EPiC Series in Computing

Volume 55, 2018, Pages 93–106

GCAI-2018. 4th Global Conference on Artificial Intelligence

Using the Winograd Schema Challenge as a CAPTCHA

Nicos Isaak[1] and Loizos Michael[2]

[1] Open University of Cyprus, Nicosia, Cyprus
nicos.isaak@st.ouc.ac.cy
[2] Open University of Cyprus, Nicosia, Cyprus
& Research Center on Interactive Media,
Smart Systems, and Emerging Technologies
loizos@ouc.ac.cy

Abstract

CAPTCHAs have established themselves as a standard technology to confidently distinguish humans from bots. Beyond the typical use for security reasons, CAPTCHAs have helped promote AI research in challenge tasks such as image classification and optical character recognition. It is, therefore, natural to consider what other challenge tasks for AI could serve a role in CAPTCHAs. The Winograd Schema Challenge (WSC), a certain form of hard pronoun resolution tasks, was proposed by Levesque as such a challenge task to promote research in AI. Based on current reports in the literature, the WSC remains a challenging task for bots, and is, therefore, a candidate to serve as a form of CAPTCHA. In this work we investigate whether this a priori appropriateness of the WSC as a form of CAPTCHA can be justified in terms of its acceptability by the human users in relation to existing CAPTCHA tasks. Our empirical study involved a total of 329 students, aged between 11 and 15, and showed that the WSC is generally faster and easier to solve than, and equally entertaining with, the most typical existing CAPTCHA tasks.

1 Introduction

CAPTCHAs are programs that can generate and grade challenge-response tests that most humans can reliably pass, but current computer programs cannot pass. Such tests have been used to prevent automated bots from performing illicit and fraudulent actions, including the degradation of the quality of a provided service [4]. The robustness of a CAPTCHA is its strength in resisting adversarial attacks, and this has attracted considerable attention in the research community [9]. On the flip side of things, CAPTCHAs have served as challenges for the AI community to try to develop automated tools that can reliably pass such tests [26].

Beyond, or perhaps due to, their primary characteristic of being solvable by humans but beyond the capabilities of current computer programs, CAPTCHAs must also be usable, robust, and friendly to humans. With the progress made by the AI community to develop automated tools that pass CAPTCHAs, some forms of CAPTCHAs can no longer be considered as having those characteristics. For instance, as OCR systems improve, so does their ability to pass text-based CAPTCHAs. As a result, text-based CAPTCHAs end up further distorting their presented text, making it less friendly to humans and decreasing the test's usability. At the end

of the day, CAPTCHAs end up being more difficult to be passed by humans than by machines [4], obviating their need for existence. Google's reCAPTCHA-V1, for example, is solvable with an accuracy of 99.8% since 2014, prompting Google to abandon this type of CAPTCHA in March of 2018. *UnCAPTCHA*, an AI-based automated system can break Google's audio-based reCAPTCHA challenges with an accuracy of 85% [19]. Other types of visual-based CAPTCHA schemes were broken with a near 100% success rate by different novel attacks [27]. A recent study designed a novel low-cost attack that leverages deep learning technologies for the semantic annotation of images, being able to automatically solve 70.78% of the new image reCAPTCHA-V2 challenges — the newest reCAPTCHA service of Google [21].

The current state of affairs might point to the need for a new type of CAPTCHA, for which AI techniques have not yet been developed to defeat it. In turn, the introduction of this new type of CAPTCHA will serve the dual role of presenting the AI community with a new challenge task. In this work we propose the use of the Winograd Schema Challenge (WSC) as a candidate for this new form of CAPTCHA.

The WSC has been proposed as an alternative to the Turing Test [14]. Each schema comprises two nearly identical sentences with clear but very different meanings (twin sentences), both sharing a definite pronoun and two potential co-referents. Due to the difference of a certain key phrase in the two sentences, the pronoun is naturally resolved to a different co-referent in each sentence. Given one of the two sentences, then, the task is to resolve the definite pronoun to the correct co-referent. To avoid trivializing the task, the co-referents are of the same gender and number, and one has to rely critically on the key phrase to determine the right answer. In the sentence *The man couldn't lift his son because he was so heavy. Question: Who was heavy? Answers: The man, The son*, the definite pronoun is "he", the two co-referents are "man" and "son", and the key phrase is "heavy"; had the key phrase been "weak" (in the other twin sentence), the answer to the question would have been different.

The WSC has the primary characteristic needed for CAPTCHAs: it is a hard problem for machines, but easy for humans. Indeed, in a competition ran by Nuance Communications during the 2016 edition of IJCAI, the best two approaches did not manage to achieve a score better than 58% [1]. On the other hand, humans who speak English fluently can score a mean accuracy of 92% in English WSC sentences [5]. The interest of the AI community, and the potential for the WSC to act as a challenge task, is already evident from the organization of competitions, and from a number of research papers published on the topic [16, 17, 20]. At the same time, and until AI research produces automated systems that can reliably pass the WSC — a task that is believed to be notoriously difficult [2, 18] — the WSC can be used as the basis for the development of CAPTCHAs for security purposes.

