GestureMeter: Design and Evaluation of a Gesture Password Strength Meter

Eunyong Cheon
Ulsan National Institute of Science and Technology
Ulsan, Republic of Korea
beer@unist.ac.kr

Jun Ho Huh
Samsung Research
Seoul, Republic of Korea
junho.huh@samsung.com

Ian Oakley
Ulsan National Institute of Science and Technology
Ulsan, Republic of Korea
ian.r.oakley@gmail.com

ABSTRACT

Gestures drawn on touchscreens have been proposed as an authentication method to secure access to smartphones. They provide good usability and a theoretically large password space. However, recent work has demonstrated that users tend to select simple or similar gestures as their passwords, rendering them susceptible to dictionary based guessing attacks. To improve their security, this paper describes a novel gesture password strength meter that interactively provides security assessments and improvement suggestions based on a scoring algorithm that combines a probabilistic model, a gesture dictionary, and a set of novel stroke heuristics. We evaluate this system in both online and offline settings and show it supports creation of gestures that are significantly more resistant to guessing attacks (by up to 67%) while also maintaining performance on usability metrics such as recall success rate and time. We conclude that gesture password strength meters can help users select more secure gesture passwords.

CCS CONCEPTS

• **Security and privacy** → **Graphical / visual passwords**; • **Human-centered computing** → *HCI design and evaluation methods*.

KEYWORDS

Gesture password, Password composition policy, Large scale study

ACM Reference Format:
Eunyong Cheon, Jun Ho Huh, and Ian Oakley. 2023. GestureMeter: Design and Evaluation of a Gesture Password Strength Meter. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3544548.3581397

1 INTRODUCTION

Mobile smart devices are unlocked as many as 200 times per day [54]. To achieve this, many users rely on explicit authentication techniques such as PIN [40] or pattern [56]. To minimize the time and maximize the accuracy of their frequent unlock attempts, users commonly select passcodes that are simple to enter [56] and easy to remember [31]. However, such passcodes are also easy to guess [42] via techniques such as constructing a dictionary of commonly selected codes [8] or generating passcode predictions from a probabilistic model [39]. The consequences of such breaches can be high: a malicious user who gains access to a user's smart device can also likely access a wide range of private information, accounts and services [17]. To address this problem, researchers have proposed alternative unlock schemes. One scheme that is particularly well-suited to input on smart devices is that of gesture passwords [64]. These take the form of a series of freely composed strokes produced by a finger or thumb sketching directly on the touch screen on a device. Researchers have suggested gesture passwords offer a range of advantages over PIN and pattern: they provide a substantially larger theoretical space of possible passcodes (up to 27.72 bits [50]) and may reduce the amount of visual attention required during authentication [43], an attractive feature for mobile usage scenarios. In addition they require little extra time and effort to enter [64].

Despite these potential benefits, recent studies of gesture passwords have highlighted underlying issues. Specifically, as with other explicit knowledge based authentication schemes, a significant proportion of user gestures can be guessed in both online [14] and offline attack scenarios [36]. These results suggest that, while the total number of unique gesture passwords that can be generated is extremely large, in practice many users select passcodes from a common, easy-to-remember and easy-to-guess subset of this space, just as they do with other knowledge based authentication techniques. To address this problem, researchers have begun to explore how existing policies and techniques for helping users select secure and unique PINs [31], patterns [15] or passwords [60] can be adapted to gesture password systems. For example, Cheon et al. [14] recently proposed the use of blacklists of commonly selected gestures and showed this could increase resistance to online guessing attacks, while Clark et al.[16] explored the use of composition policies—such as explicit instructions to stroke rapidly or randomly—and suggest these successfully prompted users to create more diverse gestures. Similarly, in the related area of graphical passwords based on selecting a series of points or strokes on a pre-defined image, Raptis et al. [45] showed how gamification of this process, in the form of providing a points-based scoring system that rewarded more random selections, could increase the diversity of passwords users chose.

Based on this literature, we argue that password composition policies will be essential to guide users toward secure gesture password selections and fully realize the potential of the technique. However, the design of such tools and systems is not trivial—existing

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581397

approaches tend to be highly specific to the passcode modalities they are based on. For example, common policies for PINs involve barring repeated or sequential digits [31], for passwords mandate use of different character types [49] and for patterns require system-defined start points [15] to alleviate bias in initial target selections. While these techniques are established and well-proven, none can be directly applied to free-form gesture passwords that involve neither explicit symbols nor fixed input locations—gesture passwords are typically preproccessed in order to be both scale and position invariant [36] in order to cater to variations in self-reproduction of gestures and reduce false rejection rates. As such, we identify a need for further research to adapt, develop and evaluate password creation tools and techniques that are specific to gesture password systems.

This paper explores the design of one such tool: a gesture password strength meter. Password meters [59] are a frequently deployed [18] and well studied technique for improving the quality of passwords users generate. They work by evaluating the security of a user's password and providing feedback such as strength ratings [22] or recommendations [58] to increase security. They typically operate via sets of heuristics about the likely strength of passwords in terms of properties such as password length (longer being associated with increased security) and the number of different character types used (e.g., upper and lower case, numbers, etc.). In addition, they can integrate formal calculations of *guessability*, in the form of data-driven predictions of the probability of selecting a given password [12]. Finally, they typically provide feedback to users in the form of both overall gauges or strength meters [59] and in terms of customized, reactive recommendations and advice. These meters can be highly effective: in a large scale study of more than 4500 participants, Ur et al. [58] showed that a combination of heuristics, probabilistic analysis and detailed feedback about both overall security level and specific weaknesses and potential improvements, led to users generating passwords with a substantial and significantly reduced guessability [41]. However, while this literature compellingly illustrates the benefits of password strength meters, we know of no prior work exploring the design of such systems for gesture passwords.

We aim to address this omission and design, develop and evaluate a gesture password strength meter. To achieve this, we conduct an online study to collect 1000 gesture password samples. We apply state-of-the-art gesture password cracking algorithms for both online (dictionary based guessing) and offline (n-gram Markov probability) attacks to create an initial password strength assessment tool. We augment this by exploring how a wide range of gesture features and qualities influence measured strength, ultimately selecting a subset of the most impactful for integration in a final meter design: gesture curvature, gesture symmetry, and the use of multiple (or compound) elements. Our final meter integrates these assessments into a visually displayed strength bar, as textual descriptions of potential areas for improvement (e.g., a recommendation to use non-symmetrical forms) and as visually presented distortions of a user entered gesture password candidate (e.g., a version with reduced symmetry based on rotating a subset of the original strokes) that illustrate and exemplify how an existing user gesture could be modified to improve its security. An online study of this system (N=600) indicates that it can increase resistance to guessing attacks by 67% compared to a baseline and that, while setup times are prolonged, it exacts only modest costs in usability during short-term recall—equivalent recall rates and median recall times of 3.8s, elevated by a median of just 1.38s over baseline. We further evaluate the usability impacts of the meter by conducting a small-scale multi-session study that indicates that use of the meter did not significantly impact recall rates or recall times over periods of up to one week. In addition, data from quantitative usability and workload measures suggest our meter requires only modest additional effort, and a summary of user comments suggests participants appreciated its abilities to increase the security of the gesture passwords they generated. Overall, the contributions of this work are the design and development of the first password strength meter for gestures and the presentation of a thorough evaluation of its performance that documents the substantial security benefits and various usability costs associated with its use. We believe the development of policies and tools to increase the security of gesture passwords will be essential to popularizing this promising technique towards real world deployment. This paper makes firm steps towards that goal.

2 RELATED WORK

2.1 Gesture passwords

2.1.1 Gestures for Authentication. As advances in touchscreen devices have enabled high-resolution input over large areas, gesture passwords have been proposed as a new approach to secure private information. Early work in this area explored authentication with predefined gestures. The underlying idea was that user performance of specific gestures would vary sufficiently (and reliably) to accurately distinguish between users. This work predominantly explored the types and forms of gestures (e.g., multi-touch) that could lead to robust performance in this task [46] or emphasized usability by seeking to extract features from the simplest possible single strokes [19]. More recently, Sherman et al. [50] proposed the use of multiple stroke free-form gestures as a memorable and secure input method and explored their use in a lab study. In this model, users select or create unique gestures as their passwords. This idea has attracted considerable interest, and recent work has examined more diverse use scenarios, such as creating gestures for smartphone unlock [14] or creating and recalling gestures over multiple days and for different accounts [64]. In general, this body of work is motivated by the idea that gestures on touchscreens combine a theoretically large password space with a high level of usability: gestures are easy to remember, and entering them is both rapid and accurate. These motivations also inspire the work we report in this paper.

2.1.2 Gesture Recognition Systems. Matching a submitted gesture password against a stored gesture template to achieve user authentication is a more technically challenging process than the simple symbolic matching process used for technologies such as PIN and text password. Gesture passwords need be matched based on stroke similarity measures such as Dynamic Time Warping (DTW) [19, 46] or inverse cosine distance [35]. While both can be effective at assessing similarity in large sets of gesture passwords [14], DTW, in general, achieves better performance [14, 37]. While diverse other

recognizers have been proposed and discussed [37], adopting novel technical approaches impedes comparisons with prior studies. As such, the work in this paper implements gesture matching using the established distance metric of DTW.

2.1.3 Gesture Usability and Security. Researchers have analyzed both usability and security to determine the effectiveness of gesture passwords. Sherman et al. [50] examined the security and memorability of free-form multi-touch gestures, ultimately presenting a series of guidelines and observations. More concretely, Yang et al. [64] examine the usability and memorability of user-chosen gesture passwords after one hour, one day, and one week on multiple user accounts. Their results indicate that gesture passwords outperform traditional text passwords in terms of the key usability metrics of creation time (by 42%) and entry time (by 22%) while remaining highly memorable.

Security results are more mixed. Early work shows strong benefits. For example, Sahami et al. [47] perform shoulder-surfing attacks in which a malicious observer attacks a user's in-air gesture password, an experiment that resulted in no successful attacks. Similarly, Liu et al. [37] perform an automated brute-force attack using algorithmically generated gesture passwords that failed to crack gestures matched via DTW (and a variety of other distance metrics). More recent work casts doubts on these initial results. In particular, work on gesture dictionaries is proving particularly effective. For example, Liu et al. [36] describe an offline dictionary based guessing attack that cracks between 47.71% and 55.9% of gesture passwords with 10^9 guesses [36]. Similarly, Cheon et al. [14] study the effectiveness of using dictionaries to crack gesture passwords in an online attack scenario. They report that 54.18% to 58.37% of a large gesture password data set can be successfully guessed using dictionaries composed of representative and frequently selected gestures. These results suggest that, despite their potentially high levels of both usability and security, gesture passwords suffer from the same tendencies as other knowledge based authentication credentials. In order to facilitate memorability and ease entry processes, users choose gesture passwords that are easy to guess. The work in this paper seeks to address this emerging issue and explore how users can be supported in creating gesture passwords that are harder to guess.

2.2 Password Heuristics

Heuristics are valuable evaluation [49, 63] and feedback tools [24, 58] for diverse password modalities. The effectiveness of using simple rules-of-thumb about password length, the number of non-overlapping symbols and the use of different character sets has been repeatedly and robustly demonstrated [33]. For PIN, for example, Kim and Huh [31] evaluate the impact of composition rules such as restricting use of consecutive digits or mandating various PIN lengths on security. Similarly, a recent study by Ur et al. [58] examines multiple heuristics and combines these into a single score to create an overall metric for the strength of a password. In terms of pattern locks, Aviv and Fichter [4] explore users' preferences across six rules and determine pattern length to be the strongest indicator of pattern strength. In contrast, we know of no work looking at heuristics for gesture security. However, we note that such rules-of-thumb have been used to analyze gestures in other contexts, such as to support stroke based user interfaces [34]. We believe that using heuristics to evaluate the security of gesture passwords is a promising approach that can improve their security and present the first work to examine this issue.

