Computers in Human Behavior 69 (2017) 421–436

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Fishing for phishers. Improving Internet users' sensitivity to visual deception cues to prevent electronic fraud

María M. Moreno-Fernández [a], Fernando Blanco [a], Pablo Garaizar [b], Helena Matute [a, *]

[a] Faculty of Psychology and Education, University of Deusto, Spain
[b] Faculty of Engineering, University of Deusto, Spain

ARTICLE INFO

Article history:
Received 4 March 2016
Received in revised form
12 December 2016
Accepted 18 December 2016
Available online 19 December 2016

Keywords:
Phishing
Internet security
Easy-to-hard effect
Human-computer interaction
Discrimination learning
Visual discrimination

ABSTRACT

Phishing is a form of electronic fraud in which attackers attempt to steal sensitive information by posing as a legitimate entity. To maintain the attack unnoticed, phishers typically use fake sites that accurately mimic real ones. However, there are usually subtle visual discrepancies between these spoof sites and their legitimate counterparts that may help Internet users to identify their deceptive nature. Among all the potential visual cues, we choose to focus on typography, because it is often hard for phishers to use exactly the same font as in the original website. Thus, Experiment 1 assessed the effectiveness of visual discrimination training to help people detect typographical discrepancies between fake and legitimate websites. Results showed higher sensitivity to differences when undergraduate students were previously trained with easier versions of the discrimination task (i.e., involving more noticeable differences in typography) than when they were trained with the difficult target discrimination from the start (*easy-to-hard effect*). These results were replicated with a broader and more representative sample of anonymous Internet users in Experiment 2. Implications for the design of strategies to prevent electronic fraud are discussed.

1. Introduction

Phishing is "a criminal mechanism employing both social engineering and technical subterfuge to steal consumers' personal identity data and financial account credentials" (Anti-Phishing Working Group, 2016, p. 2). Although phishers can use different strategies to reach their goals, in a typical scenario they pose as a reliable entity (e.g., trustworthy companies, acquaintances or even public bodies) and use e-mails as lures for driving Internet users to fraudulent websites. Deceitful websites are specifically designed to resemble the legitimate version, causing users to remain unaware of the fraud, and increasing their probability of being tricked.

Although not new, phishing has become an increasing threat for cyber-security. According to the Anti-Phishing Working Group (the leading international consortium of business, regulators, and agencies that monitor phishing attacks worldwide and attempt to coordinate responses to such attacks), the number of unique phishing websites detected during the 1st quarter of 2016 increased by 250% compared to the last quarter of 2015 (Anti-Phishing Working Group, 2016). Moreover, during the first semester of 2016 alone, a total of 102,573 submissions of suspected phishing attacks were verified as real (valid phishes) by the PhishTank community (OpenDNS, 2016). Given the scope of this threat and its consequences, a growing body of research has begun to explore how to prevent Internet users from being phished.

An extensive number of anti-phishing strategies have been developed, covering all stages of the phishing attack process, and using complementary approaches that range from technical to legal interventions (for a review, see Jakobsson & Myers, 2007; Mohammad, Thabtah, & McCluskey, 2015; Purkait, 2012). For example, a phishing attack can be detected at a very early stage (before it actually starts) by monitoring the registration of potential spoof domains, or by controlling unusual patterns of access to the legitimate website. The rationale of the latter approach rests on the necessity of phishers to repeatedly access the legitimate website to download and copy relevant contents to create the illegitimate version. The specific analysis of IP addresses associated with these unusual download activities may help to detect and react against an

* Corresponding author. Faculty of Psychology and Education, University of Deusto, Avda. Universidades 24, 48007 Bilbao, Spain.
E-mail address: matute@deusto.es (H. Matute).

http://dx.doi.org/10.1016/j.chb.2016.12.044

imminent phishing attack (Emigh, 2007). However, although predicting and blocking phishing activities at this early stage should be the optimal solution, it is not always possible.

When phishers succeed in launching the attack, the ideal strategy should be to prevent users from being exposed to the subsequent threat. With this aim in mind, a number of automatic detection strategies have been developed, ranging from blacklists of phishing domains (such as the Anti-Phishing Working Group blacklist, or the Google Safebrowsing service), to heuristic-based methods that can recognize phishing websites by analyzing their visual features (e.g., Liu, Deng, Huang, & Fu, 2006; Liu, Guanglin, Liu, Zhang, & Xiaotie, 2005; Maurer & Herzner, 2012; Medvet, Kirda, & Kruegel, 2008; Zhang, Liu, Chow, & Liu, 2011). But once again, although these technical approaches can be regarded as a good first line of defense against phishing, to date there is no strategy can completely prevent phishing attacks. Therefore, training users to detect fake websites and to protect themselves is currently a key component in cyber-security.

1.1. Human behavior and user-oriented approaches

The critical role of human behavior in the success of phishing attacks has encouraged the development of strategies aimed at promoting safer decisions across all the stages of the phishing attack flow in which human performance is involved. Some of these approaches have focused on teaching users to identify deception cues in phishing attack vectors such as e-mails whilst providing, at the same time, security tips (e.g., Anti-Phishing Phyllis™, Wombat Security Technologies, 2016; or PhishGuru, Kumaraguru et al., 2007). However, in addition to emails, phishers may currently use a wide range of strategies to lure users (e.g., messages posted on social media, phone calls, or SMSs). Therefore, designing preventive strategies to help users at successive stages of the attack (that is, once the illegitimate site is visited) becomes essential.

Client-side strategies such as security indicators (toolbars, warnings, or browser indicators) have been developed to signal trustworthiness or to alert users about potentially dangerous sites. Recent research has shown that warnings can effectively reduce people's likelihood of disclosing sensitive information on legitimate websites, although this reduction depends on the warning word used and on the identity information targeted (see Carpenter, Zhu, & Kolimi, 2014). Unfortunately, research on phishing also highlights the limited effectiveness of security indicators because people do not use them as expected. For example, Dhamija, Tygar, and Hearst (2006) carried out a laboratory study to assess the ability of Internet users to detect fraudulent websites, as well as the strategies that they used for judging website legitimacy. Participants were asked to categorize websites as legitimate or not, rating their confidence in their responses, and explaining the reasons underlying their choice. Results showed that even in a non-natural environment where participants were warned and primed about the possibility of being fooled, they could not distinguish accurately between spoof and legitimate websites (40% of participants' choices were incorrect). But what is probably more surprising is that browsers' warning cues such as address bars, status bars, or security indicators (e.g., lock icons in the address bar), went unnoticed by many participants.

Alsharnouby, Alaca, and Chiasson (2015) replicated and extended previous results in a more recent study using eye tracking. The authors used a procedure similar to the one used by Dhamija et al. (2006) but, in addition to behavioral measures and participants' self-reports, they included eye-tracking measures to obtain additional information about the user's attention to security cues. Their results confirmed that participants were not able to reliably identify fraudulent sites, spending most of the time examining the content of the website and paying little attention to security indicators (for similar results, see also Aburrous, Hossain, Dahal, & Thabtah, 2010; Lin, Greenberg, Trotter, Ma, & Aycock, 2011; Whalen & Inkpen, 2005; Wu, Miller, & Garfinkel, 2006). These studies reveal the essential role of human behavior in phishing success, and they highlight the relevance of considering human vulnerabilities when designing preventive strategies.

One main aspect of this vulnerability is the users' knowledge about security and security indicators. Users may not have enough information about these technical resources. For example, Wogalter and Mayhorn (2008) asked a group of participants to rate the extent to which they would trust the information of a website based on trustworthiness signals (i.e., domain suffixes, organization domain names, and quality seals that actually can be used as indicators of website reliability). The authors found that the reported trust on the website contents was related to these three indicators, but, surprisingly, participants showed limited abilities to discriminate between real and fictitious quality seals and organizational domain names. The lack of human competence at this level has raised the interest in educational approaches.

Educational strategies are primarily concerned with teaching the general concepts of cyber-security and phishing by using exercises to reinforce concepts, or by employing specific guided training protocols. However, whilst recent research has pointed out the value of these educational interventions (Kumaraguru et al., 2009, 2007; Sheng et al., 2007), there are other factors that may hinder the use of security indicators even when users do have enough knowledge about them. One of these factors is directly related to users' motivations when using the Internet and the awareness of the possibility of being tricked.

When using the Internet, users are mainly dedicated to their primary goals, that is, browsing web pages, trying to find a product on an e-commerce site, or just replying to their e-mails. Security is rarely their main goal, and consequently it is usually set aside. This "*unmotivated user property*" (Whitten & Tygar, 1999), together with other limitations imposed by human cognitive capacities that might affect decision-making (see Jones, Towse, & Race, 2015 for a review), pose a great challenge for web security as they may restrict the use of security tools.

Phishing attacks commonly profit from human confidence and the cognitive limitations of Internet users (see Dhamija & Tygar, 2005). Thus, scammers usually promote trust beliefs and judgments about legitimacy by simply creating websites that look as similar as possible to the originals, a strategy that becomes effective because of peoples' tendency to overlook security warnings (as discussed above). In this situation, it is important to develop additional strategies that take into account the flaws in human cognition, and their potential interaction with the effectiveness of anti-phishing measures.

A potential option is to increase alertness by improving users' sensitivity to visual deception cues whenever subtle differences exist between an original website and a fake site. If this were possible, websites requiring higher security measures, such as banks or health companies, could train their users to increase their ability to discriminate the original website from potential fakes. Although there are other levels of inconsistency that users might be trained to detect besides perceptual discrepancies (for example, on a procedural level users may be trained to detect credential login inconsistencies); this paper will explore the former approach to help Internet users protect their security by taking advantage of well-known research principles of human visual discrimination learning.

1.2. Improving Internet users' sensitivity to deception cues: Visual discrepancies between legitimate and spoofed websites

Visual design is a crucial aspect of branding. It aims, among other goals, to provide instant recognition of companies and to stimulate users' trust. Unfortunately, security may be easily compromised because of this confidence in visual cues, if phishers manage to recreate a visually similar version of the legitimate website. Accurate replication of the visual features of the original website is key to the success of a phishing attack. However, even though spoof sites usually resemble the target website, they are rarely a perfect copy of the legitimate version.

Spoof websites usually present misspellings and grammatical errors, as well as visual discrepancies related to, for example, typeface or design layout. These deviations from the legitimate version are useful cues for detecting deception, as they could be used to raise suspicions about website legitimacy. In fact, previous research has shown that spotting typographical errors on legitimate websites may damage their credibility (Fogg et al., 2001). Therefore, if typographical errors are detected, suspicions about illegitimacy may be raised, and users may then be motivated to further evaluate authenticity by actually checking the available security warnings (e.g., toolbars, icons) that they would otherwise ignore.

The lack of research in this area seems surprising given the relevance of human perception and behavior in the final decision of trusting a website, and the negative outcomes that may result from erroneous categorization. Thus, the aim of the present research was to develop an evidence-based strategy that could help users increase their sensitivity to visual deception cues, that is, their ability to detect perceptual discrepancies between a given website and the legitimate version of that site.

1.2.1. Visual cues of deception

The first step for developing a new strategy that could increase users' sensitivity to visual deception cues is to state which visual features can be used as deception cues. As mentioned above, detecting typographical errors may damage the credibility of a website even if it is actually legitimate. However, using typographical errors for increasing suspicions about legitimacy remains an unsuitable approach, as it requires literacy skills that may not be present in all users. Therefore, other visual features (instead of typographical errors) should be explored as potential deception cues.