To lay a foundation for WSC-based CAPTCHAs, we compare how human performance, usability, and time-needed for solving a WSC-based CAPTCHA relates to how humans perform on other types of CAPTCHAs. The rest of the paper is structured as follows: we first describe our empirical study, we then analyze and discuss our findings, and we finally review the implications of our results, along with potential directions for future research.

2 Method of Study

A request was sent to the principal of a secondary education school in Cyprus to recruit participants. The necessary permissions were obtained from the school's principal and the Cyprus Pedagogical Institute, which is responsible for research in public schools in Cyprus.

The recruitment process sought to recruit a representative sample of participants that were not familiar with the WSC, based on the fact that this is the first such study undertaken in

	Grade A	Grade B	Grade C
males	62	52	54
females	33	56	55
10-11	5	-	-
12-13	89	101	8
14-15	1	7	101

Table 1: Demographic of participants.

Cypriot schools. The participants were familiar with the use of computers, having taken two hours of computer lessons per week as part of their education, and having been active in using the Internet, social networks, and blogs. The participants had been exposed to CAPTCHA challenges at least once in the past. The survey was run in the school's computer science labs (each holding up to 16 students), during a 40-minute period and under supervision by a school teacher.

Although alternative recruitment processes could have also been adopted, resulting in a different demographic of the participants, the approach that we have followed was chosen for the following reasons: 1) Availability: One of the authors is a teacher at a secondary school, and after acquiring the necessary permissions was able to involve teenagers in the empirical study. 2) Monitoring: Compared to a crowdsourcing solution (which is not really built for human-centered studies, but for the completion of tasks), in the in-class study that we undertook we were able to monitor participants closely and measure their response times more accurately. 3) Complementarity: Studies with adults have already been reported in the literature, and the present study sought to complement those studies by examining a distinct population, and develop a new corpus that might be useful to the research community. 4) Developmental: Adults could be argued to have reached a plateau in their WSC (and close to 100% accuracy, raising issues with the statistical analysis), so that age differences would not play an important role in the reported accuracy. For teenagers, we expected that age differences would yield different results worth analyzing.

2.1 Participants

A total of 329 students volunteered and participated in a study held between November 2017 and December 2017. Participants were teenagers, residents of Cyprus who speak Greek fluently. All of them were students at a single 3-grade gymnasium school, and they were between 11 and 15 years old (see Table 1). Participants reported that they were not aware of having any kind of vision problem that hampered their effort to identify colors, shapes, or patterns. Out of 329 students who attempted the task, 17 did not finish the task, while nine students did not volunteer to participate to the survey. Participants were offered a candy costing €0.30 as a compensation for their time. Also, they were promised that at the end of the study, the group of students with the best overall results would enter a lottery for a gift. Although the study was anonymous, each group of participants was identified with a unique group number.

2.2 Design

Participants were asked to complete a survey using LimeSurvey, consisting of a set of questions recording demographic information about the participants, and five parts that included different types of CAPTCHAs. The study was designed to record user performance, perceptions, and reaction times on the various types of CAPTCHAs. For each type of CAPTCHA, a five-point

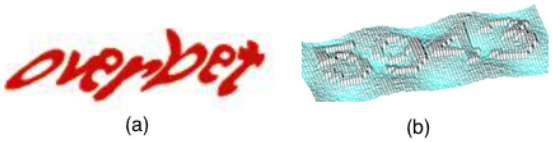

Figure 1: A distorted-text CAPTCHA and a 3D-text CAPTCHA used during the study.

numeric *Likert* scale was used to rank the level of difficulty from "1: very easy" up to "5: very difficult". At the end of the survey, participants were asked to select the type of CAPTCHA that they considered to be the most entertaining, using a pull-down selection widget.

Among the various types of CAPTCHAs that we could have used [10], we have chosen a representative sample, including ones that were based on text, images, and math. According to the school's Computer Science teachers, the school's students were mostly familiar with distorted-text CAPTCHAs and the Google image CAPTCHA challenges (reCAPTCHA v2). A variant of the distorted-text CAPTCHA that uses 3D-text CAPTCHAs was also used. The math-based CAPTCHA was included as it requires users to use their cognitive abilities [11], which relates to the need for cognitive processing that is also present when solving the WSC. Along with the WSC-based CAPTCHAs, we ended up with five different types of CAPTCHAs, each having 20 instances in the survey.

2.2.1 Text-based CAPTCHAs

The text-based CAPTCHA is the simplest type of CAPTCHA that has been invented and implemented in email services. It is still very commonly used, even though there are bots that can easily solve this type of CAPTCHAs [10].

Two text-based CAPTCHA mechanisms were developed, using available open-source software (see Figure 1). The first one is a distorted-text CAPTCHA that generates friendly CAPTCHAs (`https://github.com/josecl/cool-php-captcha`), and the second one is a 3D-text CAPTCHA (`https://github.com/qmegas/captcha-3D`). We requested and saved locally 20 instances from each of the two CAPTCHA services, along with their corresponding correct answers. Each set of instances was used to populate one part of the survey. Participants were expected to type in their answer, which was then compared against the correct answer. Each distorted-text word contained an average of 7 characters (generated using the following parameters for the software: random_word_generation=True, Yperiod=12, Yamplitude=14, Xperiod=11, Xamplitude=5, maxRotation=8, image_scale=2, blur=false). Each 3D-text word contained 4 integers (generated using the following parameters for the software: startX=random(0,35), startY=random(0,80), angle_of_camera_moved_up=35).