2.3 Password Strength Meters

2.3.1 Password Composition Policies. Password composition policies are an important tool to help users avoid creating guessable passwords [62]. However, it is important to note that extremely stringent policies may provide additional burdens to users [49]. The trend has been observed in diverse password schemes. Komanduri et al. [32], for example, suggest that security improvements enabled by password policies are often correlated with decreases in usability. For example, although users can create hard-to-guess PINs under a stringent 6-digit composition policy, they find these PINs relatively difficult to remember compared to those created under less stringent policies [31]. Another example is that while system suggested random patterns show high entropy, key usability metrics such as recall success rate are reduced [15]. These effects can also occur when only ambiguous security benefits are achieved. Clark et al. [16], for example, propose policies that request users to create gesture passwords from strokes that are fast, random or use multiple fingers. Evaluations show unclear security improvements and negative effects on usability. More positively, a recent study on gesture passwords suggests improved security can be achieved by restricting users from selecting common gesture passwords, although this comes at the cost of reduced recall success rates [14]. We identify a need to carefully design policies, such as Cho et al.'s mandated pattern start points [15], in order to balance improvements in security against any costs to usability. This paper presents work to explore this design space for gesture passwords.

2.3.2 Strength estimation. Password strength meters are a widely used technique that assess the strength of candidate user generated passwords and provide informative aids to help users improve their selections. Ur et al. [59] examine various meter designs and conclude users choose longer passwords when a meter is presented compared to a system without such feedback. Egelman et al. [22] emphasize the effectiveness of password strength meters when users are asked to create passwords for important accounts. A key quality of passwords meters is that they accurately measure the strength of entered passwords [18]. This is a challenging task: numerous studies suggest that simple estimation techniques lead to poor quality assessments of password strength [12, 18, 62]. To improve matters, recent work by Ur et al. [58] uses an artificial neural network to score passwords on multiple heuristics and shows improved performance. Additionally, tools such as n-gram Markov models [12] and Probabilistic Context-Free Grammars (PCGF) [29, 39, 61] show excellent performance in estimating text password strength. Such techniques have also been shown to be effective tools for assessing the strength of graphical password systems, such as the Picture Gesture Authentication scheme introduced in Microsoft Windows 8 and studied by Zhao et al. [65]. In this system, which is based on users drawing one of three gestures (point, line or circle) over a preselected background image, probabilistic password guessing models were able to guess up to 48.8% of user generated passwords. Multiple techniques can also be combined: Galbally et al. [26] introduce

multimodal approaches to effectively score passwords. Similarly, Song et al. evaluate pattern password security by combining three different measures [53]. While work on assessing the security of passwords is substantial, we note there is very limited work in this area on gesture passwords, an omission this paper seeks to address.

2.3.3 Password Meter Feedback. Password meters traditionally present a visual score bar or rating. Textual recommendations are also important and prior work has reported on the positive impact of deploying these types of feedback together [59]. The quality of the displayed contents also matters. Furnell et al. [25] varied the level of feedback in a text password strength meter and empirically determined that more detailed feedback led to users generating passwords that were significantly more secure. Similarly, a recent study by Ur et al. [58] suggests that feedback systems that provide detailed, rule-based and actionable feedback to users are more effective. The work in this paper seeks to apply these perspectives to the design of feedback in a gesture password meter and provide users with specific, concrete recommendations for how to improve the security of the gesture passwords they generate.

3 GESTURE PASSWORD METER DESIGN

We followed a data-driven approach in order to develop an effective password strength meter. We first identified the threat model it seeks to protect users against. After establishing this, we conducted a large-scale online gesture password collection study and applied state of the art techniques to assess the security of this user-created gesture password set: the probability calculated via an n-gram Markov model and a clustering based dictionary match score. We then evaluated the gesture set in terms of a wide variety of gesture properties (e.g., curvature, symmetry). For each property, we split the gestures into two subsets and calculated the subset crack rates. We used these results to select stroke properties associated with more secure gesture passwords. We then designed a gesture strength meter which visually indicates the strength of gesture passwords and feedback that assists users in designing more secure gestures with both textual and graphical recommendations relating to the potentially problematic gesture properties we identified.

3.1 Threat model

The threat model in this work focuses on device lock. We assume an online guessing attack scenario [14, 15] in which attackers have gained access to a users' device, but do not have their lock code or any other personal information. Attackers also have a limited number of chances to unlock the device (e.g., the 20 attempts allowed with Android pattern lock [3, 51]). Reflecting these constraints, an attacker's goal is to guess a genuine user's gesture password within 20 attempts by sequentially submitting the set of the most commonly selected gesture passwords. In order to make reasonable guesses, we assume the attacker has access to a relevant leaked gesture password data set that can be used to identify the most common gesture passwords.

3.2 Large Scale Gesture Password Study

We conducted a large scale online data collection study ($n = 1000$). We consider a mobile device usage context where users create and recall gesture passwords in order to lock and unlock their phones. Reflecting this scenario, gesture input was constrained to single one-finger strokes on a small screen region similar to that used when entering a pattern. The ethical aspects of the study were approved by our university's institutional review board (IRB). Additionally, during online recruitment, participants were made aware that we were collecting gesture passwords in order to conduct research on their security.

3.2.1 Study design. The study was implemented as a mobile-only website. We selected DTW as a distance metric to match gestures as it has been recommended by prior authors [37] and followed prior work [14] in using a permissive distance threshold (a DTW distance of 18.52) that would offer a very low false rejection rate in order to capture a wide range of valid user gestures. One thousand study participants were recruited from Amazon Mechanical Turk (MTurk). After signing up for the study on MTurk, participants entered a link or scanned a QR code on their phones to access the study site. They then read instructions and were asked to provide informed consent. They were informed no identifiable information would be collected and that they were free to terminate the study at any time. They were also provided with contact details they could use if they had questions or concerns. Participants who opted to continue then completed basic demographics and the study began.

It followed the following steps. First, participants were asked to create and then confirm (re-enter) a gesture password on their phone. Instructions emphasized the need to create a secure and memorable gesture and, in order to reduce the likelihood that participants would simply reuse their existing credentials (e.g., the patterns used to unlock their phones), we emphasized that the entered gestures would be logged and analyzed in detail by the research team. If a participant's confirm gesture did not match their creation gesture, they were required to restart the creation process. They were also able to cancel and restart the creation process at any time by clicking a button. After successfully confirming a gesture password, participants practiced their selected gesture ten times. They then completed a simple tile matching memory game, an activity designed to provide distraction and ensure gesture passwords were not being simply being retained in working memory. The game featured six pairs of differently colored tiles arranged in a three by four grid on the phone screen. All tiles were by default upside down, with their colors hidden. Touching one tile flipped it to reveal its color; touching a second did the same and also matched the two flipped tiles against one another. If the tiles had matched colors, they remained flipped, otherwise, they turned back upside down. By remembering the locations of previously flipped tiles, participants could successfully match all six pairs and complete the game. The mean completion time for this game was 38s (SD: 25.81s) a figure similar to the 30 second distraction task used in a prior study of graphical passwords [23]. The final stage of the study involved participants recalling their gesture password one final time within a maximum of five attempts to do so. The study took a median of 115 seconds to complete and participants were compensated with 0.75 USD, corresponding to an median hourly rate of 23 USD.

3.2.2 Participants. The majority of participants reported they were white (45%), Asian (37.1%), Hispanic (6.9%) or black / African American (6.7%). They were aged between 18-24 (17.3%), 25-34 (49%),

Table 1: Usability data from large scale gesture password study in terms of means (μ), standard deviations (σ) and medians ($\bar{\mu}$).

Setup Cancels (#)			Match Failures (#)			Setup Time (s)			Recall Time (s)			Recall Attempts (#)		
μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
0.32	1.09	0.00	0.17	0.65	0.00	24.97	20.89	18.27	3.68	4.86	2.47	0.15	0.75	0.00

35-44 (23.3%) and 45 or older (10.4%). Education levels were predominantly college (57%), post-graduate (22.9%) or high school (19.1%). Participants worked in highly diverse fields including computers and mathematics (13.2%), business and financial (10%) and management (8.8%).

3.2.3 Usability Metrics. We logged usability data from both setup and recall phases. In setup, we recorded the number of intentional *setup cancels*, the number of *match failures* between setup and confirm gestures and, the *setup time*, the overall time to create a gesture, including successfully confirming it. We note data from the setup measures closely follow results reported in prior work [14]. Specifically, setup cancels and match failures are infrequent, and the median setup time observed in this study was 18.27s, a figure broadly comparable to the 17.26s that has been previously reported. During recall, we logged the recall rate, or proportion of participants who successfully entered their gesture password within five attempts: 98%. This figure again closely matches prior studies (between 98% [14] and 98.9% [64]). In addition, we logged the number of *recall attempts* required to achieve a successful match, and the *recall time*, the total time taken for this process. The median recall time of 2.47s was again broadly similar to the 2.13s observed in prior work [14]. The full set of usability data is shown Table 1.

3.2.4 Security Metrics. We characterized the False Rejection Rate (FRR), or the proportion of genuine authentication attempts that are inappropriately rejected, by matching each participant's setup gesture against their confirm and recall gestures against a wide range of DTW threshold values. We then characterized the False Acceptance Rate (FAR), or the number of imposter authentication attempts that are inappropriately accepted, by matching each participant's setup gesture against the setup gestures generated by all other participants. The Equal Error Rate (EER), was then determined as the threshold value at which FRR and FAR are equal: 2.99% at a DTW distance threshold of 16.7 for our data. These figures are closely aligned with those in prior work (e.g., 2.28% EER at a DTW distance threshold of 16.24 [14]) suggesting our data set is representative of prior examples that appear in the literature.

3.3 Stroke Features

Comprehensible heuristics that are easy to assess, such as the use of multiple character types (e.g., lower and upper case), are widely deployed in password strength meters to both measure the strength of user chosen passwords [49] and to provide strategic feedback that may improve password selection [58]. In order to apply these same techniques to gesture passwords, we first need to understand which stroke features are associated with stronger passwords. To do this, we defined the following set of stroke features candidates based on a review of gesture features studied in prior work [34], gesture password classifications presented by prior authors [36][14] and a close examination of the gesture examples in our own data set. The full set of features considered are summarized in Table 2 and described in detail below. Additionally, Figure 1 shows two representative gestures for each feature illustrating how each is, in practice, expressed.

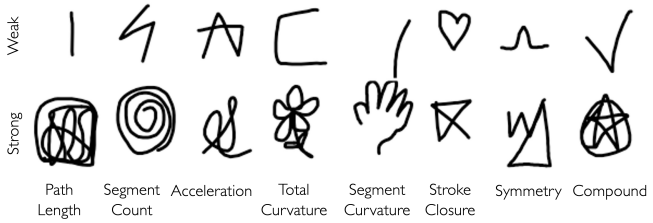

Figure 1: Representative samples of gestures showing the variability captured by the eight gesture features studied in this work.

Gesture length. Password length is one of the most common password strength estimation techniques—longer passwords are, in general, harder to guess or crack. It is likely the same holds true for gesture passwords. We break down length into two sub-features: the *path length*, or sum of distances between all points in a gesture and the *segment count* in a gesture after applying Douglas-Peucker (DP) line simplification [21].

Acceleration. Prior work has suggested that requesting users draw gestures rapidly can increase the diversity of their proposals [16]. In addition, gestures with high acceleration may be practiced and fluent, properties that may be associated with execution of more unique and complex stroke sequences. We calculate the mean acceleration for each gesture from the positions and timestamps recorded for each of its touch points.

Curvature. Gestures composed with a greater proportion of curved strokes, rather than straight lines, may be more unique and harder to replicate. We calculate the mean curvature of gestures from the angles between sequential pairs of points [38]. We define two-features for curvature. *Total curvature* is the mean angle between all pairs of points; *segment curvature* is calculated in the same way, but after first applying DP line simplification.

