Gaming Tasks as a Method for Studying the Impact of Warning Messages on Information Behavior

Alexandra Haddad, James Sauer, Jeremy Prichard, Caroline Spiranovic, and Karen Gelb

Abstract

New and nontraditional approaches are required to effectively tackle the global problem of cybercrime. Online warning messages offer the unique potential to influence information behavior at the exact point of user decision-making. This research assessed the prevention effect of differing components of warning messages. Thirty-five male participants, aged 18–43, participated in a behavioral-compliance task comprising messages received when visiting websites likely to contain malware. Participants also rated messages on believability, severity, and effects on intention to comply. The components of messages tested were as follows: three "signal words" (*warning*, *hazard*, and *stop*), two levels of message explicitness (high, low), and two imagery conditions (eyes, no eyes). Contrary to expectations, explicitness was the only message component to yield a significant preventative effect on self-rated and behavioral responses. Participants not only perceived the explicit messages as more believable, severe, and likely to increase intention to comply but also demonstrated, through their behavioral-compliance data, a preventative effect from more explicit messages. The implications of these findings for designing messages to prevent cybercrimes are explored.

Introduction

Regardless of metrics and data sources used, cybercrime is clearly pervasive and worldwide (Furnell, Emm, and Papadacki 2015). Scholars have theorized that the worldwide boom in cybercrime has been facilitated in part by criminogenic features of the online environment, including users' perceptions of anonymity, ease of offending (Wortley and Smallbone 2012), and psychological detachment from the moral quality of behavior

LIBRARY TRENDS, Vol. 68, No. 4, 2020 ("Censorship and Information Behavior in the Age of Big Data," edited by Paul A. Watters), pp. 576–98.

(Demetriou and Silke 2003). The facilitative effects of the Internet on cybercrime are seen in (a) "cyber-dependent" crimes (McGuire and Dowling 2013), such as hacking and the use of ransomware and malware (e.g., Maimon and Louderback 2019; Sarre, Yiu-Chung Lau, and Chang 2018); and (b) "cyber-enabled" crimes (McGuire and Dowling 2013), such as fraud and identify theft (Finch 2003) and offenses involving child sexual-exploitation material (CSEM) (Balfe et al. 2015; Wortley and Smallbone 2006).

The increasing scale of cybercrimes and technological sophistication of offenders (Harkin, Whelan, and Chang 2018) combined with the borderless nature of the Internet (Sarre, Yiu-Change Lau, and Chang 2018; Cross 2019) present major challenges to law-enforcement agencies. The restrictions impeding law-enforcement responses to cybercrime highlight the need for research into new and innovative ways of preventing cybercrime.

It is difficult to study cybercrime-prevention strategies, such as online warning messages (Prichard, Krone, et al. 2019). Two main categories of research methods have been utilized to date: (a) self-reported data from surveys and interviews; and (b) covertly observed human behavior on artificial online environments, such as a fake website or "honeypot." Both approaches have strengths and limitations. For instance, self-report studies are convenient and provide important information about demographic variables (e.g., Ullman and Silver 2018) but may not yield accurate information on embarrassing or stigmatizing subjects (e.g., Seto, Reeves, and Jung 2010). Honeypots can provide valuable data about human behavior in situ, but they are technically demanding and can encounter difficulties identifying demographic variables (Testa et al. 2017).

We present findings from a study that attempted to measure the influence of online warnings on information behavior using a novel hybrid method that incorporated survey methods with an artificial online environment.

Previous Research on the Effect of Warning Messages on Human Behavior

A large body of literature has investigated warning effectiveness in offline contexts (e.g., Braun and Silver 1995; Kalsher and Williams 2006; Wogalter, DeJoy, and Laughery 1999). In the online context, the most robust empirical evidence that messages have the potential to influence decision-making emanates from the field of human-computer interaction (HCI). Akhawe and Felt (2013) analyzed Internet users' decisions based on 25.4 million impressions of browser security warnings from Mozilla Firefox and Google Chrome. They demonstrated that some warnings prevented up to 70% of potential visits to unsecure websites (see further Reeder et al. 2018).

Smaller-scale studies have also shown that messages can reduce po-

tentially harmful noncriminal behaviors in online environments. For instance, some warning messages appear to be effective in reducing problematic online gambling, unsafe disclosure of personal information, and accessing pro-anorexia sites (Carpenter, Zhu, and Kolimi 2014; Gainsbury et al. 2015; Martijn et al. 2009).

Much less research has been able to examine the influence of messages on deviant behavior and crime (e.g., Decker 1972; Green 1985). Offline, postal letters have shown modest potential to reduce tax evasion (Coleman 2007) and insurance fraud (e.g., Blais and Bacher 2007). Relatively little research has explored the efficacy of warning messages to reduce deviancy and cybercrime. Google and Microsoft Bing trialed brief warning banners in combination with a blocking system to reduce CSEM-related queries in search engines (Steel 2015). The two strategies appeared to decrease CSEM searches by 67% over twelve months, although the degree to which this was due to the warning banners or the blocking system remains unclear (Steel 2015).

Good evidence has been collected from self-report studies. These have shown that messages have the potential to reduce music piracy (Ullman and Silver 2018) and the accessing of legal pornography by minors (Zaikina-Montgomery 2011). Their findings suggest that compliance with warnings appears to be influenced by message design, including symbols, signal words, and message explicitness. This is consistent with literature from the communication-human information processing (C-HIP) field, discussed below.

These studies indicate that self-report methods can provide person-centric data that are valuable for understanding demographic factors. For example, self-reports indicated that a warning message about legal pornography would increase likely compliance with adult participants, but reduce it with adolescents (Zaikina-Montgomery 2011). Self-report studies are also comparatively cheap, efficient, and convenient and do not require highly technical input from IT experts.

However, self-report methods typically cannot study whether a warning message actually results in a change of human behavior ("actual compliance")—only participants' *reported intended compliance*. Self-reported intention can differ from actual behavior as situational variables and subconscious influences also affect an individual's final actions (Kalsher and Williams 2006; Wogalter and Dingus 1999). Using self-report methods to study criminal or deviant behavior involves additional complexities. For crimes that require some type of specialist skill, such as computer hacking, recruiting sufficient numbers of participants would be difficult, and it is feasible that active hackers would have a motive to deliberately misinform researchers about their activities. Participants' responses could feasibly be influenced by social-desirability bias (e.g., Carr and Krause 1978; Krumpal 2013) when asked about messages to reduce crime, particularly those

relating to stigmatizing or sensitive topics, such as viewing CSEM (Seto, Reeves, and Jung 2010) and illicit drug use (Harrison 1997).

Honeypot methods have the potential to circumvent these problems. They involve creating an artificial online environment that can act as a proxy for the situations in which cybercrimes or deviant behaviors occur. Honeypots can involve simple covert observation of naïve participants' *actual* online behavior, such as their preparedness to access pirated material or pornography (e.g., Demetriou and Silke 2003). More sophisticated approaches incorporate interventions—such as the appearance of warning messages—and measure behavioral responses.

Maimon and colleagues (see Maimon, Alper, et al. 2014; Testa et al. 2017; Wilson et al. 2015) tested the effectiveness of a warning to deter computer hacking against a purpose-built honeypot online computer system (see Maimon, Alper, et al. 2014, 41; Testa et al. 2017, 700). The warning did not immediately reduce cyberattacks, but it did influence hackers' behavior, including by reducing time spent trespassing (see Maimon, Alper, et al. 2014) and altering computer commands entered during longer duration trespassing events (see Wilson et al. 2015).

However, data collection for such approaches can be slow, resource intensive, and involve complex ethical considerations (Prichard, Krone, et al. 2019). Additionally, these methods can encounter difficulties identifying demographic variables and excluding nonhuman behavior from bots (Testa et al. 2017).

Optimizing the Prevention Effect of Warnings on Cybercrime

We now explore what can be done to optimize the effectiveness of warning messages online to prevent cybercrime. Different theories have been drawn on to understand the influence of messages on criminal or deviant decision-making, including restrictive-deterrence theory (e.g., Wilson et al. 2015) and situational crime-prevention theory (Prichard, Krone, et al. 2019). We adopt the C-HIP model as our theoretical lens. This model has been used extensively to study human interaction with warnings in general. Its value in examining the influence of messages on deviant behavior has only recently been demonstrated (Prichard, Krone, et al. 2019; Ullman and Silver 2018; Zaikina-Montgomery 2011).

