Examining Security and Usability Aspects of Knowledge-based Authentication Methods

L. Bošnjak* and B. Brumen*
* University of Maribor, Faculty of Electrical Engineering and Computer Science/Institute of Informatics, Maribor, Slovenia
leon.bosnjak@um.si, bostjan.brumen@um.si

***Abstract* - Graphical passwords are considered to be one of the promising alternatives to conventional textual passwords. However, while offering potential theoretical improvements over their textual counterparts, it is important to evaluate how these authentication methods would fare in practice. In this study, we were interested in the user-generated passwords from the security and usability perspective. We conducted an experiment in which the participants were tasked to create and memorize three types of passwords: a textual password, a chess-based graphical password, and an association-based hybrid textual-graphical password. Two weeks after the initial registration, the users were prompted to login using their previously created passwords. By comparing the authentication methods, we showed that despite the graphical passwords' advantages, the user-created chess passwords were the weakest, and the users had the most difficulty remembering them after the two-week period. On the contrary, the association-based passwords were just as strong and memorable as the textual passwords. The conclusions drawn from this paper are therefore two-fold: firstly, alternative authentication methods should be evaluated and compared against textual passwords in real-life scenarios to determine their practical value; and secondly, association-based approaches have the potential to augment both the security and memorability of the existing and novel authentication mechanisms.**

Keywords - password security; password usability; textual passwords; graphical passwords

I. INTRODUCTION

Password literature is bloated with proposals of novel authentication methods, aiming to eventually replace or at the very least complement textual passwords. The main goal of these papers is often concentrated on showing the new method's superiority, which can introduce bias in the evaluations. In particular, the authors might be tempted to assess the method's advantages, while neglecting to focus on the other aspects, as well. Other studies may design the experiment in a way that highlights their method's strong points or choose a theoretical approach to the assessment of some aspects. Too often, a lack of empirical evaluation is replaced by descriptive argumentation of the method's advantages. Finally, very few studies ever consider textual passwords in their comparisons, despite the fact that they are ultimately the authentication method that the proposed methods are competing against.

While these new ideas are desired from an exploratory perspective, it is thus crucial to consider and evaluate their practical value. Between the memorability, security, input times, and ease of use, a potential contender for the spot of the dominant authentication method must prove itself in multiple aspects. Unless it is generally superior to textual passwords, while remaining at least as good in its weakest aspects, the new authentication scheme will not see users readily switching. It is therefore important for the studies to provide fair and comprehensive evaluations of schemes to further our understanding of what characteristics affect certain aspects. Furthermore, more studies should examine the existing authentication methods to verify and expand on the evaluations conducted in the original proposals. In our work, we endeavor to fill that gap by comparing the security and the usability of two existing authentication methods to textual passwords in a two-fold experiment. In section II, we rationalize our choice of these methods. We describe our experimental setup in section III and discuss the results in section IV. We conclude the paper with final remarks and provide some directions for future work.

II. AUTHENTICATION METHODS

The choice of the individual authentication methods to be included in the comparison was made by considering two constraints. First, we selected methods that have not yet been examined and compared. The exception to that rule are textual passwords that were included to allow for benchmark comparisons. Second, we considered methods vastly different in terms of design and use (textual, hybrid, and graphical). The chosen methods are described below.

A. Textual passwords

Traditional textual passwords have remained the most often used authentication method for the last four decades, due to their many advantages, such as the convenience of use and simple implementation. However, technological advances and users' continuous bad management practices [1] have allowed for an increase in security breaches over the last decade, prompting researchers and innovators to begin exploring possible alternatives to textual passwords.

Bonneau et al. conducted an extensive comparison of existing authentication methods, evaluating them against textual passwords in terms of security, deployability, and usability [2]. They found out that no current method outperforms the textual passwords in all three categories. In a study exploring the future of authentication, we

further elaborate on this finding, arguing that textual passwords will remain in widespread use until an overall superior authentication method is invented, and the users are motivated to make a shift to the new method [3]. Until then, it is vital to include textual passwords as a benchmark in the comparisons of novel and existing authentication methods.

