DualCheck: Exploiting Human Verification Tasks for Opportunistic Online Safety Microlearning

Ryo Yoshikawa, Hideya Ochiai, and Koji Yatani, *The University of Tokyo*

https://www.usenix.org/conference/soups2022/presentation/yoshikawa

This paper is included in the Proceedings of the Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022).

August 8–9, 2022 • Boston, MA, USA

978-1-939133-30-4

Open access to the Proceedings of the Eighteenth Symposium on Usable Privacy and Security is sponsored by USENIX.

DualCheck: Exploiting Human Verification Tasks for Opportunistic Online Safety Microlearning

Ryo Yoshikawa
The University of Tokyo
ryo@iis-lab.org

Hideya Ochiai
The University of Tokyo
ochiai@elab.ic.i.u-tokyo.ac.jp

Koji Yatani
The University of Tokyo
koji@iis-lab.org

Abstract

Learning online safety and ethics is becoming more critical for the general user population. However, they do not receive such learning opportunities regularly, and are often left behind. We were therefore motivated to design an interactive system to provide more frequent learning opportunities to the general user population. This paper presents our explorations on the integration of opportunistic microlearning about online safety and ethics into human verification. Our instantiation of this concept, called DualCheck, asks users to respond to questions related to online safety and ethics while human verification would be executed in a similar manner to reCAPTCHA v2. In this manner, DualCheck offers users microlearning opportunities when they use online services. Our 15-day user study confirmed the positive learning effect of DualCheck. The quantitative and qualitative results revealed participants' positive experience with attitude toward DualCheck, and also found its significantly higher perceived usability than text-based CAPTCHA and picture-based reCAPTCHA.

1 Introduction

As many general users enjoy online services and communication regularly, understanding online safety and ethics is becoming an essential and critical literacy. However, they do not necessarily have sufficient opportunities to learn online safety and ethics. According to recent surveys conducted by Information-technology Promotion Agency (IPA) in Japan [6], only 17.9% of smart device users claimed that they had taken explicit training on online ethics. Furthermore, such training typically occurs at school or workplace, and the frequency is also limited. This suggests that general users may not have constant opportunities for learning online safety and ethics.

We were therefore motivated to design an interactive system to provide more frequent learning opportunities to users. More specifically, we were interested in how we can exploit existing interactions which users are already familiar with that purpose. In this work, we exploit human verification tasks which are commonly seen in online forms. For instance, CAPTCHA [17] and its variants are widely used and well recognized. As human verification is common in many online services, an integration of learning opportunities would increase the frequency of such training in an opportunistic manner. Our research questions in this work are, therefore, 1) how the integration of opportunistic learning on online safety and ethics into human verification can support people's learning; and 2) how the user experience of such a system would be different from existing human verification tasks.

This paper presents our investigations on integrating opportunistic microlearning of online safety and ethics into human verification tasks to answer these two research questions. We develop DualCheck as a proof of our concept (Figure 1). Users see DualCheck as a human verification task at the end of online forms. They then read the question and answer by clicking one of the five choices. The system presents users the correct answer and explanation for their learning. It then enables a button to move to the next page 5 second after users' responses to the question. The system does not consider any information about whether their responses are correct or not for human verification. Instead, human verification is expected to be performed in a similar manner to the checkbox-based reCAPTCHA v2. In this manner, DualCheck can achieve reliable human verification while offering microlearning of online safety and ethics in an opportunistic manner.

Our evaluation through a 15-day deployment study confirmed significant improvements on the accuracies (correct answer rates) for 9 of the 10 questions used

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

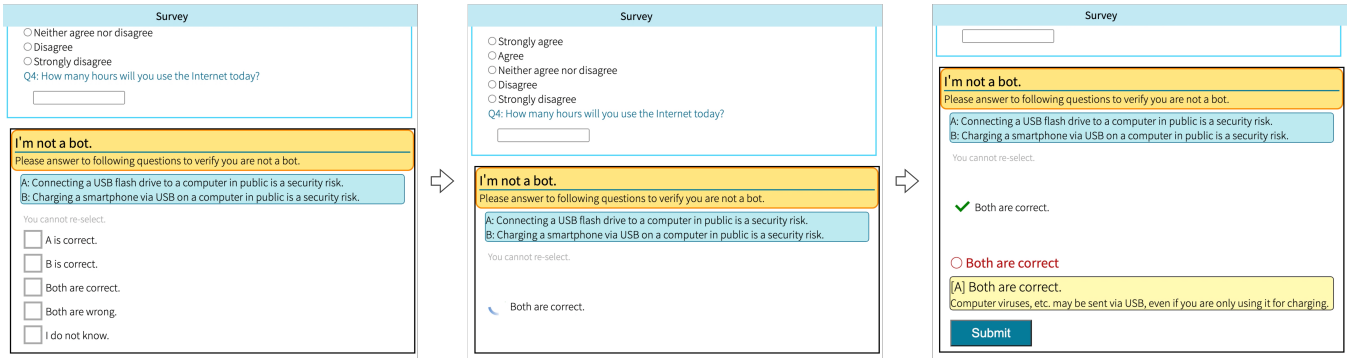

Figure 1: The DualCheck interface. Left: DualCheck can be integrated into online forms as a human verification mechanism. We note that our current prototype does not implement the human verification mechanism because our primary objective of this work is to validate the effect of opportunistic microlearning through DualCheck instead of evaluating the robustness of human verification. Middle: Users choose one of the choices after reading the question presented by DualCheck. The current implementation simply pretends to be performing human verification like reCAPTCHA v2. Right: DualCheck presents the correct answer and explanation about the given question. The system enables the submit button five seconds after it shows the correct answer and explanation. In this manner, users have an opportunity to read them. The system does not consider whether users have chosen the correct answer or not for human verification. Instead, it is expected to perform human verification through a mechanism like reCAPTCHA v2.

throughout the deployment study. In addition, our participants exhibited significantly higher accuracies on 5 of another 10 questions about online safety and ethics than general Internet users who do not use DualCheck. The perceived usability of DualCheck was significantly higher than text-based CAPTCHA and picture-based reCAPTCHA. Our qualitative results support participants' positive attitudes toward DualCheck.

The primary contributions of this work are two-folded:

- Development of DualCheck, our proof of concept of the integration of opportunistic microlearning about online safety and ethics into human verification tasks; and
- Evaluation of DualCheck through a 15-day deployment study, confirming its positive learning effect and user experience.

2 Related Work

2.1 Online Safety and Ethics Learning

Learning online safety and ethics is critical for general users as they now have multiple computer devices and access various media and online social platforms. However, people lack learning opportunities of such knowledge, often being left at risk. School curricula in different countries now include learning about online safety and ethics, but they are not necessarily effective. According to surveys conducted by Information Technology Promotion Agency (IPA) in Japan in 2019 [6, 7], only 38.0% of teenagers explicitly responded that they had online ethics training. The percentage of such people becomes even lower in older generations; for example, the number becomes only 9.6% of the survey respondents in their 70s. Furthermore, such training occurred at school or workplace for 76.8% of the respondents who claimed they had such training. Furthermore, their survey [7] also found that most of the respondents did not possess even basic knowledge concerning information security. For example, only 28.5% of the respondents were aware of the concept of "malware," and only 13.6% answered all three questions about malware correctly. A study by Grimes et al. [5] in the United States showed similar results; they found that older adults have lower awareness of online safety. Their study showed that older adults possess considerably less knowledge and awareness of Internet security hazards than university students.

These survey results indicate that learning opportunities are limited outside schools and workplaces, and thus people's knowledge about online safety and ethics is also constrained. In particular, people do not have learning opportunities regularly. Reinheimer et al. investigated the effectiveness of an awareness and education program on phishing [12]. They found that participants' phishing discrimination capabilities were maintained up to four months after the education program, but degradation occurred after that. Their result thus confirms that regular training is critical.