We used text-based CAPTCHAs with Latin characters as the majority of Greek-speaking students are known to use Greeklish (writing Greek words using Latin/English characters) on a daily basis [23]; even Google has developed a new virtual keyboard in *Gboard* that transforms Greeklish into Greek. The familiarity with the Latin characters was confirmed by a simple pre-study test, administered two weeks before the study, where 94% of the participants answered "Latin" to the question *Which character set do you find more easy to use on a QWERTY keyboard? Latin, or Greek?*.

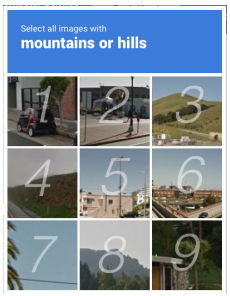

Figure 2: An image-based CAPTCHA used during the study.

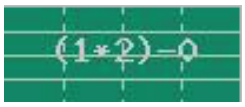

Figure 3: A math-based CAPTCHA used during the study.

2.2.2 Image-based CAPTCHAs

Image—based CAPTCHAs require the participants to select images from a set of such that have a certain characteristic [10] (e.g., select those images that show street signs).

An image CAPTCHA mechanism (https://demo.codeforgeek.com/google-captcha/) was used to implement the new Google reCAPTCHA-V2 service (see Figure 2). Even though Google advertises this service as invisible to humans and visible only to bots [25], students very often have to solve this type of CAPTCHAs. As with text-based CAPTCHAs, we requested and saved locally 20 instances from the CAPTCHA service, along with their corresponding correct answers. The set of instances was used to populate one part of the survey. Participants were expected to select the right images by clicking on them, and their choices were then compared against the correct answer. The 20 instances did not share any images, and the images used were as follows: i) 3 CAPTCHA image sets (2 x 4) based on store fronts; ii) 7 CAPTCHA image sets (3 x 3) based on cars, mountains or hills, bridges, and apartment buildings; iii) 10 CAPTCHA image sets (4 x 4) based on vehicles and street signs.

2.2.3 Math-based CAPTCHAs

Math-based CAPTCHAs ask users to solve a mathematical equation in order to pass a test. The difficulty level of the equation varies across implementations [11].

Among the several available implementations, and considering the age and educational level of our participants, we opted for choosing an implementation that produced relatively easy tests (https://www.hscripts.com/scripts/php/math-captcha.php) that use simple arithmetic operations (see Figure 3). As with other types of CAPTCHA, we requested and saved locally 20 instances from the CAPTCHA service, along with their corresponding correct answers. The set of instances was used to populate one part of the survey.

2.2.4 WSC-based CAPTCHAs

For the WSC-based CAPTCHA we have developed our own service and deployed it on our research lab's server. A client can request a WSC-based CAPTCHA instance as follows: i) The

client registers via an email account, and receives an access key. ii) A call to the API with the correct key returns a WSC sentence with a question (or a pronoun target) and the two possible answers. iii) The client submits a selected answer and receives a response on whether it is correct.

Via an optional argument, the client can request instances in different languages. The served WSC-based CAPTCHA instances are based on WSC sentences that were developed by the authors, research collaborators, or taken from various published corpora [2, 7, 12, 17]. In fact, we have designed our own tools and platform to help with the gathering and evaluation of multilingual CAPTCHAs (`http://cognition.ouc.ac.cy/ws_builder`). Currently, our service consists of 3000 schemas and our registered users can add Greek, French, and English WSC schemas. The schemas are currently checked manually by ourselves for consistency with the WSC rules, and then added to the service's database.

As an example interaction with the service, when requesting an English WSC-based CAPTCHA, the service might be called with `.../load.php?key=trial&lang=en`, and return the following: *ID: 1537 Sentence: George scored against Thomas in the shootout, so he won the game. Pronoun: he Answers: George, Thomas.* To check a proposed answer for that instance, the service might be called with `.../check.php?id=1537&answer=George`, and return the following: *Result: correct.*

To cater for the study's Greek-speaking participants, the WSC instances used in the study were developed by a Greek Literature teacher, and added to the service's database. Beyond ensuring that the instances followed the rules of the WSC in terms of having co-referents of the same gender and number, the teacher was asked to develop Greek WSC that are comparably challenging as those found in the literature [16, 17, 20]. According to the teacher, the developed schemas were judged to be of a medium degree of difficulty, considering that each sentence included two verbs.

As with other types of CAPTCHA, we requested and saved locally 20 instances from the CAPTCHA service, along with their corresponding correct answers. The set of instances was used to populate one part of the survey. Participants were expected to select the right answer by clicking on a radio-button.

2.3 Procedure

Participants were instructed to answer each question quickly but without sacrificing accuracy. Also, they were told that surfing the WWW was not allowed, nor talking to each other or asking questions to the teachers. Although all participants faced the same CAPTCHA instances, the five types of CAPTCHAs were presented to the participants in different order to counterbalance any ordering effects.

The participants participated in the survey by visiting the school's web page and following the provided links. Each student was able to see one CAPTCHA on each page, with directions written in Greek, and their progress was displayed at the top of the page. Once an instance was completed, it was not possible to revisit and edit the answer.

