Clark, Gradeigh; Lindqvist, Janne; Oulasvirta, Antti

Composition policies for gesture passwords: User choice, security, usability and memorability

Published in:
2017 IEEE Conference on Communications and Network Security (CNS)

DOI:
10.1109/CNS.2017.8228644

Published: 01/01/2017

Document Version
Peer reviewed version

Please cite the original version:
Clark, G., Lindqvist, J., & Oulasvirta, A. (2017). Composition policies for gesture passwords: User choice, security, usability and memorability. In *2017 IEEE Conference on Communications and Network Security (CNS)* (IEEE Conference on Communications and Network Security). IEEE. https://doi.org/10.1109/CNS.2017.8228644

Composition Policies for Gesture Passwords: User Choice, Security, Usability and Memorability

Gradeigh D. Clark
Rutgers University

Janne Lindqvist
Rutgers University

Antti Oulasvirta
Aalto University

***Abstract*—Research on gesture passwords suggest they are highly usable and secure, leading them to be proposed as a strong alternative authentication method for touchscreen devices. However, studies demonstrate that user-chosen gesture passwords are biased towards familiar symbols, increasing the risk of guessing. Prior work on gesture elicitation focuses on creating sets with high overlap, but gesture passwords require solving an inverse problem: minimal overlap between different users. We present the results of the first study (N = 128) of composition policies for gesture passwords, wherein we compare four policies derived from unique properties of gesture passwords. Our main result is that implementing a policy changes user choice, security, usability, and memorability compared to a control group and that the strength of those changes depend on the policies. We report trade-offs among the instruction policies while showing that simple policies cause users to choose stronger and diverse gesture passwords.**

I. Introduction

This paper studies authentication on mobile devices, which is rapidly becoming the dominant means to access the internet, banking, and other privacy-sensitive data. However, we still lack an authentication method with acceptable usability, security, and memorability [1]. The most prevalent authentication methods on mobile platforms, text passwords and PINs, are legacies of other terminals and are beset by a number of issues [2] – some examples being: high rates of password reuse, weak password choices, and confusing policies. Text passwords are incompatible with mobile interaction, which allows only short and intermittent bursts of visual attention [3]. No acceptable solution to text passwords exists without flaws. The (Android) pattern unlock has weaker security than 3-digit PINs [4] and click-based graphical passwords are susceptible to attack algorithms [5]. Biometric methods suffer from revocation problems: having a fingerprint stolen has serious ramifications outside of smartphone privacy; spoofing attacks, to which the iPhone fingerprint scanner is dangerously susceptible [6]; and high dependence on the original recording conditions – the iPhone scanner does not work well in humid environments.

There has been considerable effort in the security and mobile computing communities to develop gestures as an authentication method [7], [8], [9], [10], [11], [12], [13], [14], [15]. We can categorize these efforts into three approaches: (1) *template-based gesture authentication* [13], [12], [16], which employs biometric features (e.g. hand size) to authenticate users with predefined gestures without any secrets. The second: (2) *continuous authentication* [7], [8], [17], analyzes behavior in real time to determine a user's identity; this is inefficient and deviations in normal behavior on the device can lead to false negatives. Finally, there is: (3) *free-form gestures* [9], [13], [18], which allows users to draw any shape or pattern on a touchscreen sensor with one or more fingers.

We believe that free-form gestures are promising and warrant further research, as they can incorporate the biometric factors of template-based authentication and the learning methods of continuous authentication while still providing the advantages of a full-blown authentication method. Gestures fit the requirements of mobile interaction [3] – minimal need for visual attention – since recognizers can be built to be scale-, rotation-, and location-invariant. Gestures also have potentially more security than passwords owing to the difference in expressive information between a text string and a continuous, multi-finger input. Moreover, free-form gestures enjoy many advantages: (1) lower accuracy required for a correct input, (2) faster entry, (3) increased memorability owing to their visual nature [19], and are easily-adoptable for users given that gestures are as familiar as drawing shapes.

This paper addresses a major problem: when just told "create something others could not guess", users are generating simple passwords with high overlap. A field study on gesture password use demonstrated that, out of 345 gesture passwords created, only 150 were actually unique (43%) and that 89% of the generated gestures could be grouped into categories like shapes, letters, lines, numbers, symbols, and words [18]. We also examined a dataset of a previous gesture password laboratory authentication study [9], and we observed that participants generated passwords that were similar to those in the field study [18]. We note that we have later conducted more systematic analysis of these data and more when implementing guessing attacks [15].

The lack of password uniqueness for gestures is a serious concern. Attackers often compromise user accounts through the use of guessing attacks, wherein leaked datasets consisting of user-chosen passwords are used as a dictionary to attack other user accounts. The preponderance of evidence from prior studies [18], [20], [21], [9] raises questions about how users can be aided in creating more unique gesture passwords. If only 43% of generated passwords are unique and 89% of them fall into common buckets [18], then guessing attacks are not limited by trying to figure out what the gesture password is but rather how to perform it the right way to pass a recognizer. Guessing attacks trained using prior user data against newly user-chosen gesture passwords have achieved crack rates of 48% [15], so changing user choice is critically important.

The goals of our study are two-fold. First, we examine the effect of composition policies on how users generate gesture passwords – can instructions change how users create gesture passwords? Second, we examine how composition policies might affect the security, usability, and memorability of gesture passwords – do instructions modify the interaction experience of using gesture passwords?

We created several plausible policies that are both actionable and span different aspects of possible variables, recruiting 128 participants and dividing them randomly into one control group and three policy groups. The policies are: shape complexity, multiple fingers, and speed. We discuss the policies in detail in the Method section, while Figure 1 shows examples of gestures created with the policies.