Existing work attempted to utilize games to motivate people's learning about online safety, Sheng et al. [14] integrated anti-phishing knowledge into a video game. They [13] further confirmed the insufficiency of people's cyber hygiene behaviors and knowledge through conducting roleplay-based phishing attacks. Although such approaches can be beneficial, further explorations on online safety and

ethics learning approaches are necessary as Drury et al. [9] suggested that attacks and threats are evolving and becoming more complex and sophisticated.

Our work exploits a human verification task seen in various online forms for online safety and ethics learning. As people often encounter such tasks during their Internet use, our DualCheck can offer more frequent learning opportunities than existing school curriculum or training at a workplace. The main objective of this work is to validate the effect of the integration of microlearning into human verification.

2.2 Opportunistic microlearning

Microlearning is a learning style where learners undergo small learning units repeatedly. Tasks in microlearning are deliberately designed to be small so that learners can complete them within a short amount of time. Another merit of microlearning is that it can be integrated into users' interactions or tasks to provide learning moments in an opportunistic manner. Prior work in the field of Human-Computer Interaction has examined the learning effect of opportunistic microlearning systems.

Many projects targeted vocabulary development through their opportunistic microlearning systems. Trusty and Truong [16] developed a browser extension that automatically translated words on a Web page in English to a foreign language users were learning. The foreign translations were thus integrated into the existing context in English, offering opportunistic vocabulary learning when users were reading Web pages. Their user study revealed that participants were able to acquire 50 new foreign words per month on average. Cai et al. [2] created a vocabulary-learning system that exploits users' waiting time during a text chat. The study showed that participants learned 57 words in two weeks on average, indicating that the system was effective for vocabulary learning. Dingler et al. [3] implemented QuickLearn to exploit mobile notifications for microlearning. It presents users vocabulary questions via mobile notifications. In this manner, QuickLearn offers lightweight access to vocabulary learning materials even if users are on the go or only have a limited amount of attention. In their experiment, participants learned 18 words per week on average.

While vocabulary development is a common opportunistic microlearning application, this work extends its scope to online safety and ethics learning. Mohammed et al. [10] conducted a study incorporating microlearning into ICT education for elementary school students. They found that microlearning with flashcards and videos increased learning ability by up to 18% compared to textbook-based education and also resulted in better retention of long-term memory. Our investigation of this work demonstrates how effective opportunistic microlearning of online safety and ethics would be in a case of the integration into human verification.

2.3 GUI-based human verification

Human verification systems distinguish users from bots to prevent malicious automated access. CAPTCHA [17], developed by Ahn et al., is one of the most widely-used human verification systems. An early version of CAPTCHA required users to correctly type a visually-skewed string. Ahn et al. also developed reCAPTCHA [18]. It provides the same function as the original CAPTCHA but can also improve OCR software. However, problems in functionality and usability were also recognized. Yan and El Ahmad [20] discussed the robustness and usability of text-based CAPTCHAs. They pointed out usability concerns owing to the degree of distortion of the text and the presence of confusing characters.

To address such usability issues, research has examined alternative forms of human verification. Yamamoto et al. [19] designed a task of reordering four-frame cartoons. Fanelle et al. designed new audio CAPTCHAs that are primarily used by users with visual impairments [4]. Their designs were superior to those of existing audio CAPTCHA in terms of accuracy and speed. Recent developments on CAPTCHA have led to more lightweight interaction for human verification. reCAPTCHA v2 only requires the user to click a checkbox. reCAPTCHA v3 does not even require any explicit interaction from users.

The objective of this work is not to propose a novel human verification task nor evaluate its robustness and usability. Our primary advantage is the integration of opportunistic microlearning into a human verification task. Tanthavech et al. [15] showed Math CAPTCHA, which asks users to solve simple calculation problems, received the highest user experience rating among the five human verification task designs. One possible reason for this is that such tasks might have served as quick brain exercises. We hypothesize that a human verification task would become more acceptable if users could perceive benefits directly from it. This work examines this hypothesis in the context of online safety and ethics learning.

3 DualCheck

3.1 System Implementation

Our system, DualCheck, provides opportunistic microlearning of online safety and ethics through the human verification task of ticking a checkbox, similar to reCAPTCHA v2. More specifically, our interface presents users with a multiple-choice question about online safety and ethics. In this manner, users can learn online safety and ethics while performing human verification tasks.

Figure 1 shows the interface implemented in a Web environment. Our interface can be easily integrated into online forms for human verification. DualCheck shows a multiple-choice question about online safety and ethics comprising two statements, and users are asked to determine

whether each statement is correct. The following five choices are provided as responses: "Only Statement A is correct," "Only Statement B is correct," "Both statements are correct," "Both statements are wrong," and "I cannot tell." While deeply investigating the learning effect of question and response formats is out of scope of this work, we decided to employ a multi-choice question for DualCheck because it is a common question style and reCAPTCHA v2 would fit this style well. After users tick one of the checkboxes, the system presents the correct answer and a short explanation to encourage them to acquire the appropriate knowledge. The system offers users 5 seconds before enabling the button to move to the next page. In this manner, it encourages users to read the correct answer and explanation.

The human verification in our system would not be based on whether users answer questions correctly. Instead, human verification is expected to be performed in a manner similar to that in reCAPTCHA v2, i.e., analyzing the cursor behavior when clicking a checkbox. We also note that reCAPTCHA v2 or equivalent human verification mechanisms are not integrated into our current prototype due to the unavailability of these codes. Moreover, our primary purpose of this work is to investigate the effect of opportunistic microlearning instead of human verification performance.

The question curation and user studies for DualCheck were executed in the local language of the authors though other languages can be accommodated. We translated the questions and answers into English for the report in this paper.

3.2 Question Curation

We created a set of questions for our demonstration and deployment study of DualCheck. We set the following two criteria to create questions: 1) questions should cover common issues and practices related to online safety and ethics; 2) questions should neither be too difficult nor too well-known. Using these criteria, we conducted a literature survey of existing online safety and ethics guidelines and learning materials designed for high school or older users. These references included materials for teaching high school students [11] and materials to educate the public on the latest knowledge of Internet hazards [8].

We initially prepared 29 questions from these resources, which covered various common online threats. One of the authors, an expert on network security systems, reviewed them to filter out questions considered too difficult or obvious. We also revised the phrasing of the questions based on their feedback. Finally, we had 25 questions.

We then conducted a crowdsourcing-based study to validate these 25 questions. The objective of this part of the study was two-fold: 1) confirming whether the questions were comprehensible and 2) observing how many participants would respond to these questions correctly. We used a crowdsourcing service available in the country of the authors. Each crowdsourcing participant was asked to answer a subset of the 25 questions in a multiple-choice format. In addition, the task included a quality control question where its answer was obvious even for the general user populations (e.g., "I posted the password to my account on an SNS."). This question curation process was approved by our institutional review board.

331 crowdsourcing participants volunteered for this study in total, and we collected 100–115 responses for each question (110 on average). 20 participants failed the quality control question, and their responses were discarded. Table A.1 in the Appendix A includes the entire set of questions and their percentage of the correct answers.

We found that 15 of the 25 questions exhibited correct answer rates greater than 80%. These questions would not be appropriate for our deployment study because people are already aware of these online safety and ethics issues. Consequently, we chose the remaining ten questions where the correct answer rates were below 80% with small modifications on their phrasing. Q1–10 in Table 6 are the final set of the ten questions and corresponding answers that we used in our deployment study.

4 Deployment Study

A 15-day user study was conducted to evaluate DualCheck. The primary objective of our study is to examine the effect of microlearning supported by DualCheck rather than its human verification performance. As explained in Subsection 3.1, our current prototype does not integrate a human verification mechanism. The following user study protocol was approved by our institutional review board.

