Evaluating the usability and security of a graphical one-time PIN system

Sacha Brostoff
Department of Computer Science
UCL (University College London)
Malet Place, London, UK.
s.brostoff@cs.ucl.ac.uk

Philip Inglesant
Department of Computer Science
UCL (University College London)
Malet Place, London, UK.
p.inglesant@cs.ucl.ac.uk

M. Angela Sasse
Department of Computer Science
UCL (University College London)
Malet Place, London, UK.
a.sasse@cs.ucl.ac.uk

ABSTRACT

Traditional Personal Identification Numbers (PINs) are widely used, but the attacks in which they are captured have been increasing. One-time PINs offer better security, but potentially create greater workload for users. In this paper, we present an independent evaluation of a commercial system that makes PINs more resistant to observation attacks by using graphical passwords on a grid to generate a one-time PIN. 83 participants were asked to register with the system and log in at varying intervals. The successful login rate was approximately 91% after 3-4 days, and 97% after 9-10 days. Twenty five participants were retested after two years, and 27% of those were able to recall their pattern.

We recorded 17 instances of failed attempts, and found that even though participants recalled the general shape of the pass-pattern in 13 of these instances, they could not recall its detailed location or sequence of cells. We conclude that GrIDsure is usable if people have one pass-pattern, but the level of security will depend on the context of use (it will work best in scenarios where repeated observations of transactions are unlikely), and the instructions given to users (without guidance, they are likely to chose from a small subset of the possible patterns which are easily guessed).

Categories and Subject Descriptors

H.5.2 User Interfaces – *Graphical user interfaces;* K.6.5 Management of Computing and Information Systems: Security and Protection

General Terms

Security, Human Factors

Keywords

Authentication usage scenarios; graphical passwords; PINs

1. INTRODUCTION

Knowledge-based authentication through passwords and PINs remains the most widely used security scheme. Most users struggle to recall their PINs and passwords, and resort to unsafe practices, such as writing them down [1], or choosing predictable passwords [12]. PINs have many of the same problems – but are usually used in combination with a physical token.

PIN authentication – commonly used for cash withdrawals at ATMs with bank cards –now also replaces signatures for credit card payments in many countries. In the UK alone, there were 805 million ATM withdrawals in the final quarter of 2008, each of them requiring the use of a PIN [3]. Many telephone and online banking schemes also use PINs- so most people have to use a PIN at least once a day. However, any static knowledge-based credential faces the risk of interception through shoulder-surfing, screen-scraping, a compromised terminal, etc. [9]. This has increased interest in one-time passwords and PINs, because capture of any specific example is not useful to the attacker. In the past, one-time password schemes were the preserve of high-secrecy domains such as national intelligence, but even highly motivated and trained operatives struggled to use them correctly [14]. More recently, we have seen developments of solutions that deliver one-time PINs to users via SMS, or from special tokens.

In this paper, we present an evaluation of a commercially-available authentication scheme which aims to provide the security of a one-time PIN without special hardware, while being easy to understand, remember, and use [10].

Section 2 reviews the usability problems and security risks inherent in existing authentication schemes, and some approaches which have been taken to address these. Section 3 presents the GrIDsure scheme. Section 4 describes the methodology of two empirical evaluations of GrIDsure performance. In the second evaluation we used instructions designed to encourage participants to chose “less obvious” patterns. Section 0 presents the results in terms of usability and security. We found the instructions improved pattern choice. We conclude with some suggestions of future possibilities for GrIDsure, in terms of its potential real-world usage and as an area for further research.

2. BACKGROUND

Secure authentication is particularly challenging whenever the communication is susceptible to interception because [2]:

1. The interface cannot be trusted - for example, when using a Web browser on a potentially compromised computer, at an ATM, or at a point of sale terminal;
2. Communication is susceptible to electronic or physical eavesdropping; or
3. Users disclose the secrets in social engineering attacks (such as “phishing” or “pretexting”).

These problems are particularly acute with PINs. They are usually 4 digits long - which gives a 1 in 10,000 probability of random guessing. The brevity and simplicity of a PIN, and the fact that it is often used in a public place - such as at an ATM or point of sale terminal - makes it particularly vulnerable to observation, whether simple shoulder-surfing or more sophisticated technical attacks.