Stroke Closure. Closed shapes, defined as strokes in which the start and end points are proximate, may serve as frequent inspiration for gesture passwords. Indeed, prior literature suggests that geometric shapes, a category including various closed forms such as circles, squares and triangles, are used in up to 44% of gesture passwords [36]. Such high usage frequency may facilitate guessing. Accordingly, we calculated *stroke closure* simply as the distance between the start and end point of a gesture password [6].

Symmetry. Prior work has observed that symmetry can be exploited in guessing attacks on gesture passwords [36]. As many shapes may be drawn off-axis, we opted to assess gesture symmetry by comparing the first half of a gesture to the second half. The

Table 2: Stroke features examined in this work in terms of how they split the gesture set captured in the first study into two groups. For each feature, we report the size of each group, the proportion of gestures cracked in each group and whether or not these crack rates differed. In addition, the rightmost column indicates how different features were clustered together to retain a final set of features. Rows in bold indicate features that were retained in the final meter design.

	Weak Subset		Strong Subset		Chi-squared	Feature
Stroke Feature	Crack Rate (%)	Cluster Size (#)	Crack Rate (%)	Cluster Size (#)	p-value	Cluster
Path length [1]	48.29%	642	20.95%	296	p < 0.001	1
Segment Count [6]	47.25%	690	18.75%	240	p < 0.001	1
Acceleration [55]	41.80%	244	37.96%	627	p = 0.333	NA
Total Curvature	**48.08%**	**599**	**25.31%**	**320**	**p < 0.001**	**1**
Segment Curvature [38]	49.13%	574	25.97%	335	p < 0.001	1
Stroke Closure [6]	34.71%	363	40%	615	p = 0.115	NA
Symmetry	**50.08%**	**597**	**19.21%**	**380**	**p < 0.001**	**2**
Compound	**50.57%**	**522**	**23.68%**	**456**	**p < 0.001**	**3**

Figure 2: Dictionary gestures selected by applying AP clustering on DTW distances between all gestures in our data set and selecting the central examples from the top 20 clusters.

intuition here is that a symmetrical gesture will be divided into two matching portions rather than by aligned along any given spatial axis. To deal with minor differences in the scale of otherwise symmetrical half-gestures, we extend this analysis and consider gestures divided at 40%, 50% and 60% of their length. For each division point we create two sub-gestures by simply splitting the original gesture. In addition, we capture different axes of symmetry by considering both forward and reversed stroke order for one of the sub-gestures [2]. In total, we evaluate gesture symmetry by creating 12 different pairs of sub-gestures (three division points by two symmetric axes by two drawing orders). We assess the similarity of each of these pairs via DTW matching (after rendering each sub-gesture scale, location, and rotation invariant) and retain the lowest DTW distance as the symmetry score.

Compound. We define compound gesture passwords as those that contain two or more distinct forms—for example, one shape drawn within or adjacent to another. Prior work has suggested that gesture passwords involving compound forms are more challenging to crack [14]. In order to detect the presence of compound forms, we examined our data set in detail. We noted that many compound gestures are divided close to the mid-point, with at least one of the sub-gestures taking the form of a closed shape, such as a circle, square, or triangle, that embellishes the other. To detect this pattern, we again split each gesture at the 40%, 50%, and 60% points, retaining the division point achieving the lowest *stroke closure* score on one of its sub-gestures. Based on the intuition that a compound gesture must be composed of two non-trivial sub-gestures, we then calculated the *segment count* for each sub-gesture as a surrogate for complexity. Finally, we assessed the *symmetry* of each sub-gesture, reflecting the idea that genuinely compound gestures will show low symmetry. To combine these three measurements into a single metric, we normalized each over the whole set of gestures, inverted closure and symmetry, then summed them.

3.3.1 Stroke Feature Evaluation. We conducted a multi-stage process in order to evaluate whether the features we study effectively discriminate between stronger and weaker gesture passwords. We first created a dictionary of commonly selected gestures from the entire data set. This approach has previously been used in a highly successful online attack on gesture passwords [14]. Secondly, for each feature, we divided the gesture set into two groups by applying k-means clustering (with k=2) to the full set of feature scores. This creates what we consider to be strong and weak gesture subsets [36], specified in terms of each of our gesture features. Finally, we applied the dictionary in an online attack scenario to each group and compared the crack rates achieved. If the two groups show markedly different crack rates, we interpret this to indicate the feature is a salient metric for determining the strength of a gesture password. We describe the different stages in this process in more detail in the sections below.

Dictionary creation: We follow prior work [14] and create a dictionary for online attack in which attackers are assumed to have a fixed cap on the guesses they make (e.g., 20 [3]). We do this by calculating the DTW distances between all gestures in our data set and applying Affinity Propagation (AP) clustering to this data. This creates a set of spatially coherent gesture clusters. We retain the largest 20 clusters, and select the most central gesture in each cluster as a representative example. These twenty gestures, shown in Figure 2, form our online attack dictionary.

Guessing attack: We first calculate and normalize all stroke features described in Section 3.3 for each gesture. For each feature we then create two subsets of scores by first excluding the dictionary gestures and removing outliers (+/- 1.5*IQR). We then apply k-means clustering (with $k = 2$) to determine a threshold for dividing the gestures into weak and strong subsets. We then apply our dictionary to each subset and record the proportion of gestures cracked. For this process we select a DTW threshold value

of 11.28, corresponding to a FRR of 10%, and representing a reasonably strict level of performance that has been previously studied in prior work [14]. We apply non-parametric significance testing (Chi-squared test of independence) to determine if there is a difference between the two subsets. We use an alpha threshold of 0.00625, equivalent to applying Bonferroni corrections over our set of eight tests. The results are summarized in Table 2 and indicate the majority of features (all bar acceleration and stroke closure) lead to subsets which vary significantly in crack rate. We interpret this to mean they can serve effectively as heuristics to assess the security of gesture passwords.

3.3.2 Feature Clustering. Given the relatively high number of features that show good ability to distinguish between more and less secure gesture subsets, we opted to explore the extent to which they were redundant, or tended to assess the same underlying properties. We achieved this by normalizing all feature scores, then cross-correlating the results. We then apply AP clustering to the correlation matrix and retain the most central features in each cluster. This process resulted in merging all length and curvature features into a single cluster best represented by *total curvature* and retaining *symmetry* and *compound* as unique features. Cluster numbers and retained features (marked in bold) are shown in Table 2.

3.4 Gesture Password Score

Password meters typically display a score or indicator of password strength. To calculate an equivalent metric for gesture passwords we combine three assessments: 1) n-gram Markov probability, 2) dictionary match score, and 3) stroke feature score. We use an n-gram Markov model due to the technique's well established use as a metric for assessing password strength [12]. We follow a process introduced in prior work to adapt this technique to gesture passwords [14]; we refer readers to this prior work for a full description of the technique. In brief, it entails first simplifying and discretizing each user-chosen gesture into all permutations of a fixed number of length (2 to 4) and angle (8 to 12) segments. These tokenized representations are then used to train n-gram Markov models. We use a bi-gram model as the size of our data set is insufficient to support higher order models. We also optimize the models by evaluating edge-case handling and add-1 smoothing. A final model is selected based on balancing three metrics: performance in a 5-fold guessing attack with top 20 probable gestures generated from the model; mean DTW distance between each user-chosen and n-gram represented gesture in the model and; the proportion of observed n-gram cases. We ultimately selected an optimal n-gram model that discretized gestures into three length and ten angle segments and used 8% edge-case handling and Laplace add-1 smoothing. We note these parameters are similar to those selected in the study that introduced this method [14]. With this model, we then calculate a probability score for each gesture. After completing this process, we normalize the probability scores based on those calculated for the full set of gestures.

Comparing user-chosen gestures with dictionary items may prevent users from selecting common gesture passwords. We create an extensive gesture dictionary by extracting a large set of representative gestures from our gesture password set. Specifically, we apply AP clustering on the match scores between all gestures. This generated 141 clusters. We choose the center of each cluster as a dictionary gesture. To derive a dictionary match score, we match a user gesture password against all items in the dictionary and select the maximum similarity value. As with the probability score, we normalized the dictionary match score with respect to the maximum and minimum dictionary match scores observed in our data. We then calculate a stroke feature score by calculating the three selected features described in Section 3.3: total curvature, symmetry and compound. We normalize each metric using distributions from the full gesture set, then take the mean to generate a final normalized stroke feature score. Finally, to create a overall score for each gesture password, we assign equal weights to each of these three metrics to create a combined total (ranging from 0-3). This score assesses a gesture in terms of its susceptibility to online (dictionary match score) and offline (n-gram Markov probability) security vulnerabilities as well as in terms of the salient features and properties it exhibits (stroke feature score).

3.5 Gesture Strength Meter Design

We designed a gesture strength meter based on three design elements: a score bar; textual advice on how to improve an entered gesture and; interactive graphical recommendations for possible changes. We note our meter did not present (or prime) users with novel gestures at any time. Rather it offered various recommendations for how users could improve the gesture candidates they themselves generated. The final meter is illustrated in Figure 3 and we describe the design elements in detail below.

3.5.1 Score Bar and Text. We provide basic feedback in the form of a visual bar: the bar fills (by quarters) and changes its color as the rated security of a gesture password increases. We presented four security levels: very weak (red), weak (orange), fair (yellow), and strong (green), each corresponding to one quartile of the gesture password score [11]. In addition, following recommendations in prior work [59], we provided text feedback explicitly highlighting these classifications (e.g., "Your password is fair").

3.5.2 Minimum Strength Requirement. Password strength meters commonly require users to create strong passwords which meet predefined criteria, such as being assessed with rankings of at least "medium" or "fair" [18]. Furthermore, meters studied in a research context typically mandate passwords meet minimum length limits or bar passcodes on a blacklist [58]. Following this practice we required that gesture passwords achieve a rating of "fair" or better (corresponding to the mid-point of our score scale) in order to progress. This choice also reflects prior work on studying gesture passwords with crowd workers. Cheon et al. [14], for example, propose several gesture password policies, but found only those that mandated compliance (such as a blacklist) were effective. We note than although we require proposed gesture passwords meet a minimum score, participants are not required to use meter functionality to achieve this goal; entirely user generated gestures that meet score requirements were also accepted.

3.5.3 Recommendations and Feedback. Password strength meters are more effective when they provide interactive and responsive feedback that can help users improve their password selections [58].

Figure 3: Screenshots showing the gesture strength meter on a mobile device. Left shows a user's entered gesture, an overall strength rating (orange bar and text) and three specific recommendations for improvement. In the next two shots, the user has tapped the third and the first recommendations. The system proposed gesture revision is shown on the gesture entry box and the meter has been updated to reflect its strength rating. In the right shot the user has entered (and matched) the recommended gesture and is ready to select this as their final gesture password.

We sought to deploy similar techniques by supporting users in adapting their gestures to achieve improved performance in terms of the three gesture features associated with improved security identified in Section 3.3.1: total curvature, symmetry and compound. For gestures that score poorly on each feature we display a textual recommendation for improvement, a specific recommended gesture improvement and a security evaluation (simply re-running our assessment procedures) of the improved gesture. As such, a user entered gesture will receive between zero and three specific recommendations for improvement that depend on a detailed assessment which of its features are potentially problematic. This overall interface can be seen in Figure 3 and three representative recommendations for a single user-generated gesture can be seen in Figure 4. We provide a detailed description of how we generated this feedback below.

