Better the Devil You Know: A User Study of Two CAPTCHAs and a Possible Replacement Technology

Kat Krol, Simon Parkin, M. Angela Sasse
University College London
E-mail: {kat.krol.10, s.parkin, a.sasse}@ucl.ac.uk

***Abstract*—CAPTCHAs are difficult for humans to use, causing frustration. Alternatives have been proposed, but user studies equate usability to solvability. We consider the user perspective to include workload and context of use. We assess traditional text-based CAPTCHAs alongside PlayThru, a 'gamified' verification mechanism, and NoBot, which uses face biometrics. A total of 87 participants were tasked with ticket-buying across three conditions: (1) all three mechanisms in comparison, and NoBot three times (2) on a laptop, and (3) on a tablet. A range of quantitative and qualitative measurements explored the user perspective. Quantitative results showed that participants completed reCAPTCHAs quickest, followed by PlayThru and NoBot. Participants were critical of NoBot in comparison but praised it in isolation. Despite reporting negative experiences with reCAPTCHAs, they were the preferred mechanism, due to familiarity and a sense of security and control. Although slower, participants praised NoBot's completion speeds, but regarded using personal images as invading privacy.**

I. Introduction

A CAPTCHA ("Completely Automated Public Turing test to tell Computers and Humans Apart") [21] is a challenge-response test where the user is tasked to recognise and enter a sequence of distorted characters. CAPTCHAs are integrated into web services with an aim to allow access only to humans and prevent automated programs (robots) from exploiting online resources (e.g., forum comment sections).

Now an established part of the online experience, CAPTCHAs are nonetheless a hurdle to users' completion of tasks. Pogue [19] says users are together wasting 17 person-years every day on solving CAPTCHAs, calling this immense effort "a disgraceful waste of our lives". There are technical solutions which remove the need for user effort, but they are difficult to implement [7].

In this study, we assess the user experience of three mechanisms. We look at (1) the traditional text-based CAPTCHAs, specifically reCAPTCHA[1] provided by Google (v 1.0), (2) an alternative called PlayThru[2], that requires the user to drag and drop elements in a simple game and (3) a technology which could mitigate the cognitive and physical demands of CAPTCHAs called NoBot[3]. NoBot is a face-recognition solution that aims to minimise user effort using capabilities of the host device (i.e., a camera) and service-based image processing. The study was designed and partially conducted before the introduction of reCAPTCHA v 2.0 that only requires a tick from the user. For NoBot in particular, we assess the user experience of repeated use, across two devices: laptop and tablet. Examining use of a tablet is important as interactions and purchases become increasingly mobile, posing significant usability challenges due to virtual keyboards [10].

Our study makes methodological advances by comparing three human verification mechanisms in a realistic scenario using comprehensive quantitative and qualitative measurements. 87 participants were tasked with using a mock-up of an event ticket-purchasing website, rather than solving CAPTCHAs. We consider it essential to engage participants with a realistic primary task since a study will not yield generalisable results unless conditions reflect real life. This is in line with the ISO 9241-11 [14] standard, which defines usability not only as *efficiency* at achieving specified goals, but also *effectiveness* and *satisfaction* in a *specified context of use*.

The remainder of the paper is organised as follows: The next section summarises the findings of previous studies of CAPTCHAs and takes a critical look at the methodologies used. We then describe the design of our study, followed by a presentation of quantitative and qualitative study findings. Findings are then discussed before concluding remarks and suggestions for future work in the area.

II. Related work

The majority of existing research on the usability of CAPTCHAs has compared different types of CAPTCHAs in terms of solvability. The usability of CAPTCHAs is narrowly defined as users' capability to decipher "squiggly" characters. In some cases qualities of user perception are identified

USEC '16, 21 February 2016, San Diego, CA, USA
Copyright 2016 Internet Society, ISBN 1-891562-42-8
http://dx.doi.org/10.14722/usec.2016.23013

[1] https://developers.google.com/recaptcha/old/intro
[2] http://areyouahuman.com/about-playthru
[3] The developer did not commercialise this product and wishes to remain anonymous.

secondary to the laudable (but restricted) pursuit of less cognitively-demanding alternatives to CAPTCHAs.

Bursztein et al. [5] conducted a large-scale study where workers from Amazon Mechanical Turk and an underground CAPTCHA breaking service solved more than 318k CAPTCHAs (from 12 image-based and 8 audio-based schemes). This demonstrated that humans find CAPTCHAs difficult with audio-based ones being particularly difficult.

Reynaga et al. [20] compared nine CAPTCHA schemes on smartphones with alternative input mechanisms that aimed to increase usability. Although participants considered traditional input mechanisms to be error-prone, they preferred them due to familiarity. The study is notable for capturing a variety of subjective user perceptions (e.g., ratings for memorability, preference) and free-text responses from participants. When discussing their findings, the authors stress that both correctness of solving a CAPTCHA and user perceptions ought to be considered during evaluation.

Gossweiler et al. [9] evaluated a CAPTCHA alternative based on image orientation, where users rotate 2D images to an upright position. Evaluation included a "Happiness Study" to determine participants' preference for either a traditional text-based CAPTCHA or the image-orientation variant. Participants were asked to solve both types of CAPTCHA, five times each. They would then identify their preferred method in a free-text box, where 11 of 16 participants preferred the image orientation variant. It was noted that "many users [of the text-based CAPTCHA] referenced feeling like they were at an eye exam while deciphering the text" [p. 849].

Ho et al. [13] proposed metrics for quantifying the usability of CAPTCHAs, engaging with participants through a game which tracks completion of CAPTCHAs as a measure of progress. Surveys were deliberately avoided, specifically as a way to manage study resourcing. The work posits that the most appropriate way to assess CAPTCHA usability is to ask many people to solve CAPTCHAs repeatedly. Metrics included completion time, typing error, and number of abandoned CAPTCHA solving attempts.

Belk et al. [2] investigated the link between users' cognitive styles (by way of a psychometric-based survey) and both performance with text- or picture-based CAPTCHAs challenges and preferred form of CAPTCHA. The motivating observation is that CAPTCHAs should minimise cognitive effort. Results are reported for 131 participants, who chose to complete either a text- or picture-based CAPTCHA (where the choice was recorded as their preference). Users overwhelmingly preferred text-based CAPTCHAs, where this is attributed to familiarity. Notably, the work implies that users can both exercise a choice and have individual qualities informing that choice, where here we explore influential factors from the perspective of user context and perceived workload.