One-time PINs are a way to overcome this problem. Most existing solutions require special hardware, which costs, may be lost or

24th BCS Conference on Human Computer Interaction - HCI2010, September 5-10, 2010, Dundee, Scotland

stolen, and relies on an issuing authority and so is only suitable where there is a strong relationship between user and issuer, as in the workplace [9]. A usability weakness is that users may forget to carry the token, and then be unable to make the transaction.

In most current solutions, the one-time PIN is used *alongside* another authenticating method, to generate a challenge-response system. If the one-time PIN can be combined with a more usable method of authentication, it would reduce users' workload. In GrIDsure, the PIN is part of a *graphical password*, combining personal knowledge and one-time PIN generation into one step.

A number of authentication mechanisms have been devised which rely on *visual*, rather than verbal, memory. These build on psychological findings that recall of pictures is better than either passwords or PIN [7, 15, 16] - the so-called *picture superiority effect*. For example, users are presented human faces in the case of Passfaces [17] and randomly-generated art in the case of Déjà Vu [6]- from which they recognize and select previously-chosen images.

All of these have an element of recognition rather than recall. GrIDsure, in contrast, is what de Angeli et al. [5] call a *drawmetric* system; it depends on recall, rather than recognition.

3. THE GRIDSURE SYSTEM

GrIDsure is a graphical password scheme. However, whereas such schemes usually use some sort of graphical interaction *instead* of a PIN or textual password, GrIDsure uses the graphical scheme to *generate* a one-time PIN; it is effectively a combination of both graphical and PIN authentication. Participants read their PIN from a 5x5 number grid (see Figure 1b) by locating the numbers displayed in 4 cells they have chosen.

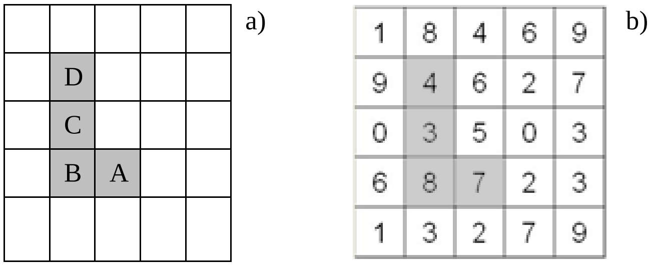
When enrolling, users are asked to "*pick a pretty pattern or 'shape' that you can remember.*" Users choose the shape (e.g. an "L" shape) and the order in which they want to read off their numbers (e.g. bottom to top). For example, a user could choose the cells we have labelled A, B, C and D in Figure 1 a. We call this shape and order of cells the user's "pattern", and it is the secret they must remember in order to authenticate.

In each GrIDsure authentication, the grid is populated by random numbers between 0 and 9, with some repetitions (example grid in Figure 1b. The user reads off the numbers that appear in her pattern, in sequence, and enters them on a separate keypad. In Figure 1b, for example, the user's PIN would be 7, 8, 3, 4 – the numbers currently occupying her cells A, B, C, and D. The next time she authenticated the PIN would be different again – whatever random numbers occupied her pattern that session.

GrIDsure supports the use of different size of grid and different numbers of cells chosen to make up the one-time PIN. In some implementations cells can be re-used (e.g. four cells the same generates a PIN with 4 numbers the same), in other implementations no re-use of cells are allowed.

4. METHODOLOGY

Two evaluations were conducted on GrIDsure, approximately two years apart. The evaluations used the same hardware and software implementation of GrIDsure with the same task, differing in the instructions given to participants about how to choose a pattern – Eval 2 having more detailed guidance about choosing strong patterns, based on results from Evaluation 1.



	D			
	C			
	B	A		

1	8	4	6	9
9	4	6	2	7
0	3	5	0	3
6	8	7	2	3
1	3	2	7	9

Figure 1. a) Enrolling in the system. User picks cells A, B, C and D.
b) Authenticating with the system. User reads off random numbers chosen cells.

Twenty five participants from Evaluation 1 were recruited again for Evaluation 2. Before embarking on Evaluation 2 they were asked to recall their patterns from Evaluation 1 – last used more than two years before. This follow-up had not been planned during the original study, and so participants did not expect to remember their patterns over such an extended duration.

4.1 Procedure

Participants performed an initial enrolment and verification and then two or three subsequent verifications at varying time intervals at their work places or homes. Each participant was asked to read and sign a consent form, was given payment and asked sign for it. They were then given a standard description of the GrIDsure scheme. Any immediate questions were clarified, and a demonstration of the operation of the scheme was given.