Our main finding is that simple composition policies change how users create gesture passwords, guiding users towards different categories. They are successful at changing user behavior and addressing issues with password choice. However, policy instructions have an adverse effect on memorability: both the Random group and the Speed group produce significantly worse reproduction scores when login trials are compared to templates. Finally, we found that the Multiple group underperforms in usability metrics compared to the Control group: the users take longer to create gestures and participants iterated more frequently to match the policy instructions.

The major contributions of this paper are as follows:

1) We present the results from the first study to apply composition policies to gesture passwords.
2) We define and examine three plausible policy designs for gesture passwords: shape complexity, multiple fingers, and speed.
3) We show that simple composition policies change how users create gesture passwords. Our policies guide users towards different categories of gestures, modify the number of fingers used, and change the velocity.
4) We compare each policy systematically on security, memorability, and usability using a variety of different metrics. We find that any policy improves the security of gestures relative to the control at the expense of memorability and usability for certain policy groups.

II. Related Work

This section places our work in context by showing the divergent focuses that gesture password research has taken: methods for gesture authentication and user studies about properties of gesture passwords.

Continuous gesture authentication has been done on smartphones by analyzing tapping and stroke behavior [22], [7] and through processing natural movement with the phone [8]. Swiping behaviors have also been shown to be used for authentication and one scheme has shown the Android Swype autofill method can be used for authentication [17]. Authentication systems have also leveraged biometric features to perform authentication by using smartphone features such as finger velocity and acceleration [11], [12] and handwriting [13]. The community has expended much focus on the development of authentication methods, but the purpose of our work is to study the way users create gesture passwords. A multi-class system, Garda [14], has been proposed that achieved the lowest recorded error rate (0.040) under targeted attack scenarios. In this study, we use an existing gesture authentication method [9] and are not proposing a new system.

A first-look study at the security and memorability of gesture passwords [9] found that, with no instructions: half of the users will create a single finger gesture and the other half will create multi-finger and that the gestures people remembered best were signatures and simple shapes. A field study of gesture passwords across multiple accounts [18] found that: participants could remember gesture passwords at the same level as text passwords, gestures were input faster on average, and that biases in the user-chosen distribution meant users were more likely to pick gestures based on simple categories like shapes, letters, and numbers. Users have also demonstrated a lack of originality with generating unique gestures for HCI tasks, often repeating known gestures or picking uninventive ones [20]. Our study is meant to act as a complement to these two studies, wherein we are proposing policies to help users generate more diverse gesture passwords.

In summary, research in the field of gesture passwords have focused on either developing new authentication methods or on exploratory user studies. Our work is the first study on trying to change the way users generate gesture passwords along with an examination of how those instructions affect user choice, security, memorability, and usability.

III. Method

The foundation of our study design is the generate-test-retest paradigm wherein participants are asked to generate a gesture, recall it after a distraction, and then recall it again after a period of 10 days or longer. This design methodology improves ecological validity since it allows us to simulate the effects of variables over time. It also gives us a more accurate picture of how the policies affect participants after they leave the laboratory. Moreover, recent studies have shown that, at least for text passwords, passwords generated in laboratories are strikingly similar to real-world cases [23].

Researchers have expended effort in trying to assist users to create more meaningful gestures in a variety of different contexts, such as motioning with a phone in space to invoke commands [24] or performing complex tasks across two devices [25]. The way participants create personalized gestures has been studied [20], only to find that explicit instructions to focus on creating unique gestures did not stop participants from already using gestures that are familiar to them. Most of the time, researchers are trying to help designers create gesture sets [26], [25], [24] and we could not find anything to assist us on what a composition policy might look like.

Research on text and graphical password policies have shown success in modifying user-choice in password selection. Text password policies are often strict, deliberately forcing users [27] to meet a requirement such as a certain password strength or a combination of characters. Graphical password policies are less explicit and more persuasive, often asking users instead of forcing them [28], [29], [30]. There is no prior work on generating gestures under composition policies,

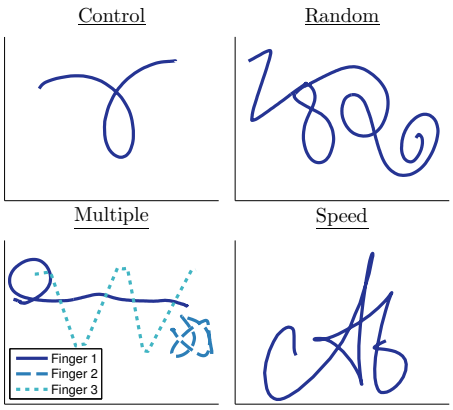


Fig. 1: Examples of gestures generated under each policy: *Control*, *Random*, *Multiple* and *Speed*. The *Random* group is told to generate more random motion, the *Multiple* group is told to create different gestures with different fingers, and the *Speed* group is told to create complex gestures at high speed.

despite the need to improve user password choice. We opted to do a composition policy in the vein of graphical passwords, where we make simple suggestions to users about how to improve their password. Forcing users is premature given that it is unknown how users will respond to instructions.

Policies in graphical passwords have asked users to click on other parts of the screen; text password composition policies have asked users to include a special combination of characters in their passwords. Composition policies for gesture passwords have multiple avenues to explore when asking users to change the way they create a password. The design of the gesture shape is an important consideration. A gesture containing many sudden changes, twists, and turns (effectively appearing more random) is extremely difficult to copy. However, this can be harder to input and may take a longer time, sacrificing usability. Gestures can also be fluidly input into the device with varying speed; this can make it harder for an attacker to obtain a correct mental image. At the same time, it could influence participants to create more fluid and flowing gestures that are easier to input. Gestures also allow for multiple fingers to draw on the screen at the same time – if different fingers drew multiple gestures then an attacker could be easily confused or misled.