B. Game Changer Password System (GCPS)

Originally proposed by McLennan et al. [4], the Game Changer Password System utilizes board games within the process of authentication. Initially, the user is presented with a panel of selected board games to choose from (similar to a movie selection screen in Netflix's interface). Upon selecting a game, an empty board is presented to the user, on which they must place the correct game pieces in the order corresponding to the user's password, as seen in Fig. 1. Even though similar game-based authentication schemes have already been proposed in the past, the main advantage of the GCPS lies in its scalability. The ability to choose between several available board games to put game pieces on allows for an exponential increase in the theoretical search space. In addition, the GCPS' graphical interface, and the users' familiarity with board games promote easier memorability, owing to the well-known Picture Superiority Effect in the memory literature [5].

The authors empirically evaluated the memorability of the GCPS in two experiments. In the first experiment, the participants were tasked to create two passwords (a chess- and a Monopoly-based one) that they thought were secure. After a delay, they were asked to re-enter their passwords again. In the second experiment, the participants created 5 different passwords, and had to re-enter their passwords a total of 24 times over the course of the next 10 weeks. For both experiments, the percentages of their password being input correctly, as well as reaction times, were measured. In their follow-up discussion of the results, the authors argue that the mean accuracies of 77% and 82% for both experiments are reasonable, considering that this is a new method. However, the experimental setup assumed creating passwords of fixed length. As it has been shown by Brumen in their follow-up analysis of the GCPS and its weaknesses, chess passwords of 4 characters in length do not provide sufficient security [6]. But if the participants were able to construct longer passwords, their accuracies would have been much lower, as well. These arguments point at the inherent problems with experimental design of the original study. Our work aims to take these points into consideration in order to re-evaluate the method under fair conditions. A further step forward is made by comparing the security and memorability of the GCPS with the other methods.

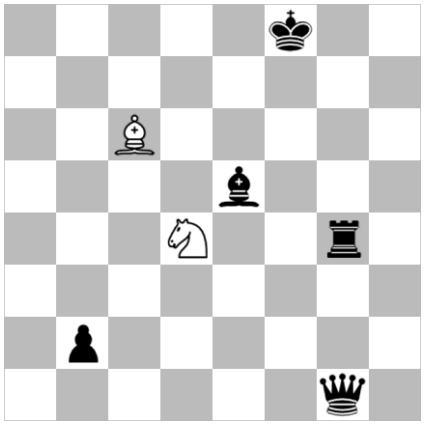

Figure 1. GCPS interface featuring a chessboard with a random, 7-character long password

C. Association Lists

The hybrid textual-graphical method Association Lists was proposed by the authors of this paper in their previous work [7]. The main idea is to employ associations between user-chosen words to boost memorability, while still maintaining a high degree of security by providing a large enough pool of available words. In the original proposal, the user is presented with three columns of ten words they have previously chosen to appear in their pool. The user's password is an indefinitely long sequence of words, with each selection made from the next column. Ideally, users would choose seemingly unrelated words to appear in the pool, and their password. That could prevent targeted guessing attacks, while keeping a high password retention due to personalized associations known only to the user.

However, there are several issues with this approach. First, the pool of available words is too small. Related to that, the users would need to create a substantially longer password to satisfy the increasingly demanding security requirements, which would in turn negatively affect the memorability. Finally, relying on the users to come up with the appropriate words for their personalized pool is not the best approach, because it hinges on the assumption that the users are security-aware and primarily motivated to protect their personal accounts. Taking these problems into account, we introduce an alteration to the original scheme. Instead of a predetermined set of words, the user is allowed to choose any word in the dictionary to appear as a given word in the password (see Fig. 2). In that sense, the proposed method can be considered somewhat similar to passphrases [8], though the system aids the user during the input by filtering and displaying a list of possible words as the user types, to decrease input time. To prevent the users from choosing predictable sequences of words, a distance algorithm should be employed in order to discard related words from the available pool. In our study, we examined the version without such restrictions, however; much like with textual passwords, and the GCPS, we were primarily interested in the strength and memorability of user-chosen passwords without any additional password requirements in place.

