Multitasking and Monetary Incentive in a Realistic Phishing Study

Haoruo Zhang, Shirish Singh, Xiangyang Li, Anton Dahbura, Meng Xie
Johns Hopkins University Information Security Institute
3400 North Charles Street, Baltimore, Maryland, 21218, USA
{zhanghaoruo, Shirish, xyli, antondahbura, mxie6}@jhu.edu

This paper introduces an empirical study focusing on task settings similar to those in the real-world that captures user behavioral information of fine granularity. In online experiments, participants recruited from the Mechanical Turk human subject pool sorted legitimate and phishing emails. Subgroups of these remote users performed a secondary question-answering task and/or were incentivized by a monetary reward based on email sorting accuracy. This web-based framework automates a complete process from the informed consent to a post-study questionnaire, which can be scaled up to a large number of human subjects. In the preliminary result analysis, the monetary incentive can positively affect users' behavior and performance, but not in a straightforward manner. Multitasking, on the other hand, has a negative effect on users' ability to correctly classify emails.

Computer Security, Phishing, User Behavior, Multitasking, Incentive.

1. INTRODUCTION

Technological countermeasures do not always protect information assets when human elements fail due to distraction or a lack of awareness (Thomson & Solms, 1998; Willison & Warkentin, 2013). Active research efforts have studied the risks of phishing and other computer security issues. Many are conducted in a lab environment that can significantly change the attitude and behavior of participants. Moreover, data collection often relies on video recording or self-reporting, which are hard to scale up or to consider as realistic scenarios in an employee's office.

Our contributions are three-fold: (1) The user study focuses on multitasking and a monetary incentive in sorting real legitimate and phishing emails; (2) the experimentation framework supports large-scale "in the wild" experiments in an automated and unattended manner similar to real-world settings; (3) the implementation enables data capture and collection of micro-level user behaviors. We developed a web-based solution using JavaScript and the LAMP stack (Lawton, 2005). It integrates Roundcube (https://roundcube.net), a webmail system, and Qualtrics (https://www.qualtrics.com), an online survey system.

2. RELATED WORK

Benítez et al (2017) proposed a web-based tool that enables researchers to manage questionnaires and visualize the data collected. Kaczmarek et al (2015) presented an unattended study of users performing security tasks like pairing wireless devices. Gajos et al (n.d.) have been conducting an online user study on multitasking with the help of Google Analytics.

Ollesch et al (2006) found no significant difference in psychometric data collected in a lab setting and its online, virtual counterpart. Many successful examples used participants remotely from Amazon Mechanical Turk for performing research-focused tasks (Bartneck et al, 2015; Kittur et al, 2008; Layman & Sigurdsson, 2013). For example, Bianchi et al (2015) utilized Amazon Mechanical Turk to disseminate noVNC clients via HTTP to end users to study Android GUI design-based attacks.

Atterer et al (2006) proposed a framework of using web technologies (e.g., JavaScript, Proxy) to track user interactions with a web page. We expanded upon their idea to track users' interaction with a webmail client.

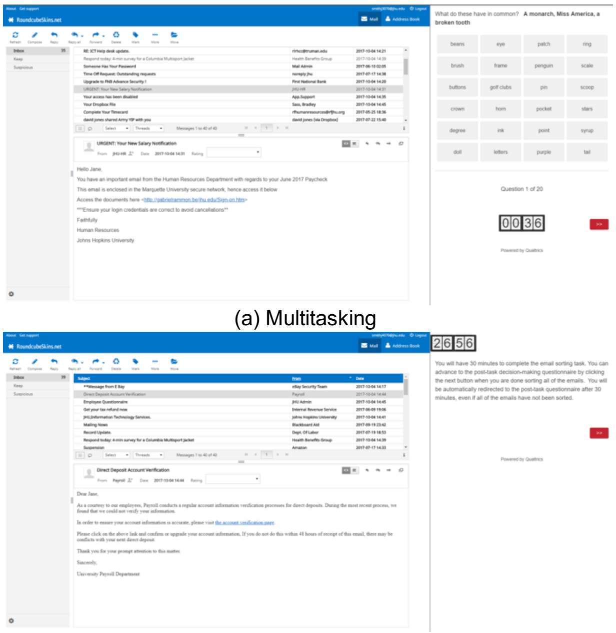
3. DESIGN OF USER STUDY EXPERIMENT

A key challenge here is how to balance uncertainty and familiarity of an email's source to participants. Participants in this user study were instructed that they were an administrative assistant working for the department chair, Dr. Jane Smith. They did not need to respond to any of the 40 emails, only sort them into either a "Keep" or "Suspicious" folder, without using the internet or other sources.