For gestures assessed to be weak in terms of total curvature (i.e., those composed predominantly of straight lines), we display an explicit recommendation to "add more curved lines in your gesture". We provide an improved gesture by applying DP line simplification, selecting a large straight stroke and replacing this with a quadratic Bézier curve with a randomly chosen intermediate control point. This replaces a straight stroke segment with a randomly curved one, while leaving the rest of the gesture unchanged. For gestures that score highly on symmetry, we recommend users "avoid using [a] symmetric gesture". We propose an improved gesture by rotating part of the original. We randomly select a large portion of the gesture (between 40% and 60%) and apply a random rotation (between 45° and 60° in either direction). Finally, for those gestures scoring weakly on our compound feature, we recommend users "combine more elements in [their] gesture". We present an improved gesture by randomly scaling, rotating and appending a commonly used gesture (from the gesture dictionary defined in Section 3.3.1) to the start or end of the user's original gesture. For each of these recommendations, we presented an associated "Click to Apply" button. Selecting this updates the gesture input region

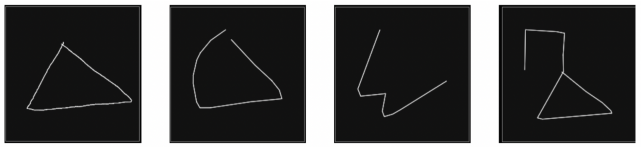

Figure 4: Gestures suggestions in the password meter. An original user-chosen gesture (left) is improved in terms of curvature (left-center), to avoid symmetry (right-center) and by integration of an additional forms to create a compound gesture (right).

with the recommended gesture and provides updated feedback (bar, text and recommendations) related to the new gesture. Users can then further revise their gestures by integrating a new set of recommendations and suggestions, return to their previous gesture via a "back" button or confirm the currently displayed gesture as their final password by tracing over it.

4 METER STUDY

We conducted an online study to evaluate the performance of our gesture strength meter. We capture and contrast the security and usability of user generated gesture passwords in a between groups design comparing a baseline condition against one involving our gesture strength meter in a large sample (n = 600). The study was approved by the local IRB and participants were aware that we were collecting gesture passwords in order to conduct research on their security.

4.1 Study design

We recruited participants from MTurk following general procedures in the first study. In addition, we logged user's IP address and used

Table 3: Usability data from the meter study in terms of means (μ), standard deviations (σ) and medians ($\bar{\mu}$).

	Setup Cancels (#)			Match Failures (#)			Setup Time (s)			Recall Time (s)			Recall Attempts (#)		
	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
Baseline	0.48	1.98	0.00	0.10	0.42	0.00	23.16	46.12	14.26	3.67	4.99	2.42	0.10	0.55	0.00
Meter	1.58	2.78	1.00	0.25	1.01	0.00	113.19	127.24	78.32	6.75	14.85	3.80	0.23	0.83	0.00

this to screen our data and prevent repeat participation [30]. Participants who signed the consent and opted to continue with the study went through broadly similar steps of completing a demographic survey then creating, confirming, practicing and recalling their gesture passwords. The study took a median of 180 seconds to complete and participants were compensated with 1 USD, corresponding to an median hourly rate of 20 USD.

4.2 Participants and Measures

Six hundred US based MTurk participated in this study: 300 in each condition. The majority of participants were white (75.5%), black / African American (9.5%), Hispanic (7.5%) or Asian (5.3%). Most were educated to college (55%), post-graduate (32.3%) or high school level (11.5%). Participants worked in diverse fields such as management (19.2%), computers and mathematics (16.2%) and business and financial operations (12.2%). They were aged between 18-24 (8.7%), 25-34 (58.5%), 35-44 (20.5%) and 45 or older (12.3%) and 95% of them were right-handed. Similar to the first study, we recorded all raw gestures, and details on the setup (number of setup cancels and setup time) and recall (recall rate, recall attempts and recall time) processes from each participant. We use the baseline data to check for equivalency with our first study and also compare results between the baseline and password strength meter conditions.

4.3 Usability Results

Usability measures are summarized in Table 3. We note that baseline results closely match those recorded in the first study, suggesting the two conditions were, in practice, equivalent. All measures show positive skews and failed normality checks, so we conduct non-parametric Wilcoxon rank sum tests to compare between baseline and password strength meter. The results for setup measures indicate that both setup cancels and setup time are significantly elevated with the password strength meter (both at $p < 0.01$). Setup time is particularly high. However, we note that substantial increases in setup time are commonly reported for password strength meters [44]. In Ur's [57] substantial study of textual passwords, for example, use of a meter triples password setup time from 19.9 seconds to 59.8 seconds. Similarly, alternative policies previously proposed for gesture passwords, such as Cheon et al.'s [14] blacklists led to similar performance hits (raising setup times from from 29.48 seconds to as high as 82.69 seconds). So while elevated setup times with our meter are not surprising, we do note their extent (means greater than four times baseline) suggest there may be scope for improving the usability of our meter design. To explore potential causes for these substantial increases, we also examined engagement with the meter's recommendation functionality, noting that 46.33% of users generated recommended gestures (each a mean of 3.53 times). The process of creating and viewing these recommended gestures likely took additional time and may also have resulted in some gestures that were ultimately deemed to be unsuitable and, thus, contributed to the increased the number of setup cancels. During recall, only recall time significantly differed between conditions, with meter requiring modestly prolonged time to enter ($p < 0.01$). Short-term recall rate remained high (at 99.3% for baseline and 98% for meter) throughout; a Chi-squared test of independence did not show a significant difference ($\chi^2 = 1.14$, $p = 0.29$) in this measure.

4.4 Security Results

4.4.1 EER. We calculated EERs for each condition following procedures in the first study. In the baseline condition, we recorded an EER of 2.8% at a DTW distance threshold of 16.42. While for meter condition, the EER was 1.82% at a distance threshold of 18.77. We note that the baseline EER closely matches that attained in the first study (2.99%) while that achieved in the meter condition is notably reduced. Additionally, this reduction is achieved at a higher DTW distance threshold (18.77 vs 16.42), further reinforcing the idea that that there is increased diversity of generated gesture passwords in the meter condition.

4.4.2 Guessing Entropy. To extend our security analysis, we evaluate the gesture sets from both conditions, and data from the first study, in terms of their entropy. As the gesture passwords generated with our meter may exhibit different probability distributions to those captured in our first study, we first optimize 2-gram Markov model parameters for data from this condition using the processes outlined in Section 3.4. The results were similar: discretization into three length and ten angle segments was optimal. As such, we retained model parameters from our first study and trained two new 2-gram Markov models with the data from our baseline and meter conditions. We then calculated the probabilities of all possible gesture passwords generated in these models in descending order and use this data to derive partial guessing entropy [7]: the number of guesses needed to crack different fractions of a password data set expressed in terms of bits of information. This data is shown in Figure 5 (left). These results indicate that the meter condition outperforms the baseline and original data set at all alpha levels (referring to the portion of the passwords that can be cracked). This effect is particularly prominent at lower alpha levels (< 0.4), suggesting that the meter condition includes fewer gestures that are easily cracked—or to use the term proposed by Liu et al. [36], a smaller *weak subspace* of gesture passwords. Additionally, at higher alpha levels the baseline condition outperforms the original study. This may be due to the smaller number of participants. To further quantify the differences between our conditions, we calculated the probabilities of all gestures from the first and second studies using the respective 2-gram Markov models, and generated kernel density estimations of this data. The results are shown in Figure 6, and clearly suggest that a greater proportion of meter gestures exhibit lower probabilities than baseline gestures from both studies. To

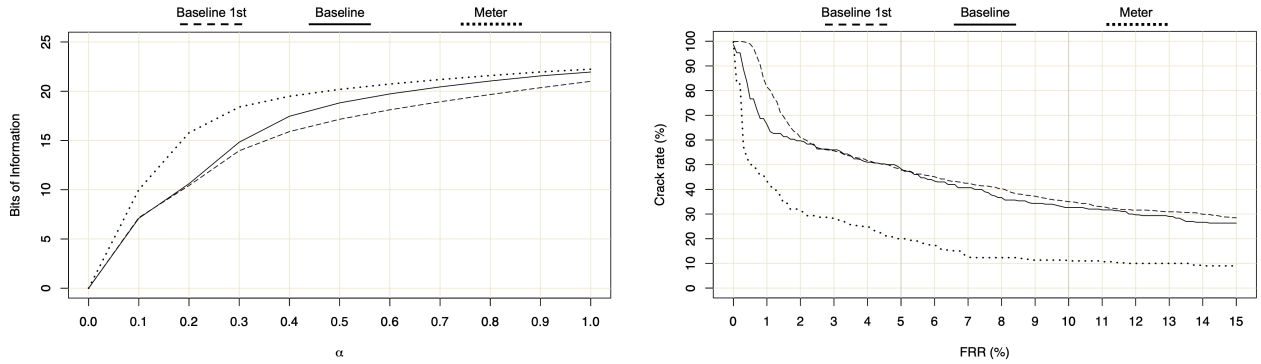

Figure 5: Left chart shows partial guessing entropy, expressed as bits of information, to crack portions of the gesture password data sets. Right chart shows online attack crack rates over FRR thresholds from 0 to 15%. Bold vertical grid-lines highlight performance for FRRs of 5% and 10%. Both charts include data from both online studies.

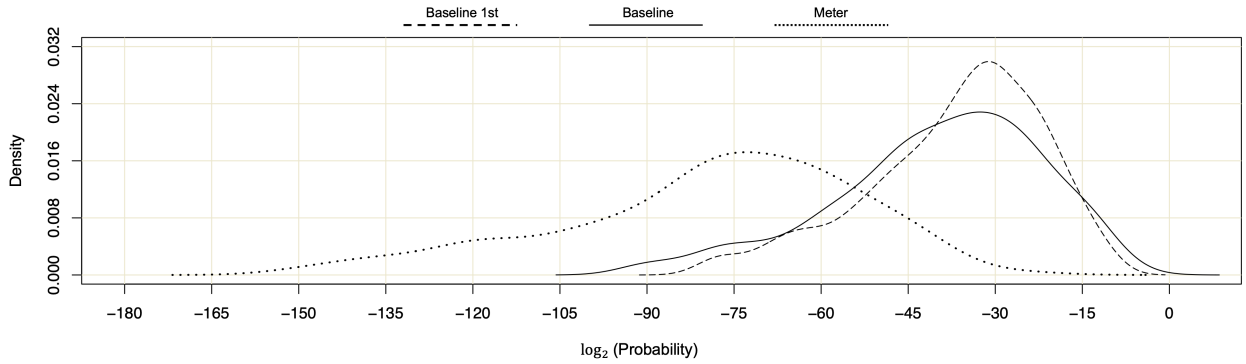

Figure 6: Kernel density estimations of the probability distributions of gestures from both online studies.

determine if these differences are statistically significant, we ran a two-sample Kolmogorov-Smirnov (K-S) test on the baseline and meter data from the second study: this assesses whether or not they are drawn from the same overarching distribution. The K-S test result ($D = 0.62333$, $p < 0.01$) indicates that they are not—that the probability distributions of gestures in our baseline and meter conditions differ significantly. This result reinforces our conclusions from the entropy data—meter gestures were harder to guess than baseline gestures.

4.4.3 Guessing attack. We followed procedures from the first study to generate new gesture dictionaries specific to each condition—see Figure 7. We then used these to conduct an online dictionary based guessing attack (see Section 3.3.1) on the gestures generated in each condition. When presenting and discussing results from this analysis, we also include results using this attack on the original study to provide context and shed light on whether the reduced sample sizes in the meter study influence the results. Full data for a wide range of FRR thresholds is shown in Figure 5 (right). To examine these data statistically, we selected two DTW thresholds, one corresponding to a fairly lenient threshold (5% FRR) that would accept a relatively high proportion of genuine user authentication attempts, and the other to a more strict threshold that would reject genuine users more frequently (10% FRR). With these threshold values, crack rates for our first study are, respectively, 47.8% and 35%. They are similar for the baseline condition in our second study—48.33% and 32.67%—but notably reduced in our meter condition—20% and 11%. We used four Bonferroni-corrected Chi-squared tests of independence to compare results from the second study baseline condition against data from both the first study and the meter condition. We record no significant differences in crack rates between the baseline and first study data ($\chi^2 = 0.01$ to 0.46, $p > 0.1$), suggesting the reduced sample size in the second study has limited impact on the validity of the results. However, the meter condition offered statistically improved resistance to guessing compared to baseline at both FRR thresholds ($\chi^2 = 40.01$ to 52.27, $p < 0.001$). The improvements are also relatively large scale—from 58.62% to 67.03% lower than baseline—suggesting that the meter enabled participants to generate substantially more secure gestures than the baseline condition.

