Our exploration in Appendix B shows that the majority of participants – 80% – were perfectly consistent across warning exposures in whether they ignored or heeded the warning. The remaining 20% were spread fairly evenly between the two poles of “always ignored” and “always heeded.” This trend of perfect consistency is interesting. It does not appear that any of our treatments from studies 1 or 3 impacted whether a participant would be perfectly consistent in how they responded to warnings. We did see, however, that some of our treatments impacted whether participants would always *heed* versus *ignore* the warnings. We are left to wonder how quickly users form mental models about how they are going to respond to a security warning.

Also, it is unknown just how set users are in these mental models – can a different security message design easily move users from one perfectly-consistent pole to another? What differences are there in people who change their responses across multiple impressions? Perhaps some felt remorse about an early “ignore” choice and sought to not expose themselves to any further risk on future impressions. This would be evidence of users engaging in context updating for each warning impression. Or perhaps some of those users heeded an early warning, after which they quickly became fatigued by their security posture, and gave up on behaving securely when they encountered future warnings. And what goes on in the minds of users who show perfect consistency in their warning choices? Perhaps the overriding desire was to avoid cognitive dissonance, so they matched their first choice in all future choices. Potentially more likely, the warning context of subsequent warnings matched the context of the first warning so closely that no additional information had been provided that would require users to update their mental models and change their behavior. We did notice a sharp decline in response times across multiple warning exposures. A more qualitative experimental design would be needed to tease out the thought processes between users with and without perfect consistency. Predictors could be identified that would classify a user as likely to be perfectly consistent or not.

As described in the Introduction, all studies touched on different stages of Wogalter's C-HIP model (2006a) (see Figure C-1). Theoretically, individuals process warnings in stages, or "gates" which must be passed before a warning can successfully prompt a desired behavior. The first two studies were motivated by research that shows that the very first requirement of the C-HIP model, attention to the warning, is oftentimes found wanting (Anderson et al. 2016b). The C-HIP model suggests that it is not worthwhile to focus on any behavioral gate beyond any failed gate. That is, attention must be resolved before motivation can be addressed. However, study 3 demonstrated that for an interruptive security message context, attention was *sufficiently* present for a motivation-focused treatment to be impactful. In study 3, the monetary penalties were predominantly motivational factors, not attentional ones. Study 3 did not find evidence of the motivational treatment differentially impacting attention. But the findings from study 3 suggest that users make decisions that drive their security choices ahead of time. Again, when the findings of all three studies are considered together, the lesson is that while manipulating the attention gate can still have an impact on users' choices, a manipulation of the motivational factor can have a yet stronger one, which means that attention to the interruptive security warnings must be at least somewhat present. This speaks hope against the fear that users mindlessly respond to all messages. There is at least some thought given, albeit the thought may not be given *in situ*. Ideally, users would carefully consider each security decision, so a focus on C-HIP's attention gate to drive up engagement with each security decision is still warranted. A better understanding of the C-HIP motivational gate as it relates to an interruptive security message context is also called for.

The slopes for reaction times over repeated warning exposures in all three studies, but especially in studies 1 and 3. were not linear – they were curvilinear, and they appeared to be approaching an asymptote. This observation is consistent with the theory of habituation (Rankin et al. 2009), which consistently shows responses to repeated stimuli trending towards and eventually reaching an asymptote. Eventually, reactions such as reaction time to a stimulus cannot get any weaker, and elaboration over a stimulus bottoms out. Given enough rest before the next exposure to a stimulus, a response will "recover", but not necessarily back up to its original high mark. What is interesting from habituation theory is that even once the asymptote is reached for a given stimulus, continued exposure to the stimulus will have a negative impact on the subsequent recovery, such that the recovery height will be even lower. That is to say, even after responses have appeared to stagnate, habituation still occurs for repeated stimulus exposure. We expected no differently, and found no differently, for our interruptive security message context.

One troubling observation from studies 1 and 3 should be noted. The Chrome warning stated that "the site ahead contains malware." There was no uncertainty communicated in the warning about the presence of the malware. That is to say, the warning did not say "the site ahead *might* contain malware", but rather, that the site ahead unequivocally *did contain malware*. Yet averaged across all treatments, over 50% of warnings were ignored. This troubling behavior happened despite the fact that participants used their own computers, outside of any lab setting where they might feel artificial protection due to being in a lab environment (Sotirakopoulos et al. 2011). Why could this be? Perhaps users mistook the warning for an older all-red version of Chrome's SSL warnings (which now has a white background to differentiate it from malware warnings), which warned of a commonly much less severe threat with higher uncertainty about whether the site ahead was dangerous. This explanation is possible. However, several participants, unprompted, offered explanations for their own poor security behavior in the free response section of the post-task survey. Reasons for ignoring the warning included ideas such as the following: “I know that Chrome protects me from malware (there was something about that in the settings),” “my antivirus protects me from threats,” “I use a Mac so I am safe,” and “Chrome gives these kinds of warnings all the time and it’s never serious.” Underlying these responses is a troublesome trend: mistrust of the browser warning, and potentially misplaced trust in users’ own judgement and secondary protections. Malware can be written for Macs (2016); antivirus doesn’t catch all malware (malware writers use sites like virustotal.com to test their strain against up-to-date antivirus definitions) (Ducklin 2015); Chrome malware warnings have extremely high accuracy (the true positive rate is so high the normal warning does not have a “Proceed anyway” button that allows users to bypass the warning). Practitioners and researchers need to work towards increasing users’ trust – or restoring trust – in browser security warnings such as malware warnings. We also need to work towards better understanding security warning mistrust and its facets.

In all, the three studies represent important steps towards better understanding and improving user interactions with interruptive security messages. All three address the domain of basic attention to and elaboration over the warnings – the very first step on the way towards warning compliance (Wogalter 2006a) – whether through manipulating the design or by questioning whether various risk tradeoff values influence in situ warning elaboration. More research is yet needed to understand and solve the problem of low attention. May it come soon, so that fewer users will inappropriately dismiss warnings, lest one such seemingly inconsequential error lead to the next high-profile organizational disaster.

BIBLIOGRAPHY

Abelson, R. P. 1981. "Psychological Status of the Script Concept," *American Psychologist*, (36:7), pp. 715-729.

Adams, A., and Sasse, M. A. 1999. "Users Are Not the Enemy," *Communications of the ACM*, (42:12), pp. 40-46.

Akhawe, D., and Felt, A. P. 2013. "Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness." Paper presented at the Proceedings of the 22nd USENIX Conference on Security, Washington, D.C.

Amer, T. S., and Maris, J.-M. B. 2007. "Signal Words and Signal Icons in Application Control and Information Technology Exception Messages—Hazard Matching and Habituation Effects," *Journal of Information Systems*, (21:2), pp. 1-25.

Anderson, A. K., Christoff, K., Panitz, D., Rosa, E. D., and Gabrieli, J. D. E. 2003. "Neural Correlates of the Automatic Processing of Threat Facial Signals," *The Journal of Neuroscience*, (23:13), pp. 5627-5633.

Anderson, B. B., Jenkins, J., Vance, A., Kirwan, C. B., and Eargle, D. 2016a. "Your Memory Is Working against You: How Eye Tracking and Memory Explain Susceptibility to Phishing," *Decision Support Systems*, (92), pp. 3-13.

Anderson, B. B., Kirwan, C. B., Jenkins, J. L., Eargle, D., Howard, S., and Vance, A. 2015. "How Polymorphic Warnings Reduce Habituation in the Brain: Insights from an fMRI Study." Paper presented at the Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), Seoul, South Korea.

Anderson, B. B., Vance, A., Kirwan, C. B., Eargle, D., and Jenkins, J. L. 2016b. "How Users Perceive and Respond to Security Messages: A NeuroIS Research Agenda and Empirical Study," *European Journal of Information Systems*, (25:4), pp. 364-390.

Anderson, B. B., Vance, A., Kirwan, C. B., Jenkins, J., and Eargle, D. 2016c. "From Warnings to Wallpaper: Why the Brain Habituates to Security Warnings and What Can Be Done About It," *Journal of Management Information Systems*, (33:3), pp. 713-743.

Authors 2016. "From Warnings to Wallpaper: Why the Brain Habituates to Security Warnings and What Can Be Done About It," *Journal of Management Information Systems*, (33:3), pp. 713-743.

Barr, D. J. 2008. "Analyzing 'Visual World' Eyetracking Data Using Multilevel Logistic Regression," *Journal of Memory and Language*, (59:4), pp. 457-474.

Böhme, R., and Köpsell, S. 2010. "Trained to Accept?: A Field Experiment on Consent Dialogs," In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, pp. 2403-2406.

Boss, S. R., Galletta, D. F., Lowry, P. B., Moody, G. D., and Polak, P. 2015. "What Do Users Have to Fear? Using Fear Appeals to Engender Threats and Fear That Motivate Protective Behaviors in Users," *MIS Quarterly*, (Forthcoming).

Bravo-Lillo, C., Cranor, L., Komanduri, S., Schechter, S., and Sleeper, M. 2014. "Harder to Ignore? Revisiting Pop-up Fatigue and Approaches to Prevent It."

Bravo-Lillo, C., Komanduri, S., Cranor, L. F., Reeder, R. W., Sleeper, M., Downs, J., and Schechter, S. 2013. "Your Attention Please: Designing Security-Decision UIs to Make Genuine Risks Harder to Ignore," In: Proceedings of the Ninth Symposium on Usable Privacy and Security, ACM, Newcastle, United Kingdom, pp. 1-12.

Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch, S. L., Buckner, R. L., Strauss, M. M., Hyman, S. E., and Rosen, B. R. 1996. "Response and Habituation of the Human Amygdala During Visual Processing of Facial Expression," *Neuron*, (17:5), pp. 875-887.

Cannon, W. B. 1932. *The Wisdom of the Body*, W W Norton & Co: New York, NY, US.

Chen, S., and Chaiken, S. 1999. "The Heuristic-Systematic Model in Its Broader Context," in *Dual-Process Theories in Social Psychology,* S. Chaiken and Y. Trope (eds.), Guilford Press: New York, NY, US, pp. 73-96.

Crossler, R. E., Johnston, A. C., Lowry, P. B., Hu, Q., Warkentin, M., and Baskerville, R. 2013. "Future Directions for Behavioral Information Security Research," *Computers & Security*, (32:1), pp. 90-101.

D'Arcy, J., and Herath, T. 2011. "A Review and Analysis of Deterrence Theory in the IS Security Literature: Making Sense of the Disparate Findings," *European Journal of Information Systems*, (20:6), pp. 643-658.

D'Arcy, J., Hovav, A., and Galletta, D. 2009. "User Awareness of Security Countermeasures and Its Impact on Information Systems Misuse: A Deterrence Approach," *Information Systems Research*, (20:1), pp. 79-98.

Daft, R. L., and Lengel, R. H. 1986. "Organizational Information Requirements, Media Richness and Structural Design," *Management Science*, (32:5), pp. 554-571.

Daft, R. L., Lengel, R. H., and Trevino, L. K. 1987. "Message Equivocality, Media Selection, and Manager Performance: Implications for Information Systems," *MIS quarterly*), pp. 355-366.

Darwin, C. R. 1859. *On the Origin of Species by Means of Natural Selection*, John Murray: London.

Dhamija, R., Tygar, J. D., and Hearst, M. 2006. "Why Phishing Works." Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montréal, Québec, Canada.