© 2018, *Haoruo Zhang, Shirish Singh, Xiangyang Li, Anton Dahbura, Meng Xie. Published by BCS Learning and Development Ltd. Proceedings of British HCI 2018. Belfast, UK.*

3.1. Condition-based User Tasks

During the pre-study survey, participants were instructed in the way they were expected to complete the experiment. Once a participant ran out of time or finished early and chose to move on, s/he was taken to the post-study survey in Qualtrics.



RE: ICT Help desk update.	rirhcc@truman.edu	2017-10-04 14:21
Respond today: 4-min survey for a Columbia Multisport Jacket	Health Benefits Group	2017-10-04 14:39
Someone Has Your Password	Mail Admin	2017-06-10 02:05
Time Off Request: Outstanding requests	noreply jhu	2017-07-17 14:38
Upgrade to FNB Advance Security !	First National Bank	2017-10-04 14:20
URGENT: Your New Salary Notification	JHU-HR	2017-10-04 14:31
Your access has been disabled	App.Support	2017-10-04 14:35
Your Dropbox File	Sass, Bradley	2017-10-04 14:45
Complete Your Timecard	rfhumanresources@rfjhu.org	2017-05-25 18:36
david jones shared Army YIP with you	david jones (via Dropbox)	2017-07-22 15:40

URGENT: Your New Salary Notification
From JHU-HR Date 2017-10-04 14:31

Hello Jane,

You have an important email from the Human Resources Department with regards to your June 2017 Paycheck

This email is enclosed in the Marquette University secure network, hence access it below

Access the documents here <http://gabrielrammon.be/jhu.edu/Sign-on.htm>

***Ensure your login credentials are correct to avoid cancellations**

Faithfully

Human Resources

Johns Hopkins University

What do these have in common? **A monarch, Miss America, a broken tooth**

beans	eye	patch	ring
brush	frame	penguin	scale
buttons	golf clubs	pin	scoop
crown	horn	pocket	stars
degree	ink	point	syrup
doll	letters	purple	tail

Question 1 of 20

0036

(a) Multitasking

Subject	From	Date
**Message from E Bay	eBay Security Team	2017-10-04 14:17
Direct Deposit Account Verification	Payroll	2017-10-04 14:44
Employee Questionnaire	JHU Admin	2017-10-04 14:45
Get your tax refund now	Internal Revenue Service	2017-06-09 19:06
JHU,Information Technology Services.	Johns Hopkins University	2017-10-04 14:41
Mailing News	Blackboard Aid	2017-09-19 23:42
Record Update.	Dept. Of Labor	2017-07-19 18:53
Respond today: 4-min survey for a Columbia Multisport Jacket	Health Benefits Group	2017-10-04 14:39
Suspension	Amazon	2017-07-17 14:33

Direct Deposit Account Verification
From Payroll Date 2017-10-04 14:44

Dear Jane,

As a courtesy to our employees, Payroll conducts a regular account information verification processes for direct deposits. During the most recent process, we found that we could not verify your information.

In order to ensure your account information is accurate, please visit the account verification page.

Please click on the above link and confirm or upgrade your account information, If you do not do this within 48 hours of receipt of this email, there may be conflicts with your next direct deposit.

Thank you for your prompt attention to this matter.

Sincerely,

University Payroll Department

2656

You will have 30 minutes to complete the email sorting task. You can advance to the post-task decision-making questionnaire by clicking the next button when you are done sorting all of the emails. You will be automatically redirected to the post-task questionnaire after 30 minutes, even if all of the emails have not been sorted.

(b) No-multitasking

Figure 1: *User Study Task Interface*

Participants were randomly assigned to one of four experimental conditions determined by two factors, multitasking/no-multitasking and incentive/no-incentive. For the multitasking condition in Figure 1(a), participants answered 20 sets of questions in Qualtrics on the right side while completing the email sorting task on Roundcube. Each question set would be presented for two minutes; participants could manually advance to the next question set after one minute had elapsed. For no-multitasking in Figure 1(b), participants only had the email sorting task and had 30 minutes to complete it. Second, participants were in either the incentive or no-incentive condition. For the incentive condition, participants could earn an additional monetary compensation, up to $8.00, based on the number of correctly sorted emails, in a tiered scheme. Those participants in the condition of incentive and multitasking, in order to be eligible for the incentive, must have correctly sorted 30 out of 40 emails (75%) and correctly answered 15 out of 20 multitasking questions (75%).