These three broadly defined qualities motivate the three policies we used in our study. We believe that these instructions exploit properties that are unique to gesture passwords, much like how text and graphical password instruction policies directly target their own unique aspects. One group, *Random*, was instructed to create gesture passwords by trying to create a confusing shape. We gave them a set of suggestions to follow when creating their password: (i) try to suddenly reverse direction, (ii) make curves sharper and more jagged, (iii) do not complete circular loops but rather diverge from the motion, and (iv) instead of traveling in a straight line, try diverting the gesture off its path. The *Multiple* group received instructions asking them to: (i) use multiple fingers, and (ii) to draw different gestures with different fingers. The *Speed* group was asked to create a complex, fast gesture they could repeat comfortably. Finally, we also had a *Control* group that were not given any specific instructions beyond creating a secure gesture password others could not guess.

A. Participants

We recruited participants with campus flyers, email lists, and through online message boards. The participants were only required to be 18 years or older and were asked to perform a study involving mobile devices, with no explicit mention of passwords or gestures anywhere in the recruitment materials. We targeted a recruitment of 128 participants, splitting them randomly into four even groups of 32 people each.

The ages of our participants ranged from 18 to 34 ($\mu = 21.53$, $\sigma = 2.86$); 82 were male and 46 were female. The educational background varies amongst the participants; 34 were pursuing a graduate degree, 92 were pursuing their bachelor's, and two had already graduated (one with a Master's, the other with a PhD). Of the 128, 124 returned for the second study – an attrition rate of about 3%. Participants were compensated $35 for completing the entire study.

B. Apparatus

We implemented the Protractor [31] algorithm to perform our gesture recognition and we extended it to work for multitouch gestures by treating each finger individually and averaging the full result. Protractor uses the inverse cosine distance between two gestures as a similarity score. We use the same configurations as in the original Protractor paper with respect to the number of sampled points and thresholding for optimally lowering the error rates [31].

C. Experiment Design

Our experiment followed a between-group design, with a repeated measurement variable of gesture repetition and a categorical variable corresponding to the policy group. Participants were asked to register (Generate), perform a distraction task, then login (Test). We performed the entire study across two sessions. The second session was held after a period of ten days following the first session.

First session: The participants began the session by being informed of: the purpose of the study, their rights as participants, and their compensation for completing the entire study. After understanding and agreeing to the aforementioned, the participants signed their consent form and began the study.

Gesture Creation (Generate): First, the participants were given a tablet and asked to familiarize themselves with the gesture acquisition application. Participants were given an outline of what the study wanted them to do after they felt they were familiar with the application. Participants were told that they need to equally weigh two factors: security and memorability; "Create a secure password, one that you think others cannot guess." Participants were urged to create a secure and memorable password that fit the policy instructions and was not a reproduction of an example from their instruction document. Participants were given the policy document after listening to the ground rules and given unlimited time to both read the policy document and to create a gesture that fits it.

Distraction Tasks: Participants were asked to perform a mental rotation worksheet and count down silently from 20 to 0 in order to flush their short-term memory.

Recall 1 (Test): Participants were asked to login successfully. Failed attempts, where the recognizer was unable to match their inputs, were recorded and kept.

Post-recall 1: We then asked them to perform one more login that we would record over their shoulder with a camera for the purposes of doing a shoulder surfing attack in the second session. They were then asked a series of demographic questions about their age, gender, education, occupation, and mobile phone usage. We then thanked them and asked them to return after 10 days or longer for the second session.

Second session: **Recall 2 (Retest):** After a gap of 10 days or more, participants returned and were asked to perform another successful login task with their passwords created from the first session.

Shoulder surfing: We followed a structure similar to other studies wherein participants attempt to replicate a gesture or graphical password from a video [12]. We asked each participant to perform shoulder surfing on four gestures, one from each one of the four groups. Each participant was presented the four gestures based on a latin square arrangement in order to mitigate any possible training effects. Participants were permitted to watch the video unlimited times and had the option to give up at any time.

D. Analysis Methods

In here, we describe the different metrics we used to evaluate the security, usability, and memorability differences between the two groups.

1) User Choice: We want to measure different ways that the composition policies affect user choice. A way to do this is to see whether the *categories* of user-chosen gestures vary between groups. Prior work [18] has shown that users cluster their choices in different categories, usually based on meaning. We have expanded these categories to highlight differences more effectively. The categories we are using are: *Numbers*, which refer to users drawing numbers; *Letters*, which refer to users drawing roman characters; *Shapes*, which refer to users drawing simplistic geometric shapes; *Words*, which refer to users drawing out specific words, *Shape Combinations*, which refer to users combining two or more geometric shapes together, *Sloping*, which refers to users drawing flowing gestures that have a high degree of curvature and little-to-no jagged edges, and *Abstract*, which refers to gestures that cannot be categorized easily. Figure 2 shows examples of each category.

Other user choices are a product of individual policies. The Multiple group, for example, has a focus on using more than one finger when creating gestures. To check whether the effect worked, the gestures can be clustered by how many fingers are used in each group and compared to each other. Multiple is expected to have a larger proportion of multiple finger gestures, but the effect is less clear on the other groups. For the Speed group, we expect gestures to have a higher velocity compared to the Control group.

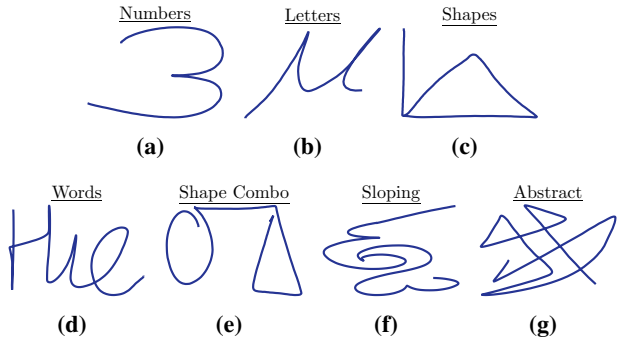


Fig. 2: Examples of all of the gesture group categories, all generated by users in our study.