They were then asked to select four cell pass-pattern, and immediately attempt to recall it on a fresh grid populated with random numbers. The results of this attempt were recorded as the *enrolment verification*. Participants subsequently performed 2 or 3 further verifications of their patterns, at varying intervals ranging 1 to 75 days. Participants' spontaneous comments during system use where written down by experimenters.

4.2 Participants

All participants were volunteers, and were paid £10 for taking part. Fifty one participants were recruited for Evaluation 1 from administrative staff in the university, and members of the public known to the experimenters. Twenty seven of them were male, 5 were post-retirement age. These participants were contacted again two years later and asked to participate in the follow-up experiment- and 25 were willing to participate in Evaluation 2, along with 29 newly- recruited participants (postgraduate students, and administrative, support and research staff at UCL, and experimenters' acquaintances external to UCL). Of the total 54 participants in Evaluation 2, 32 were male, 1 was post-retirement age.

4.3 Apparatus

An early demo of the proposed PIN-replacement scheme obtained from GrIDsure was installed on a number of PDAs (touch-screen iPaq Pocket PC running Microsoft Windows Pocket PC 2003 Premier). The use of PDAs allowed us to take the trial to participants, rather than have them attend a lab.

The implementation of GrIDsure that we tested used a plain (no colour) 5x5 grid and pass-patterns of exactly 4 squares (no repetition allowed). *A priori*, this seems to offer a reasonable balance between memorability and the size of the pattern space; a study of visual memory has found that short-term recall of dots in a pattern decreases rapidly once the grid size is at least 5x5 and the number of dots exceeds 4 [11].

Simple paper-based observation sheets were used by the experimenter to record the chosen pass-pattern and the results of each attempted recall. Experience in Evaluation 1 led to the redesign of the sheets for Evaluation 2, to also record the patterns used in failed verification attempts.

4.3.1 Evaluation 1

After the enrolment verification, subsequent verifications took place at intervals varying from a few hours to 11 weeks. The participants were divided roughly into two equal cohorts, one with short (less than one week) and one with longer intervals between verifications.

4.3.2 Retesting after an Extended Period

Evaluation 1 took place during October, 2006 to February, 2007. After a much extended period, we were interested to investigate whether the patterns, based as they are on visual memory, would be recalled after a long time. If they could be, then GrIDsure could be useful in situations where passwords or PINs are used infrequently.

25 participants from Evaluation 1 were still available and willing to participate. They attempted to recall their patterns from two years earlier.

4.3.3 Evaluation 2

In Evaluation 1, we observed that users tended to select patterns from a restricted pattern set (see “Patterns chosen by users” below), with implications for the guessability of the pass-pattern. We performed a further evaluation, to see if we could encourage greater diversity in the pass-patterns chosen through careful re-wording of the instructions given to participants.

Instructions are crucial; it has been shown that they impact on the password-choosing behaviour of experimental participants [20], and they also simulate, in a controlled way, the real-world password advice given to people.

In Evaluation 1, the instructions to participants did not include any particular guidance about the pattern, beyond the instruction to choose “*a pretty pattern or ‘shape’ that you can remember*” and reminding participants to consider both the pattern and the sequence.

In Evaluation 2, the instructions were changed to include the sentences:

Make any pattern or shape that you can remember, but that you think other people will find it hard to guess. Here are a couple of simple examples, although it is recommended that you choose less common patterns that would be harder for an attacker to guess.

Since we were particularly concerned about the *order* of the cells chosen:

There is a tendency for people to select patterns that start at the top left and run towards the bottom right of the grid in the same way that we read from left to right and top to bottom. Again this makes your pattern easier to guess, so try to avoid this behaviour when choosing your squares, always ensuring that you can recall both the shape and order of the squares used.

The actual examples given in the instructions were the same as before, and it is worth noting that, in contrast to the words which they illustrate, these consisted entirely of adjacent cells and symmetrical patterns.

5. RESULTS

5.1 Errors

During Evaluation 2, and for the very extended (2 year) period of recall, we recorded not only the number of errors but also their type, by asking participants after each unsuccessful authentication attempt what they *thought* their pattern was.

Over short periods, the shape of the pattern was recalled most frequently, and order of cells in the pattern least frequently. Over longer periods of time, the most common problem was forgetting the *position* of the pattern on the grid.