3.2. Email Design and Phishing Cues

The 40 emails were presented in a random order for each participant. Twenty phishing emails were derived from a semi-random sample of emails in Cornell University's "Phish Bowl" database (it.cornell.edu/phish-bowl). The 20 legitimate emails were derived from emails received by the research team. Their selection and design for experimental use considered 14 different phishing cue categories, including Sender's Display Name, URL Hyperlink, and Spelling and Grammar Errors, among others.

3.3. User Self-reported Information

Like other user studies, we were interested in acquiring demographics and other self-reported information on participants' experience to better interpret experiment results. For example, in the post-study survey, each participant reported whether s/he took a network or cybersecurity course/certificate before and estimated the number of correctly sorted emails. Moreover, during email sorting, the participants' confidence of classifying each email was collected by selecting a rating (1 - not confident at all, to 10 - extremely confident).

4. THE EXPERIMENTATION FRAMEWORK

The solution supported user tasks conducted remotely, managed concurrent user experiments, and logged participants' actions in experiment and responses to surveys in real time.

4.1. System Workflow

The system consists of four major components: residing on the client-side web browser, JavaScript-Based Data Capturer to collect participants' input and AJAX-Based Data Sender to communicate the captured data to the server, and, residing on the server side, PHP Listener to receive the data sent from AJAX and a Logger to log the data. The Qualtrics view is embedded as a HTML Inline Frame (IFrame) in Roundcube's interface.

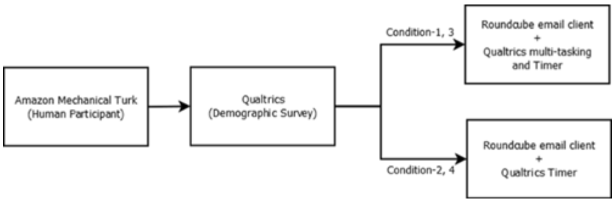


Figure 2: *User Study Workflow*

As in Figure 2, participants at Amazon Mechanical Turk were led to an online pre-study survey for demographic information and the informed consent, powered by Qualtrics. They were then redirected to the modified webmail client with user account information passed as URL parameters. The post-study survey including questions on participants' experiment experience varies according to their assigned experimental conditions.

4.2. User Interface Design

We disabled some of Roundcube's functionalities to prevent unexpected operations, such as the "New" and “Reply” menus for email editing. We added new elements including the “Keep” and “Suspicious” mail folders and a “Rating” drop-down list menu for reporting classification confidence.

For data collection, we identified UI artifacts in the source code of Roundcube and added listeners on their respective components, focusing on behaviors that a user commonly performs:

- Click - An event triggered when the user left-clicks on an object;
- Hover - An event triggered when the mouse hovers over certain interactive objects;
- Scroll - An event triggered when the user scrolls the mouse in the email body view;
- Mouse Movement - Mouse cursor coordinates recorded.

4.3. Data Collection

Tables 1-3 shows a sample of the information collected on the Roundcube webmail client and the online Qualtrics survey system.

Table 1: *Features Collected on Roundcube*

Feature Name	Description
Clicking on Buttons on Menu Bar	Recording timestamp of user clicking of Menu Bar buttons
Clicking on Email Item	Recording timestamp of user clicking on email
Hovering in Sender's displayed Address	Recording timestamp of user hovering in the Sender's Address
Hovering out Sender's displayed address	Recording timestamp of user hovering out the Sender's Address
Hovering in embedded URLs	Recording timestamp of user hovering in one URL in email
Hovering out of embedded URLs	Recording timestamp of user hovering out one URL in email
Clicking on embedded URLs	Recording timestamp of user clicking on one URL in email body
Rating Confidence Levels	Recording rating and timestamp of User selecting a confidence level
Classifying Emails	Recording classification and timestamp of classifying an email

Table 2: *Features Collected on Qualtrics*

Feature Name	Description
Hovering in Question Field	Recording timestamp of user hovering in a Qualtrics question
Hovering out Question Field	Recording timestamp of user hovering out a Qualtrics question
Hovering in Choice Field	Recording timestamp of user hovering in a Qualtrics choice
Hovering out Choice Field	Recording timestamp of user hovering out Qualtrics choice
Clicking on Choice Field	Recording timestamp of user clicking on Qualtrics choice

Table 3: *Information Collected of User Operations Switching Between Roundcube and Qualtrics*

Feature Name	Description
Entering Qualtrics Interface	Recording timestamp of user hovering in Qualtrics
Leaving Qualtrics Interface	Recording timestamp of user hovering out Qualtrics
Entering Round Cube Interface	Recording timestamp of user hovering in Roundcube client
Leaving Round Cube Interface	Recording timestamp of user hovering out Roundcube client

5. PRELIMINARY RESULT ANALYSIS

As in Table 4, we performed the experiments in batches of 40 participants at a time for 177 participants in total. They averaged 34 years of age. Sixty participants were female and 117 were male. Sixteen participants were students. One participant noted that English was not his/her first language.