Dimoka, A., Pavlou, P. A., and Davis, F. D. 2011. "Research Commentary—NeuroIS: The Potential of Cognitive Neuroscience for Information Systems Research," *Information Systems Research*, (22:4), pp. 687-702.

Ducklin, P. 2015. "Malware-as-a-Service "Fully Undetectable" Operators Busted," *Naked Security* (available at https://nakedsecurity.sophos.com/2015/11/30/cops-arrest-2-for-operating-fully-undetectable-malware-service/). Accessed 4/21/2017.

Ducklin, P. 2016. "Malware Museum Shows How It Was "before It Was All About Money"," *Naked Security* (available at https://nakedsecurity.sophos.com/2016/02/08/malware-museum-shows-how-it-was-before-it-was-all-about-money/). Accessed 4/21/2017.

Ebner, N. C., Riediger, M., and Lindenberger, U. 2010. "FACES--a Database of Facial Expressions in Young, Middle-Aged, and Older Women and Men: Development and Validation," *Behavior Research Methods*, (42:1), pp. 351-362.

Egelman, S., Cranor, L. F., and Hong, J. 2008. "You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings." Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy.

Ekman, P., and Friesen, W. V. 2003. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*, ISHK.

Felt, A. P., Ainslie, A., Reeder, R. W., Consolvo, S., Thyagaraja, S., Bettes, A., Harris, H., and Grimes, J. 2015. "Improving SSL Warnings: Comprehension and Adherence." Paper presented at the Proceedings of the Conference on Human Factors in Computing Systems, Seoul, South Korea.

Felt, A. P., Ha, E., Egelman, S., Haney, A., Chin, E., and Wagner, D. 2012. "Android Permissions: User Attention, Comprehension, and Behavior," In: Proceedings of the Eighth Symposium on Usable Privacy and Security, ACM, New York, NY, USA, pp. 3:1-3:14.

Felt, A. P., Reeder, R. W., Almuhimedi, H., and Consolvo, S. 2014. "Experimenting at Scale with Google Chrome's SSL Warning," In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, pp. 2667-2670.

Gauthier, I., Skudlarski, P., Gore, J. C., and Anderson, A. W. 2000. "Expertise for Cars and Birds Recruits Brain Areas Involved in Face Recognition," *Nat Neurosci*, (3:2), pp. 191-197.

Ghysels, E., Santa-Clara, P., and Valkanov, R. 2005. "There Is a Risk-Return Trade-Off after All," *Journal of Financial Economics*, (76:3), pp. 509-548.

Good, N., Dhamija, R., Grossklags, J., Thaw, D., Aronowitz, S., Mulligan, D., and Konstan, J. 2005. "Stopping Spyware at the Gate: A User Study of Privacy, Notice and Spyware," In:

Proceedings of the 2005 Symposium on Usable Privacy and Security, ACM, New York, NY, USA, pp. 43-52.

Goodin, D. 2017. "Feds Deliver Fatal Blow to Botnet That Menaced World for 7 Years," *Ars Technica* (available at https://arstechnica.com/tech-policy/2017/04/feds-deliver-fatal-blow-to-botnet-that-menaced-world-for-7-years/). Accessed 4/21/2017.

Gray, J. A. 1987. *The Psychology of Fear and Stress*, (2 ed.) Cambridge University Press: New York, NY US.

Guo, K. H., Yuan, Y., Archer, N. P., and Connelly, C. E. 2011. "Understanding Nonmalicious Security Violations in the Workplace: A Composite Behavior Model," *Journal of Management Information Systems*, (28:2), pp. 203-236.

Harris, E. A. 2014. "Data Breach Hurts Profit at Target," 2014/02/26/ (available at http://www.nytimes.com/2014/02/27/business/target-reports-on-fourth-quarter-earnings.html). Accessed 2014/02/28/22:16:10.

Hibbeln, M., Jenkins, J. L., Schneider, C., Valacich, J. S., and Weinmann, M. 2016. "How Is Your User Feeling? Inferring Emotion through Human-Computer Interaction Devices," *MIS Quarterly*, (forthcoming).

Hitlin, P. 2016. "Research in the Crowdsourcing Age, a Case Study," *Pew Research Center: Internet, Science & Tech* (available at http://www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/). Accessed 4/21/2017.

Hofmann, S. G., Ellard, K. K., and Siegle, G. J. 2012. "Neurobiological Correlates of Cognitions in Fear and Anxiety: A Cognitive-Neurobiological Information-Processing Model," *Cognition and Emotion*, (26:2), pp. 282-299.

Jenkins, J. L., Anderson, B. B., Vance, A., Kirwan, C. B., and Eargle, D. 2016. "More Harm Than Good? How Messages That Interrupt Can Make Us Vulnerable," *Information Systems Research*, (27:4), pp. 880-896.

Jensen, A. R. 2006. *Clocking the Mind: Mental Chronometer Individual Differences*, Elsevier: Amsterdam, Netherlands.

Johnston, A., Warkentin, M., and Siponen, M. 2015. "An Enhanced Fear Appeal Rhetorical Framework: Leveraging Threats to the Human Asset through Sanctioning Rhetoric," *MIS Quarterly*, (39:1), pp. 113-134.

Johnston, A. C., and Warkentin, M. 2010. "Fear Appeals and Information Security Behaviors: An Empirical Study," *MIS Quarterly*, (34:3), pp. 549-566.

Kahneman, D., and Tversky, A. 1979. "Prospect Theory: An Analysis of Decision under Risk," *Econometrica: Journal of the Econometric Society*), pp. 263-291.

Kock, N. 2004. "The Psychobiological Model: Towards a New Theory of Computer-Mediated Communication Based on Darwinian Evolution," *Organization Science*, (15:3), pp. 327-348.

Kock, N. 2009. "Information Systems Theorizing Based on Evolutionary Psychology: An Interdisciplinary Review and Theory Integration Framework," *MIS Quarterly*, (33:2), pp. 395-418.

Kock, N., Chatelain-Jardon, R., and Carmona, J. 2008. "An Experimental Study of Simulated Web-Based Threats and Their Impact on Knowledge Communication Effectiveness," *Professional Communication, IEEE Transactions on*, (51:2), pp. 183-197.

Krebs, B. 2016. "Ransomware Getting More Targeted, Expensive," *Krebs on Security* (available at https://krebsonsecurity.com/2016/09/ransomware-getting-more-targeted-expensive/). Accessed 4/21/2017.

Lowry, P. B., and Moody, G. D. 2015. "Proposing the Control-Reactance Compliance Model (CRCM) to Explain Opposing Motivations to Comply with Organizational Information Security Policies," *Information Systems Journal*, (25:5), pp. 433-463.

Lundqvist, D., Esteves, F., and Öhman, A. 2004. "The Face of Wrath: The Role of Features and Configurations in Conveying Social Threat," *Cognition and emotion*, (18:2), pp. 161-182.

Malhotra, D., Loewenstein, G., and O'Donoghue, T. 2002. "Time Discounting and Time Preference: A Critical Review," *Journal of Economic Literature*, (40:2), pp. 351-401.

McDonnell, J. V., Martin, J. B., Markant, D. B., Coenen, A., Rich, A. S., and Gureckis, T. M. 2012. "Psiturk (Version 1.02) [Software]," New York University (available at https://github.com/NYUCCL/psiTurk).

Morales, A. C., Wu, E. C., and Fitzsimons, G. J. 2012. "How Disgust Enhances the Effectiveness of Fear Appeals," *Journal of Marketing Research*, (49:3), pp. 383-393.

Motiee, S., Hawkey, K., and Beznosov, K. 2010. "Do Windows Users Follow the Principle of Least Privilege?: Investigating User Account Control Practices," In: Proceedings of the Sixth Symposium on Usable Privacy and Security, ACM, New York, NY, USA, pp. 1:1-1:13.

Mourad, H. 2015. "Sleeping Your Way out of the Sandbox," SANS Institute (available at https://www.sans.org/reading-room/whitepapers/malicious/sleeping-sandbox-35797). Accessed 11/10/2016.

Nomura, M., Ohira, H., Haneda, K., Iidaka, T., Sadato, N., Okada, T., and Yonekura, Y. 2004. "Functional Association of the Amygdala and Ventral Prefrontal Cortex During Cognitive Evaluation of Facial Expressions Primed by Masked Angry Faces: An Event-Related fMRI Study," *NeuroImage*, (21:1), pp. 352-363.

Öhman, A., Lundqvist, D., and Esteves, F. 2001. "The Face in the Crowd Revisited: A Threat Advantage with Schematic Stimuli," *Journal of Personality and Social Psychology*, (80:3), pp. 381-396.

Ohman, A., and Soares, J. J. 1994. ""Unconscious Anxiety": Phobic Responses to Masked Stimuli," *Journal of Abnormal Psychology*, (103:2), pp. 231-240.

Osman, A., Barrios, F. X., Osman, J. R., Schneekloth, R., and Troutman, J. A. 1994. "The Pain Anxiety Symptoms Scale: Psychometric Properties in a Community Sample," *Journal of Behavioral Medicine*, (17:5), pp. 511-522.

Phillips, M. L., Williams, L. M., Heining, M., Herba, C. M., Russell, T., Andrew, C., Bullmore, E. T., Brammer, M. J., Williams, S. C. R., Morgan, M., and others 2004. "Differential Neural Responses to Overt and Covert Presentations of Facial Expressions of Fear and Disgust," *Neuroimage*, (21:4), pp. 1484-1496.

Rankin, C. H., Abrams, T., Barry, R. J., Bhatnagar, S., Clayton, D. F., Colombo, J., Coppola, G., Geyer, M. A., Glanzman, D. L., Marsland, S., McSweeney, F. K., Wilson, D. A., Wu, C.-F., and Thompson, R. F. 2009. "Habituation Revisited: An Updated and Revised Description of the Behavioral Characteristics of Habituation," *Neurobiology of Learning and Memory*, (92:2), pp. 135-138.

Reed, T. 2016. "First Mac Ransomware Spotted," Malwarebytes Labs (available at https://blog.malwarebytes.com/cybercrime/2016/03/first-mac-ransomware-spotted/).

Riedl, R., Mohr, P. N. C., Kenning, P. H., Davis, F. D., and Heekeren, H. R. 2014. "Trusting Humans and Avatars: A Brain Imaging Study Based on Evolution Theory," *Journal of Management Information Systems*, (30:4), pp. 83-114.

Rogers, R. W. 1983. "Cognitive and Physiological Processes in Fear Appeals and Attitude Change: A Revised Theory of Protection Motivation," in *Social Psychophysiology: A Sourcebook,* J. T. Cacioppo and R. E. Petty (eds.), Guilford: New York, pp. 153-176.

Rozin, P., and Fallon, A. E. 1987. "A Perspective on Disgust," *Psychological Review*, (94:1), pp. 23-41.

Schechter, S. E., Dhamija, R., Ozment, A., and Fischer, I. 2007. "The Emperor's New Security Indicators," In: Security and Privacy, 2007. SP'07. IEEE Symposium on, IEEE, Berkeley, CA, pp. 51-65.

Schneier, B. 2004. *Secrets and Lies: Digital Security in a Networked World*, Wiley: Hoboken, NJ.

Schneier, B. 2011. "Schneier on Security: Details of the RSA Hack," *Schneier on Security* (available at https://www.schneier.com/blog/archives/2011/08/details_of_the.html).