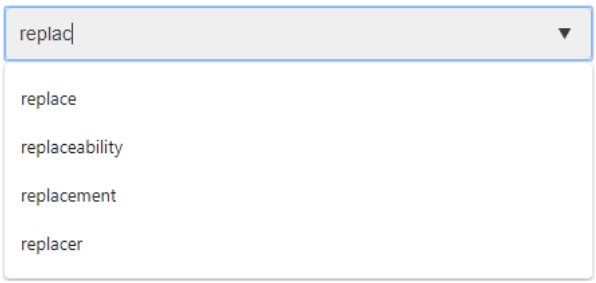

Figure 2. Word filtering and selection based on user input in Association Lists

III. RESEARCH METHODOLOGY

To examine and compare the security and usability of the three selected authentication methods, an experimental procedure was devised. We recruited a total of 58 students (36 male and 22 female) to participate in the study. All participants were between 18 and 27 years of age (mean = 20.5 years) and were undergraduate students of Computer Science or Psychology. To increase the motivation, the experiment was incorporated into the lectures concerning password security, and memory retention. The participants were presented with the experiment as an opportunity to assess their security awareness and were promised their individual results after the conclusion of the user study for future reference.

During the lectures, the students were educated about the concept of graphical passwords on the example of the GCPS, and the Association Lists. They were encouraged to ask questions about the schemes and had the chance to familiarize themselves with them by trying them out prior to the start of the experiment. For the sake of clarity, we broke down the experiment into three stages. To prevent any possible violations of internal validity, the participants were notified and briefed about each stage immediately prior to the start of it. Each stage is henceforth described separately.

In the first stage, the participants were presented with a real-life scenario in which they were required to protect their own account containing personal information. For that purpose, they were tasked to create three passwords: a textual password, a chess password in the GCPS, and an association password using Association Lists. While they were informed about the importance of secure passwords, and the ability to memorize them, they were not educated about *how* both secure and memorable passwords should be formed. With such approach, we ensured they created their passwords to the best of their own ability. During the registration phase, we measured the registration times and recorded the created passwords for all three authentication methods.

Immediately after the registration, the participants were tasked to login using their newly created passwords. Considering the participants were not aware of the second stage at the start of the experiment, the chances of them rehearsing the passwords after creation to keep them in their short-term memory were minimal. For every method, the participants had up to three attempts to login, after which their account would lock down, preventing any further attempts with the same authentication method. On the contrary, a successful attempt displayed a welcome message. Passwords and login times for all attempts were recorded. When computing the metrics, we considered the means across all attempts.

Two weeks after the registration and the initial login, the students were asked to participate in the final stage of the experiment. In this stage, the participants were tasked to login into their accounts with all three passwords that they created at the beginning of the experiment. In order to minimize the effect of any other external variables, the setup of the third stage mimicked the second stage's: three login attempts were allowed for every password, and all password attempts as well as login times were saved. Only 46 out of the initial 58 students completed the final stage and were included in the final analysis.

To estimate strength of the user-generated passwords, we used two metrics. First, we considered entropy, which is a common metric used in password research for the estimation of password strength. The advantage of the metric is that it takes both the length of the password and the character pool into account. In spite of its popularity, certain studies have pointed out the metric's weaknesses, however [9]. In particular, passwords do not follow any statistic distributions (such as languages), and guessing an entire password is not dependent on the characters that we have already guessed. For that reason, we also calculated the number of combinations for an exhaustive brute-force attack, which gave us a more accurate representation of how strong a given user-chosen password actually is.

Usability was also estimated using two metrics. Time was measured during the registration and login process to determine how long it would take the user to authenticate using a chosen method. The timer started when the user opened the login or registration screen and stopped when they pressed the submit button. Alternatively, we could measure the password input time alone, though we wanted to take the password recall time into account as well. The second metric was the number of attempts necessary for the user to input the correct password. We decided to limit the attempts to a maximum of three; if the participant was unable to input their password correctly on their final attempt, we would mark their number of attempts as four.

IV. RESULTS

To report the results in a comprehensive and organized manner, we broke the result section down into two parts, with each part focusing on one of the considered aspects: security and usability. Every aspect was initially evaluated by analyzing the corresponding metrics for every method. Then, we conducted pairwise comparisons between the authentication methods to determine whether there are statistically significant differences between them in terms of a given metric. Finally, we conducted a time analysis, in which we observed how the metric scores for every method changed over a period of time. The results of both statistical analyses are reported in Tables 1 and 2, respectively. Considering the participants had to create and re-enter all three passwords, a related samples Wilcoxon signed rank test was used for the statistical comparisons.