There were a total of 17 pattern recall errors in Evaluation 2. Of these, 2 participants had problems recalling their pattern immediately after registering it – experiencing 1 and 5 errors on their first session. Three participants had 3, 2 and 6 errors respectively, at the second session (after an interval of 3 to 4 days). In 13 of the 17 errors, the *shape* was correctly recalled, but other details were wrong, such as the pattern *placement* in the grid or the *order* of squares in the pattern (**Error! Not a valid bookmark self-reference.**).

Table 1: Errors in recalling patterns in Evaluation 2

Error type	**Error combinations**							**Row Total**
Shape					X	X	X	4
Order	X	X	X			X		13
Placement		X					X	6
Density		X	X	X				7
Count	**6**	**5**	**1**	**1**	**2**	**1**	**1**	17

Thirty six errors were observed during the very long recall of patterns. In 24 of these errors (67%) the *shape* of the pattern was conserved, but other details were wrong (Table 2). *Placement* of the pattern on the grid appeared to be the least remembered (15 of the errors – 42%), followed by the *order* of squares within the pattern (3 errors), although combinations of placement and order errors were also frequent (5 errors). There was only one instance recorded of a *density* problem - the squares in the pattern being too spread apart, but correct in all other respects.

Table 2: Errors in pattern recall over an extended period.

Error type	**Error combinations**							**Row Total**
Shape					X	X	X	12
Order		X	X				X	12
Placement	X	X			X		X	29
Density				X			X	5
Count	**15**	**5**	**3**	**1**	**5**	**3**	**4**	36

These problems have an interesting parallel to a findings reported in [18] with traditional character-based passwords: people often remember the general shape of the password, but not details such as upper and lower case characters.

5.2 Patterns chosen by users

5.2.1 The effective pattern space is smaller than the theoretical space

On a 5x5 grid, there are 25×24×23×21 = 303,600 possible unique patterns. However, the *effective* password space might be much smaller than this: if users chose only from a sub-set of patterns the guessing difficulty of patterns in much lower than that. A parallel example is the “beauty bias” observed with participants in studies of Passfaces [16]. To analyse the *guessing difficulty* of GrIDsure, we analysed the chosen patterns.

Each pattern was categorised independently by two researchers, and a taxonomy of patterns developed (see Table 3). . We found that patterns were indeed chosen from a smaller subset.

Table 3: Occurrences of basic shapes in Evaluations 1 and 2

Basic shape	**Count**	
	Evaluation 1	**Evaluation 2**
Corners	5	1
Diamond	4	6
Square	7	11
Diagonal	6	1
Line	16	11
Split Diagonal	4	2
Split pair	8	5
Uncategorisable	1	17
Total	51	54

The taxonomy (Table 3) is simple, but reflects the patterns chosen. Of the 51 participants we tested first in Evaluation 1, all bar one selected a pattern in the taxonomy.

We also checked account for the *order* of traversing each pattern (Table 4). Order also influences the size of the pattern space, since any four-cell pattern can be traversed in 24 unique ways. In our analysis, we categorize the traversal order in various “natural” ways; versus “cross-wise”, meaning that the pattern has been traversed in some other way.

Table 4: Traversal orders in Evaluations 1 & 2

Traversal Order	**Evaluation 1**	**Evaluation 2**
Clockwise	7	13
Anti-clockwise	3	3
Top-bottom/Left-right	22	7
Bottom-top/Right-left	5	5
Top-bottom/Right-left	3	2
Bottom-top/Left-right	1	4
Cross-wise	10	20
Total	51	54

Table 4 shows that 41 of 51 participants in Eval. 1 chose patterns traversed in one of the “natural” orders, and that there is a tendency towards clockwise over anti-clockwise, and top-bottom or left-right over other directions. Thus, the order of traversal increases the effective size of the pattern space by far less than is potentially possible.

Finally, we noted some simple “variants” to the basic shapes (Table 5); in these patterns, just one cell is offset from a basic type, which could be thought of as “three of a kind”. For example, an “L”, “T”, or “J” shape is a line with a variant - the cell at one end is moved from the line to another, adjacent, cell.

Straight lines are by far the most often varied of all the basic shapes. In principle all possible single-cell variants multiply the effective pattern space by at least 21 × 4 = 84. However, if in practice variants are mainly applied to lines, and only in predictable ways, then the pattern space remains much smaller than the possible space.