2) Security: We are using two measures for the security of a gesture password. The first is True Positive Rate (TPR), the ratio of honest users' authentications to the total number of true authentications. The second is False Positive Rate (FPR), the ratio of attackers authentications to the total number of rejected authentications.

3) Memorability: We use the gesture recognition scores of the participant as a way of measuring the memorability between groups for a given session. The recognition scores are defined as the inverse cosine distance measured between a gesture and a template [31]. An exact match would yield a score of infinity, since the cosine distance between two exact vectors of data is zero – the reciprocal of this would be infinity. However, two gestures are rarely so closely matched; scores typically do not go much higher than 20 and most are clustered in the range between 1 and 4.

Short-term memorability can be examined by comparing the scores of the groups for the first session as a whole. If the similarity scores between groups differ significantly during short-term recall then we can say that certain policy instructions can make gestures more difficult to recall. Long-term memorability can be assessed between groups through use of the second session scores, spaced out at a minimum of 10 days after the first session.

4) Usability: Factors that affect the usability of gesture passwords and gesture policies are those that can be irritating or cumbersome to both the creation process and the authentication procedure. The usability factors we examine are: duration of a gesture, time to create a gesture, and the amount of thought put into creation. Duration affects how long it takes a user to bypass their lock and reach their main activity. Creation time can affect user experiences with trying to craft gestures that fit policy instructions and can lower the overall security and memorability of the generated password. The amount of thought can be an indicator of the policy's ability to engage a user in meaningful attempts at password creation.

IV. Results

In this section we report on descriptive and inferential statistics. The majority of our statistical testing is done using Kruskal-Wallis (KW), a non-parametric test. We use this as an alternative to one-way ANOVA since our variables are not normally distributed. KW does not assume normally distributed data, is more conservative than parametric one-way ANOVA, is not sensitive to differences in sample sizes between the groups, and is robust to outliers. For categorical data, we use Fisher's exact test. We report Holm-Bonferroni corrected p-values for the post-hoc Mann-Whitney U test (MWU) in the event an omnibus test is statistically significant.

A. Effects on User Choice

Gesture Categorization: Table I shows how each policy group was examined to match gestures to categories. A Fisher's exact test shows statistically significant differences for percentage of the gesture categories by policy group ($p = 0.019$). Each policy group has a different category making up a plurality of the total gestures generated. The Control group's largest category is Shapes (34.38%), the Random group's largest category is Abstract (31.25%), the Multiple group's largest category is Shape Combinations (40.63%), and the Speed group's largest category is Abstract (9.38%).

Finger Count: Table II shows the number of fingers for each gesture in each group. A Fisher's exact test showed a statistically significant difference for the percentage of the fingers by policy group ($p < 0.001$). Most policy groups have single finger gestures making up the majority of the generated gestures, at least in the cases of the Control (53.13%), Random (50.00%), and Speed (78.13%) groups. The Multiple group stands out singularly, with only 12.50% of generated gestures having a single finger and 87.5% having more than one finger. The Multiple group is also the only group that generates double-digit percentages of three-finger gestures (15.63%) and is the only group to generate any four-finger gestures (6.25%).

Gesture Velocity: The last user choice change was the effect of the Speed policy on the velocity of generated gestures compared to the control group. KW showed a statistically significant effect of the policy on the velocity [$\chi^2(3,124), p = 0.02$]. Post-hoc analysis using MWU showed that a statistically significant effect exists between the Control and Speed groups

Category	Control	Random	Multiple	Speed
Numbers	6.25%	3.13%	6.25%	3.13%
Letters	15.63%	6.25%	15.63%	9.38%
Shapes	34.38%	18.75%	3.13%	18.75%
Words	9.38%	15.63%	6.25%	18.75%
Shape Combos	15.63%	15.63%	40.63%	12.50%
Sloping	12.50%	9.38%	15.63%	28.13%
Abstract	6.25%	31.25%	12.5%	9.38%

TABLE I: Breakdown of the data per category and policy group, where each policy group contains $N = 32$. A Fisher's exact test showed a statistically significant effect on category by policy group (p = 0.019). Each group skewed towards a particular category: Control has the most Shapes, Random has the most Abstract, Multiple has the most Shape Combinations, and Speed has the most Sloping.

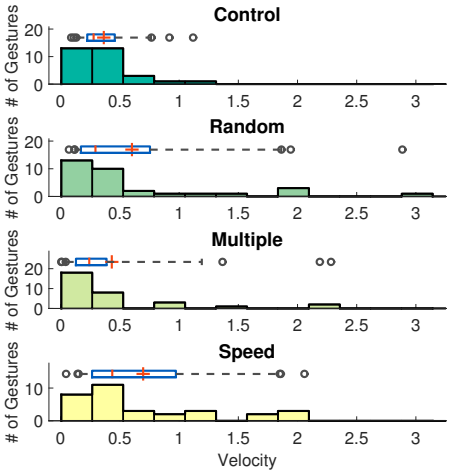


Fig. 3: Velocity histograms for each policy group. There is a statistically significant effect of composition policies on the velocities. Post-hoc analysis showed that there is a significant effect between the Control and Speed groups, with the Speed group being much faster than the Control group.