The Communication Human Information Processing (C-HIP) Model

The C-HIP model is a popular theoretical framework used to organize and conceptualize the various aspects of warnings that might influence their effectiveness (Wogalter, DeJoy, and Laughery 1999). C-HIP divides the warning into source (e.g., security software or a browser), channel (in this case online), and receiver components. The "receiver" component of a warning is further divided into factors relating to (a) *attention* and *notice-*

ability, (b) *memory* and *comprehension*, (c) *attitudes* and *belief formation*, and (d) *motivation* and *behavior*.

In line with these "receiver" components, an established body of literature has considered the compliance effect of the following warning characteristics: (a) severity (which prompts motivation and behavior); (b) believability (which impacts on attitudes and belief formation as well as motivation and behavior); (c) signal words (which affect attention and noticeability as well as motivation and behavior); (d) warning explicitness (which increases memory and comprehension as well as motivation and behavior); and (e) the presence of "eyes" (which increases attention and noticeability as well as attitudes and belief formation). We discuss the importance of each of these characteristics below.

Severity. Increased severity of consequences for noncompliance has been consistently linked to higher behavioral intention and compliance (Wogalter, Conzola, and Smith-Jackson 2002; Wogalter, Young, et al. 1999; Zaikina-Montgomery 2011). Hazard severity perception is an important factor in the risk-assessment process (Wogalter, Young, et al. 1999) and a strong predictor of subsequent behavior (Kalsher and Williams 2006; Wogalter, Brelsford, et al. 1991). Increasing the explicitness of a warning message (i.e., including clearly stated consequences of noncompliance) increases perceived severity and accordingly results in higher rates of behavioral compliance (Heaps and Henley 1999).

Believability. Belief formation is an important step in the processing of warning information (Wogalter 2006b). Warnings considered less believable are unlikely to have an effect on behavior because they are dismissed as false information (Riley 2006). The credibility of warnings in the online environment has been raised as critical because Internet users are often highly discerning and alert to the existence of false information online (Selejan et al. 2016; Wathen and Burkell 2002; Wogalter and Mayhorn 2008; Zaikina-Montgomery 2011).

Signal words. Signal words are highly salient words used to draw a receiver's attention to a warning and convey a certain level of hazard severity (Hellier and Edworthy, 2006; Wogalter and Silver 1990). Through the mechanisms of enhanced noticeability and higher perceived hazard severity, signal words have been associated with increased self-reported intention to comply (Cheatham and Wogalter 1999; Hellier and Edworthy 2006). There is clear evidence that the use of any signal word results in a more effective warning than an absence of a signal word; however, signal words also vary in level of perceived severity, which in turn leads to varying levels of behavioral compliance (i.e., higher severity signal words lead to higher levels of compliance: Cheatham and Wogalter 1999; Jensen and McCammack 2002; Smith-Jackson and Wogalter 2000). The signal word *stop* has been found to convey a high level of self-reported hazard severity in a study aimed at deterring minors from accessing pornography

(Zaikina-Montgomery 2011) and in a behavioral compliance study (Braun and Silver 1995). In one of the few behavioral studies conducted online, Carpenter, Zhu, and Kolimi (2014) tested participants' willingness to disclose personal information on websites selling insurance in the presence of warnings with different signal words. The word *hazard* was found to convey a high level of severity and effectively increased compliance (Carpenter, Zhu, and Kolimi 2014). This is in comparison to the signal word *warning*, which has been found to convey a moderate level of severity and accordingly moderate level of behavioral intention to comply (ANSI 2016; Drake, Conzola and Wogalter 1998; Hellier and Edworthy 2006).

Warning explicitness. There is a consensus in the literature that an ideal warning message should include information about the hazard itself, consequences of noncompliance, and safety instructions to allow the user to avoid the hazard (Braun and Shaver 1999; Laughery and Smith 2006; Martin 2000). The concept of explicitness has been raised as a key factor in increasing comprehension of this information (Laughery and Smith 2006; Young and Wogalter 1998). A message that is explicit is clear, precise, and detailed (Laughery and Smith 2006). Increasing message explicitness increases comprehension and memory for warnings (Trommelen 1997). These are critical stages of information processing (Wogalter, Young, et al. 1999), as warnings that are not well understood or remembered are unlikely to influence behavior (Trommelen 1997; Wogalter and Laughery 1996). Increasing explicit information about consequences of noncompliance most effectively increases perceptions of hazard severity, believability, and intention to comply (Heaps and Henley 1999; Laughery and Smith 2006; Laughery, Vaubel, et al. 1993).

The effect of eyes. Using images (such as basic symbols designed to convey potential threat) can effectively increase warning noticeability and hazard perception (Argo and Main 2004; Smith-Jackson and Wogalter 2000), both of which are crucial early stages of the process of warning perception and eventual compliance (Wogalter 2006b). Faces are uniquely effective in drawing attention and in eliciting emotions (Eastwood, Smilek, and Merikle 2003; Gliga et al. 2009), and eyes may be particularly significant in conveying social information and eliciting emotional responses (Jarick and Kingstone 2015).

Images of eyes are thought to emulate the sense of being watched, which triggers concern for social reputation and a subsequent increase in prosocial behavior (Bateson, Nettle, and Roberts 2006; Ekström 2012). Research has linked the use of images of eyes to increased prosocial behavior, both in laboratory and real-world settings (Bateson, Nettle, and Roberts 2006; Ernest-Jones, Nettle, and Bateson 2011; Francey and Bergmüller 2012; Nettle, Harper, et al. 2013). The "eyes effect" is also thought to reduce individuals' sense of anonymity by engaging the psychology of surveillance (Nettle, Nott, and Bateson 2012)—akin to using a warning

banner on a computer to notify the user that a system is under surveillance and activity is being monitored (Maimon, Alper, et al. 2014; Wilson et al. 2015). The "eyes effect" may, therefore, be particularly useful in the online environment, where anonymity is a key situational motivation or facilitator for CEM offenders (Wortley and Smallbone 2006).

The presence of imagery in general (e.g., symbols conveying hazard information) in combination with warning messages has been associated with increased perceptions of hazard severity, noticeability, and understandability (Laughery, 2006; Braun and Shaver 1999; Zaikina-Montgomery 2011). However, the "eyes effect" has not previously been tested in combination with other warning components.

Limitations of Research on Optimizing Warning-Message Effectiveness: Intent to Comply versus Actual Compliance

As noted, studies of components that optimize warning effectiveness rely predominantly on self-reports of intention to comply (Silver and Braun 1999). Certain visual and content-related components of warnings have been shown to have significant effects on a warning's noticeability, perceived severity, understandability, and on behavioral *intention to comply* (Frascara 2006; Laughery 2006; Wogalter 2006a; Zaikina-Montgomery 2011). However, far fewer studies have investigated (actual) behavioral compliance.

Since self-reported behavioral intention can differ from actual behavior, directly measuring behavior is critical to evaluate the efficacy of warnings as prevention strategies. Thus, in addition to testing behavioral intention to allow comparison with previous literature (e.g., Zaikina-Montgomery 2011), this study also directly measured behavior, thereby addressing a significant gap in existing knowledge.

Aims of the Present Study

This study aimed to experimentally test the compliance effect of differing components of messages within an online gaming environment whereby participants were motivated to avoid websites associated with malware. This will provide guidance on how to optimally design messages to prevent cybercrime. We also aimed to address a key limitation in the evidence base on the effectiveness of various message features by investigating not only behavioral intention but also actual behavioral compliance.

Specifically, we tested the following three hypotheses:

1) Compared with messages containing the signal word *warning*, messages containing the signal words *stop* and *hazard* will result in higher self-report ratings of warning effectiveness, including intended compliance, and a lower frequency of visits to malware websites (i.e., behavioral compliance).

2) Messages high in explicitness will result in higher self-report ratings of effectiveness, including intended compliance, and a lower frequency of malware website visits (i.e., behavioral compliance) than will messages low in explicitness.
3) Messages with eyes present will result in higher self-report ratings of effectiveness, including intended compliance, and a lower frequency of malware website visits (i.e., behavioral compliance) than will messages in which eyes are absent.