Table 5: Variants on the basic shapes chosen in Evaluations 1 and 2

	Count	
Variant	**Eval.1**	**Eval.2**
Horizontal and vertical lines with variant	11	7
Diagonal lines with variant	0	0
Squares with variant	1	2
Diamonds with variant	1	2
“Corners” with variant	2	0
Split pairs with “perpendicular”	0	3
Split diagonals with “perpendicular”	0	1
“Sparse” lines - must have a variant	2	2
“Sparse” diagonals - must have a variant	0	0
No variation or “uncategorisable”	31	37

In conclusion, the *effective* pattern space is smaller than the *possible* space, even when the order traversal and common variants on simple shapes are taken into account, and consequently the security of GrIDsure is reduced.

5.2.2 Better instructions increase guessing difficulty

Recall that one of the findings from Evaluation 1 was that patterns can almost all be categorised using a rather simple taxonomy, and that we re-worded our instructions to participants as a first step to overcoming this problem.

Using the same categorizations as before, we found that the frequencies of shapes chosen changed with as a result of revised instructions (Table 3). A chi-squared test supports this finding with a statistically significant association between “Evaluation” and “Basic shapes” (χ^2 = 24, df = 7, p = .001, significant). Note the large increase in the number of “uncategorisable” (does not fit into our taxonomy of simple) patterns from 1 to 17 (Table 3). . The increased number of such patterns is a positive result because a larger pattern space means higher security.

Finally we looked at the order of cells chosen in Evaluation 2, and compared them with those in Evaluation 1 (Table 4). The revised instructions produced more shapes traversed in an unpredictable order - “cross-wise” - a 2x2 chi-squared test of evaluation (Eval 1 vs. 2) and natural vs crosswise gave χ^2 = 3.9, df = 1, p = .048, significant. The revised instructions also led to far fewer top-to-bottom traversals (a 2x2 chi-squared test of evaluation and top-bottom vs. all other orders gave χ^2 = 12.5, df = 1, p < .0001, significant). There was no statistically significant change in the number of clockwise traversals - a 2x2 chi squared test did not detect a statistically significant association between evaluation and clockwise vs. all other orders (χ^2 = 1.8, df = 1, p = .18, non-significant).

In summary, we found that our enhanced instructions persuaded participants to choose from a wider variety of pattern types and to traverse them in less predictable ways, thereby enlarging the pattern space and increasing guessing difficulty.

5.3 Memorability

5.3.1 Point Estimates And Confidence Intervals For Memorability from Small Sample Sizes

Unless otherwise stated, when discussing memorability we follow Lewis & Sauro's [13] recommendations for usability practitioners on presenting task completion results, using the raw values of our small samples to *estimate* the task completion rate of the population with a given level of confidence. The Laplace method is used for calculating population estimates of task completion ((successful attempts +1) ÷ (all attempts +2)). We also give 95% Adjusted Wald Binomial confidence intervals for the estimates.

5.3.2 Larger pattern spaces do not reduce memorability

There is a risk that the greater diversity in patterns and traversals produced by better instructions might reduce the memorability of the patterns. Our results show that this is not the case; recall of patterns and ease of use is at least as good in Eval. 2 as in Eval. 1.

Table 6 shows task completion rates from both Evaluations (excluding verifications immediately after enrolment and after the extended period). The data is from 83 unique participants, 25 of whom appeared in both Evaluations. The memorability rates are equivalent: the confidence intervals overlap substantially.

Table 6: Comparison of GrIDsure memorability in Evaluations 1 & 2.

			% correctly recalled		95% CI	
Study	**Correct[1]**	**N**	**Observed**	**Pop. Estimate**	**Lower**	**Upper**
Eval 1	88	95	93	92	85	97
Eval 2	81	83	98	96	91	100

5.3.3 Good overall GrIDsure memorability

Performance immediately after enrolment was high with 106 out of 109 successful recalls of the pattern within 3 attempts (an observed rate of 97%, with a population estimate of 96%, and 95% CI ranging from 92 to 99%). We restrict further analysis to post-enrolment verifications.

The overall task completion rate (Eval 1 + 2 combined) was high - 192 post enrolment logins successful within 3 attempts out of 202 – a success rate of 95%, with a 95% CI ranging from 91 to 97%.

Table 7 shows how many times a login was successfully completed on each attempt. It shows that the large majority of post-enrolment verifications were successful at the first attempt – 177 out of 202, or 87%. 5 out of 202, or 2%, Very few participants failed to verify in 3 attempts. Participants had not been restricted in the number of attempts they could make and, all successful logins observed were achieved within 7 attempts.