Sharek, D., Swofford, C., and Wogalter, M. 2008. "Failure to Recognize Fake Internet Popup Warning Messages," In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Sage Publications, New York, New York, pp. 557-560.

Siponen, M., and Vance, A. 2010. "Neutralization: New Insights into the Problem of Employee Information Systems Security Policy Violations," *MIS Quarterly*, (34:3), pp. 487-502.

Sotirakopoulos, A., Hawkey, K., and Beznosov, K. 2011. "On the Challenges in Usable Security Lab Studies: Lessons Learned from Replicating a Study on SSL Warnings." Paper presented at the Proceedings of the Seventh Symposium on Usable Privacy and Security (SOUPS), Menlo Park, CA.

Squires, N. K., Squires, K. C., and Hillyard, S. A. 1975. "Two Varieties of Long-Latency Positive Waves Evoked by Unpredictable Auditory Stimuli in Man," *Electroencephalography and Clinical Neurophysiology*, (38:4), pp. 387-401.

Steelman, Z. R., Hammer, B. I., and Limayem, M. 2014. "Data Collection in the Digital Age: Innovative Alternatives to Student Samples," *MIS Quarterly*, (38:2), pp. 355-378.

Straub, D., Boudreau, M.-C., and Gefen, D. 2004. "Validation Guidelines for IS Positivist Research," *Communications of the Association for Information Systems*, (13:24), pp. 380-427.

Sunshine, J., Egelman, S., Almuhimedi, H., Atri, N., and Cranor, L. F. 2009. "Crying Wolf: An Empirical Study of SSL Warning Effectiveness," In: Proceedings of the 18th conference on USENIX security symposium, Montreal, Canada, pp. 399-416.

Tom, S. 2016. "Fbi Quietly Admits to Multi-Year Apt Attack, Sensitive Data Stolen," Threatpost (available at https://threatpost.com/fbi-quietly-admits-to-multi-year-apt-attack-sensitive-data-stolen/117267/). Accessed 11/10/2016.

Trumbo, C. W. 2002. "Information Processing and Risk Perception: An Adaptation of the Heuristic-Systematic Model," *Journal of Communication*, (52:2), pp. 367-382.

van Hooff, J. C., Devue, C., Vieweg, P. E., and Theeuwes, J. 2013. "Disgust- and Not Fear-Evoking Images Hold Our Attention," *Acta psychologica*, (143:1), pp. 1-6.

Vance, A., Anderson, B. B., Kirwan, C. B., and Eargle, D. 2014. "Using Measures of Risk Perception to Predict Information Security Behavior: Insights from Electroencephalography (EEG)," *Journal of the Association for Information Systems*, (15:10), pp. 679-722.

West, R. 2008. "The Psychology of Security," *Communications of the ACM*, (51:4), pp. 34-40.

Willison, R., and Warkentin, M. 2013. "Beyond Deterrence: An Expanded View of Employee Computer Abuse," *MIS Quarterly*, (37:1), pp. 1-20.

Wogalter, M. S. 2006a. "Communication-Human Information Processing (C-HIP) Model," in *Handbook of Warnings,* M. S. Wogalter (ed.), Lawrence Erlbaum Associates: Mahwah, NJ, pp. 51-61.

Wogalter, M. S. 2006b. "Purposes and Scope of Warnings," in *Handbook of Warnings,* M. S. Wogalter (ed.), Lawrence Erlbaum Associates: Mahwah, NJ, pp. 3-9.

Wu, M., Miller, R. C., and Garfinkel, S. L. 2006. "Do Security Toolbars Actually Prevent Phishing Attacks?," In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, Montreal, Quebec, Canada, pp. 601-610.

APPENDIX A

FMRI TECHNICAL DETAILS FOR STUDY 2

A.1 EQUIPMENT

MRI scanning took place at a university MRI research facility with the use of a Siemens 3T Tim-Trio scanner. For each scanned participant, we collected a high-resolution structural MRI scan for functional localization in addition to a series of functional scans to track brain activity during the performance of the various tasks. Structural images were acquired with a T1-weighted magnetization-prepared rapid acquisition including a gradient-echo (MP-RAGE) sequence with the following parameters: TE = 2.26 ms, flip angle = 9°, slices = 176, slice thickness = 1.0 mm, matrix size = 256 × 215, and voxel size = 1 mm × 0.98 mm × 0.98 mm. Functional scans were acquired with a gradient-echo, echo-planar, T2*-weighted pulse sequence with the following parameters: TR = 2000 ms, TE = 28 ms, flip angle = 90°, slices = 40, slice thickness = 3.0 mm (no skip), matrix size = 64 × 64, and voxel size = 3.44 mm × 3.44 mm × 3 mm.

A.2 PROTOCOL

Participants were given a verbal briefing about the MRI procedures and the task, and were then situated supine in the scanner. Visual stimuli were viewed using a mirror attached to the head coil reflecting a large monitor outside the scanner that was configured to display images in reverse so that they appeared normal when viewed through the mirror. We first performed a 10-second localizer scan, followed by a 7-minute structural scan. Following these, we started the experimental task (see Figure A1). We used E-Prime software to display the stimuli and synchronize the display events and scanner software. Total time in the scanner was 55 minutes.

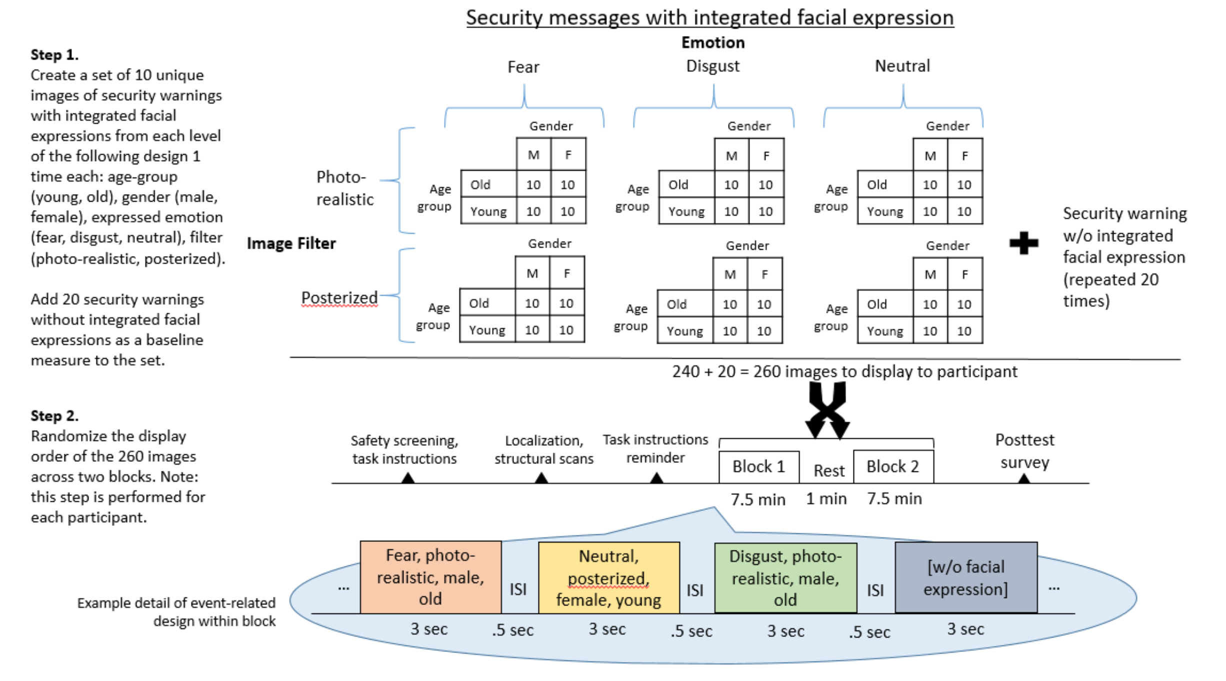

Security messages with integrated facial expression
Step 1.
Create a set of 10 unique images of security warnings with integrated facial expressions from each level of the following design 1 time each: age-group (young, old), gender (male, female), expressed emotion (fear, disgust, neutral), filter (photo-realistic, posterized).
Add 20 security warnings without integrated facial expressions as a baseline measure to the set.
Emotion
Fear
Disgust
Neutral
Image Filter
Photo-realistic
Posterized
Gender
M
F
Age group
Old
Young
10
Security warning w/o integrated facial expression (repeated 20 times)
240 + 20 = 260 images to display to participant
Step 2.
Randomize the display order of the 260 images across two blocks. Note: this step is performed for each participant.
Safety screening, task instructions
Localization, structural scans
Task instructions reminder
Block 1
Rest
Block 2
Posttest survey
7.5 min
1 min
7.5 min
Example detail of event-related design within block
...
Fear, photo-realistic, male, old
ISI
Neutral, posterized, female, young
Disgust, photo-realistic, male, old
[w/o facial expression]
3 sec
.5 sec

Figure A-1 – Graphical representation of fMRI experimental procedure for Study 2. ISI = Interstimulus interval

A.3 ANALYSIS

MRI data were analyzed with the Analysis of Functional Images (AFNI) suite of programs (Cox 1996). Briefly, functional data were slice-time corrected to account for differences in acquisition time for different slices of each volume; then, each volume was registered with the middle volume of each run to account for low-frequency motion. Data from each run were aligned to the run nearest in time to the acquisition of the structural scan. The structural scan was then co-registered to the functional scans. As in previous studies (e.g., Motley and Kirwan 2012), spatial normalization was accomplished by first warping the structural scan to the Talairach atlas (Talairach and Tournoux 1988) followed by warping to a template brain with Advanced Neuroimaging Tools (ANTs). The ANTs transformation resampled all voxel dimensions to 3mm x 3mm x 3mm. For the single-subject ("first level") analyses, we performed multiple regression of the form $y=\beta_0+\beta_1x_1+\beta_2x_2\ldots\beta_nx_n+\varepsilon$, where "y" is the observed fMRI timecourse for each voxel and each "x" term is a vector regressor representing either conditions of interest (e.g., stimulus type or repetition number) or a nuisance regressor (e.g., motion or scanner drift). We created separate behavioral regressors coding for each cell in the fully-crossed factorial design: emotion (disgust, fear, neutral), filter (photo-realistic, posterized), gender (young, old), and age (young, old), giving us a total of 24 task regressors, in addition to nuisance variables coding for motion (3 rotations and 3 directions of translation) and scanner drift (4 polynomial regressors for both of the scan runs). The security malware image without integrated facial expressions, shown 20 times, served as the implicit baseline for the regression analysis (see "Study 2 – Design"). Stimulus events were modeled using a stick function convolved with the canonical hemodynamic response. Resulting parameter estimates (beta values) were blurred with a 5-mm FWHM Gaussian kernel. Parameter estimates for the conditions of interest were then entered into group-level analyses, such as ANOVAs or *t*-tests (see "Study 2 – Analysis and Results " for detailed descriptions of group-level analyses), which were used to determine functional regions of interest (ROIs). Once functional ROIs were identified, we extracted mean parameter estimates within each ROI for further investigation in order to characterize the direction and strength of interactions and other effects. All whole-brain voxel-wise tests were corrected for multiple comparisons using a spatial extent threshold of 24 contiguous voxels (648 mm^3).

Two linear mixed model regression analyses were conducted on the extracted betas. In the first, our face-integrated warnings were grouped by the displayed emotion. In the second, the face-integrated warnings were grouped by image filter.