Method

Participants

We tested a sample of 35 male undergraduate students at an Australian university (age *M*=27; *SD*=7). We were aiming to trial the effectiveness of certain message elements on young males, since males are overrepresented in some categories of cybercrime, such as CSEM, and research has shown that young males are readily exposed to easy opportunities for cybercrime (e.g., as per Prichard, Watters, and Spiranovic's 2011 analysis of CSEM content on a mainstream P2P website). Almost two-thirds of participants reported English as their native language (63%). On average, participants reported using the Internet for 3.8 hours a day (*M*=3.82; *SD*=2.98), excluding Internet usage for work or study purposes. Eighty percent of participants reported prior experience with "malware" (i.e., malicious software). Our reasons for placing the task in the malware context are explained below under *Procedure and Materials: The experimental online context.*

Design

The effectiveness of signal words, warning explicitness, and eyes were investigated for their influence on a warning's believability, severity, behavioral intention to comply, and behavioral compliance in an online environment. We used a 3 (signal words: *stop*, *warning*, *hazard*) × 2 (explicitness: high or low) × 2 (eyes: present, absent) within-subjects design. Behavioral compliance was measured by willingness to engage with potentially risky websites.

Procedure and Materials

The experimental online context. Ethical constraints precluded directly testing the effects of warning messages on cybercrime in situ. Thus, we investigated risk-taking in the use of fake websites likely to contain malware as an analogue for risks undertaken in accessing websites potentially containing illegal content (e.g., CSEM). *Malware* is an umbrella term for various intrusive and destructive software programs designed with malicious intent (Furnell 2010). We used this context to test the effectiveness of various message components in preventing risky online behavior. While accessing illegal online content and exposure to malicious software both involve

risk, we acknowledge the disparity in potential consequences and the differing motivations between someone actively seeking high-risk illegal material and a user seeking to avoid the risk of harm from malware.

The experimental tasks: behavioral compliance and self-reported ratings. There were two components to this study: a behavioral-compliance task and a self-report-ratings task. After participants provided consent to participate, the experimenter left the room to avoid the possibility of participants feeling a sense of surveillance that might interfere with the "eyes effect" (Ernest-Jones, Nettle, and Bateson 2011). Participants first completed the behavioral-compliance task.

The behavioral-compliance task was designed for this study. Participants were asked to imagine they were shopping online for fitness-related products in a game-like task in which they accrued points. They were informed that some of the websites would have good deals while others would be associated with malware. They viewed a series of relevant websites—12 blocks of trials with 20 trials per block—and their willingness to engage with the sites was tested after seeing different warning messages relating to malware websites. Combinations of the various warning components were counterbalanced across blocks of trials, followed by a series of screenshots of websites. A random 20% of trials were associated with malware. Participants could then choose whether to visit each website.

The number of websites visited within blocks was the key measure of behavioral compliance. Lower engagement with malware websites indicated increased compliance with warnings. Response times to websites were also measured and assessed for each warning as a screening measure to ensure participants viewed warnings for a sufficient length of time.[1]

For each website, participants had to choose whether to interact with it (by pressing the *y* key) or avoid it (by pressing the *n* key). If they visited a site that did not have malware, they received a message: "Congratulations, you got a great deal! You receive 100 points!" If they visited a site that did have malware, they received the message, "The website you just visited had malware—your computer is now infected. You lose 100 points!" If they chose to avoid a website by pressing the *n* key, they moved immediately to the next trial with no loss or gain in points.

The behavioral-compliance task for this study was designed to create some external motivation for participants to interact with the websites in a way that involves the possibility of personal risk and reward, as is the case for types of cybercrime (e.g., Wortley 2012).

The self-reported-ratings task immediately followed the behavioral-compliance task; participants completed ratings for the various warnings. Warnings were displayed on the screen in random order, and participants were asked to rate each in terms of believability, severity, and how likely they felt they would be to comply with the warning if encountered online (behavioral intention). Ratings were made on an 8-point Likert scale,

conforming with previous research to facilitate comparison (e.g., Zaikina-Montgomery 2011).

Warning materials. The various warning components were created for this study based closely on previous research and guidelines and on current online warnings used by web browsers, search engines, and virus protection software (ANSI 2016; Egelman, Cranor, and Hong 2008; Zaikina-Montgomery 2011). Twelve warnings were created in total (i.e., by fully crossing the experimental design). The target stimuli comprised 240 screenshots of unique websites. The image of eyes was sourced from the Radboud Faces Database (Langner et al. 2010). Previous research indicates that male, "stern" eyes tend to be most effective in producing behavioral compliance, and thus the eyes used in this study matched these criteria (Bateson, Nettle, and Roberts 2006; King et al. 2016). The experimental programs for both the behavioral-compliance task and self-report data were programmed in E-Prime (see http://www.pstnet.com).

Data Analysis

Analyses comprised both mixed-effects models as well as the traditional approach of repeated measures ANOVAs. Mixed-effect models were created using the package lme4 (Bates, Maechler, and Bolker 2011) in R (Bunn and Korpela 2020). Mixed-effects models have many advantages over ANOVAs, including the ability to partial out random variance (i.e., noise associated with participant and stimulus variance) and the use of individual trials as data points rather than aggregating means and proportions. (For a more detailed discussion of the advantages of mixed-effect modelling over ANOVAs, see Baayen, Davidson, and Bates 2008; Jaeger 2008). Partialing out random noise associated with individual differences (receiver characteristics) is important in the present study as these individual differences can moderate warning effectiveness (Smith-Jackson 2006). Similarly, some website stimuli might simply appear more threatening than others. Removing this noise from the analyses is advantageous.

Logistic mixed-effects models present the estimated log odds of interacting with a website and can be interpreted as per standard regressions. The essential difference is that, unlike standard regressions, mixed-effects models allow intercepts to vary across individual participants and stimuli. Each model provides an intercept, a beta coefficient (reflecting magnitude of the effect), a p value, and 95% confidence intervals (to determine if the manipulation had a meaningful effect). Repeated measures ANOVAs were considered sufficient for the self-report data.

Results

The results are presented separately for each of the two tasks: the behavioral-compliance task first, followed by the self-report task. Within each section, results are presented for each of the three hypotheses in turn.

Compliance

Based on the literature, we expected that warnings containing the signal words *stop* and *hazard* would result in a lower frequency of website visits (i.e., behavioral compliance) than warnings containing the signal word *warning* (hypothesis 1). We also hypothesized that that presence of eyes (hypothesis 2) and high explicitness (hypothesis 3) in warnings will result in a lower frequency of website visits.

To test the effects of different warning components on the proportion of websites visited, we constructed a logistic mixed-effects model with the predicted log odds of visiting a website as the outcome variable and participant and stimulus as random effects. The effect of signal word was first added to the model; however, contrary to hypothesis 1, this did not significantly improve the fit of the model, $\chi^2(2) = 3.29$, $p = 0.19$. Second, we added the effect of warning explicitness. Consistent with hypothesis 2, adding warning explicitness significantly improved the fit of the model, $\chi^2(1) = 12.92$, $p < .001$, and indicated participants were less likely to interact with websites after seeing the more explicit warnings. Third, we added the effect of eyes, which, contrary to hypothesis 3, did not improve the fit of the model, $\chi^2(1) = 0.17$, $p = 0.68$. We tested more complex models (i.e., including all two-way and three-way interactions), but including these higher-order effects did not improve the fit of the model to the data, $\chi^2 < 1.81$, $p > .404$.

Figure 1 plots model-estimated probabilities (panel A) and log odds (panel B) of visiting a site, with associated model coefficients reported in table 1. Note that, for comparability between measures, the probabilities in panel A were computed from the model-estimated log odds, rather than being based on the raw data. Thus, like the model-estimated log odds, these probabilities represent effects after accounting for the random effects of participant and stimulus. The signal word *warning*, low explicitness, and the "absent eyes" conditions were set as the referents for all models. Beta values therefore indicate the predicted change in outcome when the level of these variables is manipulated. For example, the coefficient for Signal Word *stop* in table 1 indicates how the likelihood of entering a site changes when the signal changes from *warning* to *stop*, and the coefficient for Explicitness indicates change in the likelihood of entering the site as the warning changes from low to high explicitness.