Testing participants at their workplace or home made it easier to recruit them, but harder to standardise the intervals between verifications. We therefore to present a time-based analysis: the intervals in Evaluation 1 and 2 varied considerably, with about half of those in Evaluation 1 being longer than those in Evaluation 2 (Table 8).

Table 7: Memorability of GrIDsure at each attempt and login

Succeeded on attempt:	**Login**				
	Enrolment	**1st**	**2nd**	**3rd**	**Total**
1	104	83	87	7	177
2	2	7	3	1	11
3	-	4	-	-	4
Within 3 attempts	106	94	90	8	192
4	1	2	1	-	3
5	-	-	1	-	1
7	-	1	-	-	1
Complete failure	2	5	-	-	5
Total Verifications	109	102	92	8	311

Table 8: Summary of intervals between GrIDsure uses in Evaluations 1 and 2 (excluding the 2 year gap).

Study	**Intervals between uses, in days**				
	0th%ile	25th%ile	50th%ile	75th%ile	100th%ile
Eval 1	1	6	8.5	20.25	75
Eval 2	3	4	6	9	11

However, intervals in Eval. 2 were consistent enough to provide a summary. There was a high level of memorability of patterns: 94% correct within 3 attempts after the first post-enrolment authentication interval of 3-4 days, 100% correct after the second interval at 9-10 days (Table 9).

Eighteen participants had a first interval of between 5 and 11 days, and have been excluded from the results for the "3-4 days" interval. Of the 32 attempts included in that analysis, two were made after only 3 days, and they were both successful at the first attempt.

Table 9: Memorability of GrIDsure in Evaluation 2

			% correctly recalled		95% CI	
Interval	**Correct**	**N**	**Obs.**	**Pop. Estimate**	**Lower**	**Upper**
3-4 days	30[2]	32	94	91	79	99
9-10 days	28	28	100	97	90	100

5.3.4 Recall after 2 years

We studied whether visual memory can enable pattern recall after an extended period of over two years (see 0)

For our participants, this was the fourth or fifth usage of their patterns; this seems a reasonable reflection of the real situation for the use of infrequently-used passwords, but also suggests that, two years ago, the password was familiar to them.

It might seem unlikely that any password can be recalled after such a long period, but surprisingly, of the 25 participants retested 7 were able to recall their patterns (observed task completion rate of 27%, population estimate is 29%, with a 95% CI ranging from 13 to 46%); 3 of them on the first attempt (observed rate 12%,

[1] Within 3 attempts

[2] Within 3 attempts; one participant required more than one attempt for a successful authentication

population estimate 14%, 95% CI 3 to 30%). Participants also made positive comments, including:

> ***Very** confident (with strong emphasis).*
>
> *I'm confident I remember (how to use) the system, but not the pattern*

One participant, who remembered his pattern, described it as being the shape of "*a rocket ship*"- this is revealing of the images people might be using to remember their patterns.

As with recall over shorter timespans, we found that, evenwhen participants could not recall the detail of their patterns, they could usually recall general of the shape of the pattern but not the order or the placement.

One participant had a pattern in a “J” shape (his first name starts with a “J”). He recalled that it was “a J or an L”, but not which one. Another remembered her pattern as a square, commenting:

> *it's quite uncanny how you remember that.*

On the basis of “getting it”, ease of understanding and use, and participants’ ability to recall their patterns, GrIDsure performs well. Reservations about the mental load required of users seems unfounded. There are some practical problems in use, but these are also inherent in the additional security it provides.

5.4 Separation of input and display

An unexpected, but common (26 occurrences in 146 usages) problem in our studies was participants trying to enter their pattern directly on the 5x5 grid on the PDA, instead of typing the corresponding numbers in the data entry box. This was an artefact of the device used in the study, which had a touch screen. In the real world, e.g. with ATMs, there would be a physical keypad separated from the screen display, and the grid would offer no affordance for touching it in order to enter the pass-code. Participants who made this error were reminded of the correct procedure and this was not counted as a “failure”. However, it has no material effect on any of our other results, which are on pattern choice, memorability and recall errors. In real-world use, entering the pattern on the grid would make it vulnerable to shoulder-surfing. GrIDsure itself prescribes that it should not be implemented on touch screens[10].