Yan and Ahmed [22] defined a framework for assessing the usability of both text- and audio-based CAPTCHAs, along dimensions of distortion, content, and presentation. Evaluation by the authors – essentially as subject-matter experts – identified a number of factors which could reduce the approachability of a CAPTCHA for humans (e.g., use of unfamiliar character strings, use of colour).

Studies on CAPTCHAs have suffered from methodological shortcomings. In their evaluation of Chimera CAPTCHAs that use merged objects, Fujita et al. [8] asked leading questions such as "Is it easy solving the CAPTCHA?" and "If you choose 1 or 2 in Question 1, please write why you think that it is not."; the participant might then rate the answer higher than 2 to avoid explaining why they did not find it easy. The question also suggests the researchers are not interested in what made the CAPTCHA easy for the participant. In an evaluation of video-based CAPTCHA, Kluever and Zanibbi [15] asked their participants: "Which task do you enjoy completing more?" suggesting CAPTCHAs are there to be enjoyed.

In summary, studies aiming to evaluate the usability of CAPTCHAs often fail to include a robust and comprehensive methodology that would assess them under realistic conditions. Nearly always participants are invited to studies to solve CAPTCHAs, which might be priming and in dissonance with reality where users go online not to solve CAPTCHAs, but to accomplish some primary tasks (e.g., contribute to a forum). Asking users about their experience has to be balanced and open-ended to learn about the user perspective; studies until now have mostly examined user performance, to confirm if CAPTCHAs – or variants – are merely solvable. A holistic user perspective is lacking, and the fit to users' every-day online tasks is not considered. Our study fills this gap and studies two CAPTCHAs and a possible replacement in a comprehensive study collecting qualitative and quantitative data. We further explore user perceptions of CAPTCHAs as initiated by previous studies.

III. Methodology

Our study looks at three human verification mechanisms: reCAPTCHA, PlayThru and NoBot. PlayThru differs from reCAPTCHA by exploiting the (unwritten) rules of a themed drag-and-drop game (e.g., putting a baseball with a baseball glove). A PlayThru game appears as a small window similar to a traditional text-based CAPTCHA, as in Figure 1. Small icons represent different objects – some of these objects move slightly to indicate that they can be selected with the mouse (or by contact with a touch-sensitive screen). The moving objects can be moved closer to a fixed object at the other side of the PlayThru window. If according to the rules a selected object relates to the fixed object, the user is verified as being a real person and allowed to continue use of the website. The theme of the host website can be integrated into a PlayThru game, however here we use example games only to assess the use of a 'game' as a challenge-response mechanism.

NoBot utilises a camera on the host device, and a remote service which processes signatures of captured images against a database of signatures. Although initially designed for authentication, here NoBot is evaluated solely as a human verification mechanism. Using NoBot does not explicitly require cognitive and physical effort, but requires that the user position their face in view of the camera. NoBot initially appears in a small window similar to a traditional text-based CAPTCHA; once the user activates the widget, NoBot enters a full-screen mode with a fixed oval outline in the centre of the screen which is overlaid on a white-on-black outline of what the device camera sees (including the silhouette of the device user). The oval outline changes colour to indicate when the user has positioned themselves correctly, at which point they click to begin image capture; a sequence of differently coloured blank screens are presented while the camera captures images. NoBot then returns to a widget embedded in the webpage; the remote service will process the images, and the widget then indicates if the user is verified as human and can continue using the website.

There were three experimental conditions in the study. In the first condition, participants were asked to complete three ticket purchases on a laptop using a different human verification mechanism each time (denoted as M3_{all}): reCAPTCHA (reC_{mix}), PlayThru (PT_{mix}) and NoBot (NB_{mix}) with the order of the mechanisms randomised for each participant. In the second condition, participants completed a NoBot verification on each occasion, using a laptop (NB_{Lap}). In the third condition, participants followed the same process but on a tablet (NB_{Tab}). PlayThru and NoBot both represent contrasts to CAPTCHAs, but NoBot does so significantly enough to warrant study on its own separate to the side-by-side comparison.

A. Study goals and hypotheses

The study aimed to learn about the user experience of three human verification mechanisms: reCAPTCHA, PlayThru and NoBot. Since in real life security is not a user's primary task, in our study we tasked participants with buying three event tickets and completing three distinct verification processes.

The following hypotheses were devised to assess the time and workload required to verify using the mechanisms.

a) Time: We hypothesise that: (H1) there will be a difference as to how fast participants verify between the three mechanisms: reCAPTCHA, PlayThru and NoBot. We predict that reCAPTCHA will be the fastest due to familiarity; (H2) There will be a difference between the time that participants take verify using NoBot on a laptop and a tablet; (H3) The time needed to verify using NoBot will decrease with more practice.

b) Workload: We hypothesise that: (H4) there will be a difference in perceived workload between different human verification mechanisms; (H5) There will be a difference in perceived workload for NoBot on different devices.

B. Procedure

The following procedure applied to conditions NB_{Lap} and NB_{Tab}. Upon participant's arrival to the laboratory, the study was briefly explained to them by the experimenter. The experimenter stressed that none of the technologies tested in the study were created by the researchers themselves but by external companies (where researchers acted as independent assessors). The participant was asked if they were sensitive to flashing lights (the NoBot capture process involves shining light on the skin of a person's face). If no sensitivity was reported, they were then asked to read through the information sheet and encouraged to raise any questions they might have, after which they would sign a consent form. The experimenter then switched on the voice recorder and the video-camera, and the participant was asked to make three ticket purchases using a mock-up ticket purchasing website.

When on the mock-up site, there were three steps: select a ticket for an event of the participant's choice, complete one of the human verification processes, and then enter the details (e.g., name, address) of a fictitious person named Adam/Anne Johnson at the checkout to complete the transaction.

The experimenter stayed in the room throughout the session, taking notes and reacting to any comments raised by the participant, where responses were kept to an absolute minimum. After the purchase of the third ticket, the experimenter switched the camera off. The participant was then asked about their experience with the mechanism they had just used, and encouraged to voice any speculations they may have about the purpose of the mechanism. The real purpose – if not deduced already – would be revealed by saying it was to replace CAPTCHAs and a print-out with various CAPTCHAs was shown to make sure the participant knew what was being referred to.