Companies invest considerable effort and economic resources in designing a brand style that is easily recognizable by users. Although some of these features can be easily copied (e.g., logos or images), the branding process makes other features particularly difficult to emulate by phishers. One of the most interesting features at this level is the typeface.

When phishers try to replicate an original website, they are likely to have problems with using exactly the same typeface. First, identifying a specific typeface among thousands of similar ones is not an easy task. There are several websites providing this identification service automatically, but their success ratio is far from perfect. For this reason, phishers tend to use typefaces that mimic original typefaces with respect to - among other features - serif (i.e., endings of strokes), but not weight (i.e., thickness of the character outlines relative to their height), counter (i.e., the partially or fully enclosed space within a character), arm/leg (i.e., upper or lower stroke that is attached on one end and free on the other), leading (i.e., how text is spaced vertically in lines, measured from the baseline of each line of text where the letters “sit”), or kerning (i.e., distance between two letters). Thus, whilst some parameters of the typeface are mimicked, others are not. Moreover, it is usual for website designers to slightly modify typefaces, obtaining as a result a new, unique typeface with which customers can identify the brand, and which makes the typeface mimicking process even more difficult (or almost impossible) as the typeface used in the original site would be protected and not easily available for download.

There are additional reasons why typefaces could be revealing about the trustworthiness of a website. Nowadays, most professional website developers use the Web Open Font Format (WOFF) specification that allows the inclusion of the typeface to be an asset of the webpage. This means that a path to the typeface file is specified either in the CSS or in the JavaScript code, so that it is accessible to the browser when it renders the page. However, phishers tend to make verbatim copies of the HTML, CSS, and JavaScript code from the original page, because it is much easier than coding it manually. Thus, they often forget to include proper relative routes to the font assets, or fail to serve those font assets from a domain under the phisher's control. As required by the implementation of same-origin policy in modern browsers, dynamically downloaded assets must be in the same domain as the website. As a result, the browser often fails to access the intended font. When this happens, browsers fail gracefully, providing a similar typeface instead of the original. This creates a visual discrepancy with the original website that can be exploited to raise suspicions about the legitimacy of a spoof website. Consequently, even if the remaining visual features of the website can be emulated, phishers often fail at using the exact typeface used in the original site, providing, at best, a similar font whose difference from the original might be spotted.

Thus, the typefaces used in web pages remain a specific feature that may allow users to recognize a trusted site. However, and before continuing, it is important to note that typefaces are just one of many visual features that may disclose the deceptive nature of a website. Visual layout or color scheme may also be used for the same purpose, although they are probably less useful for security because of their dynamic nature. For example, layout is usually determined by screen proportions or website content (one of the most changing aspects of a website). In sum, we have reasons to choose typeface discrimination as our target feature to test our training strategy, as it is a good example of a difficult — and useful — skill that users should acquire. We assume that some companies might prefer to target a different aspect of their website, but we decided to start by using typeface as the target dimension in our experiments. If our procedure proved to be effective, then other perceptual inconsistencies could be trained in similar ways. For example, it may help to prevent homograph spoofing attacks in which phishers use non-ASCII Glyphs or numbers to create domain names that looks very similar to the legitimate ones (Gabrilovich & Gontmakher, 2002).

The question to be addressed now is: How could users learn to detect perceptual deviations in typeface between legitimate and spoof sites if these differences are usually very subtle and difficult to verbalize?

1.2.2. Training users to detect visual cues of deception

In practical situations, training is usually considered an effective strategy for improving discriminative skills, and therefore it could be a valuable tool in applied contexts where perceptual distinction is required. Consequently, this may be a suitable approach to increase sensitivity to visual deception cues. However, the effectiveness — or even the viability — of this strategy remains unexplored in the context of Internet security.

Long before the creation of the Internet, discrimination learning was established as a fruitful research area in Psychology laboratories, and for many years it has been applied to a large variety of

learning situations. Among the training procedures that may produce an enhancement of discriminative abilities, one has been shown to boost these abilities over simple discrimination training in several domains. This procedure is known as *transfer along a continuum* (Lawrence, 1952, 1955), and it essentially consists of starting the training program with an easier version of the discrimination task that overemphasizes the perceptual differences between the stimuli that are to be discriminated (i.e., training starts with two stimuli that are clearly discriminable). Additionally, progressively increasing the difficulty by including intermediate steps has been shown to be effective (Lawrence, 1952; Liu, Mercado, Church, & Orduña, 2008; Moreno-Fernández, Ramos-Álvarez, Paredes-Olay, & Rosas, 2012), but even the inclusion of just one easy condition before the difficult discrimination stage has been shown to facilitate learning (e.g., Arriola, Alonso, & Rodríguez, 2015; Scahill & Mackintosh, 2004; Suret & McLaren, 2003).

The facilitation derived from starting the discriminative training with an easier discrimination is known as the *easy-to-hard effect*, a reliable learning phenomenon that has been demonstrated across various species and sensorial modalities in many laboratories worldwide (e.g., Liu et al., 2008; Scahill & Mackintosh, 2004; Suret & McLaren, 2003). This reliability, together with the undemanding way in which the general procedure can be transferred to applied contexts, makes this training regime a potentially valuable tool for increasing discriminative abilities in practical situations. In fact, easy-to-hard protocols have already been used in applied contexts, such as language learning (Jamieson & Morosan, 1986, 1989; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002; Tallal, Miller, Bedi, Wang, & Nagarajan, 1996) and flavor evaluation (Moreno-Fernández et al., 2012), and they have also proven to be quite effective in enhancing discriminative abilities with complex stimuli in the visual domain (e.g., Suret & McLaren, 2003). This paper aims to explore the potential use of this procedure to the area of cyber-security.

1.3. Changes in sensitivity and strategy of response

So far, we have discussed the ability to detect differences between fake and original websites. However, as shown by Signal Detection Theory (SDT, Green & Swets, 1966), the tendency (or bias) to categorize a new site as fake or original could also be of interest. For example, in a laboratory situation, participants are primed to think about security, which may increase alertness about the possibility of being tricked, thus encouraging more cautious responses. Participants may adopt a biased response strategy to minimize the possibility of categorizing a fake website as an original one. This strategy may encourage a "safer" performance (in terms of avoiding being phished), but it will not reflect an actual improvement in the ability to detect spoofed sites.

In this regard, previous research has shown that anti-phishing educational strategies can not only affect the ability to identify phishing websites, but can also have an impact on the response strategy of the users (e.g., Kumaraguru, Sheng, Acquisti, Cranor, & Hong, 2010; Sheng et al., 2007). For example, Sheng et al. (2007) carried out a study to assess the effectiveness of an online game (*Anti-Phishing Phil*) aimed at teaching users how to identify phishing URLs, how to look for Web browsers' cues, and how to use search engines to avoid fraudulent sites. The authors compared the ability to discriminate between fake and legitimate websites in three groups of participants: A group that had played the game, a second group that only read the training material used in the game (without actually playing it), and a third group that read other tutorials about spoof e-mails and phishing. The results showed that all participants were more accurate in identifying fraudulent web sites after the intervention, but, interestingly, by using SDT analysis, the authors also found changes in response strategy, with participants being more cautious after reading training material but not after playing the game. These results illustrate that anti-phishing strategies may increase sensitivity (i.e., discriminative abilities), but they can also promote changes in the response strategy. Moreover, the fact that both sensitivity and response strategy can be affected by training stresses the relevance of carefully evaluating the contribution to each process, and the need to use a bias-free measure of discriminative abilities.

We are interested in knowing whether our easy-to-hard training strategy can improve sensitivity independently of the response strategy. Thus, we will use an analysis based on SDT to evaluate how training may affect discriminative abilities (i.e., sensitivity to perceptual differences that define the original and fake websites) independently of response strategy (i.e., the response tendency to select between categories), ensuring that our measures of discriminability are not contaminated by strategic/decisional changes in responding.

2. Experiment 1

Experiment 1 was designed with the main goal of assessing the effectiveness of a visual discrimination training protocol based on the easy-to-hard effect to improve website identification. Given the lack of previous research evaluating the easy-to-hard effect in this context, we chose a highly-controlled laboratory situation to assess the effect. This initial strategy allows for controlling methodological details (e.g., stimulus presentation, resolution, or external interferences while performing the task) and it can be understood as an appropriate preliminary step before assessing the effect in a more ecologically valid setting and with a broader sample (see Experiment 2).

2.1. Method

2.1.1. Participants

As stated previously, we decided to run the Experiment on a highly-controlled laboratory situation. This strategy forced us to use a sample of volunteers that would attend to the experimental session on campus. We chose a sample composed of undergraduate Psychology students. The effects of the easy-to-hard training on human performance have been frequently studied with undergraduate students (Church, Mercado, Wisniewski, & Liu, 2012; Pashler & Mozer, 2013; Suret & McLaren, 2003), and results with this collective usually parallel those reported with other human samples (e.g., May & MacPherson, 1971; McCandliss et al., 2002); and with non-human animals such as rats (e.g., Arriola et al., 2015; Lawrence, 1952), pigeons (Mackintosh & Little, 1970), rabbits (Haberlandt, 1971), octopuses (Sutherland, Mackintosh, & Mackintosh, 1963), or even honeybees (Walker, Lee, & Bitterman, 1990). Consistency of the easy-to-hard training effect on discriminative abilities among different samples, species and procedures confirms that it is a general and reliable effect, so we did not expect sample characteristics to modulate it with our procedure. However, it is important to note that this sample still represents a reduced portion of the target population (see Landers & Behrend, 2015 for a review of the topic). For this reason, Experiment 2 will use a different sample of Internet users.

A review of the few previously published papers that used visual stimuli and a procedure similar to the one we planned to use here (see Suret & McLaren, 2003; Pashler & Mozer, 2013) revealed that, overall, the beneficial effect of using an easy-to-hard training regime is of a large size (Cohen's d between 0.937 and 2.58) with samples that usually did not exceed $N = 40$. Therefore, we fixed the minimum simple size of our sample at $N = 40$.

Forty-two first-year Psychology students at the University of Deusto (33 women and 9 men, $M_{age} = 19$, age range: 18–30 years) volunteered for this experiment as an optional activity and in return for course credits. No exclusion criteria were used, and all data were included in the study.

Participants were tested in a large computer room, seated at least 1.5 m away from each other, and were randomly assigned to one of two groups. This resulted in 23 participants in the Easy-to-Hard (hereafter, ETH) group, and 19 participants in the Hard-to-Hard (hereafter, HTH) group.

2.1.2. Apparatus and stimuli

The experiment was presented as a web application based on World Wide Web Consortium (W3C) standards (i.e., HTML, CSS, and JavaScript). Stimuli were all presented in full-screen mode on 17-inch TFT displays, to ensure that all participants saw the stimuli at the same size.

We created an "original" bank website template with visual characteristics (general layout, color scheme) inspired by those of a real bank website, but without using logos, images, texts or any material that could resemble the real website. We then generated three different fake sites (F_1, F_2, and F_3) from the original site (O) by changing exclusively the typeface. The original website was written in Verdana (a sans serif font designed for computer screens and widely used on the web) and the fake sites were typed in Tahoma, Capriola, and Times New Roman.