($p = 0.02, r = 0.283$), but not for the Control and Random group ($p = 0.926$, $r = 0.011$) or the Control and Multiple groups ($p = 0.686$, $r = 0.118$) We observe, then, that the users of the Speed group generated much faster gestures than those in the Control group.

B. Effects on Security

We note that there were four gestures that were not subjected to shoulder surfing trials in the second session due to participant attrition. However, the experimental setup compensated for this by having each participant attack four gestures (one from each group) – this left each group missing just one gesture each. As such, Figure 4 represents histograms of 31 gestures per group for a total of 124 (as opposed to 32 per group, totaling 128).

Number of shoulder surfing attacks: Figure 4 shows the total number of attacks against a given gesture, an aggregate of all four attackers' attempts. As noted in the Method section, we used a strong attack method: the shoulder surfers were allowed to attack repeatedly until they either cracked the system or gave up. This means there are an unequal number of attacks per

Fingers	Control	Random	Multiple	Speed
One	53.13%	50.00%	12.50%	78.13%
Two	43.75%	40.63%	65.63%	18.75%
Three	3.13%	9.38%	15.63%	3.13%
Four	0.00%	0.00%	6.25%	0.00%

TABLE II: The number of fingers per policy group, where each policy group contained $N = 32$. A Fisher's exact test showed a statistically significant effect on category by policy group ($p < 0.001$). Most groups are composed primarily of single-finger gestures, except for the Multiple group which has a majority of more than single-finger gestures and the only group to contain gestures with more than three fingers.

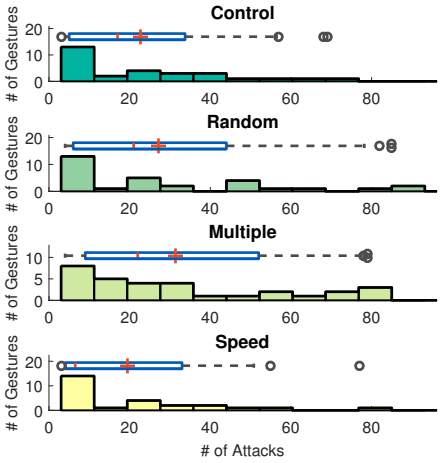


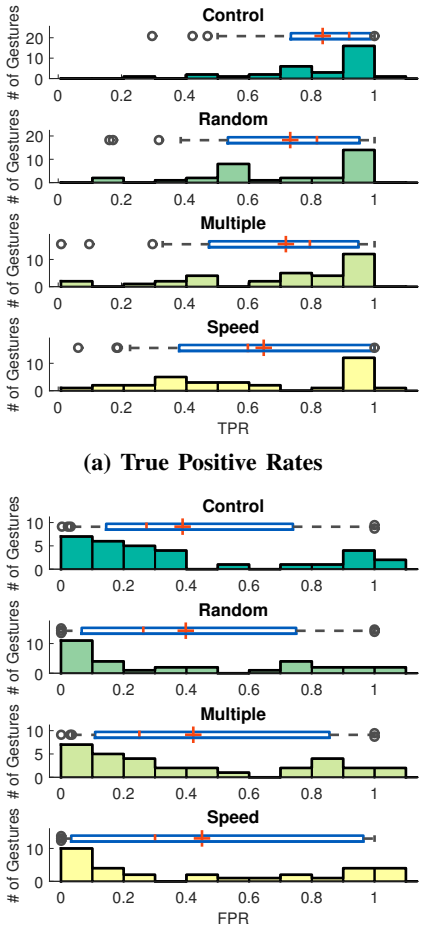
Fig. 4: Histograms displaying the frequency of attacks experienced by gestures in each group. Most attacks fall in the range between 4 (one successful attack per four attacks with no rejections) and 20. Some amount of gestures in each group were difficult for attackers, requiring up to 80 attempts.

gesture; some only have four attack attempts (the minimum, broken immediately by the four attackers) while some have over 50 attempts from dedicated attackers. A KW test on the effect of policy on the number of attacks [$\chi^2(3,124) = 0.42$, $p = 0.93$] was statistically non-significant. We note that most values fall in the range of four attempts (each of the four attackers authenticates successfully with one attack) and 20 attempts. Each group contains a set of gestures that posed particular challenges for attackers, taking upwards of 20 attempts to break through.

True Positive and False Positive Rates: Figure 5 shows the TPR and FPR distributions. Recall that the TPR is the proportion of true users authenticating relative to the total number of true login attempts and that FPR is the proportion of false users authenticating relative to the total number of false login attempts. A KW test on policy relative to the TPR [$\chi^2(3,124) = 4.69$, $p = 0.196$] and the FPR [$\chi^2(3,124), p = 0.22$] demonstrated statistically non-significant results. Though the results are statistically non-significant, this does not mean there are no differences to extrapolate from the distributions. First, a reminder that an ideal system would have all gestures with an FPR of 0 and a TPR of 1. The spread of TPR values less than 1.0 is a consequence of the long-term time gap. When participants returned after the ten-day period (with no training or recall time in between), some percent in each group could not properly log-in, resulting in the variance shown by the error bars. The FPR values are a consequence of the unlimited attempts by attackers, though it is encouraging that they are mostly held between 0 and 0.2.