APPENDIX B

EXPLORATION OF WARNING ADHERENCE TRENDS

Because our field study task employed repeated measures on a binary dependent variable (whether or not the warning was dismissed), as a preliminary analysis we explored the ratio of ignored warnings for each participant and for each treatment group. We first calculated the ratio of warnings ignored for each participant, and then we set a dummy variable marking whether a participant was perfectly consistent in their warning behavior (whether their perfect consistency was to always ignore or always heed). An ANOVA did not find any significant differences for the ratio ignored averaged across exposures among Study 1's emotion treatment groups (F(3,499)=1.641, p=.179) or among filter treatment groups (F(2,500)=1.058, p=.348), or among Study 2's penalty treatment groups (F(3,499)=1.641, p=.179), so only the overall differences are reported in Figure B-1 and Table B-1. We note that, for 503 participants, many participants either ignored all of them (*n*=193), or heeded all of them (*n*=230), totaling 423 participants and leaving 80 who showed differential behavior. Of the remaining 80, 37 ignored 3 out of 4 warnings, 20 ignored 2 out of 4 warnings, and 15 ignored 1 out of 5 warnings. We cannot say whether the participants who showed the polarized behavior (all or nothing) always behave this way when they encounter security warnings, or if their behavior was particular to our experimental manipulations. But the lack of significance among treatment groups for the overall rate of warnings ignored may suggest that these trends reflect participants' everyday behavior, regardless of the warning design. Interesting is the observation that differences were observed in the predicted direction among treatment groups when only considering the first warning impression. This finding highlights the quick habituation rates that likely occurred because of the number of repeated measures over such a short time frame, and emphasizes the need for improved experimental designs that can better test habituation rates to security warnings with more realistic gaps between warning impressions (such as a one week gap).

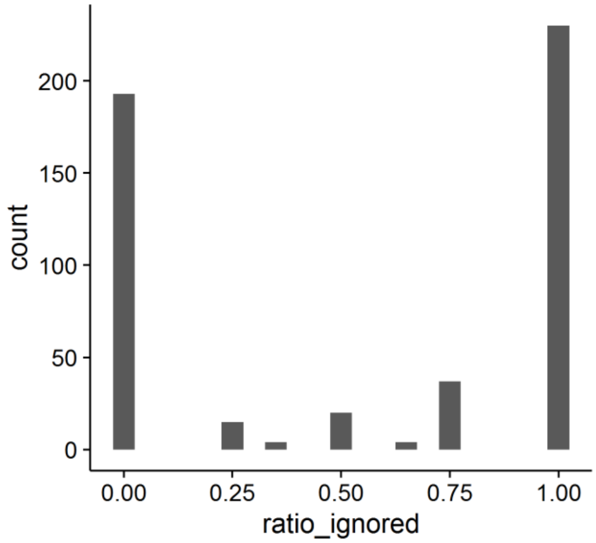


Figure B-1. Overall trends of ratio of warnings ignored.

Table B-1. Overall trends of ratio of warnings ignored.

Ratio ignored	***n***
0	193
0.25	15
0.333333	4
0.5	20
0.666667	4
0.75	37
1	230
Total	**503**

APPENDIX C

SUPPLEMENTARY FIGURES

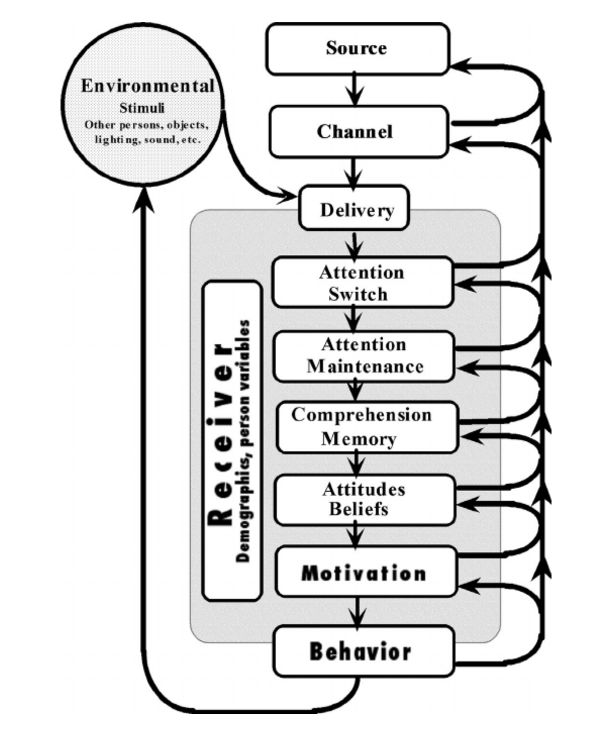
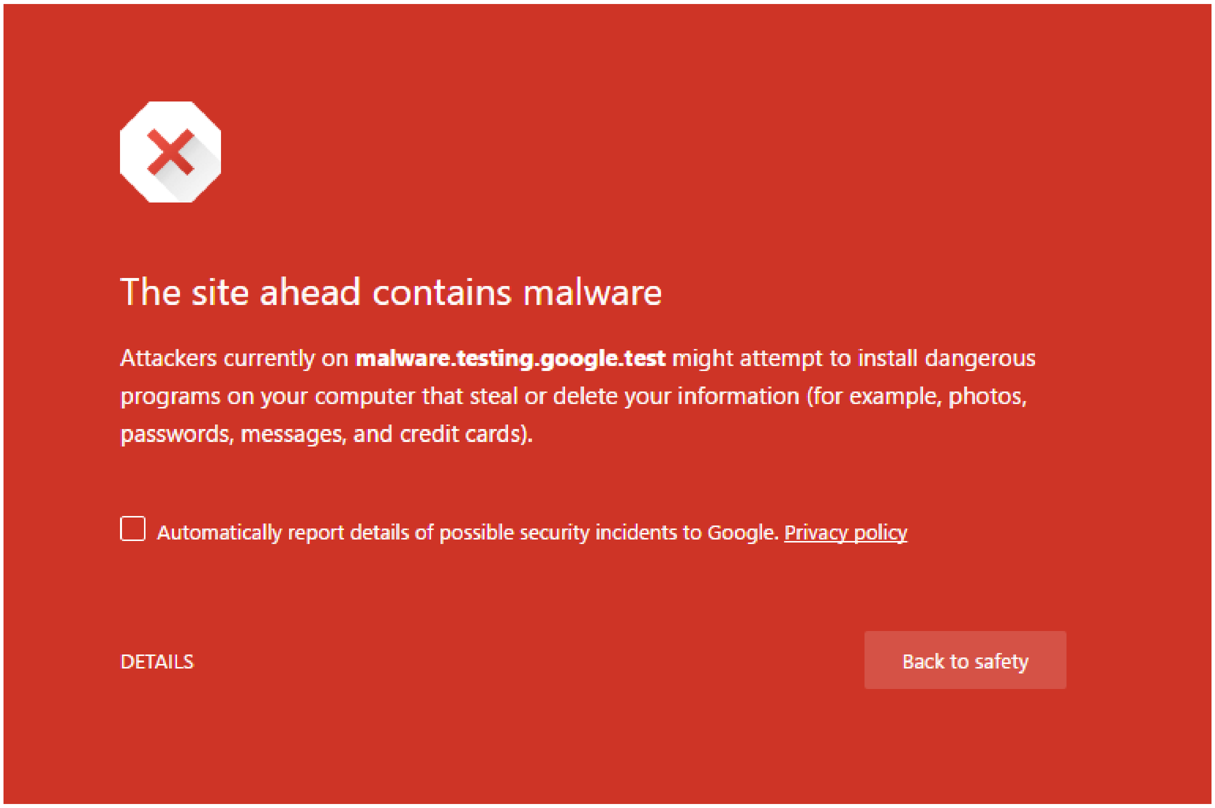


Figure C-1. Communication-Human Information Processing Model (adapted from Wogalter 2006a)



The site ahead contains malware

Attackers currently on **malware.testing.google.test** might attempt to install dangerous programs on your computer that steal or delete your information (for example, photos, passwords, messages, and credit cards).

☐ Automatically report details of possible security incidents to Google. Privacy policy

Figure C-2. Google Chrome browser malware warning, build 51.0.2704.63 m

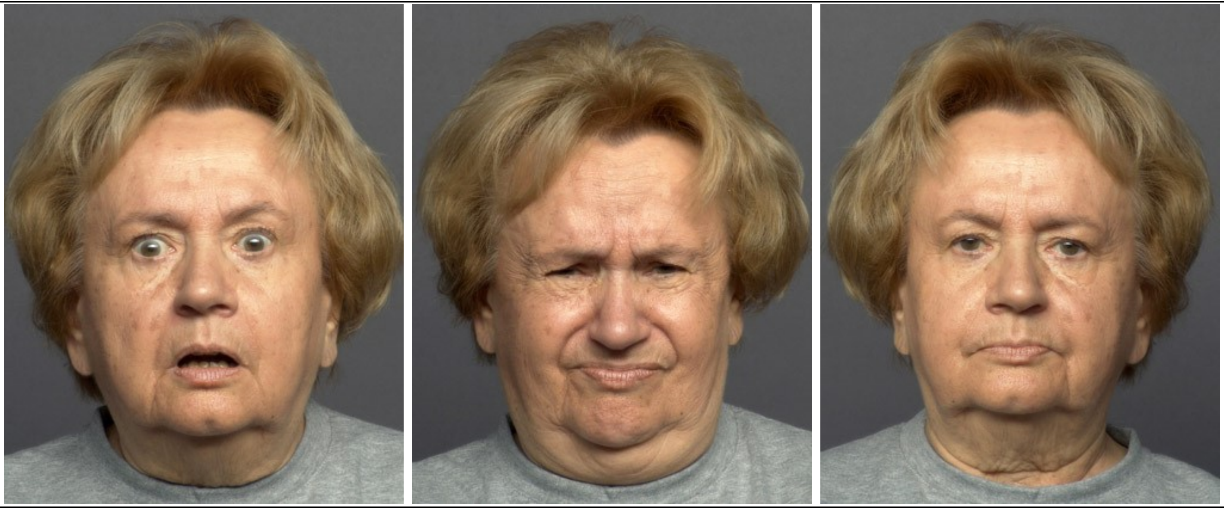
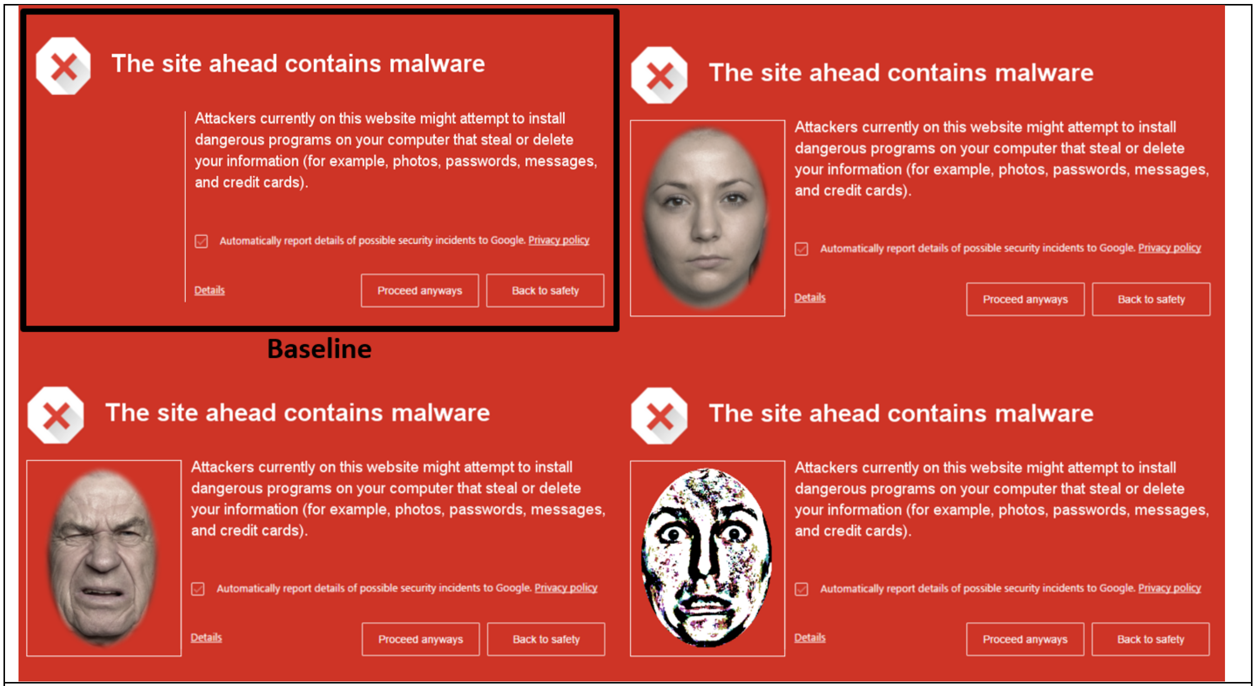