Figure 7: Dictionary gestures selected by applying AP clustering on DTW distances between all gestures in the baseline (top) and meter (bottom) conditions in our second study. Selected gestures are the central examples from the top 20 clusters from each data set.

5 MULTI-SESSION STUDY

We conducted a multi-session lab study to complement the results from our online study; it sought to capture aspects of performance beyond the scope of the online study. Specifically, we designed this study to capture gesture password recall performance after between one and seven days. This data will shed light on the impact of our meter on memorability over more protracted periods. In addition, we report on both quantitative data (in terms of usability and workload) and qualitative comments (by summarizing user feedback from interview sessions) to improve our understanding of participants' experiences with our meter.

5.1 Participants

We recruited 20 users (ten males, ten female) for this study from the local student body: 15 were undergraduates and five were graduate students. They majored in engineering (17), management (1), physics (1) and design (1). They were between 19 to 29 years old (mean: 22.4, SD: 3.1) and all right-handed. All were Asian. They reported using various (and in some cases multiple) methods to secure their smartphones: fingerprint (10), four digit PIN (8), Face ID (6) and pattern lock (6). They were, in general, heavy and experienced smartphones users: they self-rated their familiarity with smartphones to be high (4.4/5.0) and, while one declined to answer, the remainder reported unlocking and using their smartphones very frequently: between 50 or fewer times (4) a day through 51-100 times (7), 101-150 times (6) and more than 151 times (2).

5.2 Study Design

The study followed a between groups design: ten participants used the meter and ten participants used the baseline gesture password system. Participants came to the lab to complete the initial enrolment session. However, to ensure results are comparable with our online studies, we used the same experimental platform and had participants use their own mobile device to access the gesture password system. Enrolment followed processes in our online studies. Participants were first asked to make gesture passwords that were both secure and memorable. They then created and confirmed a gesture password, practiced it ten times, completed a memory game and then an immediate recall session. After this session, participants completed the NASA Task Load Index (TLX) [28] and System Usability Scale (SUS) [10] questionnaires to assess their experiences during enrolment. Additionally, we conducted an in-person semi-structured interview to capture qualitative aspects of their experiences. We then conducted three recall sessions: after one day, two days and seven days. Each recall session involved sending participants links to the study which they then completed on their own device and at their convenience. They did not return to the lab for these sessions. The recall sessions were structured identically to the immediate recall session that took place directly after enrolment. If a participant failed to recall their password in any session (after five attempts) their participation in the study was terminated. After the final recall session, participants again completed the NASA TLX and SUS questionnaires to assess their experiences during recall. Participants were compensated with 3.6 USD (in local currency) for each session in this study (max of 14.4 USD for all sessions).

5.3 Security Results

We examined crack rates for the gestures generated in this study. To calculate these we used the gesture dictionaries and processes from the meter study and report crack rates at the key FRR thresholds of 5% and 10%. For baseline these are 30% at both thresholds. For meter, they are, respectively, 20% and 10%. While sample sizes in this study are too small to support statistical analysis, we note these figures are broadly aligned with those from the online meter study (see Figure 5) and support the claim that our meter helps users to create more secure gestures. In addition, baseline may be performing modestly better than in our meter study, a trend also observed with other password schemes: credentials generated in offline studies [13] tend to be more secure than those generated in otherwise similar online studies [15].

5.4 Usability Results

Quantitative usability results are summarized in Table 4 for the enrolment session and Table 5 for the multi-day recall sessions. One participant in the baseline condition dropped out of the final (7-day) recall session. Beyond that, we note that all participants successfully completed all recall sessions (indeed, there was only a single failed authentication attempt across the whole study). All measures failed normality checks due to positive skews, so we used non-parametric Wilcoxon rank sum tests to examine differences between baseline and meter. The only measures in which significant differences occur are setup time and setup cancels: both are increased in the meter condition (at $p < 0.01$). These differences mirror the two most prominent variations in the online meter study and support our prior conclusions that it takes longer for participants to create gestures in the meter condition, at least in part because they cancel or otherwise revise their gestures more frequently. The results

Table 4: Usability results from enrolment session and initial (immediate) recall in the multi-session study. Data reported in terms of mean (μ), standard deviation (σ) and median ($\tilde{\mu}$).

	Setup Cancels (#)			Match Failures (#)			Setup Time (s)			Recall Time (s)			Recall Attempts (#)		
	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$
Baseline	0.00	0.00	0.00	0.00	0.00	0.00	29.06	20.51	24.91	3.36	1.55	3.93	0.00	0.00	0.00
Meter	2.90	2.51	2.50	0.10	0.32	0.00	130.50	54.51	139.91	3.81	0.99	3.97	0.00	0.00	0.00

Table 5: Recall time (s) and recall attempts (#) during follow-up recall sessions after one day, two days and one week. Data is reported as mean (μ), standard deviation (σ) and median ($\tilde{\mu}$).

	Day 1 Recall Time			Day 2 Recall Time			Day 7 Recall Time			Day 1 Recall Atts.			Day 2 Recall Atts.			Day 7 Recall Atts.		
	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$
Baseline	2.95	1.40	2.64	3.71	2.33	2.79	2.58	1.45	2.31	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Meter	4.26	1.94	3.10	3.60	1.37	2.97	3.57	1.65	2.91	0.1	0.32	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 6: NASA TLX data after enrolment and final recall sessions in terms of mean (μ), standard deviation (σ) and median ($\tilde{\mu}$).

		Mental			Physical			Temporal			Performance			Effort			Frustration			Overall		
		μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$	μ	σ	$\tilde{\mu}$
Enrol	Baseline	4.4	4.1	3.5	2.9	2.1	2.0	3.1	2.5	2.0	2.9	2.0	2.5	4.5	3.7	4.5	1.8	1.3	1.0	3.3	2.0	2.7
	Meter	7.3	2.5	7.5	5.4	3.7	6.0	4.5	2.5	4.0	5.8	4.2	4.0	8.5	4.4	6.0	4.9	3.4	4.0	6.1	2.8	5.8
Recall	Baseline	1.8	0.8	2.0	1.2	0.7	1.0	1.2	0.4	1.0	1.3	0.5	1.0	1.8	0.8	2.0	1.5	1.4	1.0	1.5	0.4	1.7
	Meter	3.0	1.6	3.0	2.9	1.7	2.5	3.1	1.5	3.0	3.0	1.5	3.0	4.0	2.5	3.5	1.8	0.9	1.5	3.0	1.3	2.8

also extend our prior data by indicating that meter and baseline gestures are equally rapidly and accurately recalled in subsequent authentication sessions. This suggests that use of our meter during enrolment does not impact subsequent performance, in terms of usability metrics, during multi-day recall.

The subjective measures corroborate these conclusions. Due to similar trends over all TLX items (see Table 6), we opted to conduct Wilcoxon rank sum tests only on overall workload scores for both enrolment and recall. Both show significant differences ($p=0.031$ and $p<0.009$, respectively), indicating that meter led to higher levels of perceived workload: meter gestures took more effort to both create and recall. However, it is also worth contextualizing the raw values. Overall workload for baseline is 3.3 for setup and 1.5 for recall. For meter these figures are 6.1 and 3. These scores are associated with very low to low levels of workload [27] and suggest that, despite the numerical differences, participants experienced few difficulties in either condition or task. SUS data confirm this. Scores of 83 (σ=13.58) and 86.94 (σ=10.74) for enrolment and recall were recorded for baseline and of 78.76 (σ=16.3) and 77.75 (σ=16.89) for meter. These levels are associated with "good" levels of usability [5]. Combined these results suggest our participants were able to operate both baseline and meter systems with ease.

5.5 Qualitative Results

Qualitative data from interviews and questionnaires with participants shed light on some of these variations. After enrolment, comments about security in the baseline condition suggested participants lacked knowledge about how to create gestures that would be secure. While six participants felt gestures would be "diverse" and "secure" as a password system, two worried that "simple or memorable shapes would be easily guessed by others". Many also referenced naive approaches such as the use of simple shapes (six) or letters (two) as a creation strategy. A key factor underlying these tensions was lack of knowledge about what would make a gesture more secure. Three participants reported simply relying on their own intuition to create a "unique shape" while three others felt "modifying the strokes of a [common] shape" was appropriate. Usability concerns also impacted their proposals. In order to ensure their gesture was memorable, three participants employed the strategy of "modifying the shape of [their] current pattern lock" and one based their gesture on the initial characters from their name. Others were motivated by efficiency or reliability, with two remarking they felt a key property of gesture passwords is that they would take "less time time to input" and a further two stating that they created gestures that would be "easily reached by the thumb". Perhaps reflecting these creation strategies, participants responded to questions about how they remembered their gestures after final recall by remarking they simply "recalled the shape" (four) or "keywords" (two) that reference it. No participant reported noting down their gesture.

The guidance presented in the meter condition strongly altered such patterns. Three participants again noted that the inherently "diverse" nature of gestures would provide strong security and two remarked the "recommended options" in the meter would likely increase this. In particular, the meter was valued as it could increase security via the strategy of "fill[ing] the bar" (three users) via refinements that increased the "length" or "complexity" of proposed gestures. In addition, three participants explicitly noted strategies of "distorting", "combining" or "aligning" their original gesture with system suggestions. One participant observed that the various suggestions presented helped ensure gestures modified in this way were "still memorable" and two noted that the revised gestures would still take "less time" to enter. After recall, participants reported using a modestly wider range of strategies to remember their

gestures. Once again, no participant indicated noting down their gesture, but one stated they "practiced repeatedly on [their] hand to memorize". Three others simply noted recalling the "dominant shapes", the "two shapes and their orders" or "keywords imagined from both shapes". These comments support our claims that feedback from our meter was able to effectively support users as they create gesture passwords.

Participants also expressed concerns and made recommendations in both conditions. Five participants worried that their may be "recognition failures" with gesture recognition authentication systems. For meter, individual participants also raised issues about the "time to memorize" and "complicated" form of the final gestures, with another noting that the meter's operating principles were hard to fathom: the "improvement [in a revised gesture] is visually unclear". They suggested various interface improvements to deal with these issues. Some related to the input surface itself and included enlarging it or augmenting it with visual aids such as a grid, or suggestions for candidate "touch start positions" during gesture creation. Others relate to memorability, including presentation of "hints or aids" to recall gestures or "numerical feedback on how successfully gestures are matched" and, in the case of meter, for "more diverse recommendations". In general these comments indicate that participants had generally positive opinions about the viability of gesture authentication and felt they were able to understand and use our meter to increase the security of the gestures they create.

6 DISCUSSION

We presented the data driven (N=1000) design and evaluation of a gesture password strength meter incorporating a feedback bar (or strength rating), textual recommendations for improvement and interactive, dynamically explorable recommended gestures. Results from a study evaluating its impact (N=600) indicate that it achieves substantial increases in security at the cost of extended enrolment times and modest decreases in recall time. Recall rates remain high. A subsequent multi-session study confirms these results for enrolment time and recall rate, but also suggests that multi-session recall times may be stable over protracted periods. We discuss the implications of these findings in detail below.

Usability data demonstrated the most extreme differences between baseline and meter conditions during setup. Most critically, in our second study, setup times were elevated from a median of 46.12s for baseline (itself somewhat greater than the 30.13s reported in closely related prior work [14]) to 127.24s for meter. Our multi-session study echoes these variations. These results clearly show that users spent longer creating gesture passwords using the meter than they did with the baseline system; such increases are commonly reported in studies of password meters [57]. However, interpreting such increases as a reduction in usability may be misleading. We note that are users were not mandated to use the meter features. As such, an alternative view of the increased set up times is that they encouraged user engagement in the gesture creation process, a potentially desirable quality associated with users selecting higher security passwords [45]. Indeed, use of the interactive meter functionality was relatively high: 46.33% of users in our second study selected at least one proposed improvement, a figure modestly greater than the 37.8% Ur at al. [58] report in their study of a similar recommendation system for text passwords. Our multi-session study suggests reasons for these high levels of engagement: participants reported iterating on their gestures to increase security, directly integrating meter suggestions and even requested that future meter designs propose a larger menu of options.