A. Security

The entropy of all user-generated passwords was on average higher than average password entropies estimated in some studies exploring the strength of human-generated passwords [10]. That indicates the participants followed our instructions and tried to create strong passwords based on their own understanding of password strength. A median of 74 bits for both textual and list passwords is relatively high, suggesting that the passwords were not only sufficiently long but also contained various types of characters. The average of 89 bits for lists was slightly higher than the average of 82 bits for textual passwords, and the distribution of entropy scores was more skewed towards higher scores for the list passwords. Primarily,

that is because character pool size is constant for all association list passwords, which means that entropy is entirely dependent on the password size. In other words, the entropy scores for textual passwords were potentially lower in cases when the participants did not use all types of characters (digits, lowercase, uppercase, and symbols). Comparatively, a median of 51 and a mean of 67 bits for chess passwords is much lower. Using the Wilcoxon signed rank test, we showed that the difference between the chess passwords and the other two methods is statistically significant for all three stages. Despite the fact that the GCPS had the highest maximum entropy, the majority of the participants chose passwords that were too short. The most plausible reason is that they were simply not aware of what constitutes a strong chess password. Whereas that could be amended with education, longer passwords would certainly have a negative impact on their memorability, thus putting their usefulness in question.

Fig. 3 displays the distributions of entropy scores for all three authentication methods during the registration and the two subsequent login stages. Means and medians of both logins were similar to those of registration for all methods, suggesting that the participants were attempting to input the same passwords they had created during the registration. The differences in entropy were caused by the incorrect password inputs. By comparing the entropies for each authentication method through time, we found there were practically no significant differences between either of the logins and the registration stage. The two reported significant differences, as seen in Table 2, are situational and do not suggest anything more than a few more errors during the logins. In general, the participants were able to at least remember (or guess) a similar size and characters appearing in their original password.

To examine resilience against brute-force attacks, we also considered the number of combinations necessary to complete an exhaustive key search. The distributions of combinations depicted in Fig. 4 appear on a logarithmic scale, and they are somewhat similar to the distributions of entropy in Fig. 3. Considering both metrics are measuring password strength, and rely on the length of the password, such result is to be expected. An important distinction lies in the fact that this number represents the most pessimistic cracking scenario. In other words, the user's password can *definitely* be retrieved in the number of tries equal to the number of all possible combinations. Consequently, if the average number of combinations is too low, the method can be considered unsafe.

Since arithmetic means are skewed by the exponential values, we focused on reporting and interpreting medians. Of the three considered authentication methods, the user-generated textual passwords had the highest median of 5.1×10^{25} combinations, followed by association lists with 1.5×10^{22} possible combinations. While the former had a higher average, the latter had a wider distribution, which means that some participants created passwords requiring considerably longer to be cracked. Regardless, a statistical test comparing both methods showed that there were no significant differences between the methods during any of the password input stages. On the contrary, the GCPS had significantly less combinations than both competing methods during registration and the first login. After the two-week period, the number of combinations of the input passwords did not significantly differ from that of textual and list passwords, suggesting that the participants were in general inputting longer passwords, which hints at poorer memorability discussed in the next section. Regarding security, a median of 2×10^{17} combinations for registered chess passwords can be considered inadequate protection against contemporary computers.

A temporal analysis once again showed no statistically significant differences between any of the stages for any of the three authentication methods. This finding supports our previous conclusion that the participants were able to recall similar lengths of their passwords after a two-week period. Albeit not significant, the differences were bigger after two weeks than immediately after the registration. It would therefore be sensible to repeat the experiment after a month to further investigate the memory retention.