Self-Report

The self-report measures of warning efficacy—measuring believability, perceived severity and intention to comply—were analyzed using separate 3 (Signal Word: *warning*, *stop*, *hazard*) x 2 (Explicitness: high vs. low) x 2 (Eyes: present vs. absent) repeated-measures ANOVAs (see table 2 for relevant inferential test statistics and indices of effect size). Warnings containing the signal words *stop* and *hazard* were expected to result in higher

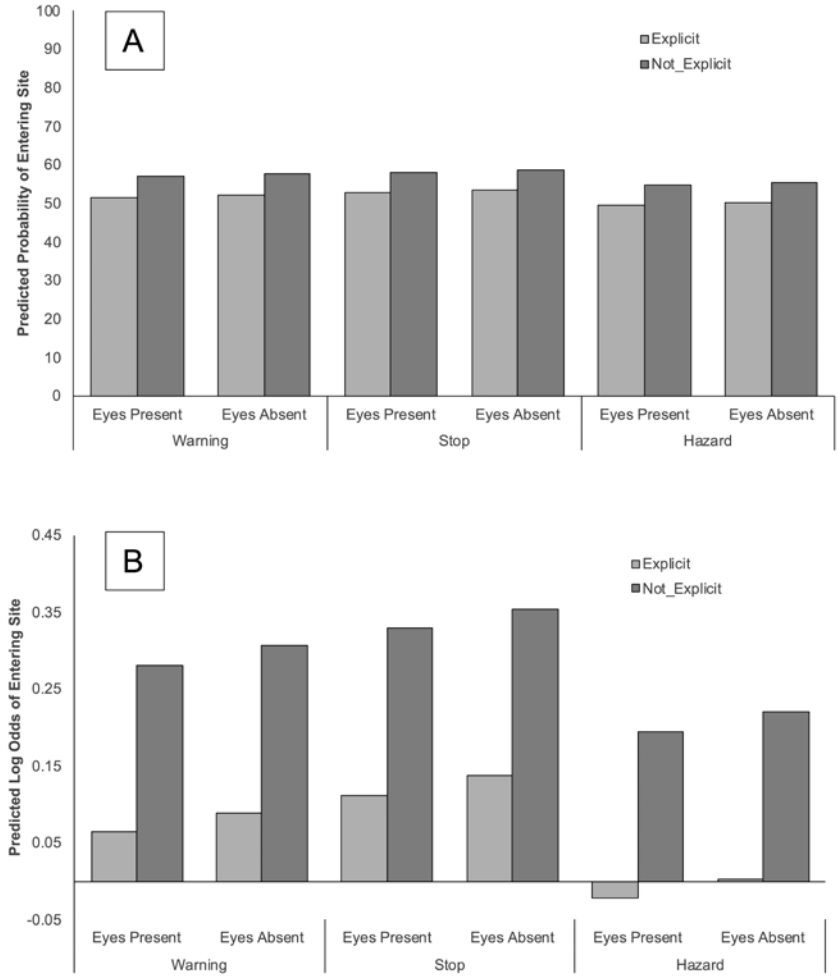

Figure 1. Estimated probability of entering a site (Panel A) and estimated log odds (Panel B) of entering a site (Panel B) by explicitness of message.

Table 1. Fixed-Effects Coefficients for Logistic Mixed-Effects Model Predicting Likelihood of Entering Websites Based on Warning Components.

Fixed effect	b	SE_b	Z	95% CI
Intercept	0.31	0.24	1.29	[–0.17; 0.78]
Signal Word *stop*	0.05	0.07	0.66	[–0.09; 0.19]
Signal Word *hazard*	–0.09	0.07	–1.19	[–0.23; 0.06]
Explicitness	–0.22	0.06	–3.62	[–0.34; –0.10]
Eyes	–0.02	0.06	–0.41	[–0.14; 0.09]

Table 2. Repeated Measures ANOVAs. Inferential Statistics for Self-Reported Data.

Measure		SS	DF	MS	F	p	η^2
Belief	Signal Word	0.93	2	0.47	0.78	.464	.022
	Explicitness	13.39	1	13.39	6.83	.013*	.167
	Eyes	112.11	1	112.12	18.07	<.000*	.347
Intention	Signal Word	0.52	2	.260	0.46	.635	.013
	Explicitness	12.00	1	12.00	6.50	.015*	.161
	Eyes	34.86	1	34.86	7.48	.010*	.180
Severity	Signal Word	1.16	2	0.58	0.90	.411	.026
	Explicitness	26.75	1	26.75	11.69	.002*	.256
	Eyes	5.95	1	5.95	1.29	.264	.036

ratings on all self-report measures of warning effectiveness than warnings containing the signal word *warning* (hypothesis 1). Furthermore, warnings with high explicitness (hypothesis 2) and eyes present (hypothesis 3) were expected to result in higher ratings on all self-report measures.

Contrary to hypothesis 1, there was no evidence for a significant effect of signal word. However, in support of hypothesis 2, the warnings high in explicitness were the most effective. Compared to warnings low in explicitness, warnings high in explicitness were rated as significantly more believable (M=5.35, 95% CIs [4.70, 6.00]; cf. M=4.99, [4.33, 5.65]), more severe (M=5.90, [5.33, 6.47]; cf. M=5.39, [4.73, 6.05]), and produced greater intention to comply (M=5.81, [5.16, 6,46]; cf. M=5.47, [4.86, 6.08]) (see table 2).

Surprisingly, and counter to hypothesis 3, the warnings with eyes were rated as significantly *less* believable (M=4.65, [3.96, 5.34]) than those without (M=5.68, [5.06, 6.30]) and produced lower intention to comply (M=5.31, [4.63, 5.99]) than did warnings without eyes (M=5.92, [5.34, 6.50]). None of the interactions were significant (all F<3.21, p>.063, η^2<.086).

In summary, there was no evidence that the signal word used produced any effect on either behavioral compliance or self-reports of warning effectiveness, contrary to hypothesis 1. Consistent with hypothesis 2, warning explicitness increased both behavioral compliance and perceptions of warning effectiveness. Finally, the use of eye imagery affected self-reports of some measures of warning effectiveness, but in the opposite direction to that hypothesized (hypothesis 3).

Discussion

We investigated which components influence warning effectiveness in the online environment. Consistent with hypothesis 2 and work by others (e.g., Laughery and Smith 2006; Laughery, Vaubel, et al. 1993; Trommelen 1997), the results of the behavioral-compliance task indicated that increasing the explicitness of warning messages significantly increases compliance (i.e., reducing frequency of visits to problematic websites).

Explicitness was also rated as significantly higher on all three self-report measures relating to warning efficacy (i.e., severity, believability, and intention to comply). However, counter to hypotheses 1 and 3, manipulations of other warning components had no effects on behavior and, in some cases, produced counterintuitive effects on self-report measures of warning effectiveness. Specifically, counter to predictions and effects reported in other warning contexts (Bateson, Callow, et al. 2013; King et al. 2016), including images of eyes with the warning lowered ratings of believability and intention to comply. Notably, however, this did not translate into any effect on behavioral compliance.

The C-HIP model offers a useful framework for integrating warning research findings into a theoretical structure and can aid in determining which stages of information processing and which specific warning attributes might be implicated in a warning's success or failure (Wogalter 2006b). The discussion below therefore applies this model in assessing how to interpret this study's findings and increase the prevention effect of warnings on cybercrime.

Signal Words

Contrary to predictions, no significant differences were seen between signal words in either self-report measures or behavioral-compliance rates. Although both *stop* and *hazard* have been proposed to convey a high level of severity (Carpenter, Zhu, and Kolimi 2014; Zaikina-Montgomery 2011), they have been subject to much less testing than the traditional signal words (e.g., *warning*: Amer and Maris 2007; ANSI 2016; Argo and Main 2004; Drake, Conzola, and Wogalter 1998; Hellier, Aldrich, et al. 2007; Hellier and Edworthy 2006; Young 1998). It is possible that these signal words are less effective at conveying high levels of severity than preliminary evidence suggested. *Stop* and *hazard* may instead convey a moderate level of severity and accordingly produce effects on behavioral compliance similar to *warning*.