In a brief interview at the end of the study, each participant was asked to elaborate on their experience with NoBot and text-based CAPTCHAs. They were asked to discuss both their advantages and disadvantages. After that, they were asked to fill in the NASA Task Load Index (TLX) questionnaire, pick three adjectives that described their experience with NoBot and indicate in what situations they would use it. When wrapping up, each participant was asked if they would like the company to keep or delete the images of their face taken in the study. They were then thanked for coming to the study and received £10 in cash for their participation.

In the third condition where participants tried all three mechanisms (M3_{all}), they were asked to fill in the aforementioned questionnaires after each ticket purchase (i.e., after having experienced each mechanism). They were also asked to rank the three mechanisms in order of preference.

C. Apparatus

1) Purchasing website: A ticket purchasing site called "SimTikats" was created for the purpose of the study. It offered different tickets for a variety of events ranging from concerts to football matches.

2) Devices: We used a Dell E5540 laptop with a screen size of 15″ and a Nexus 7 tablet (2013 model) with a screen size of 7″.

3) Recordings: While participants were purchasing each ticket, the time it took them to complete the verification process was recorded via the website. A voice recorder was used to capture participants' reactions to the mechanisms and their responses in the interviews. Participants' interactions with the devices were video-taped over their shoulder. For participants in the mixed condition where there were questionnaires between purchases, the camera was switched off and switched back on for the next purchase.

4) Questionnaires: Participants were asked to fill in the following post-task questionnaires:

NASA TLX. A NASA TLX questionnaire [11] was used to assess participants' perceptions of the workload involved in the human verification process. A pen and paper version of NASA TLX was chosen since Noyes and Bruneau [18] found that it required less cognitive effort than processing the information on a screen. Participants were asked to complete the full NASA TLX with cards for pairwise comparisons of the different aspects of workload, to capture perceived importance of the workload factors.

Adjectives. Participants were asked to pick three adjectives from a list of 24 different adjectives that best described their interaction with the mechanism. The population of the adjective list was informed in part by the work of Benedek and Miner [3], and partly by the adjectives suggested in the pilot study. The adjectives were displayed in two columns – positive on the left and negative on the right – to shorten the reading time that was needed.

Different contexts. Since usability and acceptance are contextual, participants were asked to indicate in which contexts they would be willing to use the mechanism(s). There was a list of seven different contexts (Table V) which were taken from real-life uses of CAPTCHAs. For each, participants could choose from three options: "Sure, no problem.", "No, no way." or "I don't do this." if they never engaged in an activity (e.g., had never contributed to an online forum).

D. Participants

Participants were recruited through UCL's Psychology Subject Pool. Age ranged from 18 to 53 years. Mean age was 25.5 (SD = 6.8). Of 87 participants, 57 were female and 30 were male. Four participants had A-Levels (UK school leaving certification), 29 some undergraduate education (no completed degree), 27 an undergraduate degree, 26 a postgraduate degree (Masters or PhD) and one person had vocational training. We had 27 participants in condition $M3_{all}$, 31 in NB_{Lap} and 29 in NB_{Tab}.

E. Ethics

Several measures were taken to protect study participants. There was a written agreement put in place with NoBot's developer that if requested by the participant, they will delete the images of participants' faces taken in the study. During the study, participants were asked to enter the details of a persona rather than their own. Participants' interactions with the website were video-taped over their shoulder so as not to record their faces. The study design and protocol was reviewed and approved by UCL's Research Ethics Committee (3615/004). We implemented requests from the committee, specifically that the information sheet inform participants that one of the mechanisms will take images of their face, and that we ask if they are sensitive to flashing lights.

IV. Results

A. Quantitative results

1) Time: The set-up of the website used for the experiment was such that the verification was on a separate page, participants had to verify using reCAPTCHA, PlayThru or NoBot and then click the 'Continue' button underneath (Fig. 1). This set-up mimics the one encountered on real ticket purchasing sites where the CAPTCHA is either on a separate page or as an overlay. When establishing how long participants took to verify, we consider the time how long they spent on the verification webpage. There are several factors that might have influenced the timings: website design (participants often pressed 'Continue' instead of verifying first and then pressing 'Continue'), speed and reliability of the Internet connection, participants making verbal comments, reading instructions for PlayThru and NoBot and any need to repeat the process (e.g., failures due to not being positioned in view of the camera).

Three mechanisms. In the mixed condition where participants tried all three mechanisms once, it took them on average 20.2 seconds (SD = 13.4) to solve a reCAPTCHA, 28.7 seconds (SD = 13.9) to complete a PlayThru game and 70.1 seconds (SD = 50.7) to verify using NoBot. A repeated measures ANOVA determined that the time taken to verify using each mechanism differed statistically significantly ($F(1.23, 35.681) = 27.076, p < 0.0001$). There were no significant order effects. H1 is therefore supported.

Laptop vs. tablet. Figure 2 shows the time needed to verify using NoBot across devices and trials. On average, across the three trials participants took 42.2 seconds on a laptop and 54.4 seconds on a tablet, this difference is statistically significant ($p = 0.022$, two-tailed t-test). H2 is therefore supported.

Practice. A repeated measures ANOVA determined that the time taken to verify on a **laptop** (NB_{Lap}) differed significantly between different attempts to verify using NoBot ($F(1.964, 54.986) = 5.673, p = 0.006$). Post-hoc tests using the Bonferroni correction revealed that participants' verification time dropped with practice from the first to second verification (50.3 vs. 41.2), which was not statistically significant (p = 0.131). Again, time to verify using NoBot dropped between the second and the third verification (41.2 vs. 35) but this difference was not statistically significant either (p = 0.549). However, there was a statistically significant difference between the first and the third verification (p = 0.011).

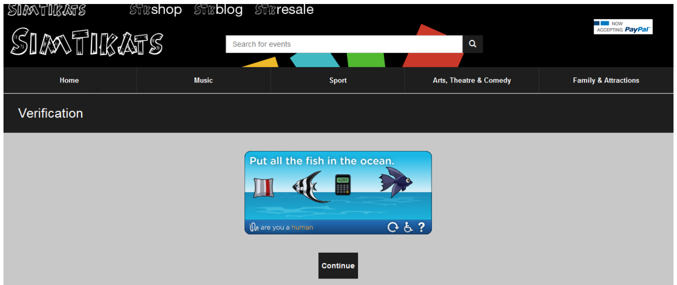

Fig. 1. A page from the study website showing PlayThru.