2.4 Hypotheses

The following null hypotheses were formulated for the purpose of this study: a) Participants cannot achieve higher accuracy on WSC-based CAPTCHAs than on other CAPTCHAs. b) There is no significant difference regarding the time needed to solve different types of CAPTCHAs. c) Participants do not think the CAPTCHAs that they complete are difficult. d) There is no general preference of participants towards a certain type of CAPTCHA.

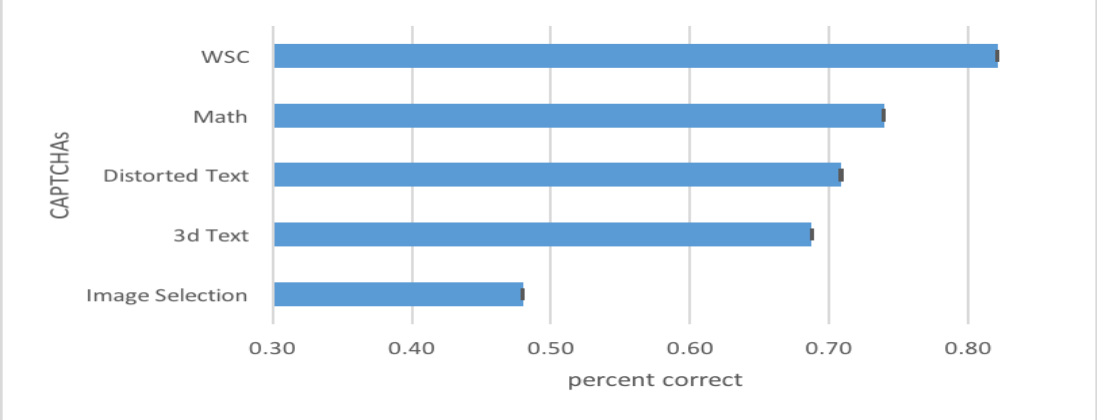

Figure 4: Distribution of scores (with standard errors) on solving different types of CAPTCHAs.

3 Results and Discussion

The 312 participants who fully completed the survey scored the highest mean accuracy of 82% ($\sigma = 0.12$) on the WSC-based CAPTCHAs (see Figure 4). Participants scored the lowest mean accuracy of 48% on the image-based CAPTCHAs ($\sigma = 0.36$), and a mean accuracy of 69% ($\sigma = 0.26$) on the 3D-text CAPTCHAs. On distorted-text CAPTCHAs they scored a mean accuracy of 71% ($\sigma = 0.24$), and on math-based CAPTCHAs — where some cognitive processing, as that for the WSC, was needed — they scored a mean accuracy of 74% ($\sigma = 0.15$), 6% below the score on the WSC-based CAPTCHAs.

The difference in scores was shown to be statistically significant using an ANOVA analysis, and the first null hypothesis was rejected with F=5.52 > Fcrit=2.46 (p=0.00048).

Thus, for WSC-based CAPTCHAs, even teenagers can achieve a significantly high degree of accuracy, whereas machines are still not able to reliably solve the task.

On the contrary, participants scored a mean accuracy of 71% on distorted-text CAPTCHAs, while machines tackle them almost to an accuracy of 100% [27]; the general strength of machines in solving this type of CAPTCHAs is an area of increasing concern [11]. For instance, unCAPTCHA can break the audio version of this challenge by 85% [19]. Also, Google engineers have defeated distorted-text CAPTCHA thanks to a Street View algorithm by 99.8% [22, 24].

On image-based CAPTCHAs, while there are systems that can solve them to an accuracy of 70.78% [21], our participants did not manage to achieve a score higher than 48%. Furthermore, on 3D-text CAPTCHAs, which do not offer more security than the traditional 2D-text CAPTCHAs [15], participants scored a mean accuracy of 69%, very close to the 71% of the distorted-text CAPTCHAs. Finally, while participants scored a mean accuracy of 74% on math-based CAPTCHAs, there are numerous articles that show how they can be handled as textual challenges that can be easily parsed and solved, using, for instance, the DeCaptcher service. Also, this type of CAPTCHA can be easily solved completely using a low-cost attack [3].

3.1 Timing

We have taken measures to ensure that timing was as accurate and meaningful as possible. Furthermore, to the best of our knowledge, this is one of the first reports of timing comparisons

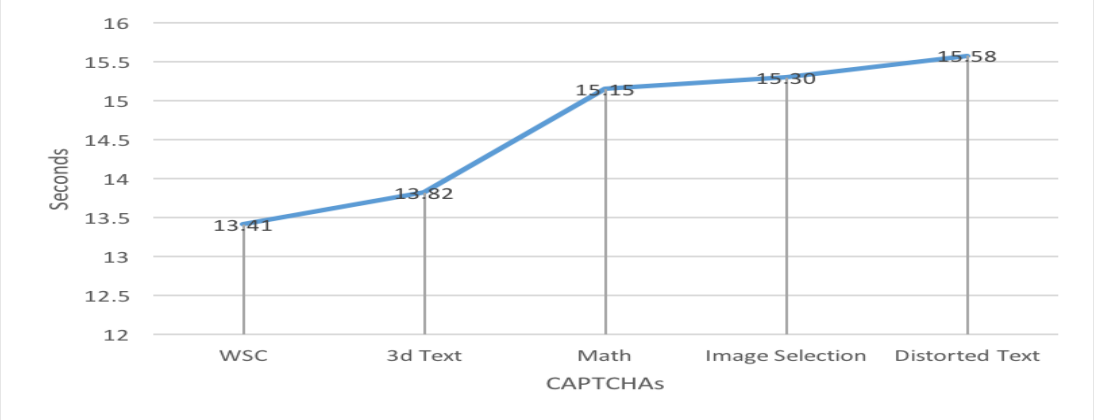

Figure 5: Distribution of response times on solving different types of CAPTCHAs.