4.1 Task Design

We designed a deception study to avoid potential bias in the evaluation of DualCheck. In contrast to other microlearning systems, DualCheck offers implicit, opportunistic learning. Thus, we designed a study similar to an experience sampling method, probing participants' Internet usage through a short questionnaire (e.g., how much time they spent on social networking sites on that day). We created four different sets of such questionnaires, and used randomly during the deployment study. We then included DualCheck at the bottom as a human verification task. The participants were then asked to respond to the questionnaire multiple times a day throughout the experiment. DualCheck showed one of the ten questions shown in Table 6. Each question was exposed to the participants four times throughout the experiment. The order of these questions was randomly shuffled, and they were presented to all participants in the same order.

4.2 Procedure

4.2.1 Day #1

We asked participants to review their consent forms and sign them. We then asked them to fill a pre-experimental questionnaire that included demographic questions. We also asked the 10 questions about online safety and ethics that were also used during the deployment study. However, we did not provide the answers to these questions. The performance on these 10 questions served as the baseline for the later analysis.

We then explained the tasks and questionnaire form used in our deployment study. The details of theDualCheck implementation was not explained to the participants, in particular DualCheck did not include an actual human verification mechanism.

4.2.2 Day #2–#14

We sent emails that included the link to our short online questionnaire three times a day between Days #2 and #13, and four times on Day #14. Our participants then filled the questionnaire and responded to the questions in DualCheck. We set two modes in DualCheck for a comparison of the learning performance: *OneTime* and *Repeat*. The *OneTime* mode indicates that DualCheck shows participants the correct answer and explanation immediately after they tick one of the choices. In the *Repeat* mode, DualCheck forced participants to respond to the question until they ticked the correct answer. When they initially ticked a wrong answer, the system showed the correct answer and explanation, and asked participants to update their responses. After they chose the correct answer, the system enabled the button to submit a form. This mode was derived from the behavior of existing CAPTCHA systems where users would need to succeed the given verification tasks to pass. We constantly monitored the participants' responses and reminded them if they had not responded an hour before the submission deadline of each questionnaire.

4.2.3 Day #15

At the end of the study, we first debriefed all participants and revealed that this was a deception study and informed them of the true objective of the study, examining how DualCheck would influence on the learning of the content of the questions. However, we did not reveal how the human verification was performed in DualCheck nor that DualCheck did not implement an actual human verification mechanism. They were then offered an explicit opportunity to withdraw themselves from this study, but none of them withdrew.

Subsequently, we asked them to complete the post-experimental questionnaire. This questionnaire comprised 2 sections. The first section contained 30 questions that gauge respondents' knowledge on online safety and ethics. 10 of the questions were the same as those used in the deployment study. Another 10 questions were similar to the first ten, and were simply paraphrased to appear different (Q1a–10a in Table 7). The remaining 10 questions were new questions that participants had never given and were used as distractors. The presentation order of the questions was randomized for each participant. The second section was designed to probe the participants' experience and perceived usability of DualCheck. For usability assessment, we used the System Usability Scale (SUS) [1]. In addition, we included free-form questions to collect opinions on DualCheck.

4.3 Participants

We recruited 34 participants (25 females and 9 males; 6, 9, 13, 4, and 2 in their 20s, 30s, 40s, 50s, and 60s, respectively) using the same crowdsourcing service used in our question curation. None of the participants participated in the study related to the question curation. We randomly split the participants into two groups to compare the effects of the presentation modes of DualCheck: 16 for *OneTime* and 18 for *Repeat*. The participants were offered approximately 22 USD in local currency for the completion of the 14-day short online questionnaires. They were additionally offered 2.6 USD in local currency for the completion of the post-experimental questionnaire. All the participants completed the entire experiment including the post-experimental questionnaire.

4.4 Hypotheses

We summarize our hypotheses to test through our deployment study below:

H1. The accuracies of the 10 questions used throughout the deployment study (Q1–10) would be higher at the post-experiment phase than the pre-experiment phase. This is because we expected participants to learn online safety and ethics through DualCheck.

H2. The accuracies of the 10 questions that are similar to Q1–10 but only shown at the post-experimental questionnaire (Q1a–10a) would be higher compared to those by general Internet users who do not use DualCheck. This is because participants would develop relevant knowledge to answer these questions correctly through DualCheck.

H3. The usability of DualCheck would be higher than that of Text-based CAPTCHA and picture-based reCAPTCHA. This is because the human verification is as simple as the checkbox-based reCAPTCHA.

H4. The accuracies of Q1–10 in the *Repeat* mode would be higher than those in the *OneTime* mode. This is because participants would learn more by responding to questions until reaching the correct answers.

H5. The usability of the *Repeat* mode would be lower than that of the *OneTime* mode. This is because participants would be forced to choose the correct answer.

5 Results

5.1 Learning Performance

The primary objective of our deployment study is to examine the effect of DualCheck on microlearning. Thus, we first investigated the improvements in the accuracy (percentage of correct answers) for the ten questions used throughout the deployment study (Q1–10 in Table 6).

The mean accuracies of the ten questions (Q1–10) in the pre-experimental and post-experimental questionnaires were 0.68 (*SD*=0.11) and 0.94 (*SD*=0.04), respectively. We then conducted a two-way ANOVA test with the factors of the experiment phase (pre-experiment and post-experiment) and system mode (*OneTime* and *Repeat*). It revealed a significant result on the experiment phase ($F(1,32)$=89.38, p<.001, generalized η^2=.50), but not the system mode ($F(1,32)$=1.12, p=.30, generalized η^2=.02). This suggests significant improvements in accuracy for the ten questions our participants had seen during the deployment study.

We then further looked into the accuracy differences of these ten questions between the pre-experimental and post-experimental phases. Table 1 shows the accuracy breakdowns for the 10 questions (Q1–10). We conducted a binomial test for each question to better understand these differences. The binomial test is a statistical test that uses the binomial distribution to determine whether the proportion of data in two categories is significantly deviated from the theoretically-expected distribution. The accuracy in the post-experiment phase would be the same as in the pre-experiment phase if DualCheck did not contribute to participants' learning effectively. Table 1 includes the 95% confidence intervals and p values derived from our binomial tests. All questions except Q2 revealed significant positive results. This result confirms strong positive learning effect of DualCheck.

We next looked into the performance of the ten questions similar to Q1–10, which were only exposed to our participants at the time of the post-experimental questionnaire (denoted as Q1a–Q10a). As these questions were not answered at the pre-experiment phase, we separately collected the reference accuracy for them through another crowdsourcing task. By taking a similar data collection method to our question curation, we recruited 50 new crowdsourcing participants (17 females and 33 males; 6, 16,17, 7, and 4 in their 20s, 30s, 40s, 50s, and 60s, respectively) who had not participated in any study related to this project.

	Pre-test accuracy	Post-test accuracy	95% CI	p	
Q1	0.79	0.97	[0.85, 1.00]	<.01	**
Q2	0.94	0.94	[0.80, 0.99]	1.00	
Q3	0.68	0.91	[0.76, 0.98]	<.01	**
Q4	0.68	0.97	[0.85, 1.00]	<.001	***
Q5	0.65	1.00	[0.90, 1.00]	<.001	***
Q6	0.56	0.94	[0.80, 0.99]	<.001	***
Q7	0.65	0.94	[0.80, 0.99]	<.001	***
Q8	0.56	0.97	[0.85, 1.00]	<.001	***
Q9	0.62	0.85	[0.69, 0.95]	<.01	**
Q10	0.68	0.94	[0.80, 0.99]	<.01	**

Table 1: The accuracies of Q1–10 observed in the pre-experimental and post-experimental questionnaire in the deployment study. In this and later tables, we also include the binomial test result for each question.