Figure C-3. Examples of validated emotive facial expressions from FACES database – 182 actors (Ebner, Riediger & Lindenberger, 2010). From this set, we randomly selection 120 actor. Then we applied an oval crop, and we made a "desaturated" copy and a "posterized" copy



The site ahead contains malware

Attackers currently on this website might attempt to install dangerous programs on your computer that steal or delete your information (for example, photos, passwords, messages, and credit cards).

☑ Automatically report details of possible security incidents to Google. Privacy policy

The site ahead contains malware

Attackers currently on this website might attempt to install dangerous programs on your computer that steal or delete your information (for example, photos, passwords, messages, and credit cards).

☑ Automatically report details of possible security incidents to Google. Privacy policy

The site ahead contains malware

Attackers currently on this website might attempt to install dangerous programs on your computer that steal or delete your information (for example, photos, passwords, messages, and credit cards).

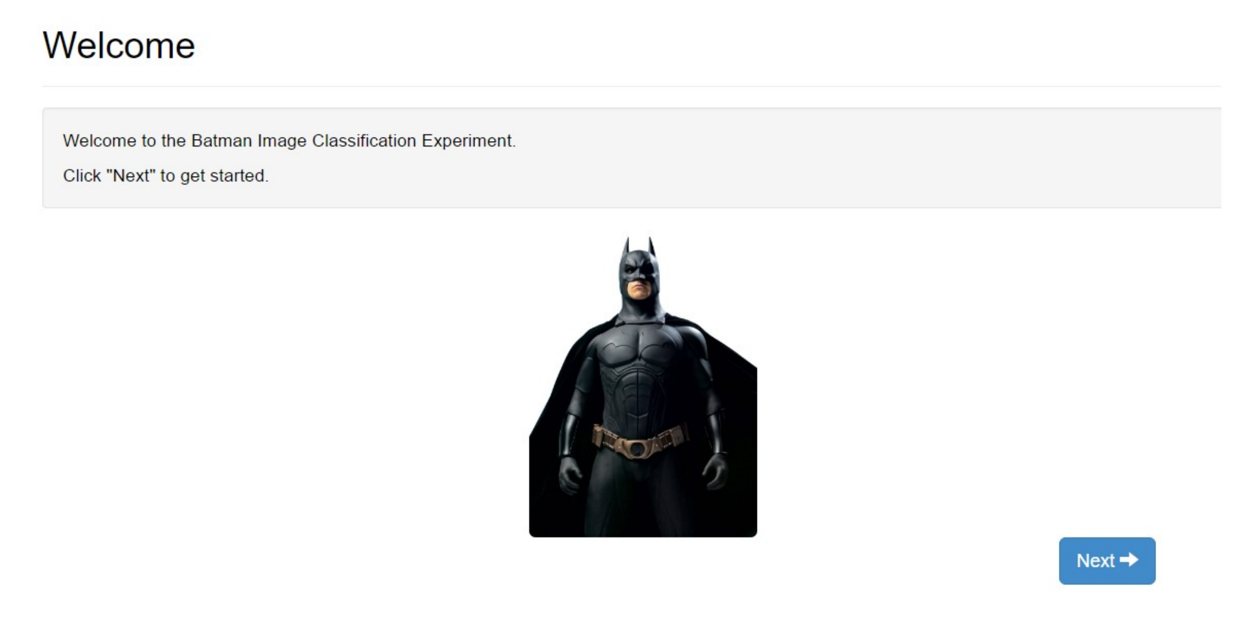
☑ Automatically report details of possible security incidents to Google. Privacy policy

The site ahead contains malware

Attackers currently on this website might attempt to install dangerous programs on your computer that steal or delete your information (for example, photos, passwords, messages, and credit cards).

☑ Automatically report details of possible security incidents to Google. Privacy policy

Figure C-4. We integrated the faces into our modification of the Google Chrome malware warning. From left to right, top to bottom, (a) blank template, (b) neutral-desaturated warning, (c) disgust-desaturated warning, (d) fear-posterized warning.

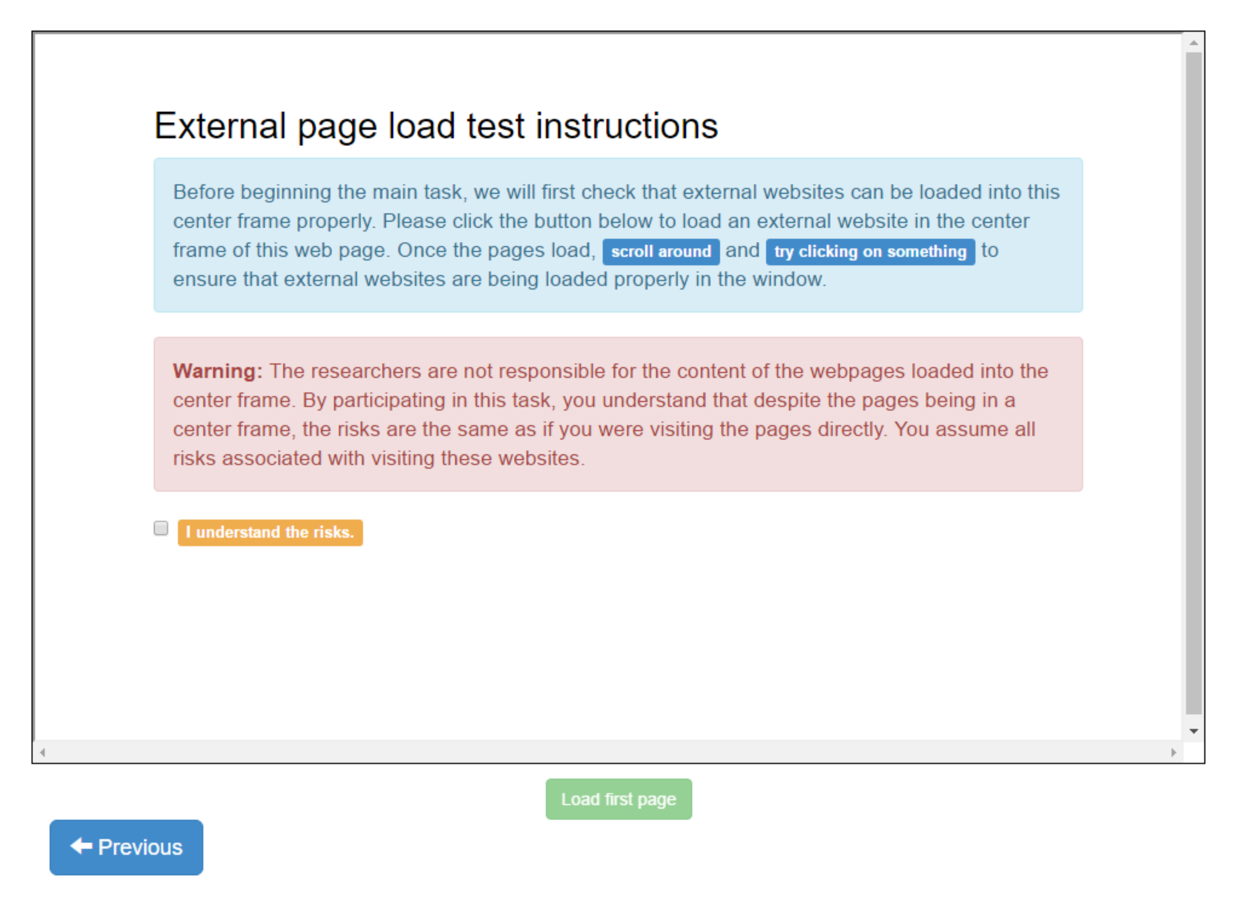


Welcome

Welcome to the Batman Image Classification Experiment.

Click "Next" to get started.