B. *Usability*

Input times are showcased in Fig. 5. As expected, the registration time was the shortest for textual passwords, with the participants spending 35.6s on average (Mdn_T = 30.8s) to create their password. Despite being graphical methods with the participants having little experience using them, the chess and list passwords did not require much more time to be created: 38.3s and 39.3s on average (Mdn_C = 33.6s, Mdn_L = 35s) for chess and list passwords, respectively. Comparing the methods using the Wilcoxon

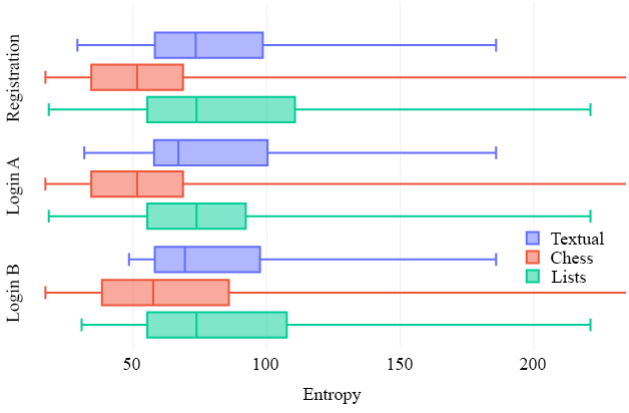

Figure 3. Entropy during registration and the logins for all three authentication methods

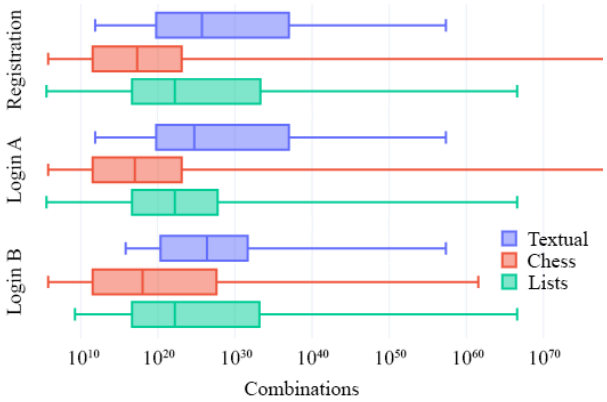

Figure 4. Number of possible password combinations during registration and the logins for all three authentication methods

TABLE I. A PAIRWISE COMPARISON OF AUTHENTICATION METHODS IN TERMS OF SECURITY AND USABILITY DURING ALL THREE STAGES

	Registration			Login A			Login B		
	T-C	**C-L**	**T-L**	**T-C**	**C-L**	**T-L**	**T-C**	**C-L**	**T-L**
Time	0.662	0.856	0.437	*0.007**	0.224	0.084	0.316	*0.01**	0.194
Combinations	*0.01**	*0.002**	0.914	*0.028**	*0.009**	0.856	0.059	0.102	0.43
Entropy	*0.001**	*0**	0.25	*0.001**	*0**	0.633	*0.027**	*0.005**	0.509
Attempt				0.52	0.061	0.284	*0.012**	*0.001**	0.168

* Statistically significant ($\alpha < 0.05$)

signed rank test found no significant differences between them. This might imply that the higher input time for graphical passwords could have been compensated by a lower time necessary to compose a new password.

Against our expectations, however, textual passwords required the most time during the first login. With 47.4s on average (Mdn_T = 43.9s), they took significantly longer than the chess method with an average login time of 32.8s (Mdn_C = 27.7s). List passwords took even less time (Mdn_L = 23.1s), though the difference was not significant due to the distribution of the login time with an average of 38.2s, which was closer to that of the textual passwords. While the shorter times for chess and list passwords might seem unusual on the first glance, they could be explained by the methods' design. It has been suggested that graphical passwords might benefit from higher memorability. Since the login time does not include only input time but recall time as well, we believe shorter times could mean that the participants were able to recall the created passwords a lot quicker. Considering it would be unlikely for input times to change to such a degree immediately after registration, such a conclusion seems appropriate.

Two weeks after the initial experiment, the login times did not change substantially. While the average login time slightly improved for textual passwords to 39.7s (Mdn_T = 37.5s), the graphical passwords' average times marginally increased to 41.2s and 34s (Mdn_C = 41.5s, Mdn_L = 31.2s) for chess and association list passwords, respectively. The slight increase for graphical passwords was expected. Since both methods were a novelty for the participants, and they had no opportunity to practice using the methods during the two-week period, their reaction times naturally increased. The same finding was reported by McLennan et al. in their GCPS proposal, in which they showed that the reaction times would decrease once again with subsequent uses of the authentication method [4].