	Reference accuracy	Pre-test accuracy	95% CI	p	
Q1	0.86	0.79	[0.62, 0.91]	.32	
Q2	0.86	0.94	[0.80, 0.99]	.22	
Q3	0.72	0.68	[0.50, 0.83]	.57	
Q4	0.66	0.68	[0.50, 0.83]	1.00	
Q5	0.68	0.65	[0.47, 0.80]	.71	
Q6	0.48	0.56	[0.38, 0.73]	.39	
Q7	0.80	0.65	[0.47, 0.80]	<.05	*
Q8	0.76	0.56	[0.38, 0.73]	<.05	*
Q9	0.62	0.62	[0.44, 0.78]	1.00	
Q10	0.82	0.68	[0.50, 0.83]	<.05	*

Table 2: The accuracies of Q1–10 observed in a separate data collection study (denoted as "reference accuracy") and the pre-experimental questionnaire in the deployment study.

	Reference accuracy	Post-test accuracy	95% CI	p	
Q1a	0.74	0.85	[0.69, 0.95]	.17	
Q2a	0.98	0.88	[0.73, 0.97]	<.01	**
Q3a	0.52	0.74	[0.56, 0.87]	<.05	*
Q4a	0.64	0.88	[0.73, 0.97]	<.01	**
Q5a	0.90	1.00	[0.90, 1.00]	<.05	**
Q6a	0.20	0.74	[0.56, 0.87]	<.001	***
Q7a	0.80	0.91	[0.76, 0.98]	.13	
Q8a	0.70	0.97	[0.85, 1.00]	<.001	***
Q9a	0.72	0.71	[0.53, 0.85]	.85	
Q10a	0.98	0.97	[0.85, 1.00]	.50	

Table 3: The accuracies of Q1a–10a observed in a separate data collection study (denoted as "reference accuracy") and the post-experimental questionnaire in the deployment study.

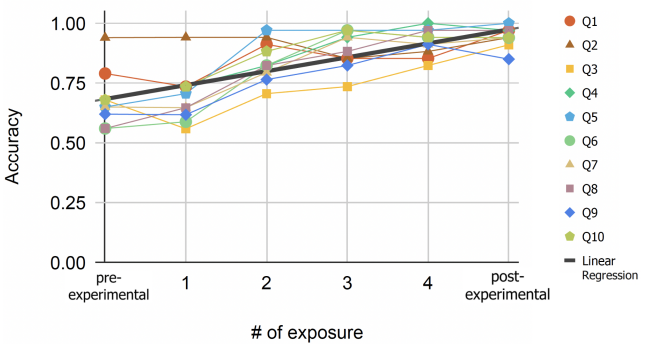

Figure 2: The accuracy transition across the number of exposure to the 10 questions. The plot includes participants' overall performance in each of the four exposures as well that in the pre-experimental and post-experimental questionnaires, which results in the six measurement points. The regression result was $y = 0.06x + 0.68$ (adjusted R^2=0.57).

They were asked to respond to the 30 questions of Q1–10, Q1a–10a, and 10 distractor questions. We then derived the accuracy rates of Q1a–10a, which we regard as the reference accuracy, ultimately summarized in Table 3. Each crowdsourcing participant was compensated approximately 1.7 USD in their local currency at the completion of the task.

The average accuracy of Q1–10 in a separate data collection study explained in the previous paragraph was 0.73 (*SD*=0.16) while it was 0.68 (*SD*=0.11) in the pre-experimental questionnaire. Our t test did not find a significant result between these two groups ($t(49,33)$=1.26, p=.21, Cohen's d=0.13). Table 2 shows the accuracy difference between the pre-experiment phase in our deployment study and a separate data collection study (denoted as "reference accuracy"). Our binomial tests confirmed significant differences in Q7, 8, and 10, where the accuracy in the pre-experimental questionnaire was lower. While we observed some accuracy differences, we did not find a significant difference in the average accuracies. We thus concluded that the performance comparison on Q1a–10a between these two groups would not be strongly biased in favor of either way.

The average accuracies of Q1a–10a were 0.72 (*SD*=0.14) and 0.94 (*SD*=0.04) in a separate data collection study and the post-experimental questionnaire, respectively. Our t test revealed a significant result between these two groups ($t(49,33)$=8.49, p<.001, Cohen's d=0.70). Table 3 presents the accuracy difference on Q1a–10a between the post-experiment phase in our deployment study and a separate data collection study. Our binomial tests confirmed significant differences in 6 of the 10 questions (Q2a, Q3a, Q4a, Q5a, Q6a, and Q8a). All these significant results except Q2a were associated with higher accuracies in the post-experiment phase in our deployment study.

We further examined how the accuracies were improved during the experiment. Figure 2 presents the accuracies across questions and the number of exposures. As explained above, participants saw each of the ten questions four times. Our linear regression analysis shows a significant effect of the number of exposure (estimated coefficient: 0.07, p<.001). The goodness of fit was .52 (adjusted R^2). Due to large variances in the accuracies we observed in the deployment study, the fitting was not very strong. However, our analysis results confirm an increasing trend of accuracies, suggesting a positive learning effect caused by DualCheck.

Verification system	Mean SUS (*SD*)
Text-based CAPTCHA	54.45 (15.46)
Picture-based reCAPTCHA	53.10 (18.82)
Checkbox-based reCAPTCHA	80.60 (13.28)
DualCheck *OneTime*	69.38 (13.02)
DualCheck *Repeat*	78.47 (13.96)
DualCheck average of both modes	74.19 (14.10)

Table 4: The mean SUS scores and their standard deviations of DualCheck and existing human verification systems.

5.2 Usability Comparison

We next examined the usability of DualCheck through the System Usability Scale (SUS) [1]. To better understand the SUS results, we conducted another data collection on the SUS scores of the existing CAPTCHA systems. They included text-based CAPTCHA, picture-based reCAPTCHA, and the reCAPTCHA Checkbox. We designed another task to collect these SUS scores in the same crowdsourcing service. All participants were offered an opportunity to participate in this data collection and a compensation of approximately 1 USD in the local currency. Consequently, 50 new participants who did not participate in our question curation or deployment study participated in this scoring task.

Table 4 presents the average SUS scores and the standard deviations of DualCheck and the three human verification systems mentioned above. A one-way ANOVA revealed significant differences in the factors of the human verification interfaces ($F(3,180)$=37.51, p<.001, generalized η^2=.63). Our Scheffe's test further showed that the SUS score of DualCheck was significantly higher than those of text-based CAPTCHA (p<.001) and picture-based CAPTCHA (p<.001). Our t test did not find a significant difference between the *OneTime* and *Repeat* modes in DualCheck($t(15,17)$=-1.96, p=0.06, Cohen's d=0.67). These statistical results confirm that the perceived usability of DualCheck was significantly higher than that of text-based CAPTCHA and picture-based reCAPTCHA.

Figure 3: The distribution of the responses about question difficulty.

Figure 4: The distribution of the responses about whether participants felt that they were able to acquire new knowledge about online safety and ethics through DualCheck.

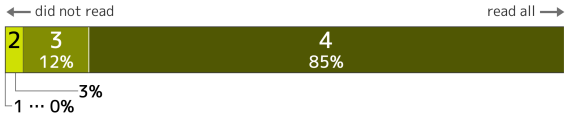

Figure 5: The distribution of the responses about how carefully participants read the questions, correct answers, and explanations.

5.3 Impressions on Questions in DualCheck

We further analyzed the participants' responses to our questions about the questions presented in DualCheck. Figure 3 shows the distribution of the participants' responses to the question about the overall difficulty of the questions they saw in DualCheck (1: The questions were too easy–5: The questions were too difficult). 20 participants (59%) considered that the questions were at the appropriate level, confirming that our question curation was properly executed. Figure 4 summarizes how strongly participants agreed that they were able to acquire new knowledge about online safety and ethics through questions provided by DualCheck. All but one participant agreed that they were able to learn through the questions. Figure 5 shows the participants' responses to the question about whether they thought they read the correct answers and explanations on a 4-point Likert scale. 29 participants (85%) responded that they read questions, correct answers, and explanations. All of these results suggest participants' positive experience with DualCheck.