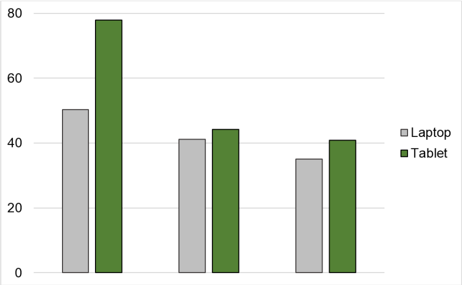

Fig. 2. Average time (in seconds) to verify using NoBot across devices and trials.

A repeated measures ANOVA determined that the time taken to verify on a **tablet** (NB$_{Tab}$) differed significantly between different attempts to verify using NoBot ($F(1.983, 53.528) = 4.686, p = 0.014$). Post-hoc tests using the Bonferroni correction revealed that participants' verification time dropped with practice from the first to second verification (78.7 vs. 44.7), which was statistically significant ($p = 0.003$). Again, time dropped between the second and the third verification (44.7 vs. 40.9) but this difference was not statistically significant. However, there was a statistically significant difference between the first and the third verification ($p = 0.009$). H3 is therefore supported.

2) Workload (NASA TLX): Table I shows NASA TLX scores for using reCAPTCHA, PlayThru and NoBot. Frustration was the only aspect of workload with a significant difference between the three mechanisms $F(1.93,\ 54.036) = 7.132$, $p = 0.002$). Post-hoc tests using the Bonferroni correction revealed that participants reported a frustration of 8.73 for reCAPTCHA and 6.31 for PlayThru, which was not statistically significant ($p = 0.167$). The difference between reCAPTCHA and NoBot (8.73 vs. 11.7) was not significant either ($p = 0.27$). However, there was a statistically significant difference between PlayThru and NoBot (6.31 vs. 11.7) ($p = 0.002$). H4 is therefore partially supported.

TLX aspect	reCAP	PlayThru	NoBot	P-value
Mental Demand	10.75	9.24	8.89	0.62, n.s.
Physical Demand	4.33	8.45	8.34	0.64, n.s.
Temporal Demand	4.62	3.08	6.47	0.59, n.s.
Performance	3.62	5.46	5.51	0.38, n.s.
Effort	8.12	8.21	6.46	0.71, n.s.
Frustration	8.73	6.31	11.7	0.002**
Overall workload	48.4	37	46.7	0.19, n.s.

TABLE I
COMPARISON OF NASA TLX SCORES FOR RECAPTCHA, PLAYTHRU AND NOBOT.

Table II presents the NASA TLX scores for using NoBot on a laptop and a tablet. Although the scores tended to be higher for the tablet, the difference was only significant for *Physical Demand*. H5 is therefore partially supported.

3) Adjectives: Tables III and IV show the top five adjectives participants selected to describe the mechanism(s) they used.

On both devices, participants found NoBot "effortless", "intuitive" and "easy to use". However, the perceptions of speed differed. For 'fast', 9 participants chose this adjective to describe NoBot on a laptop and 7 participants on the tablet. For 'slow', 1 participant picked the adjective to describe NoBot on a laptop and 9 participants on the tablet. This difference is statistically significant ($p = 0.037$, Fisher's exact test.)

TLX aspect	Laptop	Tablet	P-value
Mental Demand	4.2	4.8	0.13, n.s.
Physical Demand	2.9	7.2	0.004**
Temporal Demand	5.6	5.1	0.85, n.s.
Performance	7.2	7.9	0.5, n.s.
Effort	5.4	7	0.18, n.s.
Frustration	5.4	5.4	0.4, n.s.
Overall workload	29.9	37.4	0.13, n.s.

TABLE II
COMPARISON OF NASA TLX SCORES FOR VERIFYING USING NOBOT ON A LAPTOP AND TABLET.

Laptop		Tablet	
effortless	13	effortless	11
fast	9	intuitive	10
intuitive	8	slow	9
weird	7	easy to use	8
easy to use	6	acceptable	8

TABLE III
TOP FIVE ADJECTIVES CHOSEN TO DESCRIBE USING NOBOT ON A LAPTOP AND TABLET.

reCAPTCHA		PlayThru		NoBot	
normal	14	acceptable	14	unpredictable	9
acceptable	13	exciting	9	weird	9
effortful	9	effortless	7	creepy	8
easy to use	8	intuitive	7	fast	8
predictable	7	great	8	effortful	6

TABLE IV
TOP FIVE ADJECTIVES CHOSEN TO DESCRIBE RECAPTCHA, PLAYTHRU AND NOBOT.

However, participants were far less positive about NoBot in the mixed condition. While they found reCAPTCHA and PlayThru to be "normal" and "acceptable", NoBot was "unpredictable", "weird" and "creepy". Interestingly, NoBot was the only mechanism that was described as "fast" within the top five adjectives.

4) Different contexts: Table V shows the percentages of participants willing to use the three mechanisms in different contexts. The results from conditions NB$_{Lap}$ and NB$_{Tab}$ were combined as NB$_{L+T}$ since there was no statistically significant difference between them. Overall, participants indicated they were most willing to use the mechanisms for ticket purchasing which is no surprise since this was the scenario used in our study.

Participants were least willing to verify when contributing to a forum using PlayThru and NoBot. For PlayThru, some of them explained the game was not serious enough in the context of a serious activity. For NoBot, six participants stressed that the nature of forums is that one wants to stay anonymous.

Context	**NB**$_{L+T}$	**reC**$_{mix}$	**PT**$_{mix}$	**NB**$_{mix}$
Contributing to an online forum	16	59	15	24
Buying tickets online	76	93	79	55
Browsing for plane tickets	50	76	55	45
Checking in for flights online	62	86	54	52
Topping up your Oyster online	52	66	69	31
Bidding on items on eBay	38	66	69	31
Logging in to Facebook from a different computer	47	66	66	34

TABLE V
PERCENTAGES OF PARTICIPANTS WILLING TO USE RECAPTCHA, PLAYTHRU AND NOBOT IN DIFFERENT CONTEXTS.

PT28[4] stated: *"I would not use it for an online forum because I don't know who's going to see that image, and in online forums, there tends to be a lot of hackers, so I wouldn't feel comfortable doing that."* This reveals a common perception that the service provider or other users would see the pictures taken by NoBot which is, to our knowledge, not the intention of the company.

5) Ranking: Towards the end of each study session, participants were asked to rank the mechanisms in order of their preference. There were five choices: reCAPTCHA, PlayThru, NoBot, no CAPTCHA-like mechanism and a different mechanism. Figure 3 shows how many participants ranked the mechanisms on each position.