Figure C-5. Study 1 protocol – welcome screen



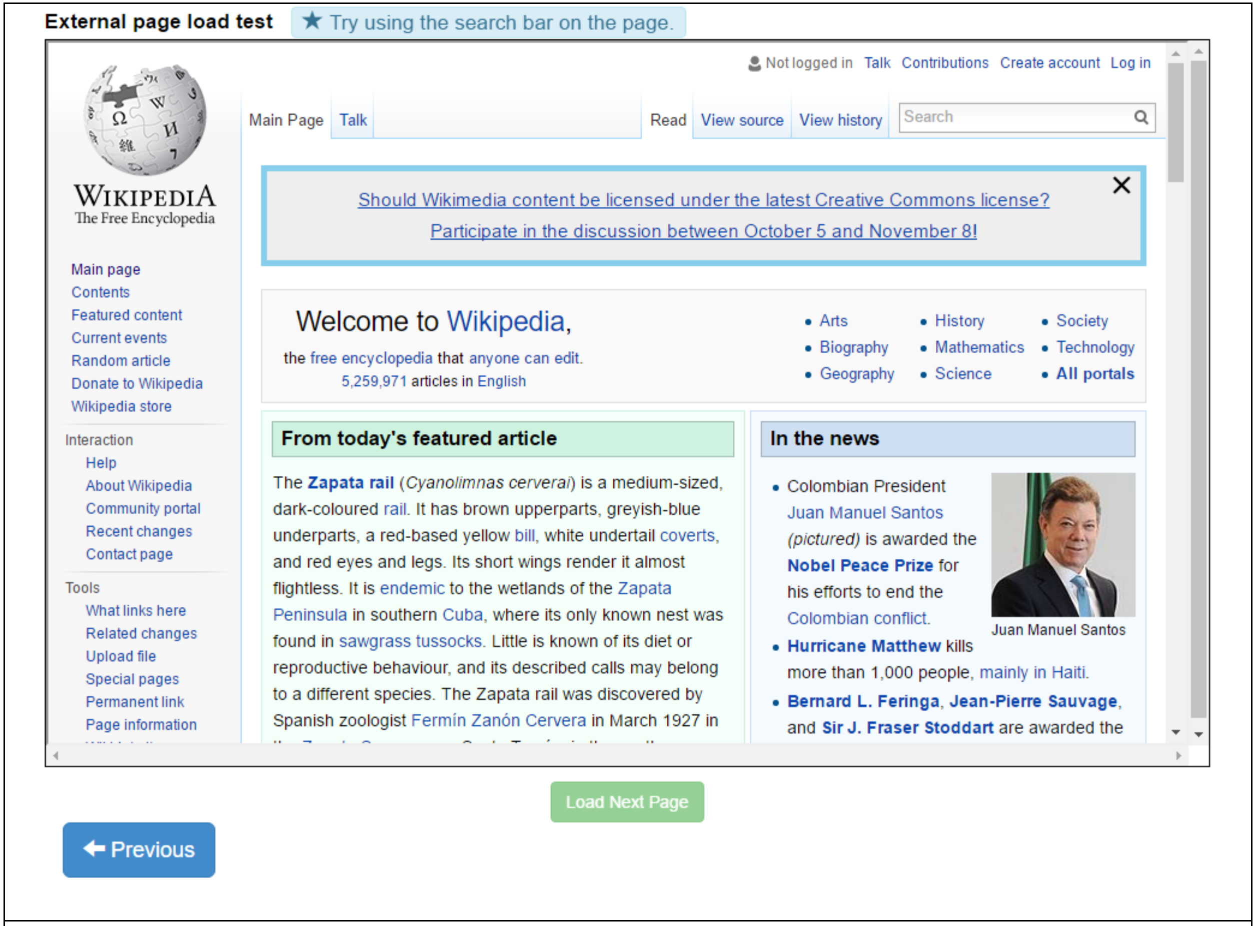
External page load test instructions

Before beginning the main task, we will first check that external websites can be loaded into this center frame properly. Please click the button below to load an external website in the center frame of this web page. Once the pages load, **scroll around** and **try clicking on something** to ensure that external websites are being loaded properly in the window.

Warning: The researchers are not responsible for the content of the webpages loaded into the center frame. By participating in this task, you understand that despite the pages being in a center frame, the risks are the same as if you were visiting the pages directly. You assume all risks associated with visiting these websites.

☐ **I understand the risks.**

Figure C-6. Study 1 protocol – external page load test instructions



Should Wikimedia content be licensed under the latest Creative Commons license?

Participate in the discussion between October 5 and November 8!

Welcome to Wikipedia,

the free encyclopedia that anyone can edit.

5,259,971 articles in English

From today's featured article

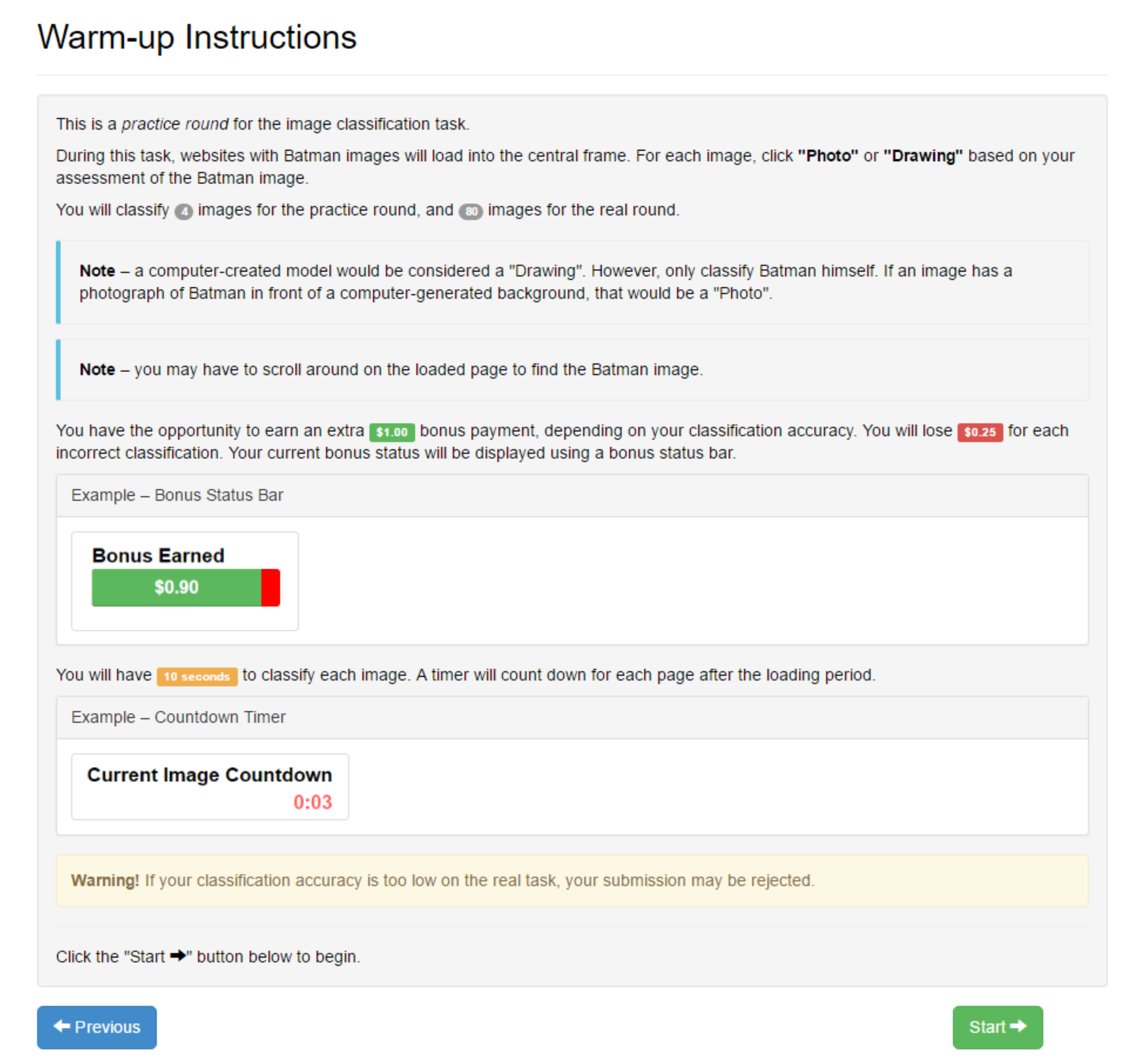
The **Zapata rail** (*Cyanolimnas cerverai*) is a medium-sized, dark-coloured rail. It has brown upperparts, greyish-blue underparts, a red-based yellow bill, white undertail coverts, and red eyes and legs. Its short wings render it almost flightless. It is endemic to the wetlands of the Zapata Peninsula in southern Cuba, where its only known nest was found in sawgrass tussocks. Little is known of its diet or reproductive behaviour, and its described calls may belong to a different species. The Zapata rail was discovered by Spanish zoologist Fermín Zanón Cervera in March 1927 in

In the news

- Colombian President Juan Manuel Santos *(pictured)* is awarded the **Nobel Peace Prize** for his efforts to end the Colombian conflict.
- **Hurricane Matthew** kills more than 1,000 people, mainly in Haiti.
- **Bernard L. Feringa**, **Jean-Pierre Sauvage**, and **Sir J. Fraser Stoddart** are awarded the

Juan Manuel Santos

Figure C-7. Study 1 protocol – external page load test



Warm-up Instructions

This is a *practice round* for the image classification task.

During this task, websites with Batman images will load into the central frame. For each image, click **"Photo"** or **"Drawing"** based on your assessment of the Batman image.

You will classify 4 images for the practice round, and 80 images for the real round.

> **Note** – a computer-created model would be considered a "Drawing". However, only classify Batman himself. If an image has a photograph of Batman in front of a computer-generated background, that would be a "Photo".

> **Note** – you may have to scroll around on the loaded page to find the Batman image.

You have the opportunity to earn an extra $1.00 bonus payment, depending on your classification accuracy. You will lose $0.25 for each incorrect classification. Your current bonus status will be displayed using a bonus status bar.

Example – Bonus Status Bar

Bonus Earned

$0.90

You will have 10 seconds to classify each image. A timer will count down for each page after the loading period.

Example – Countdown Timer

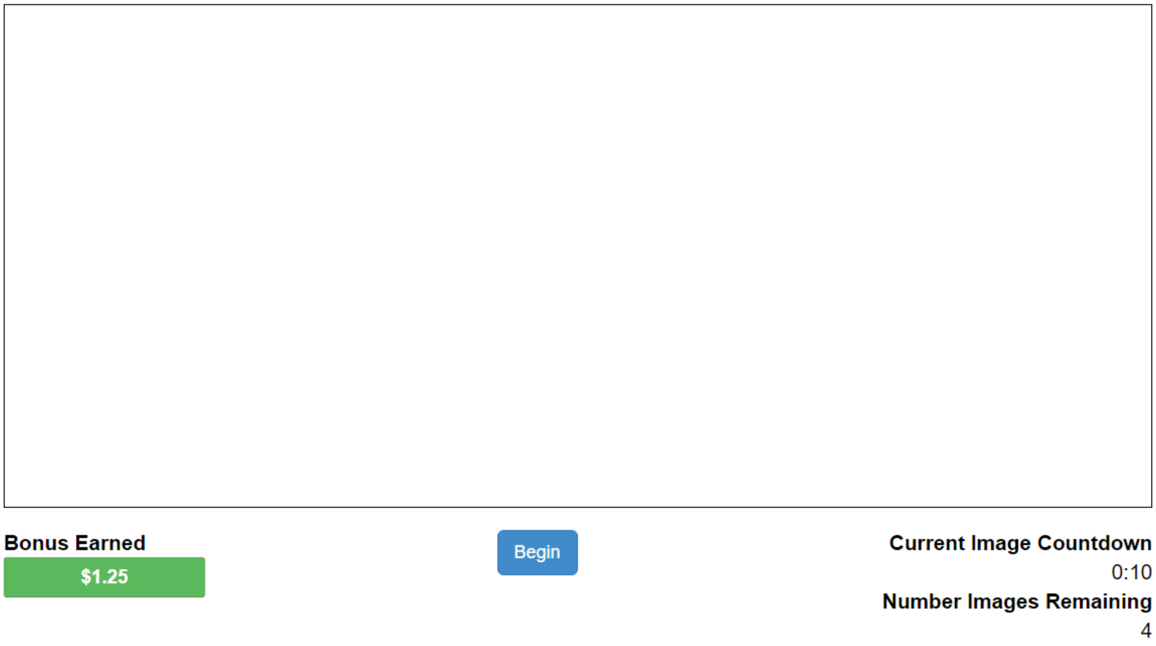
Current Image Countdown

0:03

Warning! If your classification accuracy is too low on the real task, your submission may be rejected.

Click the "Start ➡" button below to begin.