However, before dismissing these signal words as less effective, it may be worth considering alternative explanations. The signal words may have influenced perception of hazard severity (although not consciously) at a level below the threshold required to influence the next stages of processing (beliefs, attitudes, motivation, and behavior—C-HIP: Wogalter 2006b). Signal words are generally considered to be effective in increasing warning noticeability and salience (which links with the first stage in the C-HIP information process—attention: Hellier and Edworthy 2006; Wogalter 2006b; Wogalter and Vigilante 2006). They are also intended to quickly convey level of hazard severity (linking to one of the subsequent stages in C-HIP: attitude and beliefs; Hellier and Edworthy 2006; Wogalter and Vigilante 2006). For the signal words tested, there may have been an effect on attention, which did not translate into the subsequent stage of pro-

cessing (belief formation; Wogalter and Vigilante 2006). In other words, the tested signal words may act purely on the earliest stage of processing (attention) but were not sufficiently powerful to influence beliefs or behavior.

Explicitness

The findings related to explicitness of warning messages—that warnings high in explicitness (cf. low explicitness) increase perceptions of severity and behavioral compliance—align with previous research, based on self-reported intention to comply and compliance in real-world contexts (e.g., Braun and Shaver 1999; Trommelen 1997). Thus, this study demonstrated that providing specific, clear details about the consequences of noncompliance with a warning produces greater behavioral compliance in the online context in addition to real-world settings (Laughery, Vaubel, et al.1993). According to the C-HIP model, the mechanism underlying the effect of explicitness is thought to be increasing the receiver's sense of hazard severity and subsequently increasing cautious behavior and compliance (Trommelen 1997; Wogalter, Brelsford, et al. 1991; Young and Wogalter 1998). It is interesting to note that the hazard severity ratings were the only self-report measure that accurately predicted behavior (i.e., only warnings that increased ratings of hazard severity—the more explicit warnings—increased behavioral compliance). This supports the proposition that increasing hazard perception is key to the process of inducing behavioral compliance (C-HIP model; Wogalter, Young, et al. 1999) and that explicit messages help to achieve this.

The fact that explicitness also increased perceived believability is noteworthy as belief formation is an important step in warning processing (C-HIP model; Wogalter, Young, et al. 1999). This has particular importance for the online context where users tend to be wary of the existence of false information and "fake warnings" (Wogalter and Mayhorn 2008). Designing a warning that is believable is crucial to have any expectation of behavioral compliance. The finding that explicitness produced a more believable warning in the online context provides a useful guideline for future warning design, including in the ultimate area of interest—deterring cybercrime.

The Eyes Effect

Including images of eyes did not affect behavioral compliance and had an adverse effect on participant ratings of warning believability and intention to comply. According to the C-HIP model, failure to successfully process information at any stage of the system can block the flow of information from reaching the next stage toward eventual compliance (Wogalter, Young, et al.1999; Wogalter and Vigilante 2006). According to the model, beliefs are an important precursor to motivation, intention, and finally behavior (Wogalter and Vigilante 2006). It is possible that the reduced be-

lief in the warnings with eyes resulted in an obstruction in the processing of the warning and subsequently reduced compliance. This possibility is supported by the significantly decreased rates of self-reported intention to comply. Anecdotally, many participants at the end of the study expressed a perceived lack of credibility in the warnings with eyes. However, some indicated they found these warnings more intimidating. It could therefore be the case that for most participants the eyes were less credible and therefore reduced compliance, but for some they had the opposite effect. The comparatively large standard deviations (e.g., believability of eyes: *SD*=2.07, cf. without eyes: *SD*=1.86; intention to comply with eyes: *SD*=2.04, cf. without eyes: *SD*=1.75, see table 2 for details) around the self-report measures is supportive of this possible influence of individual variance.

The effectiveness of eyes shown in previous research lies in producing a sense of surveillance, reducing a sense of anonymity, and increasing prosocial behavior. These mechanisms may be less relevant to the current task context (trying to avoid malware). Thus, future testing of the effects of eye imagery in a context more closely related to cybercrime may be useful.

Limitations

A key limitation of this study was the pragmatic choice to use the risk of malware as a testing context in place of cybercrime. Clearly, the motivations and outcomes for individuals considering accessing illegal materials are quite different to those driving participants in this study. Future research in this area should move closer to the context of interest, such as the "barely legal" genre of adult pornography. However, the primary aim of this study was to gather data on the effectiveness of warning-message components in preventing unwanted online behavior. The extant literature on warning effectiveness has relied primarily on self-report data relating to perceived severity and *intention* to comply. Although this is a potentially informative measure, the disconnect between self-reported intention and actual behavioral compliance is well-documented in a variety of settings (Hassan, Shiu, and Shaw 2016; Kalsher and Williams 2006). Thus, obtaining some behavioral data on warning effectiveness in online environments represents a significant contribution to the literature.

Further, the utility of the self-report measures used in this study may have been compromised by using only a single item to asses each construct (i.e., severity, believability, and intention to comply). This is standard practice in warning research; however, it has been suggested in recent literature that a more reliable and comprehensive measure might involve multiple questions for each construct (Zaikina-Montgomery 2011). This is a limitation of this study and warning-research practices in general.

Despite limitations on the applicability of our findings to the ultimate context of interest (cybercrimes such as CEM), our results provide useful behavioral data. The analytical approach used—mixed effects model-

ling—enhanced the generalizability of findings by attenuating random effects associated with differences between participants and stimuli (Westfall, Kenny, and Judd 2014). Additionally, we have demonstrated the efficacy of a novel hybrid approach that combined self-report methods with analysis of actual behavioral compliance through artificial online environments. This hybrid method provides more options for researchers attempting to study online messages as a crime-prevention strategy.

Conclusion

Explicit warnings increase perceptions of severity and behavioral compliance in relation to risky online behaviors when measured using self-reported intention to comply and actual behavioral compliance. This result can be used to better design Internet warning messages, to improve compliance with cyberlaws, and to reduce problematic information behaviors.

Acknowledgements

This research was supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (DP160100601). The views expressed herein are those of the authors and are not necessarily those of the Australian Government or Australian Research Council. We thank Richard Wortley for his feedback on earlier versions of the manuscript. We also wish to thank Prof. Paul Watters (Professor in Cyber Security, Comp Sci & Info Tech, La Trobe University) for his assistance in the early stages of this project.

Note

1. The response time data are not reported in this paper because response time was used for screening purposes only rather than as a variable of relevance to the study hypotheses.

References

Akhawe, Devdatta, and Adrienne Porter Felt. 2013. "Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness." In *Proceedings of the 22nd USENIX Security Symposium,* 257–72. https://www.usenix.org/system/files/conference/usenixsecurity13/sec13-paper_akhawe.pdf.

Amer, Tarek S., and Joe-Mae B. Maris. 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): 1–26. https://doi.org/10.2308/jis.2007.21.2.1.

ANSI (American National Standards Institute). 2016. *American National Standard Design Principles for Environmental/Facility Safety Signs and Product Labels* (ANSI Z535.X-2016). National Electrical Manufacturers Association. https://doi.org/10.1371/journal.pone.0082055.

Argo, Jennifer J., and Kelley J. Main. 2004. "Meta-analyses of the Effectiveness of Warning Labels." *Journal of Public Policy & Marketing* 23 (2): 193–208. https://doi.org/10.1509/jppm.23.2.193.51400.

Baayen, R. Harald, Debra J. Davidson, and Douglas M. Bates. 2008. "Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items." *Journal of Memory and Language* 59 (4): 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Balfe, Myles, Bernard Gallagher, Helen Masson, Shane Balfe, Ruairi Brugha, and Simon Hackett. 2015. "Internet Child Sex Offenders' Concerns about Online Security and Their