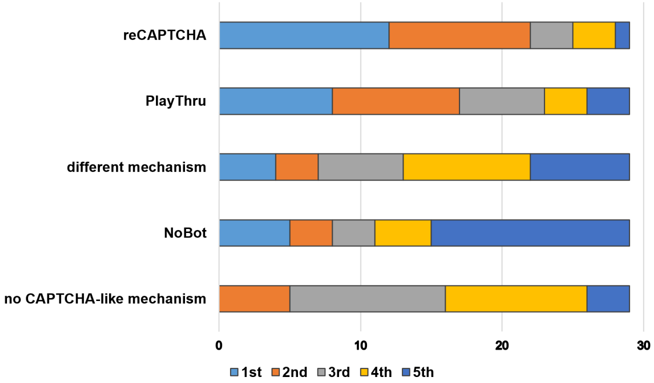

Fig. 3. Numbers of participants in M3$_{all}$ providing rankings for each mechanism and hypothetical alternatives. Mechanisms ordered from highest to lowest average rank.

reCAPTCHA had the highest average rank followed by PlayThru, a different mechanism and NoBot. "No CAPTCHA-like mechanism" seemed to be the least desirable choice, it had the lowest average rank and was never ranked as the top choice. PM06 motivated their choice of ranking it low by saying: *"I expect there is probably some benefit to all of us to having some kind of a security mechanism on some websites and some purchasing processes. I definitely wouldn't want to get rid of it altogether, not that I know in much detail what benefits they have."*

PM00 explained their ranking: *"I have always being using that, and it's a habit although it's sometimes quite difficult to read the letters but I feel cool with it. The PlayThru, it's like a game, it's not very serious so I don't… but I still quite like it though. If we don't have the reCAPTCHA, I'm OK with the PlayThru.* [Why?] *It's easy, it's fun and then you don't really put effort to do that."* Similarly, PM03 stressed the importance of being used to reCAPTCHA: *"It* [PlayThru] *disturbed my usual routine, it required some kind of effort which actually I didn't exhort that much, I just guessed through it and it was correct what I did. If not I imagine, it would have been very frustrating. So this one* [reCAPTCHA] *is fine and I'm more comfortable with doing it."*

B. Qualitative results

Any comments participants made about the mechanisms tested were transcribed and coded by two researchers. Each researcher coded half of the transcripts and both coded a subset of six transcripts (two from each condition) and they discussed any differences. The following sections present the results of the thematic analysis [4] of the data.

1) Views on text-based CAPTCHAs: Throughout the study sessions, participants shared their experiences with text-based CAPTCHAs, which were overwhelmingly negative. "Annoying" (44 mentions), "frustrating" (20) and "hate" (10) were the most common words participants used to describe their experiences with CAPTCHAs. Out of 87 participants, 64 stressed they found CAPTCHAs to be hard to read. 13 participants could not tell if the CAPTCHA characters were case-sensitive. 28 participants stated their coping strategy for reCAPTCHAs was just to press the 'Refresh' button and try another one. PL27 demanded something had to be done: "[CAPTCHAs] *are just annoying, they should find a better system."*

When considering CAPTCHAs and alternatives, it is also important to be sensitive to how this technology might impact users and their wellbeing. As shown by previous research (e.g., [1]), security technologies can cause embarrassment and make users feel bad about themselves. CAPTCHAs have this potential too, PM08 stated: *"I struggle several times to give the right answer, I'll search the Internet, I feel like an idiot, then I abandon it and find another website that does the same thing."* PL08 explained in a similar vein: *"I actually hate all of these because sometimes I really can't see properly, I'm not a robot but I just can't see, you know, and I keep refreshing the CAPTCHA. It's quite troublesome."*

2) Perceptions of time and effort: We observed a discrepancy between actual and perceived length of the verification processes. To investigate this, we compared the actual timings of mechanism use with comments about mechanism speed

[4] The first letter of the participant number indicates which condition they were in: "M"–M3$_{all}$, "L"–NB$_{Lap}$and "T"–NB$_{Tab}$.

as made by participants in the comparative condition. To illustrate, PM10 stated *"for me I'd want to do things much quicker, so I'd prefer NoBot as it's faster so I can get the tickets I want."* This participant's actual timings were 43 seconds for reCAPTCHA, 19 for PlayThru and 53 for NoBot meaning NoBot was actually the slowest. Overall, we identified six such discrepancies. In four cases, NoBot was perceived to be faster than it actually was. One can speculate why this is the case, PM14 explained: "[PlayThru] *was quick but it didn't really matter because it was a game, it was nice, [it could even take longer because it's nice], you could get sidetracked a little bit but it's nice."* It seems one does not notice the passage of time when the activity is entertaining.

What we believe might explain this difference in perceptions could be the fact that the effort expended was coloured by the type of effort required, for NoBot there was less physical and cognitive effort and mostly time. PM01 stated: *"It's the first time the screen just recognises my face and don't need to do anything, just stay here, effortless."* The participants further explained what made this experience so fast, saying: "[For PlayThru] *I still need to read the sentence to know what they want me to do. I need to think and for NoBot, I just stay here and it's done."*

Participants had the tendency to rationalise why the verification effort required of them was justified. Some stressed by completing CAPTCHAs they contributed to the services remaining available online, otherwise they would need to physically go to a box office to purchase tickets. The inconvenience of CAPTCHAs was perceived as small relative to the convenience of being able to buy tickets over the Internet. Additionally, two participants stressed CAPTCHAs were a good thing for society as they were used to digitise books.