Typefaces were selected taking into account their physical features. Thus, although Verdana has some features that are slightly dissimilar to Tahoma, which has smaller counters and reduced letter spacing, both typefaces are fairly similar as they belong to the same family, variant, and author. This ensures high perceptual resemblance between both websites but still allows for discriminating between them. The two additional fake sites were created with typefaces that are more dissimilar to Verdana than Tahoma. Like Verdana, Capriola is a sans-serif typeface, but it has moderate differences in the anatomy of the characters aimed at emulating handwriting (e.g., arms/legs). These differences are particularly noticeable in large sizes of the glyphs "G", "a", "g", "k", "e". Lastly, Times New Roman is by far the most distinct typeface and, unlike the others, it is the only one with serif, which involves significant differences in the whole anatomy of all characters. Therefore, in our procedure we had three fake websites with typefaces that differ along a continuum of similarity to Verdana: from the very similar Tahoma, to the very different Times New Roman. Before selecting the final materials, three anonymous judges were asked to rate the perceptual similarities between the different stimuli. Their ratings confirmed that Tahoma was the font most difficult to discriminate from Verdana, whereas Times New Roman was the easiest.

Screenshots of the login page of each website were used as stimuli. Panel A of Fig. 1 depicts an example of how the screenshots appeared on the computer screen. Additionally, the figure includes a larger scale copy of the text presented on each stimulus (panel B).

2.1.3. Procedure and design

Participants were informed that their involvement was voluntary and anonymous. We did not ask participants for any data that could compromise their privacy, nor did we use cookies or software to obtain such data. The stimuli and materials were harmless and emotionally neutral, the goal of the study was transparent, and the task involved no deception. The ethical review board of the University of Deusto examined and approved the procedure used in this research.

Volunteers were first asked to provide some basic demographic information (gender and age), and received a brief explanation about phishing. They were also advised that even if spoofed sites can be quite similar to their legitimate counterparts, they usually contain errors. After this general explanation, the main goal of the task was presented and participants were asked to categorize a series of screenshots as either belonging to an original website or to a fake site (detailed instructions translated from Spanish are presented in Appendix A). Participants were then advised to sit 50 cm away from the screen with their back laid on the chair to appropriately perceive the screenshots, and to click on a "Next" button (colored in blue and placed on the bottom right of the screen) to start the first trial.

The trial structure is depicted in Fig. 2. Each trial started with a fixation cross, presented in the center of the screen and replaced, after one second, by one website screenshot, either the original, O, or a fake, F. After three seconds, the image disappeared and the question "Do you think that the picture belongs to the original or to a fake website?" was presented. Participants had to answer by clicking on one of two buttons that appeared below the question, labelled as "Original" or "Imitation". The trial ended with a feedback screen in which participants received information about their response, and the actual category of the picture:

> [Original feedback] "CORRECT/INCORRECT, the picture belongs to the ORIGINAL website"
> [Imitation feedback] "CORRECT/INCORRECT, the picture belongs to an IMITATION"

The design of the experiment is presented in Table 1. The experiment had two phases: training and test. The training phase was organized into two blocks. Each block consisted of 12 trials, half of them showing the picture belonging to the original website (O) and the other half showing the picture belonging to a fake site (F). Trials within each block were presented in random order. To prevent fatigue, a break was provided after each training block. During the break, the sentences "Preparing more pictures. You can take a few seconds to rest" appeared at the center of the screen for 20 s. Participants were required to click on a "Next" button to continue after the break. The test phase maintained the same structure (two blocks of 12 trials each) but with no breaks between blocks.

Once the experiment had finished, participants were asked to complete some questions about Internet usage, online banking experience, and prior knowledge about phishing (see Appendix B).

The two groups differed only in the stimuli used as fake websites during the Training Phase. For participants in the ETH group, the Training Phase started with the two stimuli with less similar typefaces, F_1 and O. Categorization was made progressively more difficult in this group by comparing pairs of stimuli that were increasingly similar in each block: O vs. F_1 (Training Block 1), O vs. F_2 (Training Block 2), and finally, O vs. F_3 (Test Blocks 1 and 2). In contrast, participants in the HTH group were trained in all blocks, and then tested, with the most difficult-to-discriminate pair of stimuli, O vs. F_3 (i.e., the target discrimination).

We expect training with easier versions of the task at the beginning of the experiment to improve discriminative abilities in comparison with simple training with the target (hardest) discrimination. As a result, participants in the ETH group are expected to categorize the materials better in the test phase than those in the HTH group, that is, they are expected to show higher sensitivity. However, it should be noted that participants in the ETH group experience a change in stimuli between phases that requires them to adapt their categorization strategy at the beginning of the test phase. For this reason, the first block of the test phase should be considered as an additional training block for this group. Consequently — and even if participants in the ETH group are expected to perform better than those in the HTH group — additional training with the new (harder) discrimination may be required before any

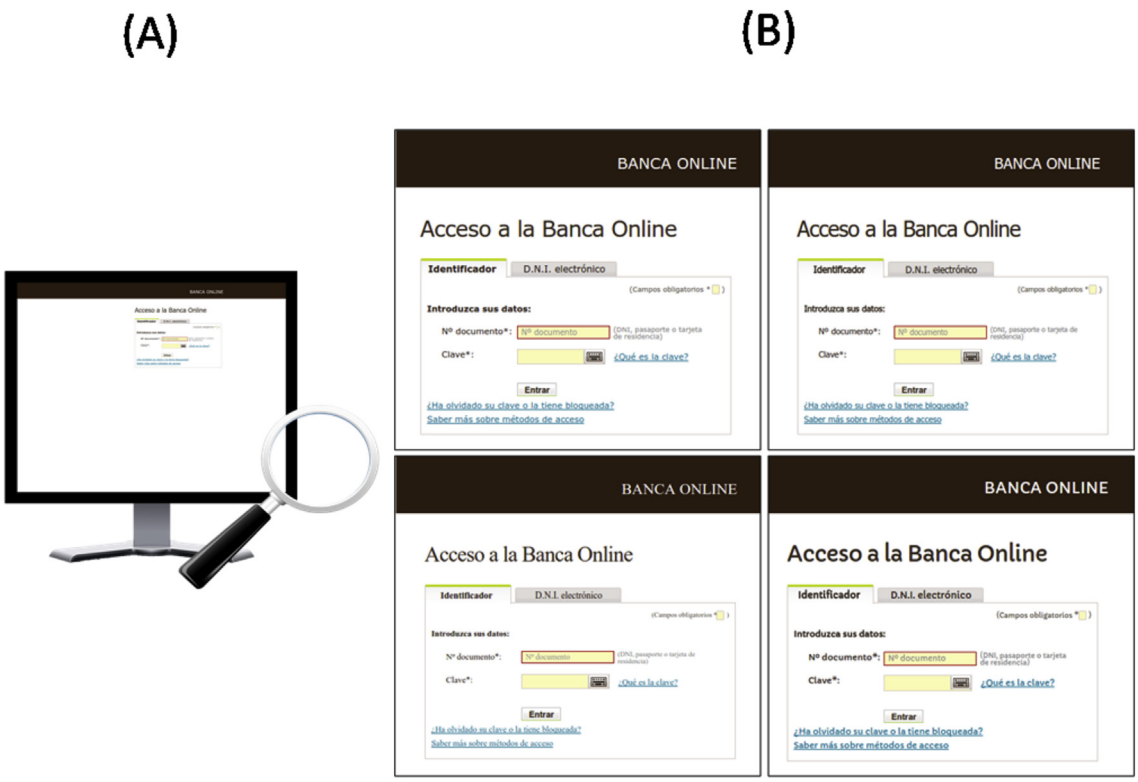

Fig. 1. Example of stimulus presentation (Panel A), and enlargement of text details from the stimulus set (Panel B, from top-left, clockwise): O (Original Verdana), F_3 (Fake Tahoma) F_2 (Fake Capriola), and F_1 (Fake Times New Roman).

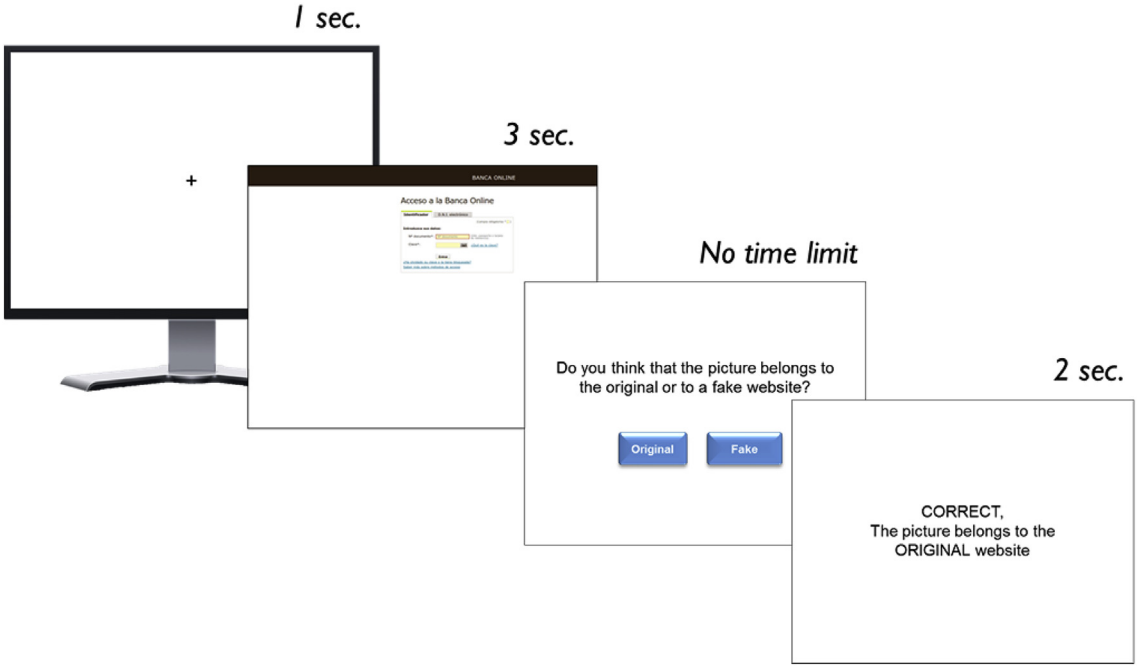

Fig. 2. Trial structure. From left to right: Fixation cross, stimulus (screenshot), question, and feedback.

improvements can be detected. Therefore, differences between groups are specifically predicted to emerge on the second block of test trials.

The main goal of this research is to assess the effects of perceptual training on discriminative abilities. However, and thanks to the SDT analytic approach, we can also evaluate the effects of each training regime on response strategy. As discussed previously, changes in difficulty are expected to affect discriminative abilities, causing groups to differ at this level (i.e., on sensitivity measures). However, it is less clear whether these changes will also

Table 1
Experimental design.

Group	Training phase		Test phase	
	Block 1	Block 2	Block 1	Block 2
Easy-to-Hard (ETH)	O, F_1	O, F_2	O, F_3	O, F_3
Hard-to-Hard (HTH)	O, F_3	O, F_3		

Note. The picture of the Original website is represented as O, and those extracted from Fake sites are depicted as F. Each block included 6 O trials and 6 F trials (i.e., 12 trials), presented in random order. Numerical subscripts refer to the discrimination difficulty, as determined by the typeface of stimulus F: Easy (F_1), Medium (F_2), and Hard (F_3). Feedback about responses and the actual category of each picture was provided on each trial.

simultaneously affect the response strategy. Measures of response strategy will allow us to describe the performance and training effects of both hard-to-hard and easy-to-hard schedules at this strategic level.