Use of Identity Protection Technologies: A Review." *Child Abuse Review* 24 (6): 427–39. https://doi.org/10.1002/car.2308.
Bates, Douglas. 2006. "[R] lmer, P-values and All That." Personal communication, May 19, 2006. https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html.
Bates, D. M., M. Maechler, and B. Bolker. 2011. "lme4: Linear Mixed-Effects Models Using S4 Classes R Package Version 0.999375-41." https://cran.r-project.org/web/packages/lme4/lme4.pdf
Bateson, Melissa, Luke Callow, Jessica R. Holmes, Maximilian L. R. Roche, and Daniel Nettle. 2013. "Do Images of 'Watching Eyes' Induce Behaviour That is More Pro-Social or More Normative? A Field Experiment on Littering." *PLoS ONE* 8 (12): e82055. https://doi.org/10.1371/journal.pone.0082055.
Bateson, Melissa, Daniel Nettle, and Gilbert Roberts. 2006. "Cues of Being Watched Enhance Cooperation in a Real-World Setting." *Biology Letters* 2 (3): 412–14. https://doi.org/10.1098/rsbl.2006.0509.
Blais, Etienne, and Jean-Luc Bacher. 2007. "Situational Deterrence and Claim Padding: Results from a Randomized Field Experiment." *Journal of Experimental Criminology* 3 (4): 337–52.
Braun, Curt C., and Eric F. Shaver. 1999. "Warning Sign Components and Hazard Perceptions." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 43 (16): 878–82. https://doi.org/10.1177/154193129904301601.
Braun, Curt C., and N. Clayton Silver. 1995. "Interaction of Signal Word and Colour on Warning Labels: Differences in Perceived Hazard and Behavioral Compliance." *Ergonomics* 38 (11): 2207–20. https://doi.org/10.1080/00140139508925263.
Bunn, Andy, and Mikko Korpela. 2020. "An Introduction to dp1R." ftp://ftp.uvigo.es/CRAN/web/packages/dplR/vignettes/intro-dplR.pdf.
Carpenter, Sandra, Feng Zhu, and Swapna Kolimi. 2014. "Reducing Online Identity Disclosure Using Warnings." *Applied Ergonomics* 45 (5): 1337–42. https://doi.org/10.1016/j.apergo.2013.10.005.
Carr, Leslie G., and Neal Krause. 1978. "Social Status, Psychiatric Symptomatology, and Response Bias." *Journal of Health and Social Behavior* 19 (1): 86–91. doi: 10.2307/2136325.
Cheatham, Deane B., and Michael S. Wogalter. 1999. "Connoted Hazard and Perceived Conspicuity of Warning Configurations." *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* 43 (16): 883–87. https://doi.org/10.1177/154193129904301602.
Clarke, Ronald V. 2013. "Seven Misconceptions of Situational Crime Prevention." In *Handbook of Crime Prevention and Community Safety*, edited by Nick Tilley, 65–96. Cullompton, Devon, UK: Willan.
Coleman, Stephen. 2007. "The Minnesota Income Tax Compliance Experiment: Replication of the Social Norms Experiment." SSRN. Accessed July 14, 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1393292.
Cooper, Al. 1998. "Sexuality and the Internet: Surfing Into the New Millennium." *Cyberpsychology and Behavior* 1 (2): 187–93. https://doi.org/10.1089/cpb.1998.1.187.
Cornish, Derek B., and Ronald V. Clarke. 2003. "Opportunities, Precipitators and Criminal Decisions: A Reply to Wortley's Critique of Situational Crime Prevention." *Crime Prevention Studies* 16:41–96. https://pdfs.semanticscholar.org/3308/389a0630bb2cdfabfc4e28b2e4f6373e1f12.pdf.
Cross, Cassandra. 2019. "'Oh We Can't Actually Do Anything about That': The Problematic Nature of Jurisdiction for Online Fraud Victims." *Criminology and Criminal Justice*. https://doi.org/10.1177/1748895819835910.
Decker, John F. 1972. "Curbside Deterrence? An Analysis of the Effect of a Slug-Rejector Device, Coin-View Window, and Warning Labels on Slug Usage in New York City Parking Meters." *Criminology* 10 (2): 127–142. https://doi.org/10.1111/j.1745-9125.1972.tb00549.x.
Demetriou, Christina, and Andrew Silke. 2003. "A Criminological Internet 'Sting': Experimental Evidence of Illegal and Deviant Visits to a Website Trap." *British Journal of Criminology* 43 (1): 213–22. https://doi.org/10.1093/bjc/43.1.213.
Drake, Kelly L., Vincent C. Conzola, and Michael S. Wogalter. 1998. "Discrimination among Sign and Label Warning Signal Words." *Human Factors & Ergonomics in Manufacturing* 8 (4): 289–301. https://doi.org/10.1002/(SICI)1520-6564(199823)8:4%3C289::AID-HFM1%3E3.0.CO;2-%23.
Eastwood, John D., Daniel Smilek, and Philip Merikle. 2003. "Negative Facial Expression

Captures Attention and Disrupts Performance." *Perception & Psychophysics* 65 (3): 352–58. https://doi.org/10.3758/BF03194566.

Egelman, Serge, Lorrie F. Cranor, and Jason Hong. 2008. "You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings." In *CHI '08: Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems*, 1065–74. New York: Association for Computing Machinery. https://doi.org/10.1145/1357054.1357219.

Ekström, Mathias. 2012. "Do Watching Eyes Affect Charitable Giving? Evidence from a Field Experiment." *Experimental Economics* 15 (3): 530–46. https://doi.org/10.1007/s10683-011-9312-6.

Ernest-Jones, Max, Daniel Nettle, and Melissa Bateson. 2011. "Effects of Eye Images on Everyday Cooperative Behavior: A Field Experiment." *Evolution and Human Behavior* 32 (3): 172–78. https://doi.org/10.1016/j.evolhumbehav.2010.10.006.

Finch, Emily. 2003. "What a Tangled Web We Weave: Identity Theft and the Internet." In *Dot.cons: Crime, Deviance, and Identity on the Internet*, edited by Yvonne Jewkes, 86–104. Cullompton, Devon, UK: Willan. https://popcenter.asu.edu/sites/default/files/problems/identity_theft/PDFs/Finch_2003.pdf.

Francey, Damien, and Ralph Bergmüller. 2012. "Images of Eyes Enhance Investments in a Real-Life Public Good." *PLoS ONE* 7 (5): e37397. https://doi.org/10.1371/journal.pone.0037397.

Frascara, Jorge. 2006. "Typography and the Visual Design of Warnings." In *Handbook of Warnings*, edited by Michael S. Wogalter, 385–403. Mahwah, NJ: Lawrence Erlbaum.

Furnell, Steven. 2010. "Hackers, Viruses and Malicious Software." In *Handbook of Internet Crime*, edited by Yvonne Jewkes and Majid Yar: 173–193. Cullompton, Devon, UK: Willan.

Furnell, Steven, David Emm, and Maria Papadaki. 2015. "The Challenge of Measuring Cyber-Dependent Crimes." *Computer Fraud and Security* 2015 (10): 5–12. https://doi.org/10.1016/S1361-3723(15)30093-2.

Gainsbury, Sally M., David Aro, Dianne Ball, Christian Tobar, and Alex Russell. 2015. "Optimal Content for Warning Messages to Enhance Consumer Decision Making and Reduce Problem Gambling." *Journal of Business Research* 68 (10): 2093–101. https://doi.org/10.1016/j.jbusres.2015.03.007.

Gliga, Theodora, Mayada Elsabbagh, Athina Andravizou, and Mark Johnson. 2009. "Faces Attract Infants' Attention in Complex Displays." *Infancy* 14 (5): 550–62. https://doi.org/10.1080/15250000903144199.

Green, Gary S. 1985. "General Deterrence and Television Cable Crime: A Field Experiment in Social Control." *Criminology* 23 (4): 629–45. https://doi.org/10.1111/j.1745-9125.1985.tb00367.x.

Hamari, Juho, Jonna Koivisto, and Harri Sarsa. 2014. "Does Gamification Work?—A Literature Review of Empirical Studies on Gamification." In *HICSS '14: Proceedings of the 47th Hawaii International Conference on System Sciences*, 3025–34. Piscataway, NJ: Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/HICSS.2014.377.

Harkin, Diarmaid, Chad Whelan, and Lennon Chang. 2018. "The Challenges Facing Specialist Police Cyber-Crime Units: An Empirical Analysis." *Police Practice and Research* 19 (6): 519–36. https://doi.org/10.1080/15614263.2018.1507889.

Harrison, Lana. 1997. "The Validity of Self-Reported Drug Use in Survey Research: An Overview and Critique of Research Methods." In *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*, edited by L. Harrison and A. Hughes, 17–36. US Department of Health and Human Services, NIDA Research Monograph 167. Rockville, MD: National Institute on Drug Abuse.

Hassan, Louise M., Edward Shiu, and Deirdre Shaw. 2016. "Who Says There Is an Intention-Behavior Gap? Assessing the Empirical Evidence of an Intention-Behavior Gap in Ethical Consumption." *Journal of Business Ethics* 136 (2): 219–36. https://doi.org/10.1007/s10551-014-2440-0.