between different types of CAPTCHA, and our timing data could be used, or refined, by other researchers, or as part of our future work.

Figure 4 shows the response time distribution on the different types of CAPTCHAs. Participants scored the lowest timing mean on the WSC-based CAPTCHAs (13.41 seconds) and the largest on the distorted-text CAPTCHAs (15.58 seconds). Although there was only a small difference in mean values across timings (only 2.17 seconds), we consider the timing differences to be a finding warranting further examination, as it suggests that the second null hypothesis might also be rejected. According to our results, not only a WSC-based CAPTCHA can be solved faster than a distorted-text CAPTCHA, but also more accurately.

One could argue that the conclusion is a direct consequence of the binary nature of the responses in the WSC-based CAPTCHAs, compared to the distorted-text CAPTCHAs where the answer needs to be typed in. On the other hand, we argue that the WSC-based CAPTCHAs require significant time to read the sentence, and additional time to reason about the answer. Additionally, current bots require multiple minutes to solve a WSC sentence; e.g., there are WSC systems that need at least 3 minutes for each WSC sentence [12]. On the contrary, other bots break existing types of CAPTCHA very quickly; e.g., unCAPTCHA breaks 450 reCAPTCHA audio challenges in under 6 seconds [6, 19].

3.2 Grade / Age

Out of the study's participants, 95 were students of the first high school grade (aged 11-12), 108 were students of the second high school grade (aged around 13), and 109 were students of the third high school grade (aged 14-15). A positive correlation was obtained between grade and accuracy, across all types of CAPTCHAs, with students of more advanced grades (and of greater age) achieving higher scores; the only exception was the second and third grades that achieved the same score on the WSC-based CAPTCHAs. The largest difference in scores was between the first and second grades, where the second grade scored 3% more on 3D-text CAPTCHAs, 8% more on distorted-text CAPTCHAs, 7% more on math-based CAPTCHAs, 4% more on image-based CAPTCHAs, and 2% more on WSC-based CAPTCHAs. Furthermore, the average timings of students of more advanced grades were smaller across all CAPTCHA types.

Although differences in value judgments rely on the knowledge that people accumulate through their experiences in the real world [5], our results might suggest that humans have

the ability to answer commonsense questions like those in the WSC from the early high school ages. On the other, this does not seem to happen with the other types of CAPTCHAs, where competence in the tasks seems to increase with age even within the high school age span.

3.3 Gender

Every participant who completed the experiment submitted also their gender (see Table 1): 165 participants were male, and 144 were female. Significant correlations were obtained between gender and accuracy, across all types of CAPTCHAs. On 3D-text CAPTCHAs the mean female score was 71% (4% more than the mean male score), on distorted-text CAPTCHAs the mean female score was 75% (8% more than the mean male score), on math-based CAPTCHAs the mean female score was 77% (6% more than the mean male score), on image-based CAPTCHAs the mean female score was 51% (5% more than the mean male score), and on the WSC-based CAPTCHAs the mean female score was 85% (6% more than mean male score). Our findings reveal higher rate of achievement on the WSC-based CAPTCHA from females.

3.4 Subjective Judgements

Figure 6 shows the participants' subjective evaluation on the *difficulty* of different CAPTCHA types. On a five-point numeric Likert scale, 54% of the participants rated WSC-based CAPTCHAs as very easy, 27% as easy, and only 3% rated it as very difficult; overall, 81% of the participants rated the WSC-based CAPTCHA as an easy type of CAPTCHA.

Ranking the different types of CAPTCHAs by the percentage of participants that gave them a "difficult" or "very difficult" score, the math-based CAPTCHA was judged as being difficult by 19% of the participants, followed by the 3D-text CAPTCHA with 13%, the distorted-text CAPTCHA with 13%, and the image-based CAPTCHA with 8%, equal to the 8% of the WSC-based CAPTCHA.

The general picture emerging from the analysis is that the participants consider the WSC-based CAPTCHA as an easy CAPTCHA and the other types of CAPTCHA as harder. With these results we can reject the third null hypothesis; participants do consider current CAPTCHAs as hard ones that seem to hamper usability and productivity.

In terms of the level of entertainment of the various types of CAPTCHAs, 22% of the participants selected the distorted-text CAPTCHA as the most entertaining type, and only 12% selected the math-based CAPTCHA as the most entertaining type. The WSC-based CAPTCHA and the image-based CAPTCHA were ranked third, with 19%, after the 3D-text CAPTCHA, which received 20%. 8% of the participants did not select any of the CAPTCHA types as being entertaining.