2.1.4. Measures

2.1.4.1. Number of correct responses. We used the number of correct responses on the categorization task as a measure of discriminative abilities (i.e., accurately assigning each picture to the category to which it belongs). This strategy has been used in previous research on the easy-to-hard effect with visual stimuli (e.g., Pashler & Mozer, 2013) as well as in phishing research (e.g., Dhamija et al., 2006). Split-Half reliability between the number of correct responses in Test Block 1 and 2 was calculated, Spearman-Brown Coefficient = 0.802.

2.1.4.2. Sensitivity. As previously noted, the number of correct responses can be influenced by strategical factors. Consequently, a purer measure of discriminative abilities, uncontaminated by these strategical aspects, should be more appropriate than the number of correct responses. Therefore, we calculated an unbiased measure of discriminative abilities using a SDT approach (Green & Swets, 1966) which has been valuable in phishing research (e.g., Kumaraguru et al., 2010; Sheng et al., 2007).

Computing SDT measures requires analyzing the performance while considering two aspects: Accuracy of each response (correct or incorrect) and trial type (in terms of SDT, whether it is a "signal + noise" trial or a "noise" trial). Easy to hard training (ETH) was aimed at improving users' abilities to discriminate specific visual features of the legitimate website (signal) among other features that are not relevant for the discrimination (noise), such as color or content; therefore, we treated screenshots from the original website as the "signal + noise" stimulus, and screenshots from the fake sites as "noise" stimulus. Note that the choice about which stimulus is considered as "signal + noise", and which one is considered "noise" is based on the particular characteristics of the procedure. Following the SDT approach, trial-wise performance was classified as hits (correctly categorizing an original screenshot), correct rejections (correctly categorizing a fake screenshot), misses (incorrectly categorizing an original screenshot as fake) and false alarms (incorrectly categorizing a fake screenshot as original), that is, we took into account participants' responses and the trial type. To avoid the effect of extreme values on the calculations, a log-linear correction was applied by adding a constant value of 0.5 to hits and false alarms rates (Brown & White, 2005).

The reduced number of trials on each block makes it difficult to meet the assumptions required to compute the SDT parametric index for sensitivity (d'), so we adopted a nonparametric SDT approach (see Moreno-Fernández et al., 2012 for a similar strategy with easy-to-hard training). Thus, the sensitivity index A′ (computationally developed by Grier, 1971; see also Stanislaw & Todorov, 1999) was calculated for each participant on each block of trials, using corrected rates of hits (H) and false alarms (F), according to Equation (1):

$$A' = 0.5 + \text{sign}\ (H - F)\frac{(H - F)^2 + |H - F|}{4\ \text{Max}\ (H, F) - 4HF} \quad (1)$$

The A′ index takes on values from 0 to 1. A value of 0.5 reflects chance-level performance (that is, null ability to discriminate between original and fake pictures), while a value of 1 reflects a perfect performance. Specifically for this task, and after correcting hits and false alarm rates, A′ values may range from 0.04 to 0.96. Split-Half reliability between A′ in Test Block 1 and 2 was calculated, Spearman-Brown coefficient = 0.731.

2.1.4.3. Response strategy. Response strategy was also measured by using a SDT approach (Green & Swets, 1966; see Kumaraguru et al., 2010; Sheng et al., 2007 for applications in phishing research). Specifically, the non-parametric index B_D (Donaldson, 1992) was calculated for each participant on each block of trials, using corrected rates of hits (H) and false alarms (F), according to Equation (2):

$$B_D = \frac{(1 - H)(1 - FA) - (HF)}{(1 - H)(1 - FA) + (HF)} \quad (2)$$

Response strategy B_D values may vary between −1 and +1, the neutral level being zero. Negative values reflect a lenient/liberal strategy (i.e., reporting a picture as belonging to the original website with the minimum evidence) while positive values show a conservative/strict strategy (i.e., reporting a picture as belonging to the original website only when the evidence is strong). This latter strategy will be interpreted as a "cautious" strategy. Specifically for this task, and after correcting hits and false alarm rates, B_D values may range from −0.99 to 0.99. Split-Half reliability between B_D in Test Block 1 and 2 was calculated, Spearman-Brown coefficient = 0.190.

2.1.4.4. Internet usage. Internet usage has been assessed in empirical research about users' trust on websites by using self-reported measures (e.g., Dhamija et al., 2006; Sheng et al., 2007). We also used this strategy, and asked participants to answer the question "How often do you use the Internet?" by using a 5-point Likert-type scale ranging from 1, (Less than once a month) to 5 (Almost every day). See Appendix B for further details.

2.1.4.5. Experience with bank sites. Experience with e-banking was also measured with a self-reported measure (see Alsharnouby et al., 2015 for a similar strategy). We asked participants to report how often they visited bank sites by using a 5-point Likert-type scale like the one used for Internet usage (see also Appendix B).

2.1.4.6. Previous knowledge about phishing. We asked participants to indicate if they knew about phishing before taking part in the research by using the following dichotomous (i.e., yes-no) question: "Before taking part in this study, did you know about phishing?" (see Dhamija et al., 2006 for a similar approach but using a semi-structured interview).

2.2. Results and discussion

Descriptive statistics and a correlation matrix of study variables appear in Table 2.

Table 2
Descriptive statistics and a correlation matrix of Experiment 1.

	M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Age	19.07	2.22	–																								
2. Gender	.21	.42	−.12																								
3. Group	.45	.50	−.12	.11																							
4. Internet usage	4.81	.51	−.01	.08	−.13																						
5. Bank site experience	1.40	.63	.36*	−.06	.10	.16																					
6. Phishing previous knoledge	.29	.46	.34*	−.07	.06	.11	.14																				
7. Correct responses (Training Block 1)	6.21	2.19	−.07	.22	−.31*	.14	.08	−.01																			
8. Correct responses (Training Block 2)	7.43	2.89	−.02	−.02	−.45**	.02	.05	−.04	.28																		
9. Correct responses (Test Block 1)	7.55	2.62	.22	−.13	−.25	−.06	.14	.03	.31*	.44**																	
10. Correct responses (Test Block 2)	7.48	2.92	.04	−.05	−.35*	−.02	.02	−.05	.45**	.67**	.67**																
11. Corrected Hits rate (Training Block 1)	.54	.20	−.11	.19	−.08	.10	.14	.02	.62**	−.01	.18	.26															
12. Corrected False Alarmas rate (Training Block 1)	.51	.25	.00	−.12	.33*	−.10	.02	.04	−.78**	−.36*	−.24	−.37*	.01														
13. Corrected Hits rate (Training Block 2)	.69	.19	.10	.01	−.46**	−.05	.14	−.07	.17	.85**	.41**	.59**	.12	−.11													
14. Corrected False Alarmas rate (Training Block 2)	.48	.27	.10	.03	.37*	−.09	−.04	.01	−.31*	−.93**	−.38*	−.61**	.10	.47**	−.59**												
15. Corrected Hits rate (Test Block 1)	.69	.20	.24	−.21	−.12	−.16	.21	.08	.19	.26	.82**	.46**	.19	−.09	.38*	−.13											
16. Corrected False Alarmas rate (Test Block 1)	.47	.24	−.14	.03	.29	.00	−.01	.02	−.32*	−.47**	−.88**	−.66**	−.13	.30	−.32*	.48**	−.45**										
17. Corrected Hits rate (Test Block 2)	.65	.22	.14	−.09	−.38*	−.14	.09	.05	.27	.53**	.44**	.83**	.24	−.16	.56**	−.41**	.36*	−.38*									
18. Corrected False Alarmas rate (Test Block 2)	.44	.27	.05	.00	.24	.05	.08	.12	−.48**	−.62**	−.69**	−.89**	−.22	.44**	−.46**	.62**	−.42**	.72**	−.47**								
19. A' (Training Block 1)	.51	.22	−.06	.22	−.28	.14	.05	−.02	.99**	.29	.27	.44**	.62**	−.76**	.16	−.32*	.15	−.30	.26	−.48**							
20. A' (Training Block 2)	.61	.26	−.06	.05	−.42**	.04	.05	−.05	.27	.98**	.39*	.67**	−.03	−.37*	.82**	−.93**	.19	−.45**	.51**	−.62**	.29						
21. A' (Test Block 1)	.63	.24	.21	−.15	−.22	−.07	.15	.02	.27	.37*	.98**	.60**	.16	−.22	.35*	−.31*	.80**	−.86**	.35*	−.65**	.24	.32*					
22. A' (Test Block 2)	.62	.25	.00	−.08	−.33*	.00	.02	−.08	.44**	.65**	.64**	.99**	.25	−.35*	.56**	−.59**	.44**	−.63**	.81**	−.88**	.43**	.65**	.58**				
23. BD (Training Block 1)	−.07	.56	.06	.04	−.22	−.02	−.06	−.02	.31*	.31*	.16	.18	−.53**	−.82**	.04	−.44**	.01	−.24	.02	−.27	.29	.33*	.14	.17			
24. BD (Training Block 2)	−.33	.41	−.27	−.11	−.07	.12	−.09	.11	.19	.26	.04	.15	−.23	−.43**	−.26	−.58**	−.18	−.21	.00	−.23	.20	.28	.01	.15	.48**		
25. BD (Test Block 1)	−.30	.38	.02	.20	−.16	.04	−.15	−.14	.14	.16	.10	.21	−.03	−.20	−.06	−.29	−.46**	−.53**	.10	−.25	.15	.21	.08	.18	.20	.31*	
26. BD (Test Block 2)	−.15	.41	−.15	.07	.10	−.08	−.14	−.21	.28	.14	.39**	.22	.02	−.34*	−.02	−.23	.24	−.41**	−.34*	−.62**	.27	.16	.41**	.23	.28	.21	.10

Note. Gender was coded 0 = Female, 1 = Male; Group was coded 0 = ETH, 1 = HTH; Phishing previous knowledge was coded 0 = No, 1 = Yes; *$p < 0.05$. **$p < 0.01$.

2.2.1. Internet usage, experience with bank sites and knowledge about phishing

First, we analyzed the data about Internet usage and habits to ensure that there were no differences between the experimental groups. Participants reported using the Internet no less than once a week (85.7% reported using the Internet almost every day, 9.5% four or five times a week, and the remaining 4.8% once or twice a week). However, they also reported having limited experience with bank sites (66.7% of them reported visiting bank sites less than once a month, 26.2% once or twice a month, and 7.1% once or twice a week), and most of them were unaware about phishing before taking part in the study (only 12 out of 42 participants reported having prior knowledge of this scam). Given the non-continuous nature of the dependent variable (i.e., Likert-type categories, and dichotomous yes/no question), we used non-parametric analyses to assess the differences between groups. No differences were detected in terms of general Internet usage, Mann-Whitney $U = 192.5$, $Z = -1.081$, $p = 0.28$, $r = -0.167$; bank site experience, $U = 248.5$, $Z = 0.915$, $p = 0.36$, $r = 0.141$, or prior knowledge about phishing, $\chi^2(1) = 0.154$, $p = 0.695$, $\Phi = 0.061$.