Heaps, Christopher M., and Tracy B. Henley. 1999. "Language Matters: Wording Considerations in Hazard Perception and Warning Comprehension." *Journal of Psychology* 133 (3): 341–51. https://doi.org/10.1080/00223989909599747.

Hellier, Elizabeth, Kirsteen Aldrich, Daniel B. Wright, Denny Daunt, and Judy Edworthy. 2007. "A Multi Dimensional Analysis of Warning Signal Words." *Journal of Risk Research* 10 (3): 323–38. https://doi.org/10.1080/13669870601066963.

Hellier, Elizabeth, and Judy Edworthy. 2006. "Signal Words." In *Handbook of Warnings*, edited by Michael S. Wogalter, 407–17. Mahwah, NJ: Lawrence Erlbaum.

Jaeger, T. Florian. 2008. "Categorical Data Analysis: Away from ANOVAs (Transformation or Not) and Towards Logit Mixed Models." *Journal of Memory and Language* 59 (4): 434–46. https://doi.org/10.1016/j.jml.2007.11.007.

Jarick, Michelle, and Alan Kingstone. 2015. "The Duality of Gaze: Eyes Extract and Signal Social Information during Sustained Cooperative and Competitive Dyadic Gaze." *Frontiers in Psychology* 6:1423. https://doi.org/10.3389/fpsyg.2015.01423.

Jensen, Roger C., and Andrew M. McCammack. 2002. "Safety and Health Sign Colors and Signal Words for Communicating Information on Likelihood and Imminence of Threat." *Safety Health & Industrial Hygiene* 26. http://digitalcommons.mtech.edu/shih/26.

Kalsher, Michael J., and Kevin J. Williams. 2006. "Behavioral Compliance: Theory, Methodology, and Results." In *Handbook of Warnings*, edited by Michael S. Wogalter, 313–31. Mahwah, NJ: Lawrence Erlbaum.

King, Dominic, Ivo Vlaev, Ruth Everett-Thomas, Maureen Fitzpatrick, Ara Darzi, and David J. Birnbach. 2016. "'Priming' Hand Hygiene Compliance in Clinical Environments." *Health Psychology* 35 (1): 96–101. https://doi.org/10.1037/hea0000239.

Krumpal, Ivar. 2013. "Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review." *Quality & Quantity* 47 (4): 2025–47. https://doi.org/10.1007/s11135-011-9640-9.

Langner, Oliver, Ron Dotsch, Gijsbert Bijlstra, Daniel H. J. Wigboldus, Skyler T. Hawk, and Ad van Knippenberg. 2010. "Presentation and Validation of the Radboud Faces Database." *Cognition & Emotion* 24 (8): 1377–88. https://doi.org/10.1080/02699930903485076.

Laughery, Kenneth R. Sr. 2006. "Safety Communications: Warnings." *Applied Ergonomics* 37 (4): 467–78. https://doi.org/10.1016/j.apergo.2006.04.020.

Laughery, Kenneth R. Sr., and Danielle P. Smith. 2006. "Explicit Information in Warnings." In *Handbook of Warnings*, edited by Michael S. Wogalter, 419–28. Mahwah, NJ: Lawrence Erlbaum.

Laughery, Kenneth R. Sr., Kent P. Vaubel, Stephen L. Young, John W. Brelsford, and Anna L. Rowe. 1993. "Explicitness of Consequence Information in Warnings." *Safety Science* 16 (5–6): 597–613. https://doi.org/10.1016/0925-7535(93)90025-9.

Lieberoth, Andreas. 2015. "Shallow Gamification: Testing Psychological Effects of Framing an Activity as a Game." *Games and Culture: A Journal of Interactive Media* 10 (3): 229–48. https://doi.org/10.1177/1555412014559978.

Maimon, David, Mariel Alper, Bertrand Sobesto, and Michel Cukier. 2014. "Restrictive Deterrent Effects of a Warning Banner in an Attacked Computer System." *Criminology* 52 (1): 33–59. https://doi.org/10.1111/1745-9125.12028.

Maimon, David, and Eric R. Louderback. 2019. "Cyber-Dependent Crimes: An Interdisciplinary Review." *Annual Review of Criminology* 2:191–216. https://doi.org/10.1146/annurev-criminol-032317-092057 .

Martijn, Carolien, Elke Smeets, Anita Jansen, Nancy Hoeymans, and Casper Schoemaker. 2009. "Don't Get the Message: The Effect of Warning Text Before Visiting a Proanorexia Website." *International Journal of Eating Disorders* 42 (2): 139–45. https://doi.org/10.1002/eat.20598.

Martin, B. Jay. 2000. "The Value of Explicit Hazard and Consequence Warnings for Products with Hidden Hazards." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 44 (27): 302–5. https://doi.org/10.1177/154193120004402705.

McGuire, Mike, and Samantha Dowling. 2013. *Cybercrime: A Review of the Evidence*. Home Office Research Report 75. London: Home Office. http://www.justiceacademy.org/iShare/Library-UK/horr75-chap1.pdf.

Nettle, Daniel, Zoe Harper, Adam Kidson, Rosie Stone, Ian S. Penton-Voak, and Melissa Bateson. 2013. "The Watching Eyes Effect in the Dictator Game: It's Not How Much You Give, It's Being Seen to Give Something." *Evolution and Human Behavior* 34 (1): 35–40. https://doi.org/10.1016/j.evolhumbehav.2012.08.004.

Nettle, Daniel, Kenneth Nott, and Melissa Bateson. 2012. "'Cycle Thieves, We Are Watching You': Impact of a Simple Signage Intervention Against Bicycle Theft." *PLoS ONE* 7 (12): 1–5. https://doi.org/10.1371/journal.pone.0051738.

Oxarart, Aimee, Jennifer Weaver, Adel Al-Bataineh, and Mohamed T. Al-Bataineh. 2014. "Game

Design Principles and Motivation." *International Journal of Arts & Sciences* 7 (2): 347–59. http://www.universitypublications.net/ijas/0702/pdf/H4V866.pdf.

Prichard, Jeremy, Tony Krone, Caroline Spiranovic, and Paul Watters. 2019. "Transdisciplinary Research in Virtual Space: Can Online Warning Messages Reduce Engagement with Child Exploitation Material?" In *Routledge Handbook of Crime Science*, edited by Richard Wortley, Aiden Sidebottom, Nick Tilley, and Gloria Laycock, 309–19. London: Routledge.

Prichard, Jeremy, Paul Watters, and Caroline Spiranovic. 2011. "Internet Subcultures and Pathways to the Use of Child Pornography." *Computer Law & Security Review* 27 (6): 585–600. https://doi.org/10.1016/j.clsr.2011.09.009.

Reeder, Robert W., Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. 2018. "An Experience Sampling Study of User Reactions to Browser Warnings in the Field." In *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. New York: Association for Computing Machinery. https://doi.org/10.1145/3173574.3174086.

Riley, Donna M. 2006. "Beliefs, Attitudes, and Motivation." In *Handbook of Warnings*, edited by Michael S. Wogalter, 289–300. Mahwah, NJ: Lawrence Erlbaum.

Sarre, Rick, Lauri Yiu-Chung Lau, and Lennon Y. C. Chang. 2018. "Responding to Cybercrime: Current Trends." *Police Practice and Research* 19 (6): 515–18. https://doi.org/10.1080/15614263.2018.1507888.

Selejan, Ovidiu., Dafin F. Muresanu, Livia Popa, I. Muresanu-Oloeriu, Dan Iudean, Anca Dana Buzoianu, and Soimita Suciu. 2016. "Credibility Judgments in Web Page Design—A Brief Review." *Journal of Medicine and Life* 9 (2): 115–19.

Seto, Michael C., Lesley Reeves, and Sandy Jung. 2010. "Explanations Given by Child Pornography Offenders for Their Crimes." *Journal of Sexual Aggression* 16 (2) 169–80. https://doi.org/10.1080/13552600903572396.

Silver, N. Clayton, and Curt C. Braun. 1999. "Behavior." In *Warnings and Risk Communication*, edited by Michael S. Wogalter, David M. DeJoy, and Kenneth R. Laughery, 245–62. Philadelphia: Taylor & Francis. https://doi.org/10.1201/9780203983836.ch11.