Recall data from our second study also shed light on user's experiences with our system. The baseline condition in our second study shows very high immediate recall rates of 99.3%, figures broadly in line with prior work (up to 98% in a similar study [14]). This reinforces the basic idea, motivating the work in this paper, that gestures can serve as highly memorable passwords. Importantly, short term recall rates for our meter (at 98%) do not significantly decrease from those in baseline, a level of performance that may be improved over that achieved in a prior study of blacklist-based gesture password policies (short term recall rates from 96.9% to 95.9%, with this latter figure representing a significant drop in performance [14]). These high short term recall rates support our assertion that users exposed to our meter have increased engagement with the gesture password creation process—high levels of engagement may be associated with increased gesture password memorability [15, 32]. Median recall times with the meter also significantly increased from 2.42s to 3.8s in our second study. While we note this could be associated with increased difficulty in entering gesture passwords in the meter condition, we also note it could also simply be due to the use of more sophisticated (or just longer) gestures. Such gestures will inevitably require more time to enter. Indeed, examining the mean path length of meter (24.3cm) and baseline (14.7cm) gestures supports this explanation. We argue that the fact that the meter's longer recall times are not associated with increased recall failures suggests the gestures generated are simply longer (or include more elements) rather than being more fundamentally challenging in terms of qualities such as memorability. Our multi-session study supports these interpretations: recall rates over one-day, two-days and one-week are unchanged (and perfect) between baseline and meter and, in this smaller sample, we do not observe variations in recall time.

Our security results also show promise. Data from the first study exhibits similar properties to that collected in prior work: partial guessing entropy levels are relatively high (compared to PIN [31] and pattern [15]) and well aligned with prior work on gesture passwords, as are other metrics such as EERs and DTW decision thresholds [14]. While crack rates in response to a dictionary attack were somewhat elevated—32% here vs 23.13% in related work [14]—this may simply reflect the smaller sample in our study (1000 vs 2594). We conclude our data set is representative of those in the literature and combines good security performance (in terms of basic metrics) with a worryingly high susceptibility to dictionary-based attacks. Our second study indicates our gesture meter successfully addressed this security concern—it demonstrated dramatically improved resistance against dictionary guessing attack. At the specific decision threshold associated with a FRR of 11.54%, a figure studied in prior work [14], a dictionary attack cracked 10.33% of gesture passwords generated using the gesture strength meter, a 67.03% improvement over the 31.33% cracked in our baseline condition. These increases in security were also associated with improvements in other security metrics such as EER (1.82% vs 2.8%) and partial guessing entropy. These improvements are competitive compared

Table 7: Subjective categorization of gestures passwords in both studies using the gesture categories from Cheon et al. [14]. Categories with no recorded gestures are omitted. Data shows mean category size from two raters and agreement calculated by Cohen's Kappa.

	Digit	Geometric Shape	Letter	Math Function	Math Symbol	Music Symbol	Compound	Cursive	Iconic	Inter-rater agreement
First Study	1.5%	26%	43%	3%	0.5%	0.5%	19.5%	3%	3%	$k = 0.46$
Baseline	8%	15.5%	33%	0.5%	0.5%	1%	26%	10%	5.5%	$k = 0.44$
Meter	4%	6.5%	0%	0%	0%	0.5%	73.5%	7.5%	8%	$k = 0.63$

to prior work—Cheon et al.'s [14] blacklist policies, for example, achieve crack rates as low as 14.93% at a similar 11.54% FRR threshold, 44.53% higher than those reported for our meter.

To shed light on how these improvements are achieved, we followed prior work [14, 36] and examined the generated gestures in detail. We first examined the sub-strokes in all gestures in terms of the distributions of their start points, their lengths and their angles—this data is shown in Appendices A and B. We note the meter is associated with an increased diversity in (position invariant) gesture start points, a quality that is associated with improved security in pattern based authentication [15]. In addition, we subjectively categorized gestures from both our studies, an approach also previously used to understand gesture diversity [14, 36]. The results, representing the mean categorizations from a pair of raters (achieving moderate to substantial agreement) are shown in Table 7. This data shows a stark contrast between the relatively high numbers of gestures categorized as letters or geometric shapes in the first study and second study baseline condition and the low numbers for these categories with the meter. These findings support our claims that our meter effectively engaged users during gesture creation, transforming this into a prolonged process that ultimately led to the generation of more complex and sophisticated, but still memorable, gesture passwords that achieved increased resistance to dictionary based attack.

There are several limitations that affect our work; these outline directions for future research. First and foremost, running larger online studies (e.g., N>1000) would offer many advantages. For example, they would support development of more accurate gesture rating algorithms and the identification of additional stroke features that are associated with more secure gesture passwords. In addition, larger samples would increase confidence in the usability and security results we present. Larger scale multi-session studies might also tease out variations in performance during multi-day recall and ones that target more frequent use, such as the tens to hundreds of times our participants' report unlocking their phones each day, might enable examination of the effects of fluent, expert gesture production. Furthermore, in order to increase confidence (and remove any lab-bias) in the qualitative results we report regarding users' experiences with our gesture password systems (in Section 5.5), future large scale studies should include capture of users' subjective opinions using formal measures such as TLX and SUS questionnaires and, additionally, open-ended text responses. In addition to such increases in scale, future work should also apply more rigorous methods to the analysis of open-ended qualitative data (e.g., thematic analysis [9]).

We also identify opportunities to improve our modeling. For example, while the DTW approach to gesture recognition we used is borrowed from closely related prior work [14, 37], alternatives such as Protractor [35], based on inverse cosine distance, and ensemble approaches such as Garda [37] may offer accuracy benefits—future work should explore and contrast how gesture strength meters perform with a range of different recognizers. Additionally, integrating more sophisticated optimization approaches such as bootstrapping [12] may lead to models that are capable of more accurately distinguishing between strong and weak gestures, or that directly support the generation of effective cracking dictionaries. We also see many avenues for improving our meter design. Feedback in our multi-session study provided numerous suggestions including improving the basic interface design (e.g., adding a grid to the drawing canvas), increasing the set of gesture recommendations for each issue detected (rather than just shoehorning users into one) and better exposing the rationale for the proposed gesture improvements, perhaps by animating them to clearly represent the changes or including explicit textual descriptions explaining how they have been improved. In addition, several users suggested increasing the size of the gesture drawing canvas to facilitate creating more detailed gestures, a design that has also been proposed in prior work [50]. While the small input area in our work was explicitly selected to support single-handed thumb-based unlock input, a formal study of the security and usability impacts of a larger gesture entry area would be of great interest. Exploring such design changes promises to further improve the security of the gesture passwords that users generate while potentially reducing the enrolment and recall times required to create and enter them.

Finally, it is worth discussing practical issues relating to how—and why—gesture passwords might be integrated into future smart devices. One current trend that impacts this is the growing use of biometric authentication techniques, such as fingerprint or Face ID, to unlock devices [20]. While practical and effective, we note such techniques do not remove the need for knowledge-based authentication schemes. Indeed, knowledge-based schemes such as PIN and pattern are still considered the primary authentication technique on devices in which biometric techniques are enabled. Knowledge-based authentication remains mandatory: it is required periodically (e.g., once every 48 hours), and used explicitly for security critical procedures such as restarting or updating the device, and changing authentication settings [48]. In contrast, due to the uncertainties about their inherent error rates and reliability, biometrics are still considered as secondary techniques to increase user convenience. As such, we see a viable future for knowledge-based device lock passwords, including those using gestures, even

in light of the rapid adoption of biometric approaches. We also note our meter has been carefully designed to be deployed in realistic future device lock scenarios. For example, we opted for a small input area to mimic the one-thumbed user experience common when operating existing pattern or PIN entry systems. In addition, we carefully modelled our enrolment processes on those present in existing smart phones—users create gestures, then feedback is given with respect to password selection policies (such as blacklists [40], or our meter), and, assuming the entered credentials are acceptable, users then confirm them. As such, we do not believe our meter raises barriers to adoption that are substantially greater than those that face any other novel authentication technique. We also note that while we study gesture passwords for phone screen lock, they may ultimately have more potential in emerging device form factors such as Head Mounted Displays (HMDs). These systems typically lack the high performance touch surfaces required to support the rapid and precise alphanumerical input required for PINs and passwords, but do integrate high performance 3D hand trackers, thus enabling 3D gesture passwords. While prior research has demonstrated the technical feasibility of such gesture password systems [52], we are not aware of any work that has acquired the large-scale data sets required to analyze their security. We identify capturing and analyzing such data sets and designing gesture selection policies to help users create better 3D gesture passwords as exciting challenges for future work.

7 CONCLUSION

This work explores the potential of gestures to serve as memorable, secure and easy to enter authentication credentials for smartphone unlock. Building on recent work that suggests that user proposed gesture passwords can be readily cracked, we present the design and evaluation of a gesture strength meter that provides both overall ratings and actionable feedback to its users. We show that this tool can help users create gestures that are substantially more secure, a process that takes longer but does not sacrifice memorability, even over periods of up to one week. The work in this paper represents a meaningful step towards understanding the strengths and weaknesses of authentication via gesture passwords and presents a system design, and evaluation data, that future researchers and system designers can build on to create real world gesture password authentication systems.

ACKNOWLEDGMENTS

This work was supported by a Korea Institute for Advancement of Technology (KIAT) grant funded by the Korean Government (MOTIE) (P0012725, The Competency Development Program for Industry Specialist).

REFERENCES

[1] Lisa Anthony, Radu-Daniel Vatavu, and Jacob O. Wobbrock. 2013. Understanding the Consistency of Users' Pen and Finger Stroke Gesture Articulation. In *Proceedings of Graphics Interface 2013* (Regina, Sascatchewan, Canada) *(GI '13)*. Canadian Information Processing Society, CAN, 87–94.

[2] Lisa Anthony and Jacob O. Wobbrock. 2012. $N-Protractor: A Fast and Accurate Multistroke Recognizer. In *Proceedings of Graphics Interface 2012* (Toronto, Ontario, Canada) *(GI '12)*. Canadian Information Processing Society, CAN, 117–120.

[3] Adam J. Aviv, Devon Budzitowski, and Ravi Kuber. 2015. Is Bigger Better? Comparing User-Generated Passwords on 3x3 vs. 4x4 Grid Sizes for Android's Pattern Unlock. In *Proceedings of the 31st Annual Computer Security Applications Conference* (Los Angeles, CA, USA) *(ACSAC 2015)*. Association for Computing Machinery, New York, NY, USA, 301–310. https://doi.org/10.1145/2818000.2818014

[4] Adam J. Aviv and Dane Fichter. 2014. Understanding Visual Perceptions of Usability and Security of Android's Graphical Password Pattern. In *Proceedings of the 30th Annual Computer Security Applications Conference* (New Orleans, Louisiana, USA) *(ACSAC '14)*. Association for Computing Machinery, New York, NY, USA, 286–295. https://doi.org/10.1145/2664243.2664253

[5] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies (JUS)* 4, 3 (2009), 114–123.

[6] Rachel Blagojevic, Samuel Hsiao-Heng Chang, and Beryl Plimmer. 2010. The Power of Automatic Feature Selection: Rubine on Steroids. In *Proceedings of the Seventh Sketch-Based Interfaces and Modeling Symposium* (Annecy, France) *(SBIM '10)*. Eurographics Association, Goslar, DEU, 79–86.