5.4 Qualitative Results

We further examined the comments we received through open-ended questions to deepen our understanding of participants' experiences with DualCheck. Two of the authors jointly conducted thematic analysis and developed six themes that categorize the quotes of comments for overall deployment study. We discarded the quotes that these two authors disagreed in categorization. As a result, all categories had the perfect agreement between the two authors. Table 5 shows our categorization and quote occurrence for each category. We note that the quotes presented below were originally written in our local language, and we translated them into English as faithfully as possible for the report in this paper.

Theme and Subtheme	# quotes
Questions	
Question difficulty	15
Issues on question presentation	9
Issues on answer explanations	5
Advantages of DualCheck	
Perceived advantages	15
Usability of DualCheck	
Positive opinions on usability	11
Issues on usability	10
Suggestions	
Possible improvements	8

Table 5: The categorization of participants' comments collected in the deployment study. We note that we only considered the comments that two of the authors agreed in their categorization and used for our analysis. Thus, all the categories above exhibited the perfect agreement.

5.4.1 Perceived Benefits of DualCheck

We observed explicit comments where participants appreciated DualCheck for offering unique microlearning opportunities. For instance, P28 and P33 offered their appreciation on DualCheck over existing human verification systems by highlighting its direct benefits to users.

> *I'm worried about phishing scams and other sophisticated scams these days, so I thought it would be good to have a lot of such problems. This is much better for learning than doing puzzles that are not easy to use, so I would like to see this implemented in general Websites.* [P33]

> *I thought it would be more interesting than a bot detection system that requires input of known illegible strings, and it would kill two birds with one stone because it would be simple and learnable.* [P28]

We further examined the participants' responses to an open-ended question about which questions were the most memorable. Fourteen participants explicitly mentioned that the question about cookies (Q6) was the most memorable. P28 and P29 shared the following comments about Q6.

> *I have gained more knowledge about information literacy in general, which I had been unclear about. In particular, I have gained accurate and clear knowledge about cookies. I also learned that I should be careful about key-marked sites, which I had blindly trusted in the past.* [P28]

I learned a lot because I knew the name of cookies, but not the details. [P29]

Q1, Q3, and Q9 were mentioned by 5, 6, and 4 participants, respectively. P6 commented on how memorable Q3 was and how it promoted awareness of the SSL presentation and the URL in a browser.

There was a lot of information that I didn't know, but the question on URLs starting with https:// left a particular impression on me. I don't usually check URLs, so I thought I'd pay attention to it from now on. [P6]

Both the quantitative and qualitative results strongly confirm the benefits of DualCheck, particularly its capability to offer microlearning opportunities.

5.4.2 Possible Improvements

Our participants suggested several improvements to DualCheck. Five participants explicitly commented that they would like to see more variation in the questions. Undo and redo features were common requests; they were suggested by two participants who were grouped into the *OneTime* condition. Our SUS comparison did not indicate statistically significant differences between the two presentation modes. Thus, the *Repeat* mode may solve these issues. Future studies should examine how to fine-tune the interface settings of DualCheck to improve the learning experience while reducing users' cognitive load. In general, DualCheck successfully encouraged participants to read questions carefully.

I thought it was very good that I could study every time. The fact that the questions are repeated every day, and that I can't re-select the options, allows me to concentrate on reading the questions and learn about things that I've only vaguely been familiar with. [P32]

The same participant also commented that the question content would substantially impact on the user experience of DualCheck. This may suggest a future research direction of personalization on topics.

I felt that it was very annoying for those who were not interested in the content, because it took a lot of brainpower to prove that I was not a bot. I was also interested in the content of this problem, so I felt I learned a lot, but if it had been a fashion problem, for example, I would have hated it. [P32]

Participants were motivated to receive more detailed explanations. The general opinion was that these improvements would not only make explanations more accessible to the general user populations but also help users learn online safety and ethics by themselves.

I remember that I always answered the same question wrong. As for the safety of the Internet, even though I understood what I should not do (such as not clicking on links unnecessarily), I did not understand the technical terms (such as domain names) properly, so I think I answered some of the questions on a hunch. It would have been nice to have a simple explanation of these IT terms. [P1]

For example, when the question is about "writing with storage services", I thought it would be good to have one or two examples of service names to show what kind of storage services are available. I was a little confused at first if it was the one I was thinking of or not. I also thought that it might be difficult to understand for people who have never used that service before. [P8]

Participants also suggested dynamic adjustment of difficulty depending on people's correct responses, more complex response styles (e.g., the "Choose all that apply" response style), and integration of gamification (e.g., awarding points for correct responses).

6 Discussion

As shown through our quantitative results, we observed positive learning effects of DualCheck. The subjective ratings and open-ended comments we obtained in the post-experimental questionnaire also support participants' positive experience in learning online safety and ethics. We conclude that our results support H1.

The accuracy of the ten questions used throughout the deployment study (Q1–10) had significant improvements except for Q2. The accuracy of Q2 was 0.94 even at the time of the pre-experiment phase, and it remained the same after the experiment. We do not have clear reasons why only Q2 exhibited such high accuracy. The accuracy of the remaining questions in the pre-experiment phase was below 80%, which were in line with our results during the question curation. We thus concluded that our question choice was appropriate in general.

5 of the 10 questions similar to Q1–Q10 and asked only in the post-experimental questionnaire (Q3a, Q4a, Q5a, Q6a, and Q8a) showed significant accuracy improvements compared to the reference accuracy. This is a promising result as participants were able to extend their knowledge to answer unseen questions correctly to some extent.

The accuracy of Q2a in the deployment study was significantly lower than the reference accuracy. This result might be related to the fact that participants did not have improvements in the accuracy of Q2. Our deployment participants were able to answer correctly from the beginning and thus might not had paid careful attention to the explanation offered by the system. This result suggests

that a future system should provide variations of the same questions (e.g., paraphrasing or converting expressions from the affirmative to the negative form) or different questions about the same topic to reinforce users' learning. In conclusion, H2 is not fully supported in this study.

Although DualCheck increases the overall performance time for human verification tasks, the usability assessment we obtained showed a higher rating for DualCheck than text-based CAPTCHA and picture-based reCAPTCHA. Our qualitative evidence also suggests that participants were able to explicitly observe the learning benefits of DualCheck, which could contribute to its higher perceived usability. Our SUS results showed that we did not have a significant result between DualCheck and checkbox-based reCAPTCHA are interpretable because interaction requested by both systems was equivalent. We thus conclude that our results support H3. This result also suggest that users could be more willing to engage in microlearning during human verification tasks because they can perceive more direct benefits to them.

Our results did not reveal strong evidence about the two presentation modes of DualCheck in terms of learning effect and perceived usability. Thus, H4 and H5 are not supported. However, other factors, such as question content, the frequency of presenting the same question, and users' personal preferences, might have influenced this result, and future work should further examine these effects.

7 Limitations and Future Work

There are several limitations to be discussed to clarify the scope and contributions of this work. We recruited our study participants through a crowdsourcing service available in the country of authors. This implies that our participants might have been more accustomed to using online services and human verification systems than the general user populations. As they could be considered active Internet users, they might be more attentive to online safety and ethics, which might have led to a positive bias toward DualCheck. Future work should conduct a wider scale of user studies to validate the effect of DualCheck.

We took the design of an experience sampling method for our deployment study to offer repeated exposure to DualCheck. In a more realistic setting, users would not see our system as frequently as our deployment study. Thus, understanding the learning effect of DualCheck in a more realistic setting requires additional studies.

While our current investigation focused on online safety and ethics questions, future work may expand the scope to other kinds of privacy and safety threats and practices (e.g., fraud in the physical world and fake news). The results of our study anticipate positive learning effects on these topics, and further examinations are encouraged.