2.2.2. Discriminative abilities after training

The main goal of the experiment was to assess whether the easy-to-hard training given to the ETH group improved discriminative abilities compared with simple discriminative training (HTH group). The critical data are those of the test phase, in which all participants were tested with the same stimuli, corresponding to the most difficult discrimination (O vs. F_3) and, particularly, data from the second block of trials.

The number of correct responses on the test phase for each group is congruent with our hypothesis, since participants in the ETH group achieved a higher number of correct responses ($Mdn = 8$ in both Test blocks) than those in the HTH group ($Mdn = 6$ in both Test blocks). However, we have already noted that by comparing the number of correct responses in each group we cannot dissociate the effects of training on discriminative abilities and response strategy. Therefore, and to truly assess the specific effect of training on discriminative ability, we used the sensitivity index A′ (Fig. 3 depicts A′ Box-Plots for each group in each block of trials).

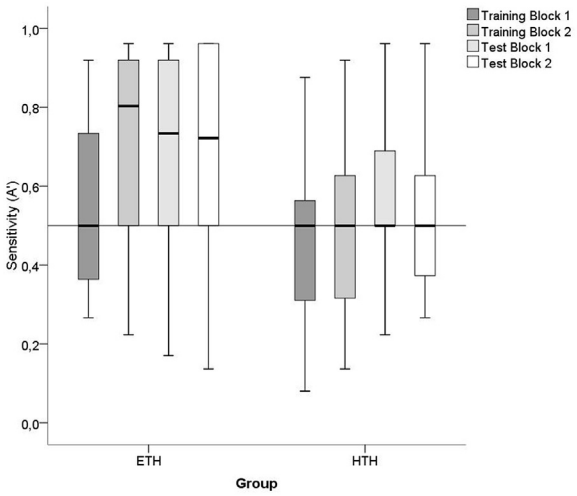

Fig. 3. Box-plots showing the sensitivity index A′ on each block of trials for participants in the ETH (left panel) group and the HTH group (right panel) in Experiment 1. Boxes represent the middle 50% of the data (interquartile range, IQR) and whiskers represent the largest and lowest cases within 1.5 times the IQR. The dark line within each box represents the median. The horizontal line is set at chance level (i.e., responding correctly on half of the trials).

Both the small sample size and the violation of the normality assumption suggest that the non-parametric approach is a more suitable option for analyzing our data. As expected, no significant differences between groups were found on the first block of the Test Phase, Mann-Whitney $U = 161.5$, $Z = -1.450$, $p = 0.147$, $r = -0.224$.

The lack of differences at the beginning of the test phase may be reasonably explained, as stated previously, by considering the change of stimulus (into a harder one) that occurs only in the ETH group. Nevertheless, differences were found on the second block of trials, Mann-Whitney $U = 137.5$, $Z = -2.073$, $p = 0.038$, $r = -0.320$ (a medium sized-effect). This indicated that ETH training worked in the way that we anticipated, with the participants showing enhanced discriminative performance after progressive training — a clear easy-to-hard effect.

2.2.3. Additional analyses

Although our main focus is on the changes in discriminative abilities in the test phase, in which the two groups are comparable, we also include additional analysis concerning the performance through the training session, and the response strategy.

For the training phase, a quick inspection of the HTH group's performance (right panel of Fig. 3) shows that this group maintained their sensitivity around the chance level throughout the entire session (note that the median score was 0.5 in all four blocks), which indicated that discrimination between O and F_3 was rather difficult. In contrast, the ETH group showed a poor performance at the beginning of the training phase ($Mdn = 0.5$); but A′ indexes became higher on the remaining blocks of trials, with data mainly distributed above chance level on the last three blocks of trials ($Mdn = 0.803$, 0.734 and 0.722, respectively for Training Block 2, Test Block 1, and Test Block 2).

At this point, it is important to mention the artificial nature of the to-be-learned categories, which is particularly relevant for interpreting the initial performance of the ETH group. At the beginning of the training phase, ETH participants are required to categorize the easiest-to-discriminate pair of stimuli. However, to perform appropriately they first need to learn which properties define each category (original or fake). Therefore (and even if stimuli on the first block of trials are easy to discriminate in this group, ETH), participants need to learn which category (either original or fake) corresponds to each screenshot. This may explain the poor performance of this group at the beginning of the session, even with the easiest discrimination.

We have already noted that some educational strategies may induce changes in the response strategy in addition to — or instead of — the changes in sensitivity. Thanks to the SDT approach, we can examine these two components separately.

Fig. 4 displays response strategy indexes (B_D) for each group on each block of trials. Participants showed a similar response strategy on the test Phase regardless of their training schedule (i.e., ETH or HTH training), and no differences between groups were detected either on Test Block 1, Mann-Whitney $U = 184$, $Z = -0.887$, $p = 0.375$, $r = -0.137$; or on Test Block 2, Mann-Whitney $U = 231.5$, $Z = 0.343$, $p = 0.731$, $r = 0.053$.

We did not find evidence of a cautious strategy in our participants (i.e. conservative response strategy), but rather the opposite. Note, for example, that response strategy indexes (B_D) were clearly distributed below zero on Test Block 1 in both groups (IQR boxes did not exceed the horizontal line set at zero, and median performance was in both cases lower that zero, −0.286 in both groups). This tendency can still be observed on the last block of Test trials, although median performance in this block was higher than in the previous block ($Mdn = 0$, in both groups). These results show that, independently from their training regime, participants started the

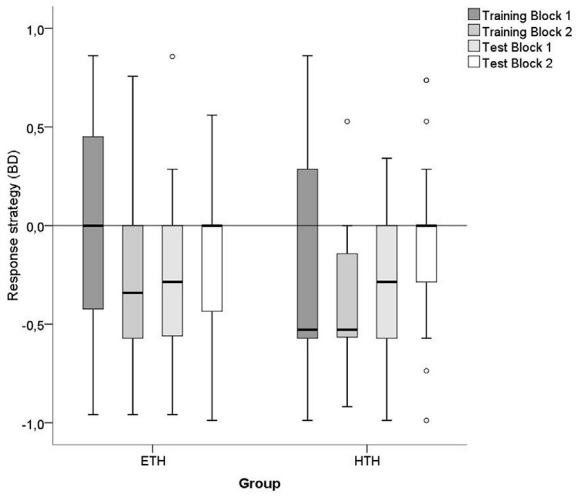

Fig. 4. Box-plots showing the response strategy index B_D on each block of trials for participants in the ETH (left panel) group and the HTH group (right panel) in Experiment 1. Boxes represent the middle 50% of the data (interquartile range, IQR) and whiskers represent the largest and lowest cases within 1.5 times the IQR. The dark line within each box represents the median, and the dots outside the boxes depict outliers (i.e., scores departing from the first/third quartile more than 1.5 times the IQR), which were neither removed, nor transformed. The horizontal line is set to represent a non-biased strategy, whilst values below this line show a lenient strategy (i.e., tendency to categorize websites as original) and values over the line show a conservative strategy (i.e., tendency to categorize websites as fake).

test phase preferentially categorizing pictures as belonging to the original website rather than to the fake site (see Canfield, Fischhoff, & Davis, 2016 for similar results on a detection task).

3. Experiment 2

Experiment 1 tested the possibility of using the easy-to-hard effect to increase sensitivity in a fake/original website categorization task. The results suggested that this strategy could indeed work to enhance participants' discriminative abilities. We assessed the differences between ETH and HTH training regimes in a laboratory situation, so that we can retain control over relevant aspects of the procedure (such as stimulus size, etc.). However, this strategy forced us to use a sample composed of undergraduate Psychology students that represent only a reduced portion of the target population (see Landers & Behrend, 2015 for a review of some limitations).

To better assess the generality and reliability of the effect, a second experiment was conducted to explore the effect of the easy-to-hard training in a less controlled situation, with a broader sample, and in a more representative context: The Internet. We expect to replicate the main results of Experiment 1.

3.1. Method

3.1.1. Participants

Data collected from anonymous Internet users are expected to be much noisier than those collected in a controlled laboratory situation. Among other important aspects, the size in which stimuli are presented cannot be controlled when running the experiment on the Internet. Screen sizes may vary considerably between users, but additionally they may even zoom in on the screen to explore the pictures. Further, they may be distracted during the task, or could perform the experiment in a noisy environment. Given that attention, noise, and perceptual characteristics of the stimuli may vary between subjects in an uncontrolled way, we decided to increase the sample size to at least 80 participants (i.e. twice as many as in Experiment 1). Additionally, and given that participants were anonymous Internet users, reporting previous participation was stablished as exclusion criteria (that is, we excluded those participants who reported having participated in the previous experiment). This strategy was used to avoid the potential effects of including data from those students that had already taken part in Experiment 1. Data collection went faster than expected, requiring less than two weeks to reach completion, eventually exceeding the sample size that we decided a priori.

One hundred and thirty eight anonymous Internet users voluntarily participated in the online experiment by accessing our virtual laboratory from an open link distributed through social media: We posted the link on Twitter, only mentioning that it was an experiment about phishing and asking for retweets (see Landers & Behrend, 2015 for some potential limitations of this sampling strategy). Data of three participants were not included in the study because they reported having taken part in a similar experiment previously, and two additional participants were also excluded for reporting ages below eighteen years. Therefore, the final sample included 133 participants (46 women and 87 men), ranging from 18 to 66 years old ($M_{age} = 38$).

As in Experiment 1, the computer program randomly assigned participants to the ETH and HTH groups, resulting in 72 participants in the former group and 61 in the latter.

3.1.2. Stimuli

Like the previous experiment, this study was also presented as a web application based on World Wide Web Consortium (W3C) standards (i.e., HTML, CSS, and JavaScript). No restrictions about hardware were set. The stimuli were the same as those used in Experiment 1.

3.1.3. Procedure and design

Procedure and design were the same as those used in Experiment 1. However, as mentioned before, Experiment 2 was carried out on the Internet (see Germine et al., 2012; Ryan, Wilde, & Crist, 2013 for the validity of web-based experiments in this context, and Vadillo, Bárcena, & Matute, 2006; Vadillo & Matute, 2011, for the validity of online experiments on associative learning very similar to the ones reported herein). We also included an additional question about previous participation, to ensure that participants of Experiment 1 did not take part in this study. Thus, after presenting the instructions and before the first trial a check box with the sentence “If after reading these instructions you remember having taken part in a similar experiment about websites, please check this box” appeared so that participants may report previous participation.

3.1.4. Measures

We used the same measures described for Experiment 1. Split-Half reliability between performance in Test Block 1 and 2 for the number of correct responses, sensitivity and response strategy indexes were also calculated, Spearman-Brown Coefficient = 0.794, 0.697, and 0.716 respectively for each measure.

3.2. Results and discussion

Descriptive statistics and correlation matrix of study variables appear in Table 3.

3.2.1. Internet usage, experience with bank sites and knowledge about phishing

Data about Internet usage and habits showed that participants used the Internet no less than four or five times a week (98.5%

Table 3
Descriptive statistics and a correlation matrix of Experiment 2.