Participants also had a good feeling about contributing to overall Internet security. When asked why they ranked "no CAPTCHA-like mechanism" the lowest, PM05 explained: *"I guess there was a reason why they had these in place. That is I may not appreciate but it probably saves a lot time compared to when a lot of machines would be entering stuff. If there was no verification process at all, it would probably take longer to filter through what's simulated and what's real."*

Some participants also expressed the view that CAPTCHAs have to be hard to give them a sense of security, PM02 explained: *"The chess one* [PlayThru] *was a bit simple, it feels like anyone can do it, it feels actually less secure, although I guess it's just checking if you are human. [. . .] And PlayThru might actually be the best out of the three because reCAPTCHA still requires a bit more effort than PlayThru, I think [. . .] I think it's quite easy to use so I feel like everyone and anyone could use it whereas with reCAPTCHA you have to go through some effort and NoBot, you have to fit your face. But PlayThru, because it's so easy and accessible, I feel like it's too easy in a way."* The experimenter then asked why CAPTCHAs had to be hard, and the participant explained that a hard CAPTCHA gave them time to think through whether they wanted to complete a transaction: *"Let's say if you're buying something and it's too easy, you'd just got through with it and if you want to change your mind or something, it's kind of too late."*

Apart from giving participants a sense of security or a moment longer to re-consider the transaction, we also saw that some participants over-attributed the security that CAPTCHAs gave them, assigning them 'powers' they do not have. When asked why they rated "no CAPTCHA-like mechanism" the lowest, PM04 stated *"If my information was stolen, my credit card and there was nothing to check that it wasn't me making the purchase.* [Do reCAPTCHAs help with this?] *But it helps fight spam bots too. If someone stole my password, my information was say Ticketmaster UK, if it was hacked and my information was stored, someone can buy a ticket. If someone can hack Sony, they can hack Ticketmaster."*

3) Security: When asked about their experience with the mechanisms, participants often expressed views on their security, rather than just usability or acceptability. Interestingly, they perceived the security of the mechanisms not based on how difficult they are for bots to solve and how well they protect the website, but they spoke about how well their transaction details are protected and how their images will be stored. 32 participants described scenarios where the security of their images would be compromised, for example if the company was hacked. Also, rather than thinking of these mechanisms as black boxes, many participants actively deliberated how they operated, with 17 participants actively questioning their resilience against attacks and thinking of ways to break them, for example through the use of pictures or masks.

PT28 explained they were not sure it CAPTCHAs were providing any security at all: *"I really really hate them, they give you a hurdle to jump over and you've been onto that site 100 times, I find them really infuriating, I don't even know if they work, I can't see the point of them. I can't read them, they're too close or the letters are at an angle, or they have these dots which sometimes you can't see."* The CAPTCHA hurdle is even more frustrating if the user is under time pressure, PM07 told us their story: *"A couple of months ago, I was ready half an hour before they* [the tickets] *even went online, I had everything ready, my information. And then I like got through and I got the ticket and it had this thing that. . . CAPTCHA box, it's called. And the laptop that I was on had an ad blocker or something, every time that I typed the word that was in it, it wouldn't allow it and I was so panicking, I had to get that ticket. . . I didn't realise I had to turn it. . . that I had to change some settings to allow it or it was or I genuinely couldn't read it, so I tried to play the audio but that was confusing and I oh. . . then I ended up losing the ticket. Basically, I did get another one but it was pretty stressful."*

Five participants were disillusioned with human verification mechanisms in general saying they were protecting the businesses that used them rather than the users. PL15 talked about NoBot: *"I'm even less inclined for people to use my photo, you know, just for the company's benefit. That is something that fully identifies you and it's actually for the company's benefit more than even my own. OK, it's security of buying my* *own ticket, so probably I would be even less inclined actually. Especially for what it's asking for to protect a company, no! [. . .] It would make me more productive because it's quick and easy, but only by a couple of minutes. . . It's being sold to me that it's actually for my protection but it's not."*

4) Privacy of images: Often after their first encounter with NoBot, participants speculated that NoBot is there for identification rather than verification that they are human. This is not surprising since the technology is using a biometric solution. In the case of our mock-up ticket-buying website, 18 participants thought that a picture of them was taken, to check if it is really them when they turn up at the event venue.

Throughout the study sessions participants raised privacy concerns with NoBot capturing and storing images of their faces. 29 participants did not like the fact that someone could see their picture, and within this group, they either worried about the service provider such as the ticket purchasing site seeing their images, or other parties such as the NoBot company or other users of the website (such as other forum members or bidders on eBay). Ten participants emphasised that they did not like the possibility that a company capturing personal images might store the images and not delete them.

When asked about what specifically invoked privacy concerns, eight participants stressed that a human face is special, by which they meant it was unique and identifying. Participants also elaborated on what kind of reassurance they would need to be able to use NoBot confidently. Ten participants would have liked a confirmation that images will be kept securely or deleted. Nine participants would have welcomed a privacy statement and two the display of security certificates. Interestingly, seven people stressed they would be more confident using NoBot if they saw that other people were using it too. Related to this, 20 participants emphasised that the technology was novel and they needed to develop a trust relationship with it.

There were also participants who said they were too ignorant to care, PM03 explained: *"Maybe some persons are not comfortable with showing their face.* [What about you?] *I am comfortable with doing that. When you are ignorant about computer systems, what they do with your data and all that, it's easier for you to give information. I'm kind of like that person so I don't really mind giving my face. I can trust the privacy statements that they give."* PM05 noted that there might be a trade-off between privacy and convenience: *"I prefer the two to NoBot because they are less invasive but equally there was a little bit more effort than the reCAPTCHA."*

5) Context of use: Throughout the study sessions, participants specified situations where they believed it would be appropriate to use the different mechanisms. This was in part prompted by them being asked to indicate their willingness to use mechanism in different contexts for which we describe the results earlier, but even before being prompted participants elaborated on when they would or would not use it and why. Six participants stressed that using NoBot would be more suitable for high-value purchases where there is more money at stake, PM06 explained: *"I guess if someone is buying something more expensive, there is kind of more risk of fraud, using someone else's details. In that case, it might be a lot more beneficial but other than that I don't think it's necessary for small amounts of money or like logging in to things, it's definitely a bit extreme."* Because online contexts vary, 9 participants emphasised they would like to be given choice which mechanism they would like to use to verify.

Some participants also explained that they would not like for the image of their face to be associated with what they were buying, PM05 explained: *"I liked it less because I don't like putting my face on the Internet; taking a picture of me and who knows what I'm buying."* This again had to do with participants' perceptions of NoBot as an identification mechanism that shares the images with the service provider.