Another important future research direction is to investigate the effect of question and response formats. Different question formats (e.g., dichotomous or free-form questions) might have different learning effect. Similarly, response methods can also influence on learning behavior. Even using the same question, users might exhibit different accuracy rates depending on the question and response formats. Our current implementation utilizes an interaction modality derived from reCAPTCHA v2 (ticking a checkbox), but advanced CAPTCHA systems does not even require explicit interaction like reCAPTCHA-v3. With such technology, a future system can completely decouple human verification and interaction for microlearning, which would allow researchers to explore different forms of microlearning. Our work serves as a foundation of such future work to integrate human verification and microlearning.

The administration of questions is necessary in a practical setting. Officers in charge of information management for organizations may take this responsibility to employ DualCheck for their members. In particular, we envision that DualCheck can complement existing learning activities at educational institutions. Future work should examine the longer-term effect of DualCheck as well as its deployment in a more practical setting.

8 Conclusion

Learning online safety and ethics is becoming more critical. However, they lack such learning opportunities and are often left behind. We introduce DualCheck, a microlearning system that is integrated into human verification tasks. Users are asked to respond to questions related to online safety and ethics while human verification would be executed in a similar manner to reCAPTCHA v2. In this manner, DualCheck offers users microlearning opportunities when they use online services. Our 15-day user study confirmed the positive learning effect of DualCheck. The quantitative and qualitative results also supported participants' positive attitudes toward DualCheck. The usability of DualCheck was rated significantly higher than those of text-based CAPTCHA and picture-based reCAPTCHA. We plan to further investigate the effect of DualCheck by expanding our studies to a wider user population and incorporating more learning topics.

	Statement and answer
Q1	A: Connecting a USB flash drive to a computer in public is a security risk. B: Charging a smartphone via USB on a computer in public is a security risk. Correct Answer: Both statements are correct.
Q2	A: On social networking sites, there is no privacy problem in sharing selfies and other information if you give limited access. B: On social networking sites, if you don't post any personal information, your identity will not be identified. Correct Answer: Both statements are wrong.
Q3	A: This is the first time I visited this Website, but I thought it was safe because it had a key symbol on my browser, so I entered my personal information. B: I entered my personal information on a Website beginning with http:// . It is risky to enter personal information on such a Website. Correct Answer: Only statement B is correct.
Q4	A: Passwords should be a combination of letters, numbers, and symbols that are difficult to remember. B: Passwords are safer if they are based on personal information, such as your hobbies, and avoid famous words that are easily guessed. Correct Answer: Only statement A is correct.
Q5	A: When the earthquake struck, local people posts the situation in the area. Even if you don't know whether it is true information, it is better to share the information quickly. B: When spreading information when an earthquake or other event occurs, it is better to only spread posts by the government or news organizations. Correct Answer: Only statement B is correct.
Q6	A: A cookie is a piece of information that sends a user's name and other personal information to a site administrator. B: Cookies are used for retargeting advertisements and other purposes. Correct Answer: Only statement B is correct.
Q7	A: Documents created with online storage services and document creation tools are not disclosed to the public. B: Documents created with online services can be seen by others through searches. Correct Answer: Only statement B is correct.
Q8	A: The procedure for requesting information about an offensive social networking account has been made easier due to a change in the law. B: Even if there is an offensive SNS account, it is difficult to identify their source address. Correct Answer: Only statement A is correct.
Q9	A: To verify that the email you received was sent from a real bank or other sources, you check the back of the @ in the source address. B: Checking the domain is one of the most important things to ensure that the URL sent to you is authentic. Correct Answer: Only statement B is correct.
Q10	A: Photos taken with a smartphone may contain location information. B: If you post a photo without the location information to a social networking site, your location will not be identified. Correct Answer: Only statement A is correct.

Table 6: The questions used in this work. Q1–10 are derived from our question curation process. They were originally written in the local language of the authors, and are translated into English as faithfully as possible.

Q1a	A: If you use a computer's USB port only to charge your smartphone, no viruses or other devices will be transferred. B: If you connect a USB flash drive to a shared computer, viruses and other malicious programs may be copied. Correct Answer: Only statement B is correct.
Q2a	A: On social networking sites, if you limit the number of people you can follow, there is no problem if you tweet personal information. B: Your identity can be identified based on your following relationship on social networking sites. Correct Answer: Only statement B is correct.
Q3a	A: Websites that start with http://... do not support encrypted communication. B: If the Website is capable of encrypted communication, it is safe to send personal information. Correct Answer: Only statement A is correct.
Q4a	A: Passwords should be a meaningless string of characters with symbols. B: It is preferable to create a password based on a hobby or something that you keep secret from others. Correct Answer: Only statement A is correct.
Q5a	A: An earthquake occurred, but there was no information from the news media or government, so I spread a post made by a person claiming to be a local. B: When the earthquake occurred, a person claiming to be a scholar on Twitter explained the situation. It is considered as credible information. Correct Answer: Both statements are wrong.
Q6a	A: The use of cookies can customize ads. B: Allowing the use of cookies is likely to leak personal information. Only statement A is correct.
Q7a	A: Documents created with online document creation tools are not likely to show up in a Web search. B: It is important to check the publication settings of documents created with online tools. Correct Answer: Only statement B is correct.
Q8a	A: It is difficult to identify the source address of an anonymous social networking account. B: You can file a request for disclosure of sender information against an offensive social networking account. Correct Answer: Only statement B is correct.
Q9a	A: Checking the domain of the URL is important to confirm whether it is genuine or not. B: I received an email claiming to be from my bank. It was the same domain as the bank's email, so I figured it was the right email. Correct Answer: Only statement A is correct.
Q10a	A: The scenery and objects in the photo could lead to the identification of personal information. B: Location information may be stored in the photo. Correct Answer: Both statements are correct.

Table 7: The 20 questions used in this work. We created another 10 questions (Q1a–10a) that are similar to Q1–10 to measure the deployment study participants' learning. They were originally written in the local language of the authors, and are translated into English as faithfully as possible.

Acknowledgements

We appreciate our lab members for giving us very helpful feedback and advice. Especially, Anran Xu offered great help on related work and thoughtful insights on this project. We also thank all participants for their help. This research is part of the results of Value Exchange Engineering, a joint research project between Mercari, Inc. and the RIISE.

References

[1] John Brooke. SUS-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[2] Carrie J. Cai, Philip J. Guo, James R. Glass, and Robert C. Miller. Wait-learning: Leveraging wait time for second language education. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3701–3710. Association for Computing Machinery, 2015.

[3] Tilman Dingler, Dominik Weber, Martin Pielot, Jennifer Cooper, Chung-Cheng Chang, and Niels Henze. Language learning on-the-go: Opportune moments and design of mobile microlearning sessions. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '17. Association for Computing Machinery, 2017.

[4] Valerie Fanelle, Sepideh Karimi, Aditi Shah, Bharath Subramanian, and Sauvik Das. Blind and human: Exploring more usable audio CAPTCHA designs. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 111–125. USENIX Association, August 2020.

[5] Galen A. Grimes, Michelle G. Hough, Elizabeth Mazur, and Margaret L. Signorella. Older adults' knowledge of internet hazards. *Educational Gerontology*, 36:173 – 192, 2010.

[6] Japan Information-technology Promotion Agency. Awareness survey on information security ethics in fy2019 report. https://www.ipa.go.jp/files/000080783.pdf, 2019. (Written in Japanese).

[7] Japan Information-technology Promotion Agency. Awareness survey on information security threats in FY2019 report. https://www.ipa.go.jp/files/000080784.pdf, 2019. (Written in Japanese).

[8] Japan Information-technology Promotion Agency. 10 major threats to information security 2021. https://www.ipa.go.jp/files/000088835.pdf, 2021. (Written in Japanese).

[9] Ulrike Meyer and Vincent Drury. Certified phishing: taking a look at public key certificates of phishing websites. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 211–223. USENIX Association, 2019.