[7] Joseph Bonneau. 2012. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy (SP '12)*. IEEE Computer Society, USA, 538–552. https://doi.org/10.1109/SP.2012.49

[8] L. Bošnjak, J. Sreš, and B. Brumen. 2018. Brute-force and dictionary attack on hashed real-world passwords. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE Computer Society, USA, 1161–1166. https://doi.org/10.23919/MIPRO.2018.8400211

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[10] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[11] Xavier De Carné De Carnavalet and Mohammad Mannan. 2015. A Large-Scale Evaluation of High-Impact Password Strength Meters. *ACM Trans. Inf. Syst. Secur.* 18, 1, Article 1 (may 2015), 32 pages. https://doi.org/10.1145/2739044

[12] Claude Castelluccia, Markus Dürmuth, and Daniele Perito. 2012. Adaptive Password-Strength Meters from Markov Models. In *19th Annual Network and Distributed System Security Symposium, NDSS 2012, San Diego, California, USA, February 5-8, 2012*. The Internet Society, Reston, VA, USA, 14 pages. https://www.ndss-symposium.org/ndss2012/adaptive-password-strength-meters-markov-models

[13] Seunghun Cha, Sungsu Kwag, Hyoungshick Kim, and Jun Ho Huh. 2017. Boosting the Guessing Attack Performance on Android Lock Patterns with Smudge Attacks. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* (Abu Dhabi, United Arab Emirates) *(ASIA CCS '17)*. Association for Computing Machinery, New York, NY, USA, 313–326. https://doi.org/10.1145/3052973.3052989

[14] Eunyong Cheon, Yonghwan Shin, Jun Ho Huh, Hyoungshick Kim, and Ian Oakley. 2020. Gesture Authentication for Smartphones: Evaluation of Gesture Password Selection Policies. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, USA, 249–267. https://doi.org/10.1109/SP40000.2020.00034

[15] Geumhwan Cho, Jun Ho Huh, Junsung Cho, Seongyeol Oh, Youngbae Song, and Hyoungshick Kim. 2017. SysPal: System-Guided Pattern Locks for Android. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, USA, 338–356. https://doi.org/10.1109/SP.2017.61

[16] Gradeigh D. Clark, Janne Lindqvist, and Antti Oulasvirta. 2017. Composition policies for gesture passwords: User choice, security, usability and memorability. In *2017 IEEE Conference on Communications and Network Security (CNS)*. IEEE Computer Society, USA, 1–9. https://doi.org/10.1109/CNS.2017.8228644

[17] Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. 2014. The Tangled Web of Password Reuse. In *21st Annual Network and Distributed System Security Symposium, NDSS 2014, San Diego, California, USA, February 23-26, 2014*. The Internet Society, Reston, VA, USA, 15 pages. https://www.ndss-symposium.org/ndss2014/tangled-web-password-reuse

[18] Xavier de Carné de Carnavalet and Mohammad Mannan. 2014. From Very Weak to Very Strong: Analyzing Password-Strength Meters. In *21st Annual Network and Distributed System Security Symposium, NDSS 2014, San Diego, California, USA, February 23-26, 2014*. The Internet Society, Reston, VA, USA, 16 pages. https://www.ndss-symposium.org/ndss2014/very-weak-very-strong-analyzing-password-strength-meters

[19] Alexander De Luca, Alina Hang, Frederik Brudy, Christian Lindner, and Heinrich Hussmann. 2012. Touch Me Once and i Know It's You! Implicit Authentication Based on Touch Screen Patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 987–996. https://doi.org/10.1145/2207676.2208544

[20] Alexander De Luca, Alina Hang, Emanuel von Zezschwitz, and Heinrich Hussmann. 2015. I Feel Like I'm Taking Selfies All Day! Towards Understanding Biometric Authentication on Smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1411–1414. https://doi.org/10.1145/2702123.2702141

[21] David H Douglas and Thomas K Peucker. 1973. Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization* 10, 2 (1973), 112–122. https://doi.org/10.3138/FM57-6770-U75U-7727

[22] Serge Egelman, Andreas Sotirakopoulos, Ildar Muslukhov, Konstantin Beznosov, and Cormac Herley. 2013. Does My Password Go up to Eleven? The Impact of Password Meters on Password Selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2379–2388. https://doi.org/10.1145/2470654.2481329

[23] Alain Forget, Sonia Chiasson, and Robert Biddle. 2010. Shoulder-Surfing Resistance with Eye-Gaze Entry in Cued-Recall Graphical Passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1107–1110. https://doi.org/10.1145/1753326.1753491

[24] Alain Forget, Sonia Chiasson, P. C. van Oorschot, and Robert Biddle. 2008. Improving Text Passwords through Persuasion. In *Proceedings of the 4th Symposium on Usable Privacy and Security* (Pittsburgh, Pennsylvania, USA) *(SOUPS '08)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/1408664.1408666

[25] Steven Furnell, Warut Khern-am nuai, Rawan Esmael, Weining Yang, and Ninghui Li. 2018. Enhancing security behaviour by supporting the user. *Computers & Security* 75 (06 2018). https://doi.org/10.1016/j.cose.2018.01.016

[26] Javier Galbally, Iwen Coisel, and Ignacio Sanchez. 2017. A New Multimodal Approach for Password Strength Estimation—Part I: Theory and Algorithms. *IEEE Transactions on Information Forensics and Security* 12, 12 (2017), 2829–2844. https://doi.org/10.1109/TIFS.2016.2636092

[27] Rebecca A. Grier. 2015. How High is High? A Meta-Analysis of NASA-TLX Global Workload Scores. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 59, 1 (2015), 1727–1731. https://doi.org/10.1177/1541931215591373 arXiv:https://doi.org/10.1177/1541931215591373

[28] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, Amsterdam, Netherlands, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[29] Shiva Houshmand and Sudhir Aggarwal. 2012. Building Better Passwords Using Probabilistic Techniques. In *Proceedings of the 28th Annual Computer Security Applications Conference* (Orlando, Florida, USA) *(ACSAC '12)*. Association for Computing Machinery, New York, NY, USA, 109–118. https://doi.org/10.1145/2420950.2420966

[30] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell, and Nicholas J. G. Winter. 2020. The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods* 8, 4 (2020), 614–629. https://doi.org/10.1017/psrm.2020.6

[31] Hyoungshick Kim and Jun Ho Huh. 2012. PIN Selection Policies: Are They Really Effective? *Comput. Secur.* 31, 4 (jun 2012), 484–496. https://doi.org/10.1016/j.cose.2012.02.003

[32] Saranga Komanduri, Richard Shay, Patrick Gage Kelley, Michelle L. Mazurek, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Serge Egelman. 2011. Of Passwords and People: Measuring the Effect of Password-Composition Policies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2595–2604. https://doi.org/10.1145/1978942.1979321

[33] Cynthia Kuo, Sasha Romanosky, and Lorrie Faith Cranor. 2006. Human Selection of Mnemonic Phrase-Based Passwords. In *Proceedings of the Second Symposium on Usable Privacy and Security* (Pittsburgh, Pennsylvania, USA) *(SOUPS '06)*. Association for Computing Machinery, New York, NY, USA, 67–78. https://doi.org/10.1145/1143120.1143129

[34] Luis Leiva, Radu-Daniel Vatavu, Daniel Martín-Albo, and Réjean Plamondon. 2020. Omnis Prædictio: Estimating the Full Spectrum of Human Performance with Stroke Gestures. *International Journal of Human-Computer Studies* 142 (05 2020), 102466. https://doi.org/10.1016/j.ijhcs.2020.102466

[35] Yang Li. 2010. Protractor: A Fast and Accurate Gesture Recognizer. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2169–2172. https://doi.org/10.1145/1753326.1753654

[36] Can Liu, Gradeigh D. Clark, and Janne Lindqvist. 2017. Guessing Attacks on User-Generated Gesture Passwords. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 1, Article 3 (mar 2017), 24 pages. https://doi.org/10.1145/3053331

[37] Can Liu, Gradeigh D. Clark, and Janne Lindqvist. 2017. Where Usability and Security Go Hand-in-Hand: Robust Gesture-Based Authentication for Mobile Systems. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 374–386. https://doi.org/10.1145/3025453.3025879

[38] A. Chris Long, James A. Landay, Lawrence A. Rowe, and Joseph Michiels. 2000. Visual Similarity of Pen Gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands) *(CHI '00)*. Association for Computing Machinery, New York, NY, USA, 360–367. https://doi.org/10.1145/332040.332458

[39] Jerry Ma, Weining Yang, Min Luo, and Ninghui Li. 2014. A Study of Probabilistic Password Models. In *Proceedings of the 2014 IEEE Symposium on Security and Privacy (SP '14)*. IEEE Computer Society, USA, 689–704. https://doi.org/10.1109/SP.2014.50

[40] Philipp Markert, Daniel V. Bailey, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. 2020. This PIN Can Be Easily Guessed: Analyzing the Security of Smartphone Unlock PINs. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, USA, 286–303. https://doi.org/10.1109/SP40000.2020.00100

[41] Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. 2013. Measuring Password Guessability for an Entire University. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer &; Communications Security* (Berlin, Germany) *(CCS '13)*. Association for Computing Machinery, New York, NY, USA, 173–186. https://doi.org/10.1145/2508859.2516726

[42] William Melicher, Darya Kurilova, Sean M. Segreti, Pranshu Kalvani, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L. Mazurek. 2016. Usability and Security of Text Passwords on Mobile Devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 527–539. https://doi.org/10.1145/2858036.2858384

[43] Antti Pirhonen, Stephen Brewster, and Christopher Holguin. 2002. Gestural and Audio Metaphors as a Means of Control for Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) *(CHI '02)*. Association for Computing Machinery, New York, NY, USA, 291–298. https://doi.org/10.1145/503376.503428

[44] Robert W Proctor, Mei-Ching Lien, Kim-Phuong L Vu, E Eugene Schultz, and Gavriel Salvendy. 2002. Improving computer security for authentication of users: Influence of proactive password restrictions. *Behavior Research Methods, Instruments, & Computers* 34, 2 (2002), 163–169.

[45] George E. Raptis, Christina Katsini, Andrew Jian-lan Cen, Nalin Asanka Gamagedara Arachchilage, and Lennart E. Nacke. 2021. Better, Funner, Stronger: A Gameful Approach to Nudge People into Making Less Predictable Graphical Password Choices. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 112, 17 pages. https://doi.org/10.1145/3411764.3445658

[46] Napa Sae-Bae, Kowsar Ahmed, Katherine Isbister, and Nasir Memon. 2012. Biometric-Rich Gestures: A Novel Approach to Authentication on Multi-Touch Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 977–986. https://doi.org/10.1145/2207676.2208543

[47] Alireza Sahami Shirazi, Peyman Moghadam, Hamed Ketabdar, and Albrecht Schmidt. 2012. Assessing the Vulnerability of Magnetic Gestural Authentication to Video-Based Shoulder Surfing Attacks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2045–2048. https://doi.org/10.1145/2207676.2208352

[48] Apple Platform Security. 2022. Face ID, touch ID, passcodes, and passwords. Retrieved Dec 14th 2022 from https://support.apple.com/guide/security/face-id-touch-id-passcodes-and-passwords-sec9479035f1/web.

[49] Richard Shay, Saranga Komanduri, Patrick Gage Kelley, Pedro Giovanni Leon, Michelle L. Mazurek, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2010. Encountering Stronger Password Requirements: User Attitudes and Behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security* (Redmond, Washington, USA) *(SOUPS '10)*. Association for Computing Machinery, New York, NY, USA, Article 2, 20 pages. https://doi.org/10.1145/1837110.1837113

[50] Michael Sherman, Gradeigh Clark, Yulong Yang, Shridatt Sugrim, Arttu Modig, Janne Lindqvist, Antti Oulasvirta, and Teemu Roos. 2014. User-Generated Free-Form Gestures for Authentication: Security and Memorability. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services* (Bretton Woods, New Hampshire, USA) *(MobiSys '14)*. Association for Computing Machinery, New York, NY, USA, 176–189. https://doi.org/10.1145/2594368.2594375

[51] Hansub Shin, Sungyong Sim, Hyukyoon Kwon, Sangheum Hwang, and Younho Lee. 2022. A new smart smudge attack using CNN. *International Journal of Information Security* 21, 1 (2022), 25–36.