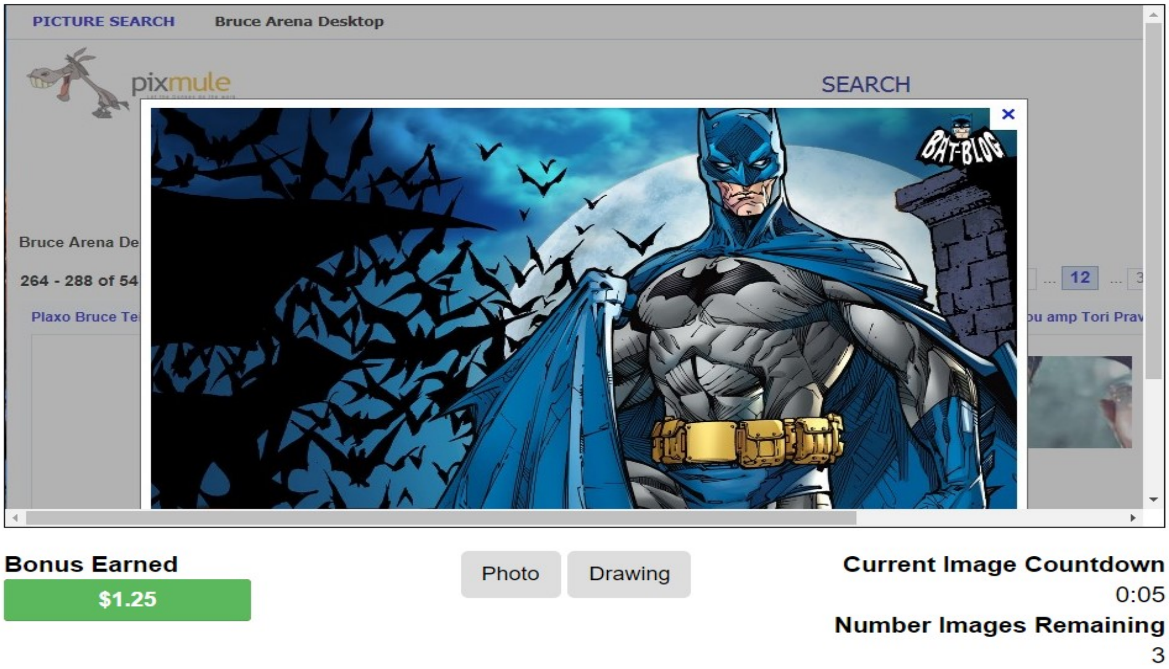
Figure C-8. Study 1 protocol – instructions



Bonus Earned
$1.25

Current Image Countdown
0:10
Number Images Remaining
4

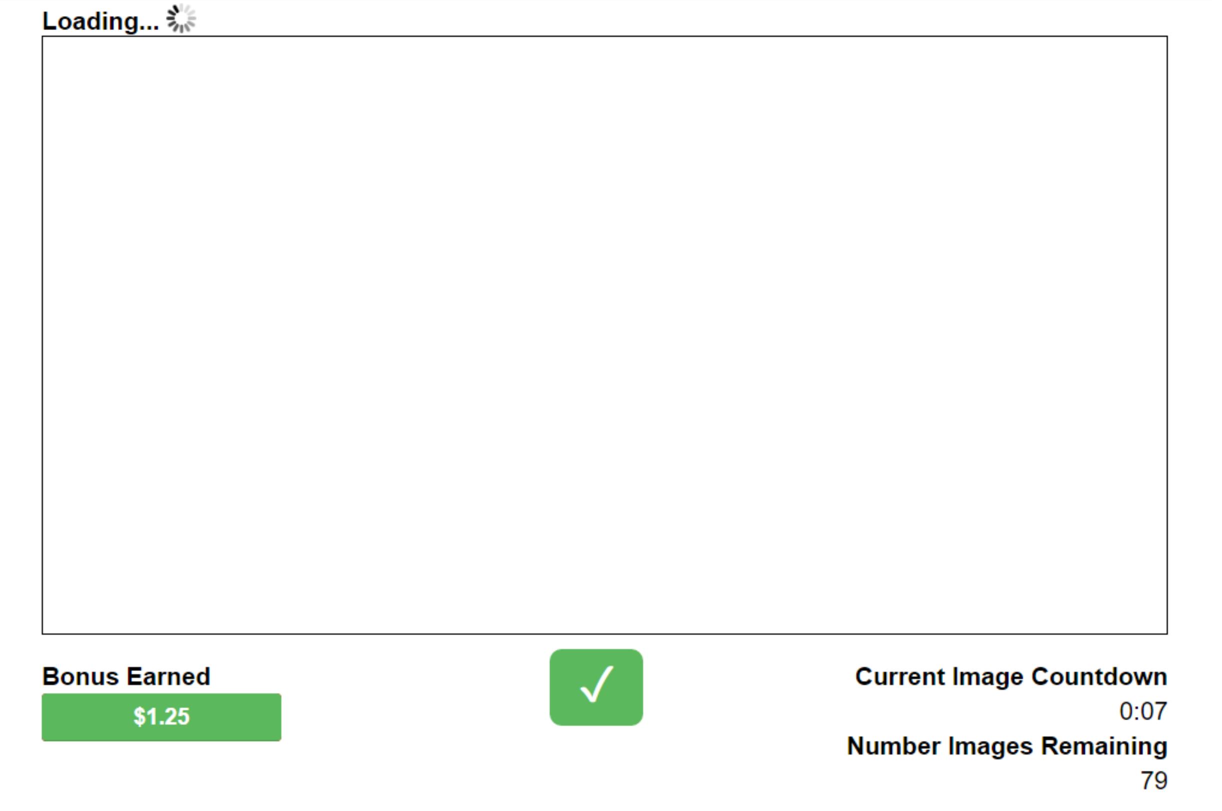
Figure C-9. Study 1 protocol – Main task, before beginning



Bonus Earned
$1.25

Current Image Countdown
0:05
Number Images Remaining
3

Figure C-10. Study 1 protocol – Main task, an image to classify

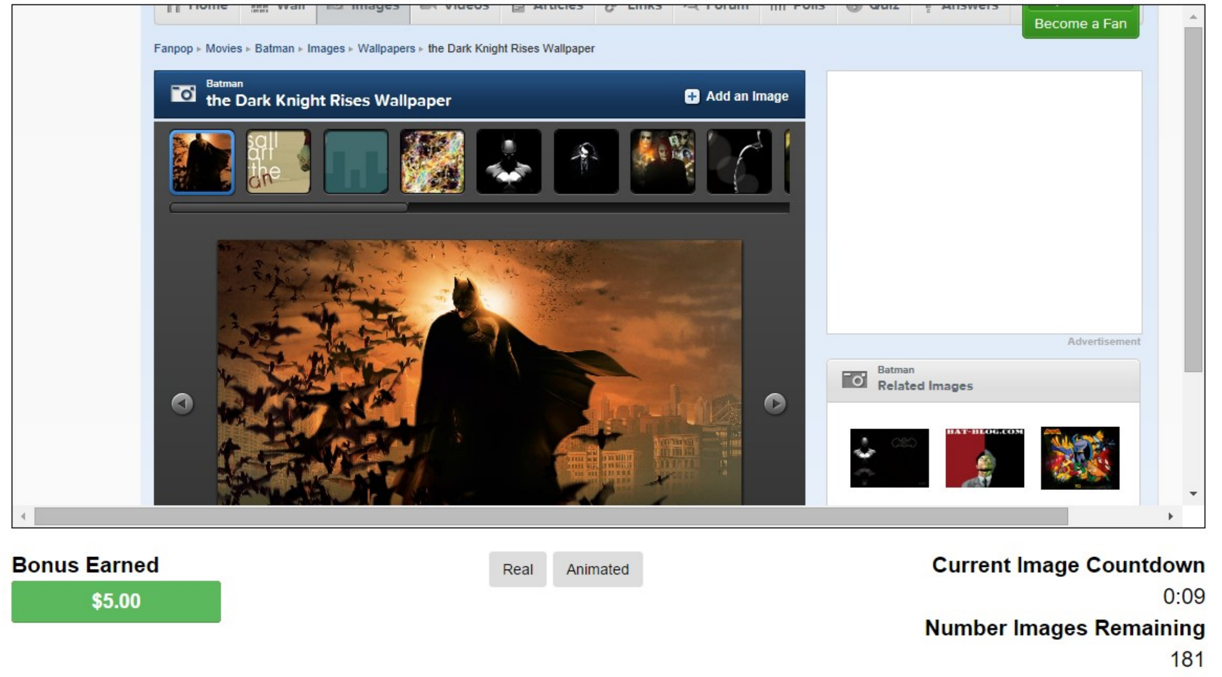


Bonus Earned
$1.25

Current Image Countdown
0:07

Number Images Remaining
79

Figure C-11. Study 1 protocol – Main task, feedback and loading icon

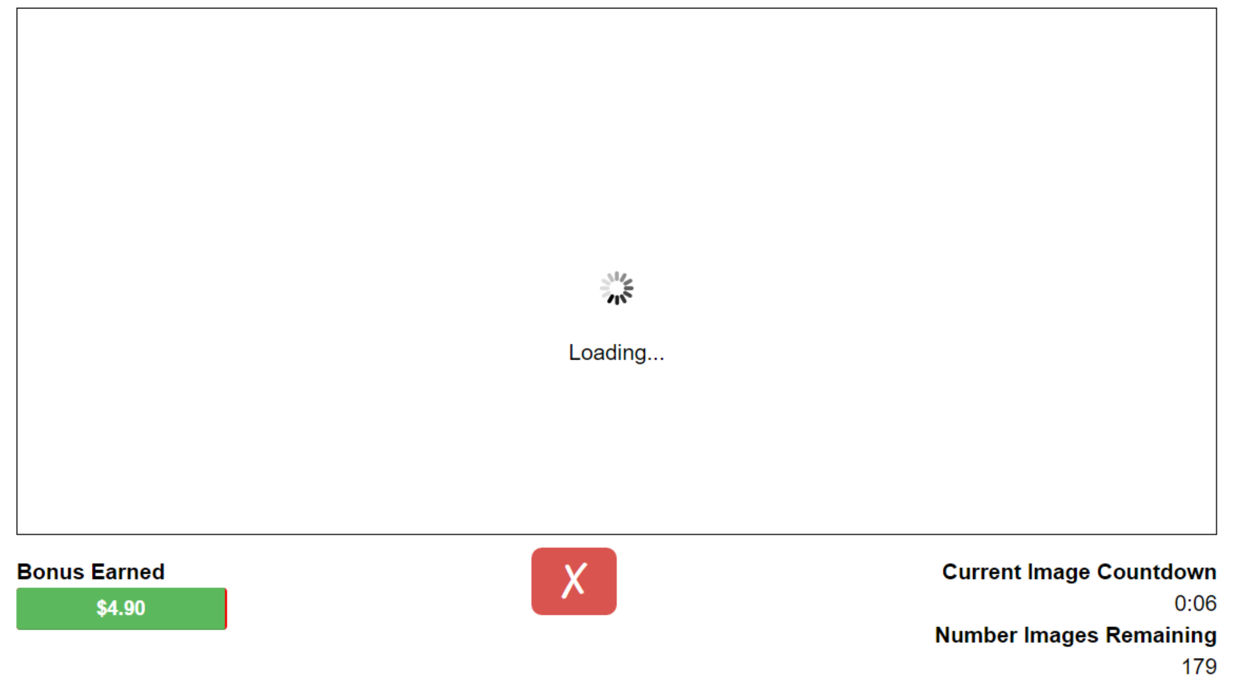

Become a Fan
Fanpop ▸ Movies ▸ Batman ▸ Images ▸ Wallpapers ▸ the Dark Knight Rises Wallpaper
Batman
the Dark Knight Rises Wallpaper
Add an Image
Advertisement
Batman
Related Images
Real
Animated