[10] Gona Sirwan Mohammed, Karzan Wakil, and Sarkhell Sirwan Nawroly. The effectiveness of microlearning to improve students' learning ability. *International Journal of Educational Research Review*, 3(3):32–38, 2018.

[11] Ministry of Education, Culture, Sports, Science, and Technology. "Information I", Chapter 1 of the teaching materials of information science for high school teachers. https://www.mext.go.jp/content/20200722-mxt_jogai02-100013300_003.pdf, 2019. (Written in Japanese).

[12] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana von Landesberger, and Melanie Volkamer. An investigation of phishing awareness and education over time: When and how to best remind users. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 259–284. USENIX Association, August 2020.

[13] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI conference on human factors in computing systems*, CHI '10, pages 373–382. Association for Computing Machinery, 2010.

[14] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd symposium on Usable privacy and security (SOUPS 2007)*, pages 88–99. Association for Computing Machinery, 2007.

[15] Nitirat Tanthavech and Apichaya Nimkoompai. Captcha: Impact of website security on user experience. In *Proceedings of the 2019 4th International Conference on Intelligent Information Technology*, ICIIT '19, page 37–41. Association for Computing Machinery, 2019.

[16] Andrew Trusty and Khai N. Truong. Augmenting the web for second language vocabulary learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 3179–3188. Association for Computing Machinery, 2011.

[17] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. Captcha: Using hard ai problems for security. In Eli Biham, editor, *Advances in Cryptology — EUROCRYPT 2003*, pages 294–311, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[18] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.

[19] Takumi Yamamoto, Tokuichiro Suzuki, and Masakatsu Nishigaki. A proposal of four-panel cartoon captcha. In *2011 IEEE International Conference on Advanced Information Networking and Applications*, pages 159–166, 2011.

[20] Jeff Yan and Ahmad Salah El Ahmad. Usability of captchas or usability issues in captcha design. In *Proceedings of the 4th Symposium on Usable Privacy and Security (SOUPS 2008)*, page 44–52. Association for Computing Machinery, 2008.

A Questions used during our Question Curation

Table A.1 had 25 questions (Qc1 – Qc25) through the question curation phase. We then collected the percentage of the correct answers. shows the questions and their accuracies.

B Distractor questions used in the pre-experimental and post-experimental questionnaire

Table B.1 shows the 10 distractor questions used in the pre-experimental and post-experimental questionnaire.

ID	accuracy	Statements
Qc1	0.87	A: Even if you post anonymously, there is a chance that you will be identified. B: I want to say something bad about my friend, but if I do so directly, it will damage our relationship, so I post it on an anonymous forum. Correct Answer: Only statement A is correct.
Qc2	0.84	A: I received an email from a web service I use that ask me to change my password. The URL contained the company's name, so I assumed it was a real site and logged in. B:You need to be careful when click websites' links because scam sites can appear higher position in web searches. Correct Answer: Only statement B is correct.
Qc3	0.98	A: A friend sent me a link to a website he recommended. It is safe because it came from a trusted friend. B: Even if the link was sent by a friend, you need to check the URL carefully. Correct Answer: Only statement B is correct.
Qc4	0.89	A: In order to get more people to watch my favorite drama, I posted a scene from that drama on social networking sites to spread the word. B: Pictures and other materials posted by individuals are not registered with the Patent Office and are not copyrighted, so they may be freely reproduced. Correct Answer: Both statements are wrong.
Qc5	0.73	A: On social networking sites, there is no privacy problem in publishing selfies and other photos as long as the account is limited public. B: On social networking sites, as long as you don't post any personal information, your identity will not be identified. Correct Answer: Both statements are wrong.
Qc6	0.94	A: The advantage of anonymous message boards is that people can post easily, and there is no problem if they post wrong things. B: Anonymous forums can be dangerous as inaccurate content may be posted. Correct Answer: Only statement B is correct.
Qc7	0.85	A: If you connect to a wireless LAN from a trusted provider, you do not have to worry about others seeing your communications. B: Before connecting to a free wireless LAN, you should thoroughly check the terms and rules of use of the wireless LAN. Correct Answer: Only statement B is correct.
Qc8	0.69	A: Connecting a USB flash drive to a computer in an Internet cafe, etc. is a security risk. B:Charging smartphones via USB on computers in Internet cafes, etc., is a security risk. Correct Answer: Both statements are correct.
Qc9	0.82	Which of the following is the correct address for Google? A: https://google-co.jp B: https://google.co.jp C: https://google,co.jp D: https://goog1e.co.jp Correct Answer: B is the correct URL.
Qc10	0.83	A: Photos taken with a smartphone may contain location information. B: If you post a photo to a social networking site, the location information are removed automatically, so your location will not be identified. Correct Answer: Only statement A is correct.
Qc11	0.95	A: I received an email I don't recognize. There was a link to unsubscribe, so I clicked on it and took the necessary steps to unsubscribe. B: While browsing a website, the message "This smartphone has been compromised" was displayed, so I followed the instructions on the screen. Correct Answer: Both statements are wrong.
Qc12	0.54	A: Passwords should be a combination of letters, numbers, and symbols that are difficult to remember. B: Passwords are safer if you avoid famous words that can be easily guessed, and create passwords based on personal things like your hobbies. Correct Answer: Only statement A is correct.
Qc13	0.61	A: When an earthquake occurred, people were sending out information about the area. Even if the authenticity of the information is unknown, it is better to spread the information quickly. B: When spreading information after an earthquake or other event, it is better to only spread posts from the government or news organizations. Correct Answer: Only statement B is correct.

ID	accuracy	Statements
Qc14	0.62	A: This is the first time I visited a website, but my browser had a key symbol on it, so I thought it was safe and entered my personal information. B: It is risky to enter personal information on a website that begins with http:// . Correct Answer: Only statement B is correct.
Qc15	0.91	A: I saw information about COVID-19 on a social networking site. Since the profile said the author was a doctor, I thought it was correct and spread it. B: Several people mentioned the information about COVID-19, so I thought it was correct and spread it. Correct Answer: Both statements are wrong.
Qc16	0.88	A: Fingerprint and face recognition are not vulnerable to being breached because only you can unlock. B: Fingerprint and face recognition enhance security when they are combined with password locks. Correct Answer: Only statement B is correct.
Qc17	0.91	A: I got a warning message while browsing a website. It instructed me to install an application in the Google Play/App Store, so I downloaded it, thinking it was safe. B: An advertisement recommended an application. It was highly rated in the app store, so I thought it was safe and downloaded it. Correct Answer: Both statements are wrong.
Qc18	0.98	A: I posted a picture I liked that I found on a social networking site, claiming it to be my own work. B: A music program I forgot to record was reprinted on a social networking site, so I downloaded it to watch it later. Correct Answer: Both statements are wrong.
Qc19	0.69	A: Documents created with online storage services and document creation tools are never made available to the outside world. B: Documents created with online services may be seen by others through searches. Correct Answer: Only statement B is correct.
Qc20	0.88	A: I posted a photo of myself with a friend under a limited public access permission on an SNS at my own discretion. B: A post such as “The train I’m on is delayed” could identify where I live, etc. Correct Answer: Only statement B is correct.
Qc21	0.47	A: A cookie is a piece of information that sends a user’s name and other personal information to a site administrator. B: Cookies are used for targeted advertisement and other purposes. Correct Answer: Only statement B is correct.
Qc22	0.61	A: Legal changes have made it easier to request information about offensive social networking accounts. B: Even if there is an offensive social network account, it is difficult to identify the source of the slander. Correct Answer: Only statement A is correct.
Qc23	0.61	A: To minimize the damage caused by ransomware, backups need to be taken regularly. B: If you are a victim of ransomware, you will only lose the use of your data, which is not a problem if you have proper backups. Correct Answer: Only statement A is correct.
Qc24	0.85	A: Two-factor authentication can be set up to reduce the risk of unauthorized login. B: Two-factor authentication may include biometrics and one-time passwords. Correct Answer: Both statements are correct.
Qc25	0.41	A: To verify that an email you receive is from a real bank or other organization, just look at the back of the @ in the source address. B: One of the most important things to make sure that the URL sent to you is authentic is to check the domain. Correct Answer: Only statement B is correct.