C. Effects on Memorability

Short-Term Memorability: We used the recognition scores for each participant as a proxy of their ability to accurately reproduce their gesture. Recall that the recognition score is the output of the recognition algorithm comparing an input trial to a stored template. KW showed statistically



(a) True Positive Rates

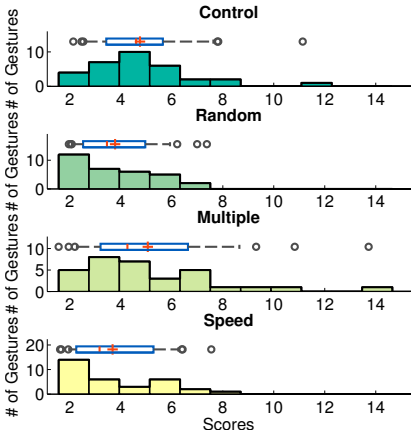
(b) False Positive Rates

Fig. 5: Histograms showing the TPR and FPR values for the four groups. Most groups are consistent in success rates irrespective of policy, and testing showed no statistically significant differences. TPR values less than one are a consequence of the long-term time gap, where a percentage of participants in each group had trouble recalling their gesture. FPR values greater than zero are a consequence of the unlimited attacks (Figure 4).

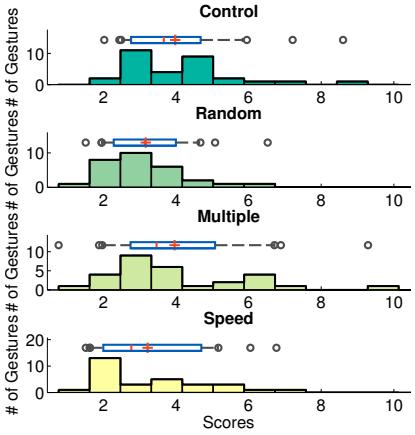
significant differences between the four groups [$\chi^2(3,124) = 10.31$, $p = 0.016$] for the recognition scores in the first session. Figure 6a shows histograms for the recognition scores for each group for the first session.

Post-hoc testing using MWU showed statistically significant differences between the Control group and the Random group ($p = 0.050$, $r = 0.279$), and the Control group and the Speed group ($p = 0.044$, $r = 0.304$). No statistically significant differences were found between the Control group and the Multiple group ($p = 0.973$,$r = 0.004$). The Random and Speed policy prescription are having a negative effect on participant's abilities to recreate gestures, lowering the recognition scores relative to the Control group.

We can measure long-term memorability by testing the recognition scores in the second session. The group sizes



(a) First Session Recognition Scores



(b) Second Session Recognition Scores

Fig. 6: Histograms showing recognition scores for each policy. The recognition score can be used as an indicator of a participant's ability to remember their gesture, as higher recognition scores indicate that the participant has more closely replicated their templates. There are statistically significant differences between the Control group and the Random and Speed groups, demonstrating that users had higher difficulty in replicating gestures created under the Random and Speed policies.

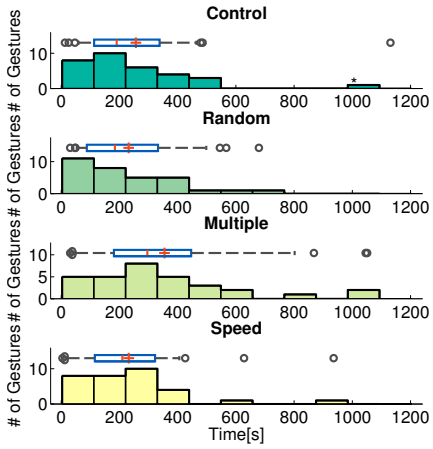


Fig. 7: The time it took participants to create a gesture. Most of the policy distributions are similar to the Control group, except the Multiple group has higher numbers of people taking more time to select a policy.

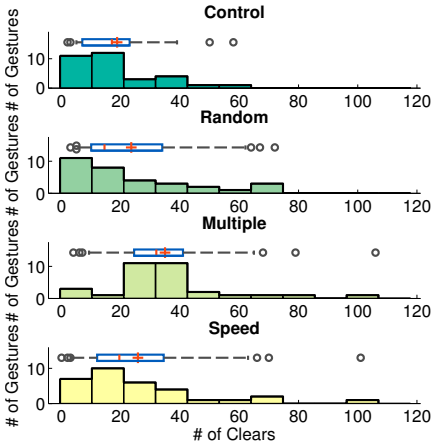


Fig. 8: Histograms of the times participants pressed 'Clear' during the first session. We are using this as a way of measuring the effort to create a gesture that fits a policy. Statistically significant differences exist between the Control and Multiple groups. The Multiple policy caused participants to iterate through many more attempts on average before selecting their final gesture.

decreased in the second session due to four participants dropping from the study. The sample sizes for the second session groups are: 28, 28, 27, and 29 for the Control, Random, Multiple, and Speed groups, respectively. These changes are small and the use of robust, non-parametric tests maintain the validity of the comparisons.

Long-Term Memorability: KW shows statistically significant differences between the four groups [$\chi^2(3,112) = 10.5$, $p = 0.015$] for the recognition scores in the second session. Post-hoc testing using MWU did not show significant differences between the Control and Random groups ($p = 0.096$, $r = 0.278$), the Control and Multiple groups($p = 0.847$,$r = 0.024$) and the Control and Speed groups ($p = 0.186$, $r = 0.2152$). The effect of the Random and Speed policy prescriptions appears similar to the effect in short-term memorability.

D. Effects on Usability

Creation Time: Figure 7 shows the time to create a gesture based on the policy document. This is an indicator of difficulty to meet the policy's expectations. We compared the time it took participants to create a gesture after they had been assigned a policy. KW test on the creation time indicates there is no statistically significant difference between the groups ($p = 0.1152$) with respect to creation time. Most of the distributions are similar, ranging from 0 to 400 seconds with occasional outliers at 1000 seconds. The Multiple group's distribution, however, shifted more towards longer creation times than all of the other groups, suggesting some difficulty with meeting policy demands.