26 participants stated that NoBot is too invasive or heavy-handed for simply checking if the user is human. PM04 explained: *"I'm trying to buy a ticket and not getting into a governmental building. It seems a bit much."* They further elaborated: *"I do support some sort of verification but I prefer getting less of myself, as in my face, facial recognition if I have to. Obviously, if that was part of my job or my livelihood, I wouldn't have a problem with NoBot, the visual confirmation, personal recognition but for purchasing tickets it's a bit too much."*. Other participants stressed that the gain in security is not worth the loss in privacy, PM06 explained: *"I can see why it would sometimes be useful in terms of fraudulent activity and purchasing things but I still don't think it's worth the breach of privacy to store an image of anyone ever who buys anything online or tried to log in to a website."*

6) Control: There was a recurring theme amongst participants of not feeling in control when using NoBot, because the outcome of the verification process depended on many factors that were outside of their influence. PM04 stressed: *"You're kind of held hostage to the quality of the camera."* Participants emphasised that they feel more in control with reCAPTCHA because it is for them to read and enter the characters and they can always 'try harder' if the system does not accept their submission. With NoBot, they can only do as much as positioning themselves in front of the camera, and if the system decides they are not human, there is little more that they can do to change the outcome. PT26 explained: "[About reCAPTCHA] *I really don't mind having several attempts actually it's not something that is annoying and it's easy to do. This* [NoBot] *can have similar potential problems, my picture couldn't be taken properly, had to try several times, three times that was annoying and I'm not sure why, could be the flashes, really testing my patience, perhaps that's no fault of my own. With CAPTCHA, I know it was somehow my fault, because I didn't interpret the* [letters] *and numbers correctly, the* [letters] *are there, I might have misread it, it works better next time around, I've a certain amount of control* [over] *the CAPTCHA system compared to this NoBot system. . .* "

Similarly, PL11 said they would not rely on NoBot for critical activities because they considered verifying using a text-based CAPTCHA to be more reliable: *"like checking in for flights, I would not use NoBot because there might be some* *error and I could not check in for my flight. Then I would use the traditional because it's easier, I can refresh the images and try again, so I think it would be faster. So if there is something urgent, I would not use it* [NoBot]. *But for every-day uses, I would advocate this."*

Similarly, PL16 expressed being worried about system failure: *"I'd be concerned that just because you know computers and machines and that have been proven to be imperfect, I'd be a little bit concerned about it being maybe not able to accurately read my face. It's just the squiggly words, CAPTCHAs, it's just the way these words sometimes, the letters are undecipherable which is a kind of quality imperfection, if you will. I'd be concerned that someday, one of these websites wouldn't be able to read my face and I can't get in. And if it's an important one, let's say self-assessment [tax submission] or checking in online or something, then I'm really screwed because of a system failure."*

Worries over NoBot's reliability also revealed participants' mental models of security, PL17 elaborated: *"I wouldn't really use it though personally because I've had fraud on my debit card before so I don't trust this kind of thing anymore. I would rather do it myself using the CAPTCHA or get it from the tickets' case, I'd just get it from some office and do it. Then I know I properly get it myself and then there is no fraud in it. Whereas with the face thing, no."* PL18 expressed they had a low level of tolerance for the inaccuracy of security measures: *"If something is to guard me, protect me and if it's not accurate, it pisses me off, you know. It's like having a guard at a gate who is a drunk."*

PL14 stressed that they would not like to be photographed when buying something sensitive but they would see the advantage of storing a picture in case a criminal successfully completed a transaction in their name: *"I'm actually not sure about the picture though, yeah. Because I wouldn't want them to know, if it's for verification purchases, then why do you need to store the picture? [. . .] Sometimes I think it's useful to store pictures though. Let's say some criminal wants to buy some. . . purchase something online and then the police can find out who bought that thing online and they can use that image to find the criminal. I don't know. But what are the chances of this happening?"*

7) Obstacles to adoption: Participants mentioned several factors that could be obstacles to the adoption of PlayThru and NoBot in particular. For NoBot, 22 participants stressed that not every computer had a camera, and 21 stressed that users' sensitivity to flashing lights might prevent them from using it. For PlayThru, some participants stated that the tasks might not be understandable to everyone, particularly where the rules of the game were culturally specific.

V. Discussion

Verification using reCAPTCHA was the fastest, followed by PlayThru, then NoBot. The level of frustration felt by participants differed across the mechanisms, with NoBot being the most and PlayThru the least frustrating. Several participants emphasised that it is hard to compare reCAPTCHA with the other mechanisms, due to a familiarity with CAPTCHAs. This is supported by the results, where "normal" and "acceptable" were the top adjectives used to describe reCAPTCHA. Participants were largely annoyed by traditional CAPTCHAs but emphasised they understood they were there to add security, although they were not always clear about how this security advantage was afforded. reCAPTCHA also had the advantage of having been around for longer which was a source of trust, PM26 explained: *"the fact that it already exists, is pre-existing, somebody approved it somewhere, it's been around a while. . . that seems fairly safe."* Where participants ranked no CAPTCHA-like mechanism as the least favourable option it indicates that security was seen as important.

In the conditions where participants completed ticket transactions with NoBot three times in succession, verification became quicker over time. Participants were significantly faster verifying using NoBot on a laptop than on a tablet, rating use on a tablet as more physically demanding and slower. Completion times on the tablet decreased rapidly after initial use of NoBot, becoming comparable to completion times on a laptop – unfamiliarity with tablets may have compounded the learning demands for some participants.

In the mixed condition, we saw a number of discrepancies between actual and perceived time to verify. Perceptions of time could have been coloured by the entertainment factor of PlayThru and NoBot, as well as their novelty. In future work, one could examine if perceptions would change as a person becomes familiar with the technology.

Participants often assumed NoBot was used to identify rather than verify them, due to the capturing of a biometric.[5] The NoBot face capture mechanism is implicit, but preparation disrupts the flow of the primary task (positioning the face in front of the device camera and remaining still, etc.). NoBot was the most time-consuming of the mechanisms, which might have lead to increased frustration.

We compared use of NoBot on a laptop three times in sequence with a single use of NoBot in the comparative condition. Participants were more critical of NoBot in the mixed condition, calling it "unpredictable", "weird" and "creepy" rather than "effortless", "fast" and "intuitive" as when used in isolation. We are careful not to draw any far reaching conclusions from this since two factors vary between the conditions: repeated vs. single use, and having direct comparison with other mechanisms and having only one mechanism. We can only speculate that repeated use might have made participants more accepting of the technology, or that direct comparison may have encouraged participants to be more critical. This would confirm an important design rule – when asked to evaluate a technology, users should be shown more than one design to be able to compare and better articulate their needs and preferences through relative assessment.