We speculate that some participants might have been drawn to select the distorted-text and 3D-text CAPTCHAs as most entertaining because of the use of color. A future study might seek to test whether WSC-based CAPTCHAs that utilize color (e.g., to highlight the pronoun and its two possible co-referents) might be viewed as more entertaining. Even though the WSC-based CAPTCHA was not first in the entertainment ranking, the difference from the first in ranking might be inconsequential, especially given that unfamiliarity with the WSC-based CAPTCHA might have negatively impacted the participants' judgement.

The participants' opinions in terms of CAPTCHA difficulty and entertainment scores show that different CAPTCHA types are evaluated differently, and that users might have preferences among different types of CAPTCHAs, which rejects the fourth null hypothesis. In terms of the WSC-based CAPTCHA, its combined low score for difficulty and high score for entertainment suggest that it might find wide acceptability among users.

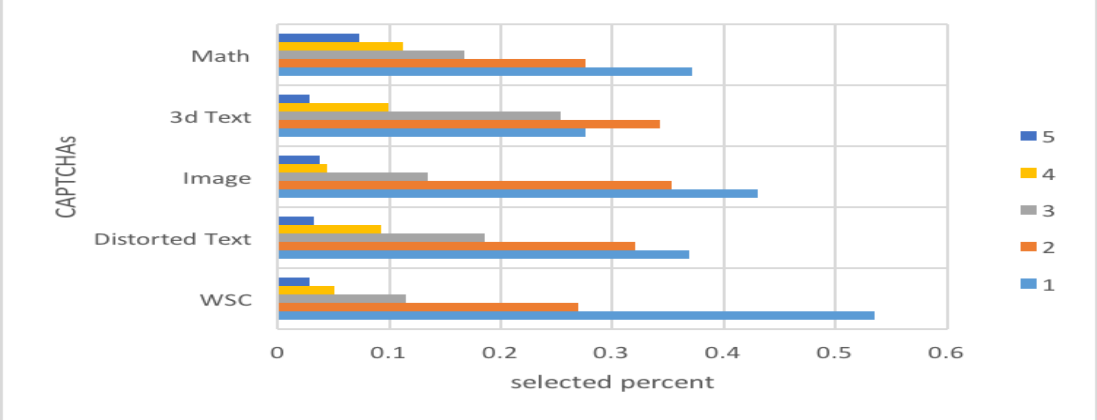

Figure 6: Distribution of participant preferences via a Likert scale that scores the difficulty of different types of CAPTCHAs.

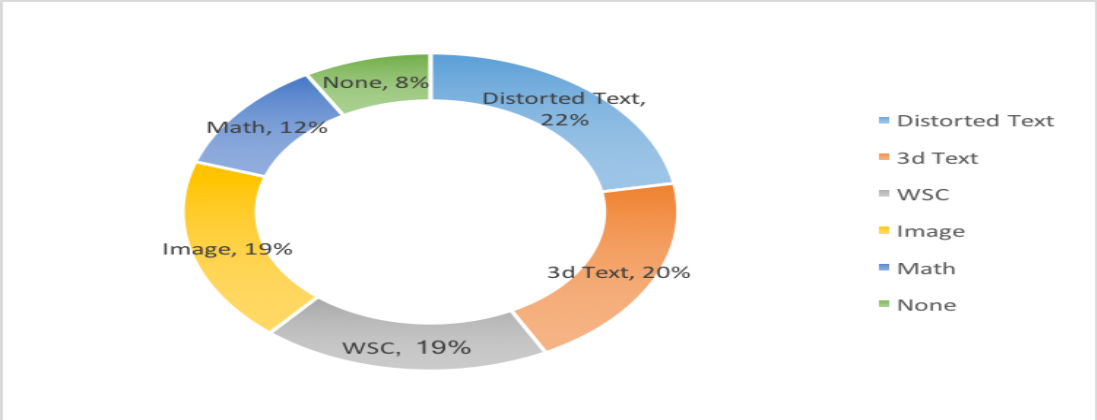

Figure 7: Distribution of participant preferences based on the most entertaining type of CAPTCHA.

3.5 Qualitative Analysis

Teachers who were responsible for monitoring the study have forwarded some remarks from the students. In the process of reviewing these remarks, we noticed several concerns or upshots, worth mentioning.

There were students, mainly from the two first grades, who asked questions about the meaning of words in WSC sentences. For example, some students did not recognize the meaning of the word *shallow* (as in "shallow water"), or *magpie* as being a specific type of bird. On another instance, students were unable to determine whether the word *unstable* was meant to characterize a girl or a chair in a sentence, with both entities being feminine in Greek; this sentence had the lowest accuracy score, with 40% of the students answering incorrectly. Perhaps unsurprisingly, students scored better on sentences that more directly related to their own experiences, with the WSC sentence mentioning, for example, a student and an exam being answered correctly by 98% of the participants.

Teachers also remarked that they observed students trying to rotate the screens to determine the text in the distorted-text and the 3D-text CAPTCHAs. Students also complained about ambiguities in the image-based CAPTCHAs, not knowing, for instance, if choosing an image that showed the wheel of a car was acceptable as an image of a car, and whether choosing an image that showed the pole of a street sign was acceptable as image of a street sign; this, indeed, seems to be a big problem in Google's new CAPTCHA service [22]. Finally, in the math-based CAPTCHAs students asked if they could use a calculator to calculate the results.