	M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Age	37.83	9.64																									
2. Gender	.65	.48	.01																								
3. Group	.46	.50	−.11	−.60																							
4. Internet usage	4.98	.12	−.03	.04	.11																						
5. Bank site experience	2.83	1.26	.06	.13	−.09	.20*																					
6. Phishing previous knoledge	.92	.28	.15	.24**	.17	.19*	.06																				
7. Correct responses (Training Block 1)	6.98	2.34	.12	.14	−.34**	−.01	.04	.03																			
8. Correct responses (Training Block 2)	7.65	2.55	−.14	.10	−.39**	.02	.02	−.02	.44**																		
9. Correct responses (Test Block 1)	7.60	2.55	−.01	−.02	−.15	.01	−.10	.08	.26**	.51**																	
10. Correct responses (Test Block 2)	8.08	3.04	−.03	.08	−.19*	.10	−.06	.08	.28**	.45**	.66**																
11. Corrected Hits rate (Training Block 1)	.61	.23	.21*	.12	−.28**	.03	.15	.05	.66**	.25**	.18*	.18*															
12. Corrected False Alarmas rate (Training Block 1)	.47	.25	.04	−.08	.19*	.06	.06	.00	−.72**	−.35**	−.19*	−.20*	.05														
13. Corrected Hits rate (Training Block 2)	.69	.23	−.01	.10	−.44**	.01	.12	−.10	.37**	.68**	.34**	.30**	.48**	−.05													
14. Corrected False Alarmas rate (Training Block 2)	.45	.27	.18*	−.05	.15	.00	.07	−.05	−.28**	−.78**	−.40**	−.35**	.07	.44**	−.08												
15. Corrected Hits rate (Test Block 1)	.67	.23	.12	−.06	−.24**	−.01	.01	−.01	.23**	.33**	.69**	.41**	.33**	.00	.55**	.01											
16. Corrected False Alarmas rate (Test Block 1)	.45	.26	.12	−.02	.00	−.07	.11	−.12	−.17	−.42**	−.79**	−.56**	.04	.26**	−.01	.56**	−.10										
17. Corrected Hits rate (Test Block 2)	.70	.24	.03	.02	−.24**	.12	.07	.00	.26**	.35**	.51**	.73**	.33**	−.04	.53**	−.02	.66**	−.14									
18. Corrected False Alarmas rate (Test Block 2)	.40	.30	.06	−.10	.08	−.10	.15	−.11	−.19*	−.36**	−.54**	−.84**	.00	.25**	.00	.49**	−.07	.68**	−.25**								
19. A' (Training Block 1)	.59	.23	.12	.11	−.31**	−.01	.04	.03	.98**	.41**	.24**	.26**	.67**	−.68**	.35**	−.25**	.20*	−.16	.24**	−.17*							
20. A' (Training Block 2)	.64	.24	−.16	.10	−.39**	.00	.03	−.04	.40**	.98**	.45**	.40**	.24**	−.31**	.68**	−.75**	.28**	−.38**	.31**	−.32**	.38**						
21. A' (Test Block 1)	.64	.23	−.01	−.04	−.16	.00	−.10	.09	.24**	.48**	.98**	.61**	.17	−.16	.34**	−.36**	.68**	−.76**	.48**	−.49**	.21*	.44**					
22. A' (Test Block 2)	.66	.26	−.02	.09	−.19*	.10	−.06	.08	.27**	.41**	.58**	.98**	.17*	−.20*	.28**	−.31**	.36**	−.49**	.73**	−.82**	.24**	.36**	.53**				
23. BD (Training Block 1)	−.12	.56	−.20*	−.02	.00	−.08	−.10	−.05	.12	.13	.00	.04	−.64**	−.74**	−.23**	−.37**	−.20*	−.17	−.15	−.18*	.08	.11	−.02	.05			
24. BD (Training Block 2)	−.24	.52	−.11	−.06	.24**	.00	−.14	.10	−.11	.12	.05	.04	−.39**	−.22*	−.61**	−.67**	−.35**	−.37**	−.32**	−.31**	−.10	.09	.03	.02	.38**		
25. BD (Test Block 1)	−.21	.50	−.15	.02	.21*	.08	−.03	.11	−.07	.05	.12	.12	−.22**	−.12	−.34**	−.35**	−.60**	−.68**	−.32**	−.42**	−.05	.06	.11	.11	.18*	.47**	
26. BD (Test Block 2)	−.16	.53	−.02	.06	.11	.03	−.17*	.11	−.05	.01	.11	.26**	−.19*	−.11	−.35**	−.31**	−.35**	−.46**	−.43**	−.71**	−.05	.00	.09	.25**	.17	.43**	.56**

Note. Gender was coded 0 = Female, 1 = Male; Group was coded 0 = ETH, 1 = HTH; Phishing previous knowledge was coded 0 = No, 1 = Yes; *$p < 0.05$. **$p < 0.01$.

reported using the Internet almost every day, and the remaining 1.5%, four or five times a week). They also reported having relatively frequent experience with bank sites (17.3% of them reported visiting bank sites almost every day, 8.3% four or five times a week, 25.6% once or twice a week, 37.6% once or twice a month, and 11.3% less than once a month), and most of them knew what phishing was before taking part in the study (only 11 out of 133 participants reported not having prior knowledge of this scam). No differences between groups were detected in terms of general Internet usage, Mann-Whitney $U = 2257$, $Z = 1.307$, $p = 0.191$, $r = 0.113$; bank site experience, $U = 1988.5$, $Z = -0.975$, $p = 0.329$, $r = -0.085$, or prior knowledge about phishing, $\chi^2(1) = 3.701$, $p = 0.054$, $\Phi = \ = 0.167$.

3.2.2. Discriminative abilities after training

The number of correct responses on the test phase showed that the ETH group performed similarly to the HTH group on Test Block 1 (*Mdn* = 7 in both groups). However, the ETH group achieved a higher number of correct responses than the HTH group on the second block of trials (*Mdn* = 8.5 and 7, respectively for ETH and HTH groups). Nevertheless, and since this measure can be contaminated by strategical components, we used the sensitivity index A′ to assess discriminative abilities in each group (Fig. 5 depicts A′ Box-Plot graphs for each group in each block).

As in Experiment 1, the relevant data are those of the test phase, in which the groups are comparable because they are presented with the same stimuli. Again, no significant differences between groups were found on the first block of the test phase (Test Block 1), Mann-Whitney $U = 1786$, $Z = -1.863$, $p = 0.062$, $r = 0.162$. However, differences appeared on the second block of trials (Test Block 2), Mann-Whitney $U = 1729$, $Z = -2.134$, $p = 0.033$, $r = 0.185$ (a small to medium sized-effect). These results replicated those from Experiment 1, and confirm that, in comparison with simple discriminative training, ETH training enhances sensitivity to visual deception cues.

3.2.3. Additional analyses

With respect to the training phase, the right-hand panel of Fig. 5 shows that participants in the HTH group improved their performance as the task progressed, with performance around chance level during the training phase (*Mdn* = 0.5 in both training blocks) but the data were mostly distributed above chance on the test phase (*Mdn* = 0.627, in both test blocks). On the other hand, participants in the ETH group showed a stable performance throughout the whole task with data mainly distributed above chance level in all four blocks (*Mdn* = 0.705, 0.803, 0.705 and 0.79, respectively for Training Block 1, Training Block 2, Test Block 1, Test Block 2).

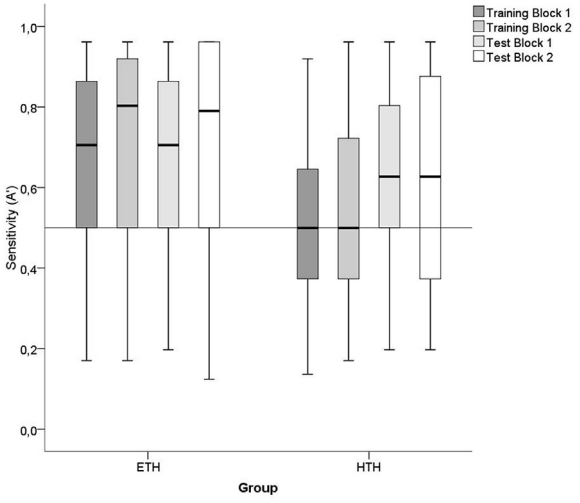

Fig. 5. Box-plots showing the sensitivity index A′ on each block of trials for participants in the ETH (left panel) group and the HTH group (right panel) in Experiment 2. Boxes represent the middle 50% of the data (interquartile range, IQR) and whiskers represent the largest and lowest cases within 1.5 times the IQR. The dark line within each box represents the median, and the dots outside the boxes depict outliers (i.e., scores departing from the first/third quartile more than 1.5 times the IQR), which were neither removed, nor transformed. The horizontal line is set at chance level (i.e., responding correctly on half of the trials).

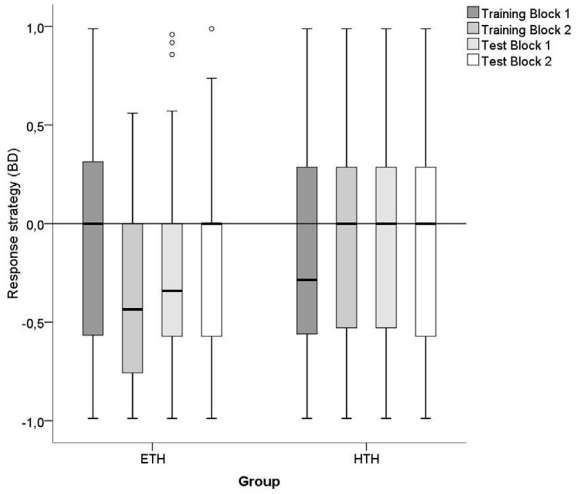

Fig. 6. Box-plots showing the response strategy index B_D on each block of trials for participants in the ETH (left panel) group and the HTH group (right panel) in Experiment 2. Boxes represent the middle 50% of the data (interquartile range, IQR) and whiskers represent the largest and lowest cases within 1.5 times the IQR. The dark line within each box represents the median, and the dots outside the boxes depict outliers (i.e., scores departing from the first/third quartile more than 1.5 times the IQR), which were neither removed, nor transformed. The horizontal line is set to represent a non-biased strategy, whilst values below this line show a lenient strategy (i.e., tendency to categorize websites as original) and values over the line show a conservative strategy (i.e., tendency to categorize websites as fake).

With regard to response strategy (B_D index), Fig. 6 shows that ETH group performance was very similar to that found in Experiment 1, with a clear deviation from neutrality on Test Block 1 (note that B_D values are mostly distributed below zero), a bias that seems to be attenuated on Test Block 2. In contrast, the HTH group exhibited a more neutral strategy throughout the whole phase, with median B_D values equal to zero on both test blocks. In fact, differences between the groups were detected on Test Block 1, Mann-Whitney $U = 2709.5$, $Z = 2.344$, $p = 0.019$, $r = 0.203$; but not on Test Block 2, Mann-Whitney $U = 2437.5$, $Z = 1.115$, $p = 0.265$, $r = 0.097$. These results show that the ETH group started the test phase with a more lenient strategy than the HTH group, a difference that disappeared as experience with the to-be-discriminated stimuli (O vs. F_3) progressed.

4. General discussion

Phishing has become a major threat to cyber-security that generates significant corporate and individual damage. In spite of the substantial growth of technological measures developed to protect users from being phished, human factors still play a critical role in security, as they are the ultimate components that determine whether or not the user decides to trust a website and to voluntarily provide sensitive information (Cranor, 2008; Jakobsson, 2007; Proctor & Chen, 2015).