A particularly important aspect to investigate was the methods' memorability (see Fig. 6). During the first login following the registration, the recall rate was expectedly high. The association lists had the highest recall rate of 94.64%, with an astounding 92.96% managing to login on their first try. Textual passwords followed with 87.72%, while the chess passwords had a recall rate of 84.48%, with 79.31% users successfully logging in on their first attempt. A Wilcoxon signed rank test found no significant differences between the three methods, however, suggesting that memorability is still suitably high for all three methods shortly after the password creation.

After the two-week period, however, the memorability of all three authentication methods dropped significantly. Exactly half of the users managed to remember their list passwords within three attempts (42.5% recalled them on their first attempt). 36.95% of the participants recalled the textual password, of which only 23.91% remembered it on their first attempt. Despite being a graphical authentication method, chess passwords had the lowest total recall rate of 15.21%, with only 10.87% remembering their password on their first attempt. A follow-up statistical test showed that the GCPS' memorability was significantly worse than either the textual or the list passwords'. With several chess pieces on the board, the participants found it difficult to remember the individual pieces and positions they placed them on, while also keeping track of the order. The longer the password was, the less likely it was for the participant to remember it; in practice, none of the longer chess passwords were remembered after two weeks. The same could

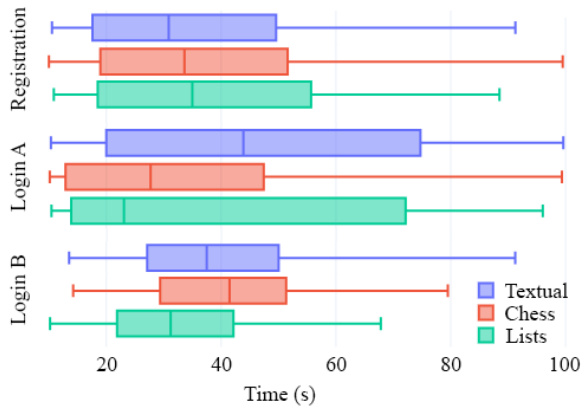

Figure 5. Password input time (in seconds) during registration and the logins for all three authentication methods

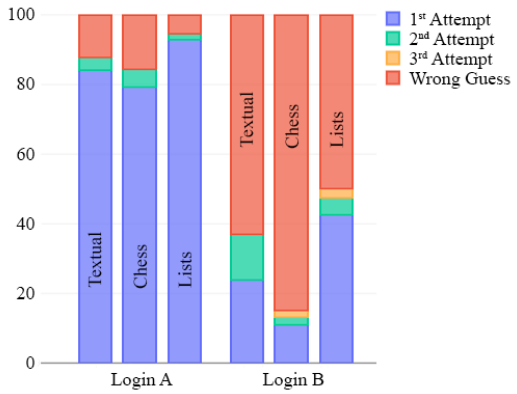

Figure 6. Password input time (in seconds) during registration and the logins for all three authentication methods

TABLE II. TEMPORAL ANALYSIS OF SECURITY AND USABILITY ASPECTS FOR THE CONSIDERED AUTHENTICATION METHODS

	Textual Passwords			**GCPS (Chess)**			**Association Lists**		
	Reg-Log$_A$	**Reg-Log$_B$**	**Log$_A$-Log$_B$**	**Reg-Log$_A$**	**Reg-Log$_B$**	**Log$_A$-Log$_B$**	**Reg-Log$_A$**	**Reg-Log$_B$**	**Log$_A$-Log$_B$**
Time	*0.019**	0.221	0.31	0.103	0.318	0.176	0.513	0.41	0.958
Combinations	0.866	0.732	0.909	0.512	0.27	0.345	0.655	0.64	0.793
Entropy	0.953	*0.047**	0.122	*0.05**	0.315	0.537	0.144	0.522	0.862
Attempt			*0**			*0**			*0**

* Statistically significant ($\alpha < 0.05$)

be claimed for textual passwords. The participants that followed our instructions and endeavored to create longer and more complex passwords, found them more difficult to remember after the two-week period. While they faced a similar problem when creating list passwords, they were generally more successful remembering them even when these passwords were a bit longer. We theorize that the associations between the selected words appearing in the password might have helped with memorability.