The two groups of students that asked the most questions were the groups that achieved the lowest mean accuracy score. These same two groups of students come from the same class, which is considered by school teachers to be the class with the lowest performing students of the school, and the one with the most students that take extra support lessons in school.

4 Other Considerations

Beyond the conclusions resulting from our study on the appropriateness of the WSC as a novel form of CAPTCHA, it is instructive to consider two other aspects of WSC-based CAPTCHAs.

4.1 Accessibility Benefits

Certain types of CAPTCHAs raise accessibility barriers especially to users with disabilities [8]; e.g., vision-impaired users have difficulties with CAPTCHAs that include images or text. Acknowledging that not all users can recognize, solve, and access a CAPTCHA, and that certain types of CAPTCHAs are inherently not adjustable to address these concerns, has led researchers to try to find other ways to control spam by bots [8].

In this work we put forward that the WSC-based CAPTCHA can offer a way out of this situation, as it can easily be adapted to adopt the Web Content Accessibility Guidelines [28] to be perceivable, operable, understandable, and robust, and provide a solution that is accessible to people with disabilities.

Although based on text, the WSC-based CAPTCHA is not predicated on the difficulty of people being able to *read* the text, as is the case in the text-based CAPTCHAs. Thus, one can easily envision extensions where the WSC sentence and question, and the user's answer are all communicated verbally. Not only this extension can cater for vision-impaired users, but it can also cater for users who might be unable to use a keyboard either because of mobility issues or because of lack of an input device. On the other hand, people unable to speak can easily choose between the two possible answers with a simple mouse click, hand gesture, or the press of a key.

Based on our earlier observation that the use of color in a CAPTCHA might have a positive impact on its usability or accessibility [28], one can consider designing WSC-based CAPTCHAs that can use images to represent the possible answers, or use colors to highlight the important parts of the sentence.

4.2 Security Enhancements

One could try to argue against the use of the WSC-based CAPTCHA on the grounds that its error rate for discriminating humans from machines is not sufficiently low. Even if humans could achieve an accuracy of 100%, machines can, at the very least, achieve 50% accuracy by chance. This argument is based on the fact that WSC relies on closed-ended questions, and that these questions have only two possible answers. Other types of CAPTCHAs, the argument goes, are

more appropriate since they either use open-ended questions (e.g., the text-based CAPTCHAs), or closed-ended questions with several possible answers (e.g., the image-based CAPTCHAs), and ensure a lower discrimination error rate.

Although the point about other existing types of CAPTCHAs being *in principle* better at discriminating humans from machines than the WSC-based CAPTCHAs is well-taken, the argument above remains mostly a philosophical one. Pragmatically, the discriminatory power of existing types of CAPTCHAs is, nowadays, worse than the WSC-based CAPTCHAs for the simple reason that machines can now solve those CAPTCHAs with an accuracy much higher than 50%, and often an accuracy comparable to that of humans.

Nonetheless, one could consider certain extensions to strengthen even further the security level of a WSC-based CAPTCHA, at the expense, potentially, of its ease of use: 1) Turn the question into an open-ended one by asking the user to identify and type in the answer among possibly multiple co-referents in the sentence. 2) Combine distorted-text CAPTCHA techniques to partially obscure the possible co-referents in the sentence. 3) Require the resolution of multiple WSC instances within a single WSC-based CAPTCHA. 4) Combine mouse movement techniques, as used in the reCAPTCHA-V2 service, to see if a human or a bot is moving the mouse to select the right answer. 5) Combine image-based CAPTCHA techniques by presenting the potential answers of a WSC-based CAPTCHA instance as images. 6) Banning and blocking of IP addresses that might repeatedly try random answers to pass WSC-based CAPTCHAs.

5 Conclusion and Future Work

We have argued that the Winograd Schema Challenge can form the basis of a new type of CAPTCHA. We have discussed the nature of this WSC-based CAPTCHA, highlighting the shortcomings of typical existing approaches, and providing motivation for a detailed WSC-based CAPTCHA design. Designing good CAPTCHAs is a tedious task, but we expect this work to be a good starting point for future designers of WSC-based CAPTCHAs.

Beyond offering a type of CAPTCHA that, given the current state of affairs, is pragmatically more secure in discriminating humans from machines when compared to existing approaches, our study has shown that this is achieved without essential compromises in usability. On a second front, we expect that the adoption and use of WSC-based CAPTCHAs will encourage more AI researchers to work on the problem of actually trying to solve the WSC, and perhaps, in the process, help towards the building of machines able to reason with commonsense knowledge. At the same time, it will also present AI researchers with the novel challenge of automating the construction of new WSC instances, or evaluating how hard they might be to humans (as pursued, for example, in [13]).

We would encourage researchers to register to use our developed WSC-based CAPTCHA service (`http://cognition.ouc.ac.cy/ws_builder`), and we would welcome suggestions for possible enhancements to increase its security level and the user interaction experience.