Table A.1: 25 questions used during our Question Curation. They were originally written in a local language where the authors curated the questions, and are translated into English as faithfully as possible.

ID	Statements
D1	A: To minimize the damage caused by ransomware, backups need to be taken regularly. B: If you are a victim of ransomware, you will only lose the use of your data, which is not a problem if you have proper backups. Correct Answer: Only statement A is correct.
D2	A: Even if you post anonymously, there is a chance that you will be identified. B: I want to say something bad about my friend, but if I do so directly, it will damage our relationship, so I post it on an anonymous forum. Correct Answer: Only statement A is correct.
D3	A: A: A friend sent me a link to a website he recommended. It is safe because it came from a trusted friend. B: Even if the link was sent by a friend, you need to check the URL carefully. Correct Answer: Only statement B is correct.
D4	A: In order to get more people to watch my favorite drama, I posted a scene from that drama on social networking sites to spread the word. B: Pictures and other materials posted by individuals are not registered with the Patent Office and are not copyrighted, so they may be freely reproduced. Correct Answer: Both statements are wrong.
D5	A: The advantage of anonymous message boards is that people can post easily, and there is no problem if they post wrong things. B: Anonymous forums can be dangerous as inaccurate content may be posted. Correct Answer: Only statement B is correct.
D6	A: I received an email I don't recognize. There was a link to unsubscribe, so I clicked on it and took the necessary steps to unsubscribe. B: While browsing a website, the message "This smartphone has been compromised" was displayed, so I followed the instructions on the screen. Correct Answer: Both statements are wrong.
D7	A: I saw information about COVID-19 on a social networking site. Since the profile said the author was a doctor, I thought it was correct and spread it. B: Several people mentioned the information about COVID-19, so I thought it was correct and spread it. Correct Answer: Both statements are wrong.
D8	A: Fingerprint and face recognition are not vulnerable to being breached because only you can unlock. B: Fingerprint and face recognition enhance security when they are combined with password locks. Correct Answer: Only statement B is correct.
D9	A: I got a warning message while browsing a website. It instructed me to install an application in the Google Play/App Store, so I downloaded it, thinking it was safe. B: An advertisement recommended an application. It was highly rated in the app store, so I thought it was safe and downloaded it. Correct Answer: Both statements are wrong.
D10	A: I posted a picture I liked that I found on a social networking site, claiming it to be my own work. B: A music program I forgot to record was reprinted on a social networking site, so I downloaded it to watch it later. Correct Answer: Both statements are wrong.
D11	A: I received an email from a web service I use that ask me to change my password. The URL contained the company's name, so I assumed it was a real site and logged in. B:You need to be careful when click websites' links because scam sites can appear higher position in web searches. Correct Answer: Only statement B is correct.
D12	A: If you connect to a wireless LAN from a trusted provider, you do not have to worry about others seeing your communications. B: Before connecting to a free wireless LAN, you should thoroughly check the terms and rules of use of the wireless LAN. Correct Answer: Only statement B is correct.

Table B.1: Distractor questions used in the pre-experimental and post-experimental questionnaire. We chose 10 questions from this set for each questionnaire. They were originally written in a local language where the authors conducted the user study, and are translated into English as faithfully as possible.

C The Experience Sampling Method Interface Used in Our Study

Figure C.1 shows the screenshot of the questionnaire we used in our survey. It consists of questionnaire for ESM and DualCheck. The ESM part asked participants to answer their recent Internet usage (hours they had spent in SNS, shopping sites, and news sites) in the example below.

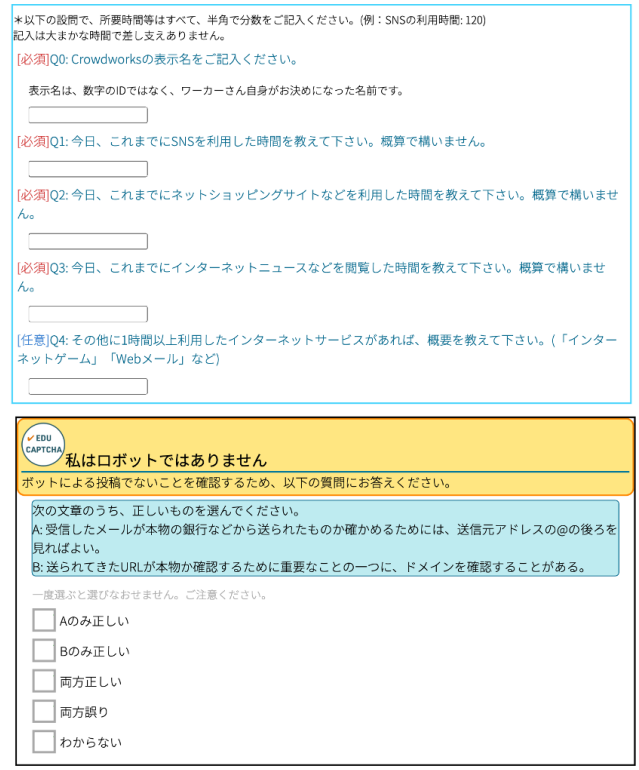

Figure C.1: The screenshot of the questionnaire we used during the deployment study.

D Post-experimental Questionnaire

We asked participants to complete post-experimental questions at the end of the study. The questionnaires consisted of two parts; the first part included 30 questions to gauge knowledge of online safety and ethics, and the second part was to probe the participants' experience and perceived usability of DualCheck. This section includes the questions we used in the second part. They were originally written in the local language of the authors, and are translated into English as faithfully as possible.

We referred DualCheck as "CAPTCHA Quiz" in this questionnaire.

- Please fill your ID of crowdsourcing service account.
- Please answer the following questions about your comfort with the CAPTCHA quiz. (We used SUS for this part.)
  - I think that I would like to use this system frequently.
  - I found the system unnecessarily complex.
  - I thought the system was easy to use.
  - I think that I would need the support of a technical person to be able to use this system.
  - I found the various functions in this system were well integrated.
  - I thought there was too much inconsistency in this system.
  - I would imagine that most people would learn to use this system very quickly.
  - I found the system very cumbersome to use.
  - I felt very confident using the system.
  - I needed to learn a lot of things before I could get going with this system.
- Please let us know if you have any feedback on the usability of the CAPTCHA quiz. Please tell us about any difficulties you had in operating the system or any points that made it easier to use. You can answer in free-form.
- Please answer the following items.
  - Overall, how difficult did you find the quiz? (1: Too easy – 5: Too difficult)
  - Do you think you gained new knowledge through this quiz? (1: Not at all – 5: Very much)
- How much did you read about the question and explanations of the CAPTCHA quiz? You can choose from the statements below.

  - I answered randomly and did not read the questions, correct answers, or explanations.
  - I read the questions, but not the correct answers and explanations
  - I read the questions and checked the correct answers, but did not read the explanations.
  - I read all the questions, correct answers, and explanations
- Please tell us about any particularly memorable content or new knowledge you learned in the CAPTCHA quiz. You can answer in free form.
- Please let us know any comments you have about the questions in the CAPTCHA quiz (e.g., They were too easy, too difficult, or any doubts about the answers). You can answer in free form.
- Were you aware that the original purpose of the survey was the experiment for CAPTCHA quiz? You can choose from the statement below;
  - I was aware that the purpose of the survey was to investigate CAPTCHA quiz.
  - I felt that there might be another purpose of the study
  - I was not aware of it.
- Please let us know if you have any comments or advice regarding the mechanism or content of the CAPTCHA quiz. We would be happy to hear any suggestions you may have, such as how we could improve the functionality or content of the quiz.