Table 4: *Participants by Condition*

Condition	Participants	Participants Sorting All Emails	Completion Rate
1. Incentivized Multitasking	46	35	76.1%
2. Incentivized No-multitasking	47	42	89.3%
3. Non-incentivized Multitasking	41	34	82.9%
4. Non-incentivized No-multitasking	43	35	81.4%
Total	*177*	*146*	*82.5%*

Only 146 participants were able to finish sorting all the 40 emails in the given time. The condition 1 group, where the participants took on two concurrent tasks of email sorting and question-answering under monetary reward, had the lowest completion rate. One potential explanation is that tasks under this condition was cognitively demanding. The condition 2 group had the highest completion rate where the participants concentrated on email sorting with the monetary incentive.

5.1. Email Sorting Accuracy

Shown in Table 5 for all participants, hypothesis tests indicated there was a significant difference in the email sorting scores between condition 1 and condition 2, and between condition 2 and condition 3, using a significance level α at 0.05. Overall multitasking significantly worsened a participant’s sorting accuracy. No-multitasking combined with the incentive helped to carry out tasks. However, the incentive alone did not make a difference in either multitasking or no-multitasking cases. Using Bonferroni correction for multiple comparisons where the number of hypotheses *m* is 6, the level of significance α drops to 0.0083. Then we did not find significance in these results.

Table 5: *Overall Sorting Accuracy for All 177 Participants*

Condition	Accuracy	T-test
1. Incentivized Multitasking	0.751±0.090	1,2: T-value=-2.219 p=0.029
2. Incentivized No-multitasking	0.793±0.092	1,2: T-value=-2.219 p=0.029 2,3: T-value=2.239 p=0.028
3. Non-incentivized Multitasking	0.738±0.133	2,3: T-value=2.239 p=0.028
4. Non-incentivized No-multitasking	0.766±0.093	

As shown in Table 6, the subset of 146 participants who sorted all 40 emails displayed similar differences for different conditions with lower *p*-

values. Using Bonferroni correction, there was a significant difference between condition 1 and condition 2.

***Table 6:** Overall Sorting Accuracy for the 146 Participants*

Condition	Accuracy	T-test
1. Incentivized Multitasking	0.742±0.088	1,2: T-value=-2.942 p=0.004
2. Incentivized No-multitasking	0.802±0.091	1,2: T-value=-2.942 p=0.004 2,3: T-value=2.511 p=0.015
3. Non-incentivized Multitasking	0.734±0.136	2,3: T-value=2.511 p=0.015
4. Non-incentivized No-multitasking	0.767±0.091	

We further analyzed the sorting accuracy for phishing emails and legitimate emails separately. There was a significant difference between the phishing sorting error rates shown for conditions 1, 2, and 3, as shown in Table 7 and Table 8. However, there was not a significant difference between the legitimate email sorting error rates among them. It shows that the condition changes had significant effects mainly on the capacity that participants had to recognize phishing emails. The incentive given for the sole email sorting task improved phishing email detection.

***Table 7:** Phishing Email Sorting Error Rate for All Participants*

Condition	Phishing Sorting Error Rate	T-test
1. Incentivized Multitasking	0.339±0.181	1,2: T-value=2.230 p=0.028
2. Incentivized No-multitasking	0.260±0.161	1,2: T-value=2.230 p=0.028 2,3: T-value=-2.172 p=0.033
3. Non-incentivized Multitasking	0.350±0.216	2,3: T-value=-2.172 p=0.033
4. Non-incentivized No-multitasking	0.306±0.164	

***Table 8:** Phishing Email Sorting Error Rate for the 146 Participants*

Condition	Phishing Sorting Error Rate	T-test
1. Incentivized Multitasking	0.366±0.186	1,2: T-value=2.620 p=0.011
2. Incentivized No-multitasking	0.261±0.161	1,2: T-value=2.620 p=0.011 2,3: T-value=-1.740 p=0.087
3. Non-incentivized Multitasking	0.337±0.209	2,3: T-value=-1.740 p=0.087
4. Non-incentivized No-multitasking	0.311±0.171	

5.2. Email Processing Time

The significance test on email processing time was done as paired tests on 40 individual emails, since phishing and legitimate emails may by their nature cause differences in processing time.

Shown in Table 9 for all 177 participants, the only difference in email processing times was between condition 2 and condition 3. The participants spent more time on each email when they could concentrate on emails while being motivated for the monetary reward. It was true even for Bonferroni correction with α=0.0083.