Creation Effort: The number of times a participant pushed 'Clear' prior to finalizing their gesture password is taken as a measure of the effort a participant had to exert to satisfy the policy requirements. Figure 8 shows the histograms for the number of clears participants performed. KW showed a statistically significant effect of the policy on the clear rate for the four groups [$\chi^2(3,124) = 15.22, p = 0.001$]. Post-hoc comparisons using the MWU test indicated that there was statistically significant differences between the Multiple group and the Control group ($p < 0.001$, $r = 0.471$), and no statistically significant differences between the Control and Random group ($p = 0.510$, $r = 0.082$) or the Control and Speed group ($p = 0.490$, $r = 0.145$). Compared to a Control, participants under the Multiple policy iterated through many different attempts at a gesture before selecting their choice; this is evidenced in Figure 8 with some participants hitting 'Clear' nearly 40 times. The other groups showed no meaningful difference compared to the Control.

V. Discussion

A. User Choice

The composition policies successfully changed the way users picked their gestures, addressing the problems observed in gesture password studies. Each instruction caused a policy group to skew towards different categories. Examining the groups, it is clear why this happened. The Control group operated similar to the other gesture password studies, with a majority of users preferring some combination of Digits, Letters, and Shapes (56.26%). This is the only policy group that has a combination of these three categories exceeding half of the dataset. This is all evidenced in Table I.

The Random group produced more Abstract gestures (31.25%) than any other group. This aligns with our policy instructions to the group – being random and unpredictable. Interestingly, this reduced Sloping to the lowest of all the four policy groups (9.38%); this is because the participants focused on sudden, sharp changes instead of curving their created gestures. They also do not adhere to just using shape combinations at a frequency any different from the Control.

The Multiple group produced more Shape Combinations than any other group (40.63%) and the reason is linked to the instruction text: creating multiple gestures with different fingers. This policy did not try to move participants away from familiar shapes, with the effect being that participants relied on combinations of these simple gestures to satisfy policy requirements. Moreover, most participants (87.51%) obeyed the instructions and used more than two fingers, as seen in Table II. All other groups split in favor of single finger gestures.

The Speed group produced more Sloping gestures than any other group, but the percentage is underwhelming compared to the pluralities seen in other policy instructions. However, there is a reason for this: the Words category is often composed of script writing of words that could also be doubly classified as part of the Sloping category. When taken together, these two groups account for 46.88% of the entire Speed group. Moreover, the Speed group succeeded in its primary objective: a higher group mean of gesture velocities compared to the Control group. None of the other groups produced significant results on this axis.

B. Security

Figure 5 gives us an idea of how the policy instructions affected the security metrics: TPR and FPR. Generally, the TPR and FPR distributions for most groups were highly similar and testing did not give any indication that they were statistically different. TPR values less than one are a result of a small percent of users in each groups having difficulty recalling their gestures. We do note that one group – the Multiple group – has a TPR distributional spread that stands out more than the other three groups. TPR values are lower in this group overall, and we expect this to be a consequence of the policy. Drawing multiple things is difficult to replicate well, and the time gap of 10 days did not make it easier.

FPR values more than zero are an effect of the very strong attack scenario, where attackers could watch the video repeatedly and attack as many times as they want. The similarity in distributions is good news for the policies – the effect from observation attacks is not worse than the control. This means that policies do not appear to make groups more (or less) susceptible to observation attacks at the cost of improving password diversity. Some gesture choices by participants are adept at resisting observation attacks despite a video, needing as many as 80 attempts as observed in Figure 4. Of course, this speaks to the strength of gesture passwords overall as an authentication technique; there has not been a comparable published figure of a pattern unlock password requiring an aggregate of 80 attempts to crack when under observation. Throttling can be employed to slow down or mitigate repeated observation attacks. Restricting guesses to ten trials, as many smartphone platforms already do, would be very effective of stopping shoulder surfing attacks with free-form gestures.

C. Memorability

Compared to the Control group, the Random and Speed group participants were less accurate at recreating their gestures. This decrease in memorability is an unfortunate side-effect of the increased burden placed by the policy instruction, and the reasons lie in Table I. The Random and Speed groups prefer Abstract and Sloping gestures, respectively. Abstract gestures are difficult, non-recognizable shapes that are harder to reproduce. Sloping gestures are also hard to reproduce, it is similar to recreating a scribble hastily on a paper twice in a row. The Multiple group has no such problem, since the participants choose gestures that are combinations of simple shapes (40.63%), which are not difficult to enter. Multiple group participants enter their gestures more slowly than the Speed and Random groups, seen in Figure 3.

D. Usability

We found that the Multiple group underperforms on usability metrics: participants in this group take more trials when trying to create their gesture, as seen in Figure 8. This fits the instructions well: different gestures with different, multiple fingers. These types of inputs are more difficult to do, requiring more hand coordination, so a more deliberate input process is

not surprising. The time it took participants to create a gesture did not significantly differ between groups, as seen in Figure 7. The Multiple group distribution does skew more towards higher times when compared with the other groups. It is very likely we lacked the sample size to detect a statistically significant effect in this instance.

VI. Conclusions

We have presented the first study on composition policies for creating gesture passwords. Our findings indicate that all of the three simple policies successfully changed the way users generate gesture passwords, moving the emphasis from simple shapes, words, and numbers to different gesture categories with more complexity. In this way, the policies effectively address the critical limitation identified in previous studies suggesting that all users do not spontaneously adopt good strategies when creating gestures. Moreover, our data details *how* the instruction policy affects security, memorability, and usability. We observed that the policies have an effect multiple ways: user choices were shifted, the memorability declined for the Random and Speed groups, and the usability for the Multiple group appeared weaker.