[52] M.A. Shukran and M.S.B. Ariffin. 2012. Kinect-based gesture password recognition. *Australian Journal of Basic and Applied Sciences* 6 (08 2012), 492–499.

[53] Youngbae Song, Geumhwan Cho, Seongyeol Oh, Hyoungshick Kim, and Jun Ho Huh. 2015. On the Effectiveness of Pattern Lock Strength Meters: Measuring the Strength of Real World Pattern Locks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2343–2352. https://doi.org/10.1145/2702123.2702365

[54] Khai N. Truong, Thariq Shihipar, and Daniel J. Wigdor. 2014. Slide to X: Unlocking the Potential of Smartphone Unlocking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3635–3644. https://doi.org/10.1145/2556288.2557044

[55] Huawei Tu, Xiangshi Ren, and Shumin Zhai. 2012. A Comparative Evaluation of Finger and Pen Stroke Gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1287–1296. https://doi.org/10.1145/2207676.2208584

[56] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. 2013. Quantifying the Security of Graphical Passwords: The Case of Android Unlock Patterns. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer &; Communications Security* (Berlin, Germany) *(CCS '13)*. Association for Computing Machinery, New York, NY, USA, 161–172. https://doi.org/10.1145/2508859.2516700

[57] Blase Ur. 2016. *Supporting Password-Security Decisions with Data*. Ph. D. Dissertation. Carnegie Mellon University.

[58] Blase Ur, Felicia Alfieri, Maung Aung, Lujo Bauer, Nicolas Christin, Jessica Colnago, Lorrie Faith Cranor, Henry Dixon, Pardis Emami Naeini, Hana Habib, Noah Johnson, and William Melicher. 2017. Design and Evaluation of a Data-Driven Password Meter. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3775–3786. https://doi.org/10.1145/3025453.3026050

[59] Blase Ur, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L. Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2012. How Does Your Password Measure up? The Effect of Strength Meters on Password Creation. In *Proceedings of the 21st USENIX Conference on Security Symposium* (Bellevue, WA) *(Security'12)*. USENIX Association, USA, 5.

[60] Kim-Phuong L. Vu, Robert W. Proctor, Abhilasha Bhargav-Spantzel, Bik-Lam (Belin) Tai, Joshua Cook, and E. Eugene Schultz. 2007. Improving password security and memorability to protect personal and organizational information. *International Journal of Human-Computer Studies* 65, 8 (2007), 744–757. https://doi.org/10.1016/j.ijhcs.2007.03.007

[61] Ding Wang, Debiao He, Haibo Cheng, and Ping Wang. 2016. fuzzyPSM: A New Password Strength Meter Using Fuzzy Probabilistic Context-Free Grammars. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE Computer Society, USA, 595–606. https://doi.org/10.1109/DSN.2016.60

[62] Matt Weir, Sudhir Aggarwal, Michael Collins, and Henry Stern. 2010. Testing Metrics for Password Creation Policies by Attacking Large Sets of Revealed Passwords. In *Proceedings of the 17th ACM Conference on Computer and Communications Security* (Chicago, Illinois, USA) *(CCS '10)*. Association for Computing Machinery, New York, NY, USA, 162–175. https://doi.org/10.1145/1866307.1866327

[63] Daniel Lowe Wheeler. 2016. Zxcvbn: Low-Budget Password Strength Estimation. In *Proceedings of the 25th USENIX Conference on Security Symposium* (Austin, TX, USA) *(SEC'16)*. USENIX Association, USA, 157–173.

[64] Yulong Yang, Gradeigh D. Clark, Janne Lindqvist, and Antti Oulasvirta. 2016. Free-Form Gesture Authentication in the Wild. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3722–3735. https://doi.org/10.1145/2858036.2858270

[65] Ziming Zhao, Gail-Joon Ahn, Jeong-Jin Seo, and Hongxin Hu. 2013. On the Security of Picture Gesture Authentication. In *22nd USENIX Security Symposium (USENIX Security 13)*. USENIX Association, Washington, D.C., 383–398. https://www.usenix.org/conference/usenixsecurity13/technical-sessions/presentation/zhao

A APPENDIX A

Figure 8 shows all sub-stroke start points from all gestures in all studies and conditions reported in this paper. The meter condition in the second study shows a markedly more even distribution of initial sub-stroke start points (basically gesture start points) compared to those generated in the baseline condition and first study.

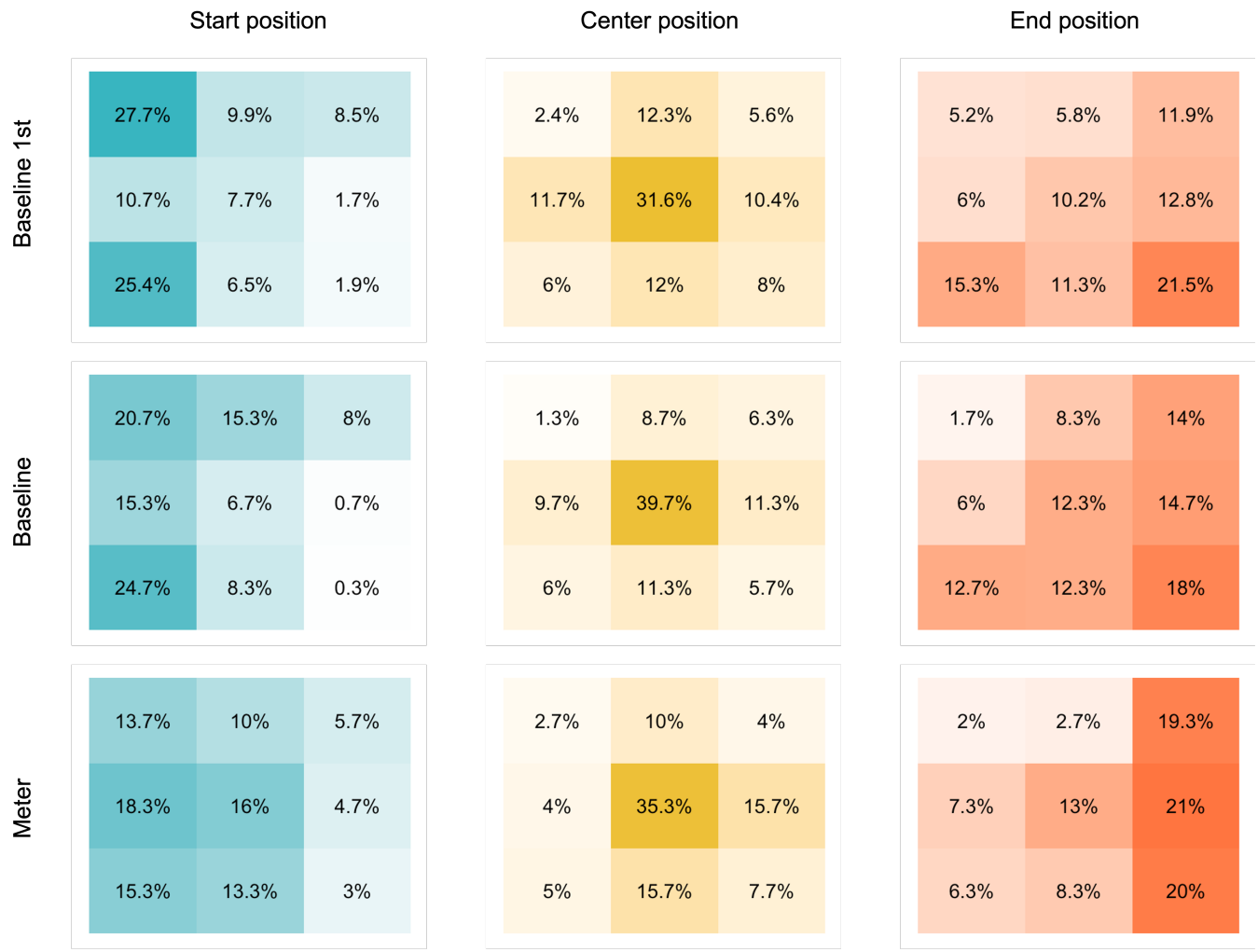

Figure 8: Distributions (in %) of all sub-stroke start points in all gestures in both studies reported in this paper, discretized into a three by three grid. Gestures are rendered scale and position invariant prior to calculating these distributions. In addition, figures show initial sub-stroke (left column), final sub-stroke (right column) and all other strokes (center column).

B APPENDIX B

Figure 9 shows all sub-stroke start lengths and angles all gestures in all studies and conditions reported in this paper. Compared to other conditions and studies, the meter condition shows a somewhat elevated use of longer initial strokes and a reduction in the use of short final strokes. In addition, initial stroke angles are more widely distributed away from the otherwise dominant left/down direction.

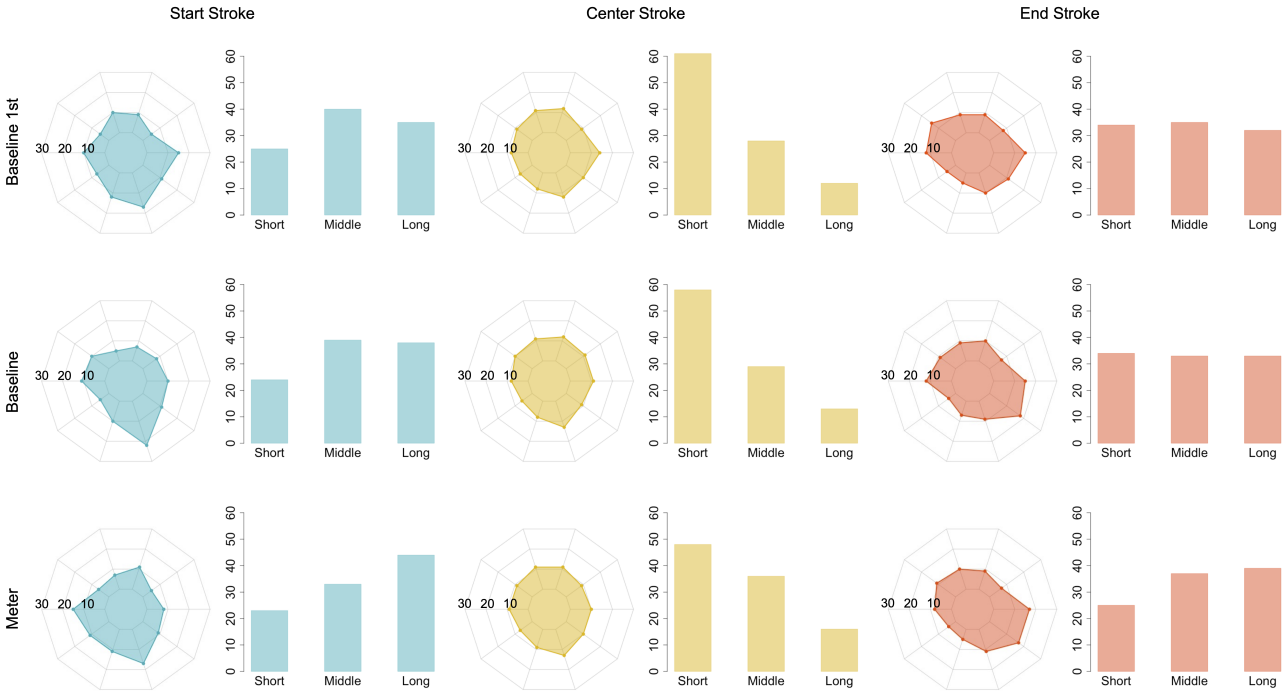

Figure 9: Distributions (in %) of all sub-stroke angles and lengths for all studies and conditions reported in this paper. The discretization into ten angles and three lengths follows that used in the n-gram models presented in this paper. We separate data for initial sub-strokes (left), final sub-strokes (right) and all other sub-strokes (center).