In our study, users reported they would make the decision of whether to use NoBot based on the context of use. Users found NoBot too heavy-handed for more frequent and low-value

[5] We describe the perception of NoBot as a biometric in [17].

transactions yet more appropriate for high-value transactions. This is similar to the findings by Krol et al. [16] where participants believed that a more time-consuming mechanism should be used for infrequent and high-value transactions. This can be related to CAPTCHAs, in that a number of participants believed that security should be difficult. To consider future work, one of the limitations of our study is that we tested our mechanisms on a sample of participants who were relatively young and well-educated, which might limit the generalisability of our findings. Testing the mechanisms in the wild with a wider range of services would be a next step. Our findings can inform development of a repeatable usability assessment framework. Future work could develop instructions for technologies which adequately explain their operation as part of the primary task, since participants have been shown to be accepting of security processes if these are explained to them [6].

VI. Conclusions

Here we conducted a study with 87 participants, examining the user experience of three human verification mechanisms: reCAPTCHA, PlayThru and NoBot. 29 participants used all three mechanisms once, 27 used NoBot three times in succession on a laptop, and 31 similarly using NoBot on a tablet. Our results show that participants were on average fastest verifying using reCAPTCHA, followed by PlayThru and NoBot. NASA TLX results showed that participants were the most frustrated by NoBot and the least by PlayThru.

For NoBot tested on two devices, we saw that using it on a tablet created significantly more physical demand than when using a laptop. Results indicated participants using NoBot on a tablet in fact reported experiencing increased physical demand.

Participants disliked reCAPTCHAs but saw them as a necessary evil, trusting that they were there for the right reasons. Participants using PlayThru thought the entertainment factor made it easy to use, but the game-based format came across as unsuitable for serious online activities. For NoBot, participants did not notice the passage of time when verifying using NoBot, but they objected to the collection and storage of their images.

Participants were more positive towards NoBot when used in isolation rather than in combination with other mechanisms. This highlights the need for the field of usable security to develop robust and repeatable evaluation methods for alternatives of technologies. Our study is a first step towards this, focusing not only on solvability but on the holistic user experience.

Our holistic study approach stands out from previous research of CAPTCHAs and usability of CAPTCHAs in several ways. Prior studies examined how the difficulty of human verification is pushed to users, without fully capturing the user's perspective. We do not see humans as CAPTCHA solving machines, but as individuals who consider the effort of security in terms of their goals and expectations [12]. Another challenge is to afford effortless security that strikes the right balance between providing reassurance and demanding effort, as users recognise a need for security.

References

[1] A. Beautement, M. A. Sasse, and M. Wonham. The compliance budget: Managing security behaviour in organisations. In *Workshop on New Security Paradigms (NSPW)*, pages 47–58. ACM, 2008.

[2] M. Belk, C. Fidas, P. Germanakos, and G. Samaras. Do cognitive styles of users affect preference and performance related to CAPTCHA challenges? In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 1487–1492. ACM, 2012.

[3] J. Benedek and T. Miner. Measuring desirability: New methods for evaluating desirability in a usability lab setting. *Proceedings of Usability Professionals Association*, 2003:8–12, 2002.

[4] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[5] E. Bursztein, S. Bethard, C. Fabry, J. C. Mitchell, and D. Jurafsky. How good are humans at solving CAPTCHAs? A large scale evaluation. In *IEEE Security and Privacy (S&P)*, pages 399–413. IEEE, 2010.

[6] S. Egelman, A. Acquisti, D. Molnar, C. Herley, N. Christin, and S. Krishnamurthi. Please Continue to Hold: An empirical study on user tolerance of security delays. In *Workshop on the Economics of Information Security (WEIS)*, 2010.

[7] C. A. Fidas, A. G. Voyiatzis, and N. M. Avouris. On the necessity of user-friendly CAPTCHA. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 2623–2626. ACM, 2011.

[8] M. Fujita, Y. Ikeya, J. Kani, and M. Nishigaki. Chimera CAPTCHA: A Proposal of CAPTCHA Using Strangeness in Merged Objects. In *Human Aspects of Information Security, Privacy, and Trust (HAS 2015), HCI International 2015*, pages 48–58. Springer, 2015.

[9] R. Gossweiler, M. Kamvar, and S. Baluja. What's up CAPTCHA?: A CAPTCHA based on image orientation. In *International Conference on World Wide Web*, pages 841–850. ACM, 2009.

[10] K. K. Greene, M. A. Gallagher, B. C. Stanton, and P. Y. Lee. I Can't Type That! P@$$w0rd Entry on Mobile Devices. In *Human Aspects of Information Security, Privacy, and Trust (HAS 2015), HCI International 2014*, pages 160–171. Springer, 2014.

[11] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[12] C. Herley. More is not the answer. *IEEE Security & Privacy*, (1):14–19, 2014.

[13] C.-J. Ho, C.-C. Wu, K.-T. Chen, and C.-L. Lei. DevilTyper: A game for CAPTCHA usability evaluation. *Computers in Entertainment (CIE)*, 9(1):3, 2011.

[14] ISO. 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs). *The International Organization for Standardization*, 1998.

[15] K. A. Kluever and R. Zanibbi. Balancing usability and security in a video CAPTCHA. In *Symposium on Usable Privacy and Security (SOUPS)*, page 14. ACM, 2009.

[16] K. Krol, C. Papanicolaou, A. Vernitski, and M. A. Sasse. "Too Taxing on the Mind!" Authentication Grids are not for Everyone. In *Human Aspects of Information Security, Privacy, and Trust (HAS 2015), HCI International 2015*, volume LNCS 9190, pages 71–82, 2015.

[17] K. Krol, S. Parkin, and M. A. Sasse. *"I don't like putting my face on the Internet!"*: An acceptance study of face biometrics as a CAPTCHA replacement. In *Identity, Security and Behavior Analysis (ISBA)*, 2016.

[18] J. M. Noyes and D. P. J. Bruneau. A self-analysis of the NASA-TLX workload measure. *Ergonomics*, 50(4):514–519, 2007.

[19] D. Pogue. Time to kill off CAPTCHAs. *Scientific American*, 306(3):23–23, 2012.

[20] G. Reynaga, S. Chiasson, and P. C. van Oorschot. Exploring the Usability of CAPTCHAS on Smartphones: Comparisons and Recommendations. In *NDSS Workshop on Usable Security (USEC)*, 2015.

[21] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford. CAPTCHA: Using hard AI problems for security. In *Advances in Cryptology (EUROCRYPT 2003)*, pages 294–311. Springer, 2003.

[22] J. Yan and A. S. El Ahmad. Usability of CAPTCHAs or usability issues in CAPTCHA design. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 44–52. ACM, 2008.