Human vulnerability has raised the interest in educational and

training approaches designed to protect the security of Internet users. However, whilst recent research has pointed out the relevance of these interventions (Kumaraguru et al., 2009, 2007; Sheng et al., 2007), Internet users do not use security information in the way that might be expected (e.g., Dhamija et al., 2006; Lin et al., 2011; Whalen & Inkpen, 2005; Wu et al., 2006). Rather, people pay little attention to security indicators and rely mainly on visual appearance to judge the legitimacy of websites (e.g., Alsharnouby et al., 2015). Fortunately, spoofed sites usually present visual deviations from their genuine equivalents, and these deviations may be used for developing additional strategies to help people reduce vulnerability in a less intrusive way, without interfering with their browsing habits. Experimental psychology, and specifically research developed on perceptual and discriminative learning, can be quite helpful in achieving this goal.

Previous research has shown that overemphasizing the differences between stimuli at the beginning of discriminative training is an effective option for improving the identification of subtle differences between them (e.g., Liu et al., 2008; Suret & McLaren, 2003). The easy-to-hard effect resulting from this training schedule has proved to be reliable and useful in applied contexts (e.g., Jamieson & Morosan, 1989). This phenomenon, therefore, clearly has potential implications for cyber-security.

In Experiment 1, we compared college students' sensitivity to differences between Original and Fake websites after discriminative training. We specifically chose deviations in typeface as the critical dimension for training, since spoofed sites usually present these discrepancies with the original ones. Therefore, one group of participants was trained with the target websites (two very similar websites that differed only in their typeface), which involves learning a very difficult discrimination. The other group was trained using an easy-to-hard procedure, which consisted of presenting pairs of stimuli that are progressively more similar until the target discrimination is reached. The results from this experiment suggest that this easy-to-hard training can be beneficial for improving sensitivity to deception cues. In particular, participants trained with this schedule showed a higher sensitivity to differences between websites in the test phase than those trained under a similar but fixed regime that included only the most difficult discrimination. This result was replicated with a more representative sample of Internet users in Experiment 2, showing the generality of the easy-to-hard effect in this context.

The SDT analytical approach has allowed us to disentangle contributions from perceptual and strategic processes (i.e., those related to response strategy), ensuring that changes in sensitivity indexes reflected only variations in discriminative abilities while obtaining additional information about the response strategy.

The experiments reported herein did not explicitly manipulate instructions, signal density, payoffs or any other factor known to directly affect the response strategy (see Macmillan & Creelman, 2005). We manipulated only the perceptual properties of the fake website, a factor that a priori is expected to affect only sensitivity. Therefore, we did not expect differences between groups in their response strategy. Our results concerning the response strategy component were not consistent between experiments. We did not find differences between training regimes in Experiment 1, but we did in Experiment 2. Differences between experiments related to response strategy could be reflecting the effect of uncontrolled variables, and they may have arisen as a direct consequence of sample differences between experiments.

Research on the easy-to-hard training strategy has been mainly focused on discriminative abilities evidencing a reliable effect across procedures, domains, samples and species (consistency between Experiment 1 and Experiment 2 in sensitivity measure also supports this idea). Therefore, we did not expect sample differences to modulate the easy-to-hard effect on discriminative abilities, but it is still possible that they do on response strategy (a component that has not been systematically explored within the easy-to-hard research). On a demographical level, there was a higher proportion of women, and a more reduced age-range in Experiment 1 that in Experiment 2; and it is possible that both samples differed in other factors such as in socioeconomic status. These differences, together with those related to previous knowledge about phishing or previous experience with bank websites may account for discrepancies in response strategy between experiments.

The reason why we measured previous knowledge about phishing, Internet usage and bank website experience was to detect potential differences between ETH and HTH groups. The concrete nature of these three aspects (i.e., they are not abstract psychological constructs) so as their secondary role for our research (our main goal was exploring ETH effect on discriminative abilities) justify the use of single item measures despite their reliability and validity limitations (see Fuchs & Diamantopoulos, 2009 for a review). However, these limitations clearly restrict a deeper exploration of these three aspects, and further research is needed to systematically evaluate the variables that would potentially modulate these strategic effects of each discriminative training regime.

Besides the discrepancies between experiments, and regarding the general response strategy of participants when facing a phishing detection task, previous research has shown that in a research context participants may be warned or primed about fraud, and this may affect their response strategy promoting a conservative strategy that minimizes the risk of being tricked (e.g., Kumaraguru et al., 2010; Sheng et al., 2007).[1] We did not find strong evidence to support this prediction, as participants did not exhibit a clear conservative strategy when asked to categorize pictures, but rather the opposite (see Canfield et al., 2016 for similar results on a detection task in this context). A possible explanation for this discrepancy is that we used an artificial categorization task, that is, the experiments were not conducted within the context of a real situation, and participants may be considering responses as equally costly (note that the experimental procedure did not favor any strategy in particular). In this situation, other strategic factors (e.g., participants' expectations about the base rate of legitimate and fake websites) may be playing a role in behavior, shaping response strategy measures. In neither Experiment 1 or Experiment 2 were participants informed about the actual base rate of the trials containing a fake website (which was 0.5 in both experiments), so it remains unclear whether participants had any assumptions regarding this issue. They may be assuming a low base rate of fake trials, as in real life (i.e., using a representativeness heuristic) that should induce a lenient strategy. Subsequent experiments in this context should evaluate the impact of these strategic factors on performance.

Nevertheless, it is important to note that regardless of the potential effect of discriminative training on response strategy, training participants with progressively more difficult discriminations systematically improved sensitivity to deception cues when compared with simple discriminative training (i.e., HTH). This result was found in both experiments with bias-free measures, which makes easy-to-hard training an appropriate approach for increasing discriminative abilities.

Despite the promising evidence, these results must be viewed as an initial step. Additional research is needed to overcome some of the limitations of this study, and to check the extensions and

[1] Note that the meaning of the terms conservative and lenient depends of the definition of signal (in our experiments the signal is the original website).

boundaries of perceptual training to ensure transferability to real world settings. Thus, further studies may address whether easy-to-hard training is still effective when other relevant features are trained. We have already discussed why typefaces can be considered a useful cue for legitimacy evaluations, but it is neither an unequivocal visual deception cue nor the only trainable feature. We previously mentioned other trainable inconsistencies such as layout, color palette or domain names as examples, and it is even possible to specifically design "secure visual features" taking into account the specific requirements of each company by jointly working with web designers.

In addition to training other visual components of websites, a further question that is worthy of future evaluation is whether the easy-to-hard effect would still be observed in a more ecologically valid setting, using real websites rather than screenshots, and when the primary task of users is not security but looking for information, or just browsing. There are reasons to believe that enhancing people's abilities to discriminate and detect phishing attacks in a way that is natural to them could prove to be highly effective if implemented in real-word anti-phishing strategies, but it still remains critical to know whether suspicions about legitimacy could actually be enhanced after this type of training in a real setting.

Finally, and as advanced before, future studies should also assess the role of individual differences in response strategy but also on the advantages provided by perceptual training in this context. Previous research has shown that the ability to detect phishing attempts may vary across individuals; for example, demographic characteristics such as age or gender, and previous experience with anti-phishing training has shown to be related to the abilities to detect and manage phishing e-mails (Sheng, Holbrook, Kumaraguru, Cranor, & Downs, 2010); and the same happens with other variables such as familiarity with computers or certain personality traits (Pattinson, Jerram, Parsons, McCormac, & Butavicius, 2012; Welk et al., 2015). The influence of these and other individual features on phishing detection, and the direct effect of perceptual training on cognitive abilities (e.g., working memory capacity, attentional control, etc.) that may affect decision-making in this context (see Jones et al., 2015 for a review) still remain unexplored.

5. Conclusions

The present research shows that progressive training (from easy-to-hard) is effective for classifying websites into 'phishing' and 'non-phishing' categories when decisions have to be made solely on the basis of perceptual evidence. This new approach could be considered as an additional strategy that may complement existing educational strategies, overcoming some of their limitations. Results of this research have also pointed out the relevance of using bias-free measures of discriminative abilities, highlighting the utility of the SDT approach in this context.

Acknowledgements

Support for this research was provided by Dirección General de Investigación of the Spanish Government (Grant No. PSI2016-78818-R).

Appendix A

Instructions (translated from Spanish)

[Instructions screen 1] "Phishing is a form of electronic fraud aimed at acquiring sensitive information from Internet users by posing as a trustworthy company. To avoid raising suspicions in their victims, scammers typically use e-mails with links to fake sites that accurately mimic real ones. However, these spoofed sites usually present peculiarities. Learning to identify these peculiarities could be quite convenient.

Next we are going to show you pictures of a fictitious website. It is not a real bank, but its website has the features of real banks. Together with these pictures, we are going to show you other pictures extracted from websites trying to imitate our fictitious bank. As in real situations, imitations look quite similar to the original one but they have errors that you can learn to identify"

[Instructions screen 2] "What is your task? Pictures will appear one at a time, and you will have a few seconds to observe them. When the picture disappears you should answer this question: Do you think the picture belongs to the original or to a fake website?

Appendix B

Internet usage, online banking experience, and prior phishing knowledge questions

How often do you use the Internet?

(1) Less than once a month
(2) Once or twice a month
(3) Once or twice a week
(4) Four or five times a week
(5) Almost everyday

How often do you visit bank sites?

(1) Less than once a month
(2) Once or twice a month
(3) Once or twice a week
(4) Four or five times a week
(5) Almost everyday

Before taking part in this study, did you know about phishing? Yes/No.

References

Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010). Experimental case studies for investigating E-banking phishing techniques and attack strategies. *Cognitive Computation, 2*(3), 242–253. http://dx.doi.org/10.1007/s12559-010-9042-7.

Alsharnouby, M., Alaca, F., & Chiasson, S. (2015). Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies, 82*, 69–82. http://dx.doi.org/10.1016/j.ijhcs.2015.05.005.

Anti-Phishing Working Group. (2016). *Phishing activity trends report, 1st quarter 2016*. Retrieved from https://docs.apwg.org/reports/apwg_trends_report_q1_2016.pdf.

Arriola, N., Alonso, G., & Rodríguez, G. (2015). Progressively increasing the difficulty of a Pavlovian discrimination in a voluntary exposure to toxin paradigm with rats attenuates the magnitude of the easy-to-hard effect. *Learning and Motivation, 49*, 6–13. http://dx.doi.org/10.1016/j.lmot.2014.12.001.

Brown, G. S., & White, K. G. (2005). The optimal correction for estimating extreme discriminability. *Behavior Research Methods, 37*(3), 436–449. http://dx.doi.org/10.3758/BF03192712.

Canfield, C. I., Fischhoff, B., & Davis, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 58*(8), 1158–1172.

Carpenter, S., Zhu, F., & Kolimi, S. (2014). Reducing online identity disclosure using warnings. *Applied Ergonomics, 45*(5), 1337–1342. doi: 0.1016/j.apergo.2013.10.005.

Church, B. A., Mercado, E., Wisniewski, M. G., & Liu, E. H. (2012). Temporal dynamics in auditory perceptual learning: Impact of sequencing and incidental learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(1), 270–276. http://dx.doi.org/10.1037/a0028647.

Cranor, L. F. (2008). A framework for reasoning about the human in the loop. *Proceedings of the 1st Conference on Usability, Psychology, and Security*, 1–15.