Smith-Jackson, Tonya L. 2006. "Receiver Characteristics." In *Handbook of Warnings*, edited by Michael S. Wogalter, 335–44. Mahwah, NJ: Lawrence Erlbaum.

Smith-Jackson, Tonya L., and Michael S. Wogalter. 2000. "Users' Hazard Perceptions of Warning Components: An Examination of Colors and Symbols." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 44 (32): 6–55.

Steel, Chad M. S. 2015. "Web-Based Child Pornography: The Global Impact of Deterrence Efforts and Its Consumption on Mobile Platforms." *Child Abuse & Neglect* 44:150–58. https://doi.org/10.1016/j.chiabu.2014.12.009.

Testa, Alexander, David Maimon, Bertrand Sobesto, and Michel Cukier. 2017. "Illegal Roaming and File Manipulation on Target Computers: Assessing the Effect of Sanction Threats on System Trespassers' Online Behaviors." *Criminology & Public Policy* 16 (3): 689–726. https://doi.org/10.1111/1745-9133.12312.

Trommelen, Monica. 1997. "Effectiveness of Explicit Warnings." *Safety Science* 25 (1–3): 79–88. https://doi.org/10.1016/S0925-7535(97)00019-2.

Ullman, Joanne R., and N. Clayton Silver. 2018. "Perceived Effectiveness of Potential Music Piracy Warnings." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62 (1): 1353–57. https://doi.org/10.1177/1541931218621309.

Wathen, C. Nadine, and Jacquelyn Burkell. 2002. "Believe It or Not: Factors Influencing Credibility on the Web." *Journal of the American Society for Information Science & Technology* 53 (2): 134–44. https://doi.org/10.1002/asi.10016.

Westfall, Jacob, David A. Kenny, and Charles M. Judd. 2014. "Statistical Power and Optimal Design in Experiments in Which Samples of Participants Respond to Samples of Stimuli." *Journal of Experimental Psychology: General* 143 (5): 2020–45. https://doi.org/10.1037/xge0000014.

Wilson, Theodore, David Maimon, Bertrand Sobesto, and Michel Cukier. 2015. "The Effect of a Surveillance Banner in an Attacked Computer System: Additional Evidence for the Relevance of Restrictive Deterrence in Cyberspace." *Journal of Research in Crime and Delinquency* 52 (6): 829–55. https://doi.org/10.1177/0022427815587761.

Wogalter, Michael S., ed. 2006a. *Handbook of Warnings*. Mahwah, NJ: Lawrence Erlbaum.

———. 2006b. "Computer-Human Interaction Processing Model." In *Handbook of Warnings*, edited by Michael S. Wogalter, 51–61. Mahwah, NJ: Lawrence Erlbaum.

Wogalter, Michael S., John W. Brelsford, David R. Desaulniers, and Kenneth R. Laughery. 1991. "Consumer Product Warnings: The Role of Hazard Perception." *Journal of Safety Research* 22 (2): 71–82. https://doi.org/10.1016/0022-4375(91)90015-N.

Wogalter, Michael S., Vincent C. Conzola, and Tonya Smith-Jackson. 2002. Research-Based Guidelines for Warning Design and Evaluation. *Applied Ergonomics* 33 (3): 219–230. https://doi.org/10.1016/S0003-6870(02)00009-1.

Wogalter, Michael S., David M. DeJoy, and Kenneth R. Laughery. 1999. "Organizing Theoretical Framework: A Consolidated Communication-Human Information Processing (C-HIP) Model." In *Warnings and Risk Communication*, edited by Michael S. Wogalter, David M. DeJoy, and Kenneth R. Laughery, 15–23. Philadelphia: Taylor & Francis. https://doi.org/10.1201/9780203983836.ch2.

Wogalter, Michael S., and Thomas A. Dingus. 1999. "Methodological Techniques for Evaluating Behavioral Intentions and Compliance." In *Warnings and Risk Communication*, edited by Michael S. Wogalter, David M. DeJoy, and Kenneth R. Laughery, 53–81. Philadelphia: Taylor & Francis. https://doi.org/http://dx.doi.org/10.1201/9780203983836.ch4a.

Wogalter, Michael S., and Kenneth R. Laughery. 1996. "WARNING! Sign and Label Effectiveness." *Current Directions in Psychological Science* 5 (2): 33–37. https://doi.org/10.1111/1467-8721.ep10772712.

Wogalter, Michael S., and Christopher B. Mayhorn. 2008. "Trusting the Internet: Cues Affecting Perceived Credibility." *International Journal of Technology and Human Interaction* 4 (1): 75–93. https://doi.org/10.4018/jthi.2008010105.

Wogalter, Michael S., and N. C. Silver. 1990. Arousal Strength of Signal Words." *Forensic Reports* 3 (4): 407–20.

Wogalter, Michael S., and William J. Vigilante, Jr. 2006. "Attention Switch and Maintenance." In *Handbook of Warnings*, edited by Michael S. Wogalter, 245–65. Mahwah, NJ: Lawrence Erlbaum.

Wogalter, Michael S., Stephen L. Young, John W. Brelsford, and Todd Barlow. 1999. "The Relative Contributions of Injury Severity and Likelihood Information on Hazard-Risk Judgments and Warning Compliance." *Journal of Safety Research* 30 (3): 151–62. https://doi.org/10.1016/S0022-4375(99)00010-9.

Wortley, Richard. 2012. "Situational Prevention of Child Abuse in the New Technologies." In *Understanding and Preventing Online Exploitation of Children*, edited by Ethel Quayle and Kurt M. Ribisl. London: Routledge.

Wortley, Richard, and Stephen Smallbone. 2006. *Child Pornography on the Internet*. Washington, DC: US Department of Justice, Office of Community Oriented Policing Services.

———. 2012. *Internet Child Pornography: Causes, Investigation, and Prevention*. Santa Barbara, CA: Praeger.

Young, Stephen L. 1998. "Connotation of Hazard for Signal Words and Their Associated Panels." *Applied Ergonomics* 29 (2): 101–10. https://doi.org/10.1016/S0003-6870(97)00038-0.

Young, Stephen L., and Michael S. Wogalter. 1998. "Relative Importance of Different Verbal Components in Conveying Hazard-Level Information in Warnings." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 42 (15): 1063–67. https://doi.org/10.1177/154193129804201502.

Zaikina-Montgomery, Helen. 2011. "The Dilemma of Minors' Access to Adult Content on the Internet: A Proposed Warnings Solution." PhD diss., University of Nevada, Las Vegas, 2011. UNLV Theses, Dissertations, Professional Papers, and Capstones. https://digitalscholarship.unlv.edu/cgi/viewcontent.cgi?article=1945&context=thesesdissertations.

Alexandra Haddad is a practicing psychologist. She completed her masters of psychology (clinical) at the University of Tasmania in 2019. Alexandra's research interests focus on psychology and its intersections with law, including in the areas of child-exploitation materials and image-based abuse.

James Sauer, PhD, is a senior lecturer at the University of Tasmania. His works focuses on (a) the intersection of psychology and law, with a particular interest in memory and decision-making in investigative settings, and, more recently, on (b) the cognitive and behavioral effects on playing videogames, particularly games that include gambling-like reward mechanisms.

Jeremy Prichard, PhD, is an associate professor of criminal law at the University of Tasmania. He collaborates with multidisciplinary teams to conduct research on (a) prevention strategies for child sexual exploitation material and (b) monitoring illicit drug markets using wastewater analysis.

Caroline Spiranovic is a senior lecturer in law at the University of Tasmania, where she is currently working as part of three separate multidisciplinary teams on Australian Research Council–funded projects focusing on public opinion on sex-offender sentencing, prediction of sexual-offender risk, and methods of preventing viewing of child-exploitation material online. Caroline completed her PhD in psychology in 2007, focusing on child sex-offender typologies. Her research and teaching interests lie broadly in the disciplines of psychology, law, and criminology, with a focus on violent (particularly sexual) offending.

Dr. Karen Gelb is a consultant criminologist and a lecturer in the Department of Criminology at the University of Melbourne, Australia. During the past fifteen years of researching courts, she has written on a range of topics, including public opinion on sentencing, sex offenders, family violence, bail and remand, specialist courts, and youth justice. In addition to consulting, Dr. Gelb teaches a masters-level course on crime and criminal justice and writes about drug policy and harm reduction for Penington Institute in Melbourne.