***Table 9:** Email Processing Time for All 177 Participants*

Condition	Time(millisecond)	T-test
1. Incentivized Multitasking	21514.00±4976.76	
2. Incentivized No-multitasking	23208.04±4813.30	2,3: T-value=3.031 p=0.003
3. Non-incentivized Multitasking	20037.48±4538.78	2,3: T-value=3.031 p=0.003
4. Non-incentivized No-multitasking	21340.37±5921.76	

Table 10 shows the results for the subset of 146 participants. Multitasking and incentive showed opposite effects on the processing time that participants spent on each email: multitasking reduced participants' time spent on email processing while the incentive did help participants invest more time processing the emails. With α=0.0083 for Bonferroni correction of multiple comparisons, there were still significant differences between several conditions.

***Table 10:** Email Processing Time for the 146 Participants*

Condition	Time(millisecond)	T-test
1. Incentivized Multitasking	19875.84±4901.76	1,2: T-value=-3.358 p=0.001 1,3: T-value=2.224, p=0.030
2. Incentivized No-multitasking	23640.44±5122.58	1,2: T-value=-3.358 p=0.001 2,3: T-value=5.745 p=1.83 x 10^{-7} 2,4: T-value=1.916 p=0.059
3. Non-incentivized Multitasking	17593.90±4250.60	1,3: T-value=2.224, p=0.030 2,3: T-value=5.745 p=1.83 x 10^{-7} 3,4: T-value=-2.846, p=0.006
4. Non-incentivized No-multitasking	21115.55±6570.42	2,4: T-value=1.916 p=0.059 3,4: T-value=-2.846, p=0.006

Spending more time on individual emails did not always guarantee better sorting accuracy. The participants in condition 1 spent more time on an email compared to those in condition 3, without increasing their sorting accuracy. Although these participants were more "careful" with their email sorting tasks, switching back and forth between two tasks might pose a challenge to them. The participants in condition 4 spent more time on emails than those in condition 3. It could simply be that they had more time at hand, without much pressure.

6. ACKNOWLEDGEMENT

This work is supported under the National Science Foundation Award No. 1544493. We would like to thank Nathan Bos and Kylie Molinaro from Johns Hopkins University Applied Physics Laboratory and Matthew Shonman from Johns Hopkins University Information Security Institute (JHUISI) for their help with the user study and data analysis.

7. REFERENCES

Atterer, R., Wnuk, M., & Schmidt, A. (2006) Knowing the user's every move: User activity tracking for website usability evaluation and implicit interaction. Proceedings of the 15th International Conference on World Wide Web, 203–212. ACM, New York, NY, USA.

Bartneck, C., Duenser, A., Moltchanova, E., & Zawieska, K. (2015) Comparing the similarity of responses received from studies in Amazon's Mechanical Turk to studies conducted online and with direct recruitment. PLOS One, 10(4).

Benítez, J. A., Labra, J. E., Quiroga, E., Martín, V., García, I., Marqués-Sánchez, P., & Benavides, C. (2017) A Web-Based tool for automatic data collection, curation, and visualization of complex healthcare survey studies including social network analysis. Computational and Mathematical Methods in Medicine, 1-8.

Bianchi, A., Corbetta, J., Invernizzi, L., Fratantonio, Y., Kruegel, C., & Vigna, G. (2015) What the App is That? Deception and Countermeasures in the Android User Interface. 2015 IEEE Symposium on Security and Privacy.

Gajos, K. (n.d.) Multitasking Test. http://multitasking.labinthewild.org/multitasking/ (Retrieved April 06, 2018)

Kaczmarek, T., Kobsa, A., Sy, R., & Tsudik, G. (2015). An unattended study of users performing security critical tasks under adversarial noise. Proceedings of 2015 Workshop on Usable Security.

Kittur, A., Chi, E. H., & Suh, B. (2008) Crowdsourcing User Studies with Mechanical Turk," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 453–456. ACM, New York, NY, USA.

Lawton, G. (2005) LAMP lights enterprise development efforts. Computer, 38(9), 18-20.

Layman, L., & Sigurdsson, G. (2013) Using Amazons Mechanical Turk for user studies: Eight things you need to know. 2013 ACM / IEEE International Symposium on Empirical Software Engineering and Measurement.

Ollesch, H., Heineken, E., & Schulte, F. P. (2006) Physical or virtual presence of the experimenter: Psychological online-experiments in different settings," International Journal of Internet Science, 1(1), 71– 81.

Thomson, M., & Solms, R. V. (1998) Information security awareness: Educating your users effectively. Information Management & Computer Security, 6(4), 167-173.

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