V. CONCLUSION

In the field of Password Security, a lot of attention has been devoted towards improving existing and developing novel authentication methods, in pursuit of a resolution to the long-lasting password security problem. However, few of these studies have employed empirical evaluations of the proposed methods, and even fewer have conducted any comparisons with other methods. In our research, we attempted to fill this gap by examining and comparing the security and usability of several password authentication schemes in a two-fold experiment. The conclusions drawn from the results have important implications for any future password evaluation studies.

An essential consideration to take in any evaluation of authentication methods should be the experimental design. In that sense, our choice of the GCPS to be included in our study was beneficial toward this conclusion. Whereas the authors had included an extensive experimental evaluation of the GCPS in their study, their obtained results have a limited usability in real-world scenarios. By compelling the participants to choose their own passwords without any limitations, we could examine both the strength and memorability of the user-chosen passwords as they would appear in the wild. We found out that without previously educating the users about what constitutes a strong chess password, we cannot expect them to create passwords of sufficient strength to withstand brute-force attacks. What is more, recall rate of these user-chosen passwords was as low as 15% after just two weeks, posing a question as to how memorable longer chess passwords would be after a certain period of time.

On the contrary, association lists proved to be a rather promising alternative to textual passwords. A close look at the differences shows that association list passwords were marginally weaker. However, they make up for that with increased memorability. While only slightly higher than the textual passwords' during the first login, the difference in recall rate increased after two weeks. It would take additional experiments to determine whether the memory retention would remain as high after a longer period of time, as well as how the memorability would be affected by further increasing the security requirements. However, rivaling textual passwords in terms of security and usability, this association-inspired method and the results of this study suggest that cognitive approaches could have the potential to extend the memorability of text-based authentication, while also maintaining a high degree of security. To draw such conclusions, we once again outline the importance of comparisons against textual passwords, which should become a necessity in any future empirical evaluations of the existing and novel methods.

ACKNOWLEDGMENT

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding no. P2-0057).

REFERENCES

[1] P. Tarwireyi, S. Flowerday, A. Bayaga, "Information Security Competence Test with regards to Password Management," In Proceedings of the 2011 Conference on Information Security South Africa, 2011.

[2] J. Bonneau, C. Herley, P. C. Van Oorschot, and F. Stajano, "The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes," IEEE Symposium on Security and Privacy, pp. 553-567, 2012.

[3] L. Bošnjak, and B. Brumen, "Rejecting the Death of Passwords: Advice for the Future," Computer Science and Information Systems, vol. 16, no. 1, pp. 313-332, 2019.

[4] C. T. McLennan, P. Manning, and S. E. Tuft, "An Evaluation of the Game Changer Password System: A New Approach to Password Security," International Journal of Human-Computer Studies, vol. 100, pp. 1-17, 2017.

[5] A. De Angeli, L. Coventry, G. Johnson, K. Renaud. "Is a Picture Really Worth a Thousand Words? Exploring the Feasibility of Graphical Authentication Systems," International Journal of Human-Computer Studies, vol 63, no. 1, pp. 128-152, 2005.

[6] B. Brumen, "Security Analysis of Game Changer Password System," International Journal of Human-Computer Studies, vol. 126, pp. 44-52, 2019.

[7] L. Bošnjak, and B. Brumen, "Shoulder Surfing: From An Experimental Study to a Comparative Framework," (unpublished) [Online] Available: https://arxiv.org/abs/1902.02501, accessed February 2019.

[8] S. N. Porter, "A Password Extension For Improved Human Factors," Computers & Security, vol. 1, no. 1, pp. 54-56, 1982.

[9] W. Ma, J. Campbell, D. Tran, and D. Kleeman, "Password Entropy and Password Quality," Fourth International Conference on Network and System Security, pp. 583-587, 2010.

[10] D. Florencio, and C. Herley, "A Large-Scale Study of Web Password Habits," Proceedings of the 16th International Conference on World Wide Web, pp. 657-666, 2007.