Acknowledgments

We would like to thank the principal of the school that participated in this study, Dr. S. Symeou, for his permission to run the study. We are also grateful to Dr. M. Sotiriou for creating the Greek WSC sentences, and to the school teachers for supervising the students.

References

[1] Evan Ackerman. Winograd Schema Challenge Results: AI Common Sense Still a Problem, for Now. *Spectrum*, 2016.

[2] Pascal Amsili and Olga Seminck. A Google-Proof Collection of French Winograd Schemas. In *Second Workshop on Coreference Resolution beyond OntoNotes*, page 24, 2017.

[3] Sandhya Tarar Anvesh Sinha. Review Paper on Different CAPTCHA Techniques. *IJCST Vol. 7*, 2016.

[4] Marios Belk, Panagiotis Germanakos, Christos Fidas, Andreas Holzinger, and George Samaras. Towards the Personalization of CAPTCHA Mechanisms Based on Individual Differences in Cognitive Processing. In *Human Factors in Computing and Informatics*, pages 409–426. Springer, 2013.

[5] David Bender. Establishing a Human Baseline for the Winograd Schema Challenge. In *MAICS*, pages 39–45, 2015.

[6] Kevin Bock, Daven Patel, George Hughey, and Dave Levin. unCaptcha: A Low-Resource Defeat of reCaptcha's Audio Challenge. In *Proceedings of the 11th USENIX Conference on Offensive Technologies*, pages 7–7. USENIX Association, 2017.

[7] Ernest Davis, Leora Morgenstern, and Charles Ortiz. Human tests of materials for the Winograd Schema Challenge 2016. 2016.

[8] Jeremy Elson, John JD Douceur, Jon Howell, and Jared Saul. Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization. 2007.

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

[10] Walid Khalifa Abdullah Hasan. A Survey of Current Research on Captcha. *International Journal of Computer Science and Engineering Survey (IJCSES)*, 7(3):141–157, 2016.

[11] Carlos Javier Hernandez-Castro and Arturo Ribagorda. Pitfalls in CAPTCHA design and implementation: The Math CAPTCHA, a case study. *computers & security*, 29(1):141–157, 2010.

[12] Nicos Isaak and Loizos Michael. Tackling the Winograd Schema Challenge Through Machine Logical Inferences. In David Pearce and Helena Sofia Pinto, editors, *STAIRS*, volume 284 of *Frontiers in Artificial Intelligence and Applications*, pages 75–86. IOS Press, 2016.

[13] Nicos Isaak and Loizos Michael. A Data-Driven Metric of Hardness for WSC Sentences. In *Proceedings of the 4th Global Conference on Artificial Intelligence (GCAI 2018)*. EasyChair, 2018.

[14] Hector J. Levesque. The Winograd Schema Challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, number SS-11-06. American Association for Artificial Intelligence, 2011.

[15] Vu Duc Nguyen, Yang-Wai Chow, and Willy Susilo. On the security of text-based 3D CAPTCHAs. *Computers & Security*, 45:84–99, 2014.

[16] Haoruo Peng, Daniel Khashabi, and Dan Roth. Solving Hard Coreference Problems. *Urbana*, 51:61801, 2015.

[17] Altaf Rahman and Vincent Ng. Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 777–789, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[18] Adam Richard-Bollans, L Gomez Alvarez, and Anthony G Cohn. The Role of Pragmatics in Solving the Winograd Schema Challenge. In *Proceedings of 13th International Symposium on Commonsense Reasoning (Commonsense-2017)*. CEUR Workshop Proceedings, 2017.

[19] By Tara Seals. unCAPTCHA Defeats Google CAPTCHA, 2017. [Online; accessed August-2018].

[20] Arpit Sharma, Nguyen H Vo, Somak Aditya, and Chitta Baral. Towards Addressing the Winograd

Schema Challenge - Building and Using a Semantic Parser and a Knowledge Hunting Module. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI*, pages 25–31, 2015.

[21] Suphannee Sivakorn, Jason Polakis, and Angelos D Keromytis. I'm not a human: Breaking the Google reCAPTCHA. *Black Hat,(i)*, pages 1–12, 2016.

[22] Technoblog.org. Google no Captcha + INVISIBLE reCaptcha – First Experience Results Review, 2017. [Online; accessed August-2018].

[23] Christiana Themistocleous. Written Cypriot Greek in online chat: Usage and attitudes. In *Proceedings of the 8th International Conference on Greek Linguistics*, volume 30, pages 473–488. University of Ioannina Ioannina, 2009.

[24] By Liam Tung. Google algorithm busts CAPTCHA with 99.8 percent accuracy, 2017. [Online; accessed August-2018].

[25] By Rob Verger. Google just made the Internet a tiny bit less annoying, 2017. [Online; accessed August-2018].

[26] Luis Von Ahn, Manuel Blum, Nicholas J Hopper, and John Langford. CAPTCHA: Using Hard AI Problems For Security. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pages 294–311. Springer, 2003.

[27] Jeff Yan and Ahmad Salah El Ahmad. Breaking Visual CAPTCHAs with Naive Pattern Recognition Algorithms. In *Computer Security Applications Conference, 2007. ACSAC 2007. Twenty-Third Annual*, pages 279–291. IEEE, 2007.

[28] 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*, pages 44–52. ACM, 2008.