Dhamija, R., & Tygar, J. D. (2005). The battle against phishing: Dynamic security skins. *Symposium on Usable Privacy and Security (SOUPS), 2005*, 77–88. http://

dx.doi.org/10.1145/1073001.1073009.
Dhamija, R., Tygar, J. D., & Hearst, M. (2006). Why phishing works. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1*, 581–590. http://dx.doi.org/10.1145/1124772.1124861.
Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General, 121*(3), 275–277. http://dx.doi.org/10.1037/0096-3445.121.3.275.
Emigh, A. (2007). Phishing attacks: Information flow and chokepoints. In M. Jakobsson, & S. Myers (Eds.), *Phishing and Countermeasures: Understanding the increasing problem of electronic identity theft* (pp. 31–63). Hoboken, NJ: John Wiley & Sons.
Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., et al. (2001). What makes web sites credible? A report on a large quantitative study. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '01, 3*(1), 61–68. http://dx.doi.org/10.1145/365024.365037.
Fuchs, C., & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Die Betriebswirtschaft, 69*(2), 195–211.
Gabrilovich, E., & Gontmakher, A. (2002). The homograph attack. *Communications of the ACM*. http://dx.doi.org/10.1145/503124.503156.
Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19*(5), 847–857. http://dx.doi.org/10.3758/s13423-012-0296-9.
Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford England: John Wiley.
Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75*(6), 424–429. http://dx.doi.org/10.1037/h0031246.
Haberlandt, K. (1971). Transfer along a continuum in classical conditioning. *Learning and Motivation, 2*(2), 164–172.
Jakobsson, M. (2007). The human factor in phishing. *Privacy Security of Consumer Information, 7*, 1–19. doi: 10.1.1.68.8721.
Jakobsson, M., & Myers, S. (2007). *Phishing and countermeasures: Understanding the increasing problem of electronic identity theft*. Hoboken, NJ: John Wiley & Sons.
Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English/ð/-/θ contrast by francophones. *Perception & Psychophysics, 40*(4), 205–215. http://dx.doi.org/10.3758/BF03211500.
Jamieson, D. G., & Morosan, D. E. (1989). Training non-native speech contrasts: A comparison of the prototype and perceptual fading techniques. *Canadian Journal of Psychology, 43*(1), 88–96. http://dx.doi.org/10.1037/h0084209.
Jones, H. S., Towse, J. N., & Race, N. (2015). Susceptibility to email fraud: A review of psychological perspectives, data-collection methods, and ethical considerations. *International Journal of Cyber Behavior, Psychology and Learning, 5*(3), 13–29. http://dx.doi.org/10.4018/ijcbpl.2015070102.
Kumaraguru, P., Cranshaw, J., Acquisti, A., Cranor, L., Hong, J., Blair, M. A., et al. (2009). School of phish: A real-world evaluation of anti-phishing training. *Proceedings of the 5th Symposium on Usable Privacy and Security*. http://dx.doi.org/10.1145/1572532.1572536, 3:1–3:12.
Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L. F., Hong, J., & Nunge, E. (2007). Protecting people from phishing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07, 905*. http://dx.doi.org/10.1145/1240624.1240760.
Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., & Hong, J. (2010). Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology, 10*(2), 1–31. http://dx.doi.org/10.1145/1754393.1754396.
Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology*, 1–23. http://dx.doi.org/10.1017/iop.2015.13.
Lawrence, D. H. (1952). The transfer of a discrimination along a continuum. *Journal of Comparative and Physiological Psychology, 45*(6), 511–516. http://dx.doi.org/10.1037/h0057135.
Lawrence, D. H. (1955). The applicability of generalization gradients to the transfer of a discrimination. *Journal of General Psychology, 52*, 37–48. http://dx.doi.org/10.1080/00221309.1955.9918342.
Lin, E., Greenberg, S., Trotter, E., Ma, D., & Aycock, J. (2011). Does domain highlighting help people identify phishing sites? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2075–2084. http://dx.doi.org/10.1145/1978942.1979244.
Liu, W., Deng, X., Huang, G., & Fu, A. Y. (2006). An antiphishing strategy based on visual similarity assessment. *IEEE Internet Computing, 10*(2), 58–65. http://dx.doi.org/10.1109/MIC.2006.23.
Liu, W., Guanglin, H., Liu, X., Zhang, M., & Xiaotie, D. (2005). Detection of phishing webpages based on visual similarity. *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, 1060–1061. http://dx.doi.org/10.1145/1062745.1062868.
Liu, E. H., Mercado, E., Church, B. A., & Orduña, I. (2008). The easy-to-hard effect in human (Homo sapiens) and rat (Rattus norvegicus) auditory identification. *Journal of Comparative Psychology, 122*(2), 132–145. http://dx.doi.org/10.1037/0735-7036.122.2.132.
Mackintosh, N. J., & Little, L. (1970). An analysis of transfer along a continuum. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 24*(5), 362–369. http://dx.doi.org/10.1037/h0082872.
Macmillan, N., & Creelman, C. (2005). *Detection theory: A user's guide*. New Jersey: Lawrence Erlbaum Associates.
Maurer, M. E., & Herzner, D. (2012). Using visual website similarity for phishing detection and reporting. In *CHI '12 extended abstracts on human factors in computing systems* (pp. 1625–1630). New York, NY, USA: ACM. http://dx.doi.org/10.1145/2212776.2223683.
May, R. B., & MacPherson, D. F. (1971). Size discrimination in children facilitated by changes in task difficulty. *Journal of Comparative and Physiological Psychology, 75*(3), 453–458.
McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective & Behavioral Neuroscience, 2*(2), 89–108. http://dx.doi.org/10.3758/CABN.2.2.89.
Medvet, E., Kirda, E., & Kruegel, C. (2008). Visual-similarity-based phishing detection. *Proceedings of the 4th International Conference on Security and Privacy in Communication Netowrks*, 1–6. http://dx.doi.org/10.1145/1460877.1460905.
Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). Tutorial and critical analysis of phishing websites methods. *Computer Science Review, 17*, 1–24. http://dx.doi.org/10.1016/j.cosrev.2015.04.001.
Moreno-Fernández, M. M., Ramos-Álvarez, M. M., Paredes-Olay, C., & Rosas, J. M. (2012). Effects of progressively increasing the difficulty of training on sensitivity and strategic factors in olive oil tasting. *Food Quality and Preference, 24*(2), 225–229. http://dx.doi.org/10.1016/j.foodqual.2011.12.001.
OpenDNS. (2016). *PhishTank. Statistics about phishing activity*. Retrieved from https://www.phishtank.com/stats.php.
Pashler, H., & Mozer, M. C. (2013). When does fading enhance perceptual category learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(4), 1162–1173. http://dx.doi.org/10.1037/a0031679.
Pattinson, M., Jerram, C., Parsons, K., McCormac, A., & Butavicius, M. (2012). Why do some people manage phishing e-mails better than others? *Information Management & Computer Security, 20*(1), 18–28. http://dx.doi.org/10.1108/09685221211219173.
Proctor, R. W., & Chen, J. (2015). The role of human factors/ergonomics in the science of security decision making and action selection in cyberspace. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. http://dx.doi.org/10.1177/0018720815585906, 18720815585906.
Purkait, S. (2012). Phishing counter measures and their effectiveness - literature review. *Information Management and Computer Security, 20*(5), 382–420. http://dx.doi.org/10.1108/09685221211286548.
Ryan, R. S., Wilde, M., & Crist, S. (2013). Compared to a small, supervised lab experiment, a large, unsupervised web-based experiment on a previously unknown effect has benefits that outweigh its potential costs. *Computers in Human Behavior, 29*(4), 1295–1301. http://dx.doi.org/10.1016/j.chb.2013.01.024.
Scahill, V. L., & Mackintosh, N. J. (2004). The easy to hard effect and perceptual learning in flavor aversion conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 30*(2), 96–103. http://dx.doi.org/10.1037/0097-7403.30.2.96.
Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L. F., & Downs, J. (2010). Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, 373–382. http://dx.doi.org/10.1145/1753326.1753383.
Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L. F., Hong, J., et al. (2007). Anti-Phishing Phil: The design and evaluation of a game that teaches people not to fall for phish. *Proceedings of the 3rd Symposium on Usable Privacy and Security*, 88–99. http://dx.doi.org/10.1145/1280680.1280692.
Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments & Computers, 31*(1), 137–149. http://dx.doi.org/10.3758/bf03207704.
Suret, M., & McLaren, I. P. L. (2003). Representation and discrimination on an artificial dimension. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology, 56B*(1), 30–42. http://dx.doi.org/10.1080/02724990244000142.
Sutherland, N. S., Mackintosh, N. J., & Mackintosh, J. (1963). Simultaneous discrimination training of octopus and transfer of discrimination along a continuum. *Journal of Comparative and Physiological Psychology, 56*(1), 150–156.
Tallal, P., Miller, S. L., Bedi, G., Wang, X., & Nagarajan, S. S. (1996). Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science, 271*(5245), 81–84. http://dx.doi.org/10.1126/science.271.5245.81.
Vadillo, M. A., Bárcena, R., & Matute, H. (2006). The Internet as a research tool in the study of associative learning: An example from overshadowing. *Behavioural Processes, 73*, 36–40.
Vadillo, M. A., & Matute, H. (2011). Further evidence on the validity of web-based research on associative learning: Augmentation in a predictive learning task. *Computers in Human Behavior, 27*(2), 750–754. http://dx.doi.org/10.1016/j.chb.2010.10.020.
Walker, M. M., Lee, Y., & Bitterman, M. E. (1990). Transfer along a continuum in the discriminative learning of honeybees (Apis mellifera). *Journal of Comparative Psychology, 104*(1), 66–70.
Welk, A. K., Hong, K. W., Zielinska, O. A., Tembe, R., Murphy-Hill, E., & Mayhorn, C. B. (2015). Will the "Phisher-Men" Reel You In?: Assessing individual differences in a phishing detection task. *International Journal of Cyber Behavior, Psychology and Learning, 5*(4), 1–17. http://dx.doi.org/10.4018/IJCBPL.2015100101.
Whalen, T., & Inkpen, K. M. (2005). Gathering evidence: Use of visual security cues

in web browsers. *Proceedings of Graphics Interface, 2005*, 137–144. Retrieved from http://dl.acm.org/citation.cfm?id=1089532.

Whitten, A., & Tygar, J. D. (1999). Why Johnny can't encrypt: A usability evaluation of PGP 5.0. *Proceedings of the 8th USENIX Security Symposium*, 169–184. Retrieved from http://dl.acm.org/citation.cfm?id=1251435.

Wogalter, M. S., & Mayhorn, C. B. (2008). Trusting the internet: Cues affecting perceived credibility. *International Journal of Technology and Human Interaction, 4*(1), 75–93. http://dx.doi.org/10.4018/jthi.2008010105.

Wombat Security Technologies. (2016). *Anti-phishing phyllis*. Retrieved from https://www.wombatsecurity.com/security-education/educate.

Wu, M., Miller, R. C., & Garfinkel, S. L. (2006). Do security toolbars actually prevent phishing attacks? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 601–610. http://dx.doi.org/10.1145/1124772.1124863.

Zhang, H., Liu, G., Chow, T. W. S., & Liu, W. (2011). Textual and visual content-based anti-phishing: A bayesian approach. *IEEE Transactions on Neural Networks, 22*(10), 1532–1546. http://dx.doi.org/10.1109/TNN.2011.2161999.