Bonus Earned
$5.00

Current Image Countdown
0:09

Number Images Remaining
181

Figure C-12. Study 1 protocol – Main task, a second image to classify



Bonus Earned

$4.90

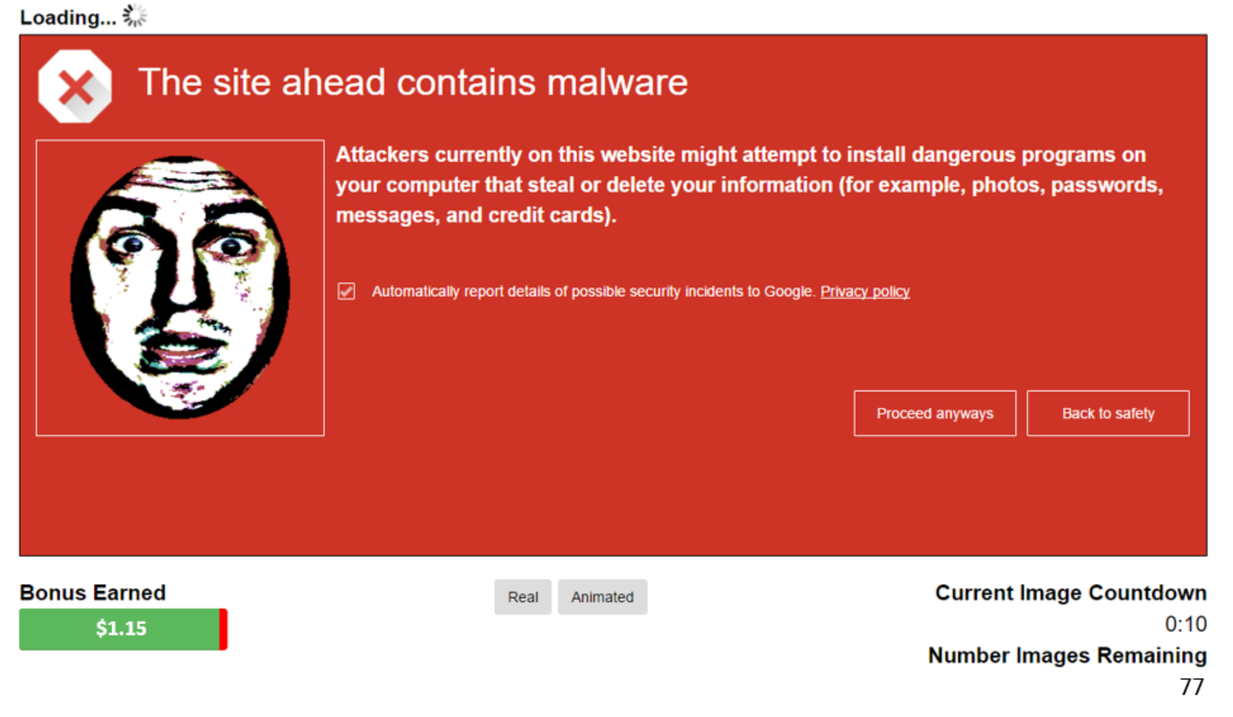
Current Image Countdown

0:06

Number Images Remaining

179

Figure C-13. Study 1 protocol – Main task, feedback for incorrect classification



The site ahead contains malware

Attackers currently on this website might attempt to install dangerous programs on your computer that steal or delete your information (for example, photos, passwords, messages, and credit cards).

☑ Automatically report details of possible security incidents to Google. Privacy policy

Bonus Earned

$1.15

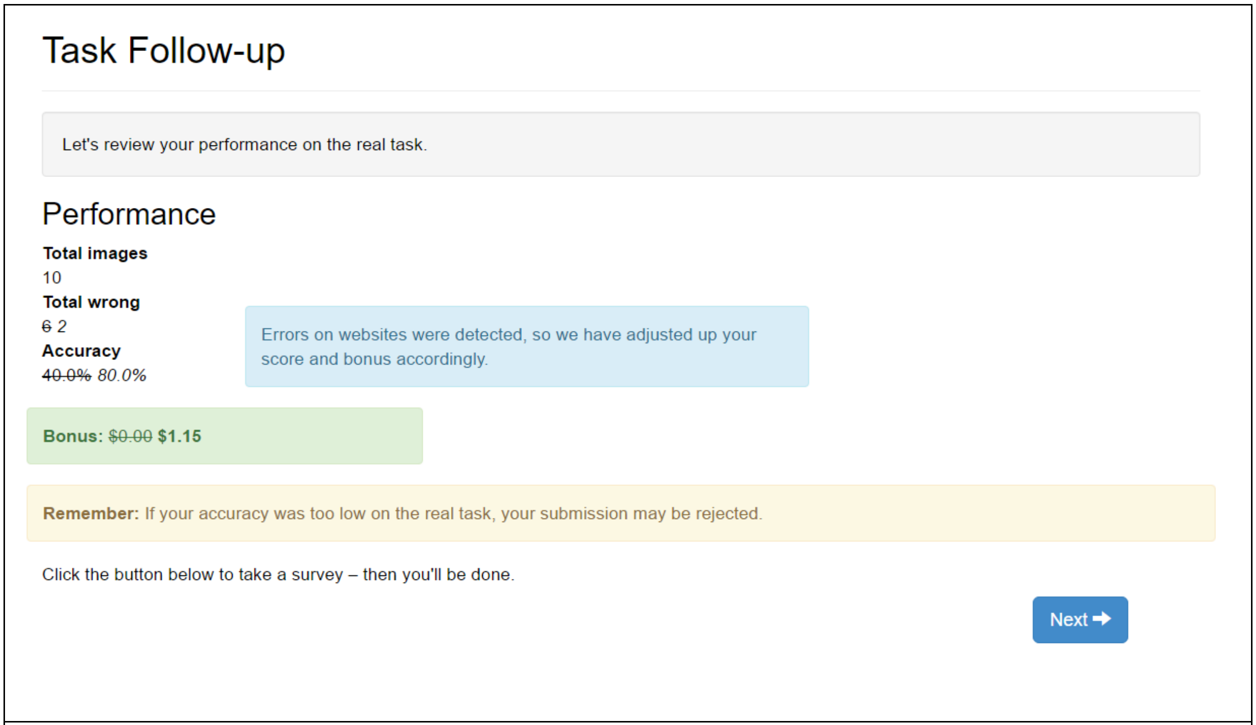
Current Image Countdown

0:10

Number Images Remaining

77

Figure C-14. Study 1 protocol – Main task, an interruptive warning



Task Follow-up

Let's review your performance on the real task.

Performance

Total images

10

Total wrong

~~6~~ *2*

Accuracy

~~40.0%~~ *80.0%*

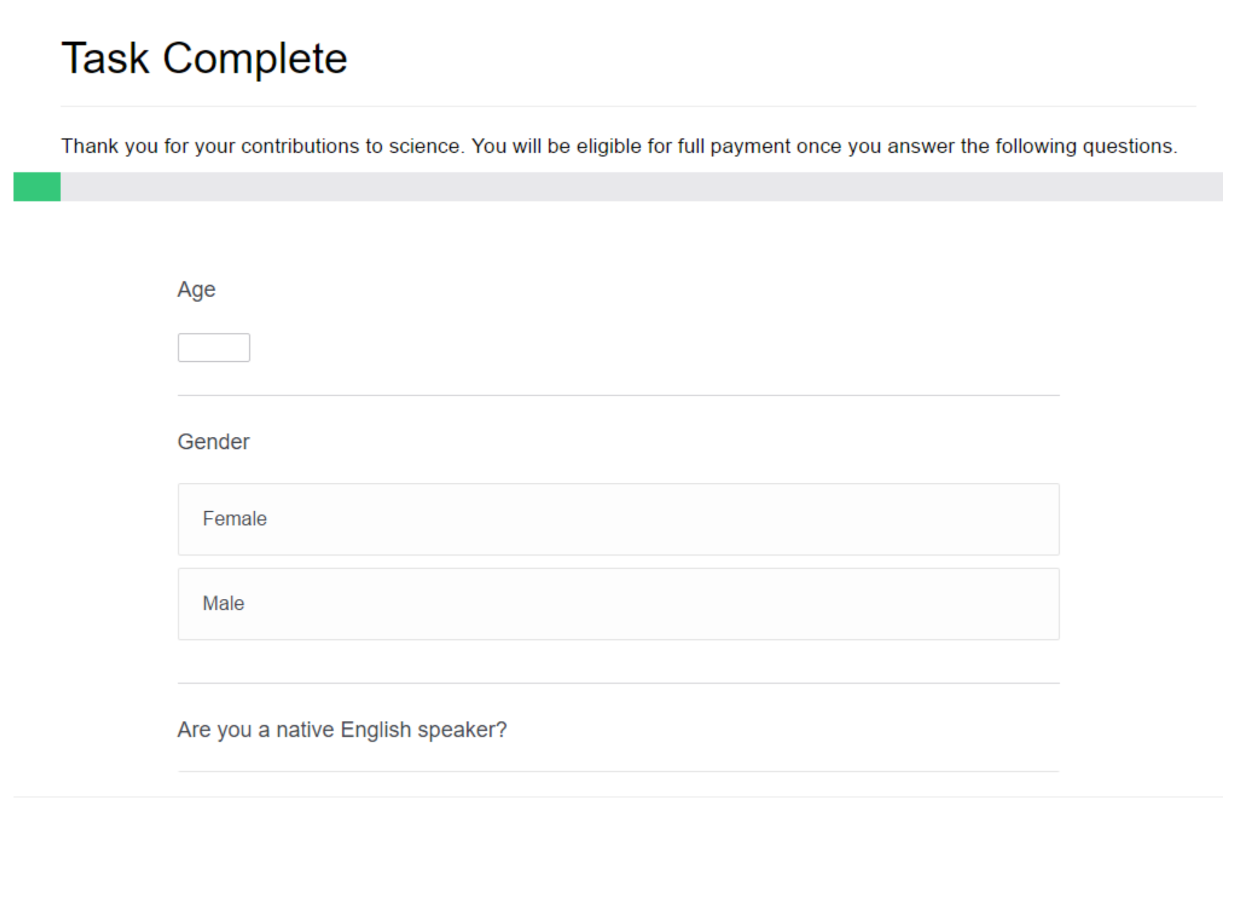
Errors on websites were detected, so we have adjusted up your score and bonus accordingly.

Bonus: ~~$0.00~~ **$1.15**

Remember: If your accuracy was too low on the real task, your submission may be rejected.

Click the button below to take a survey – then you'll be done.

Figure C-15. Study 1 protocol – Main task follow-up



Task Complete

Thank you for your contributions to science. You will be eligible for full payment once you answer the following questions.

Age

Gender

Female

Male

Are you a native English speaker?

Figure C-16. Study 1 protocol – Embedded Qualtrics survey

Cox, R. W. 1996. "AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages," *Computers and Biomedical Research*, (29:3), pp. 162-173.

Motley, S. E., and Kirwan, C. B. 2012. "A Parametric Investigation of Pattern Separation Processes in the Medial Temporal Lobe," *The Journal of Neuroscience*, (32:38), pp. 13076-13084.

Talairach, J., and Tournoux, P. 1988. *Co-Planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System: An Approach to Cerebral Imaging*, Thieme: Stuttgart.

Wogalter, M. S. 2006a. "Communication-Human Information Processing (C-HIP) Model," in *Handbook of Warnings,* M. S. Wogalter (ed.), Lawrence Erlbaum Associates: Mahwah, NJ, pp. 51-61.