Forcing users to follow our policies would have resulted in better outcomes given that there is evidence of some deviation from instructions (e.g. 12.5% of the Multiple group used a single finger). A follow-up study could examine enforcement of the composition policies. However, a more open-ended approach is better against guessing attacks, a specific set of rules invites attackers to adjust to the rules. Overall, we shown positive results about the usability, security, and memorability of gesture passwords while demonstrating how the important problem of user choice can be improved with simple, easily-implemented composition policies.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Numbers 1228777 and 1541069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Gradeigh D. Clark was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

References

[1] A. De Luca and J. Lindqvist, "Is secure and usable smartphone authentication asking too much?" *IEEE Computer*, vol. 48, no. 5, pp. 64–68, May 2015.

[2] J. Bonneau, C. Herley, P. C. v. Oorschot, and F. Stajano, "The quest to replace passwords: A framework for comparative evaluation of web authentication schemes," in *Proc. Oakland '12*.

[3] A. Oulasvirta, S. Tamminen, V. Roto, and J. Kuorelahti, "Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile hci," in *Proc. CHI '05*.

[4] S. Uellenbeck, M. Dürmuth, C. Wolf, and T. Holz, "Quantifying the security of graphical passwords: The case of android unlock patterns," in *Proc. CCS '13*.

[5] Z. Zhao, G.-J. Ahn, J.-J. Seo, and H. Hu, "On the security of picture gesture authentication," in *Proc. SEC'13*, Berkeley, CA, USA.

[6] SRLabs, "Spoofing fingerprints," Feb. 2015, https://srlabs.de/spoofing-fingerprints/.

[7] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song, "Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication," *Trans. Info. For. Sec.*, Jan. 2013.

[8] C. Bo, L. Zhang, X.-Y. Li, Q. Huang, and Y. Wang, "Silentsense: Silent user identification via touch and movement behavioral biometrics," in *Proc. MobiCom '13*.

[9] M. Sherman, G. Clark, Y. Yang, S. Sugrim, A. Modig, J. Lindqvist, A. Oulasvirta, and T. Roos, "User-generated free-form gestures for authentication: Security and memorability," in *Proc. MobiSys '14*.

[10] G. D. Clark and J. Lindqvist, "Engineering gesture-based authentication systems," *Pervasive Computing, IEEE*, vol. 14, no. 1, Jan 2015.

[11] N. Sae-Bae, K. Ahmed, K. Isbister, and N. Memon, "Biometric-rich gestures: A novel approach to authentication on multi-touch devices," in *Proc. CHI'12*.

[12] M. Shahzad, A. X. Liu, and A. Samuel, "Secure unlocking of mobile touch screen devices by simple gestures: You can see it but you can not do it," in *Proc. MobiCom '13*.

[13] J. Tian, C. Qu, W. Xu, and S. Wang, "Kinwrite: Handwriting-based authentication using kinect." in *Proc. NDSS '13*.

[14] C. Liu, G. D. Clark, and J. Lindqvist, "Where usability and security go hand-in-hand: Robust gesture-based authentication for mobile systems," in *Proc. CHI '17*.

[15] C. Liu, G. D. Clark and J. Lindqvist, "Guessing attacks on user-generated gesture passwords," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 1, pp. 3:1–3:24, Mar. 2017.

[16] N. Zheng, K. Bai, H. Huang, and H. Wang, "You are how you touch: User verification on smartphones via tapping behaviors," Tech. Rep., Oct 2014.

[17] U. Burgbacher and K. Hinrichs, "An implicit author verification system for text messages based on gesture typing biometrics," in *Proc. CHI'14*.

[18] Y. Yang, G. D. Clark, J. Lindqvist, and A. Oulasvirta, "Free-form gesture authentication in the wild," in *Proc. CHI'16*.

[19] A. Paivio, T. Rogers, and P. C. Smythe, "Why are pictures easier to recall than words?" in *Psychonomic Science*, ser. 11, 1968, pp. 137–138.

[20] U. Oh and L. Findlater, "The challenges and potential of end-user gesture customization," in *Proc. CHI'13*, 2013.

[21] C. Valdes, D. Eastman, C. Grote, S. Thatte, O. Shaer, A. Mazalek, B. Ullmer, and M. K. Konkel, "Exploring the design space of gestural interaction with active tokens through user-defined gestures," in *Proc. CHI'14*.

[22] J. Sun, R. Zhang, J. Zhang, and Y. Zhang, "Touchin: Sightless two-factor authentication on multi-touch mobile devices," in *Proc. CNS'14*.

[23] S. Fahl, M. Harbach, Y. Acar, and M. Smith, "On the ecological validity of a password study," in *Proc. SOUPS'13*.

[24] J. Ruiz, Y. Li, and E. Lank, "User-defined motion gestures for mobile interaction," in *Proc. CHI'11*.

[25] C. Kray, D. Nesbitt, J. Dawson, and M. Rohs, "User-defined gestures for connecting mobile phones, public displays, and tabletops," in *Proc. MobileHCI'10*.

[26] J. O. Wobbrock, M. R. Morris, and A. D. Wilson, "User-defined gestures for surface computing," in *Proc. CHI'09*.

[27] M. Weir, S. Aggarwal, M. Collins, and H. Stern, "Testing metrics for password creation policies by attacking large sets of revealed passwords," in *Proc. CCS'10*.

[28] Y. Meng and W. Li, "Evaluating the effect of user guidelines on creating click-draw based graphical passwords," in *Proc. RACS'12*.

[29] J. Thorpe, M. Al-Badawi, B. MacRae, and A. Salehi-Abari, "The presentation effect on graphical passwords," in *Proc. CHI'14*.

[30] S. Chiasson, A. Forget, R. Biddle, and P. C. van Oorschot, "Influencing users towards better passwords: Persuasive cued click-points," in *Proc. BCS-HCI '08*.

[31] Y. Li, "Protractor: A fast and accurate gesture recognizer," in *Proc. CHI '10*.