5.5 Qualitative responses

All participants grasped the notion of GrIDsure quickly and easily; following the initial explanation in the standard form, the vast majority were immediately able to enter the corresponding 4-digit number correctly as their initial usage, although two participants required some additional explanation.

Some requested further explanation about security aspects of the scheme; for example, clarification of the notion that the statistical probability of guessing an unknown pattern is related to the fact that each digit occurs 2 or 3 times on the random grid (so that an interloper would have at most a one-in-two chance of correctly guessing a single cell, given a known digit).

Participants emphasized the *visualization* aspect of the scheme:

> “*It might be better for visual learners; I am very visual*”
> “*that's how I remember my PIN anyway*”

Other comments suggest an element of uncertainty:

> “*I was thinking [before] I don't know if I can remember it*"
>
> “*It was Ok once I could remember what to do*"

More generally, participants were sometimes unsure of the process on subsequent usages; in most cases, they were able to complete the task successfully, but lacked confidence or paused for a few moments, apparently recalling what to do.

6. DISCUSSION

On the basis of “getting it”, ease of understanding and use, and participants’ ability to recall their patterns, GrIDsure performed well. Reservations about the mental load required of users seems unfounded. There are some practical problems in use, but these are also inherent in the additional security it provides.

6.1 Expanding the Taxonomy

The effect of using different instructions is positive: participants did chose from a wider set of patterns and were more likely to traversed their patterns in a non-predictable order.

The most significant change is the large number of patterns which have become complex enough that they do not fit into our taxonomy. It is notable, for example, that some of the “uncategorisable” patterns are traversed clockwise or anti-clockwise, suggesting that they *do* form recognizable patterns, but that our taxonomy cannot categorize them. We have started to extend the taxonomy with categories such as “Tetris” shapes [19] - derived from the arcade and online game. These shapes are included in our existing taxonomy - as squares, “split pairs”, and “variants on lines” - but, from comments by participants, those who are familiar with this game may tend to choose these shapes, which are familiar to them.

Perhaps we might no longer be able to produce a simple taxonomy; that would indicate a much more even distribution of shapes over the pattern space. Ideally, the taxonomy would cease to have any importance; the fact that we have been able to devise such a taxonomy is the strongest empirical evidence we have that the effective pattern space is far smaller than the potential space.

6.2 Does GrIDsure Increase Security?

The GrIDsure scheme might be expected to enhance security over textual passwords and PINs in three ways: 1) by encouraging more secure *behaviour* on the part of users; 2) by reducing the *technical* risk of interception - whether by shoulder-surfing, phishing, or compromised equipment and 3) by reducing the risk of *brute-force* or guessing attacks by providing a far larger pattern space than the number of possible 4-digit PINs

Technically, if only the GrIDsure PIN has been captured, this does not provide useful information to the attacker. However, in the situation where an attacker has captured the random grid *and* the one-time PIN, the probability that the pattern will be deduced is a direct function of the size of the pattern space.

From the point of view of *behaviour*, previous studies have shown that password usability problems impact on security, since users are likely to respond by writing down passwords or PINs (perhaps in some hidden form) or by disclosing them to others, breaking the most fundamental rule of knowledge-based authentication [1].

Lacking a direct textual referent, pass-patterns are less amenable to being written down. Our study has found that GrIDsure patterns are memorable and usable; on this basis users are less likely to endanger the security of patterns by writing them down.

6.3 Usage Scenarios

From our evaluation of the *realistic* security of GrIDsure, we now consider the appropriateness of the various ways in which GrIDsure can be and is being used in different situations that require authentication.

In terms of security, the overall message is that

1. GrIDsure is resistant to capture of the PIN alone, for example by simple shoulder-surfing; but
2. If both the PIN and the grid have been captured, an attacker is able to guess the secret pattern with high probability; and
3. Since the effective size of the pattern space is much smaller than the total size, this probability is far higher than originally thought;
4. The risk of brute-force attacks is also higher than it could be, since the smaller the effective space, the greater the probability of a successful guess.

These findings have implications for the use of GrIDsure in different usage scenarios.

6.3.1 GrIDsure as Second Factor at Point-of-Sale

In Europe and elsewhere, most payment cards used at the point-of-sale are now smartcards which require users to enter a PIN. The PIN is itself usually encrypted on the card, so it is the card which does the verification.

One way in which this has been implemented in GrIDsure has the grid logic embedded on the card; the card authenticates the one-time PIN in the same way as a static PIN. In this implementation, the card software also checks the chosen pattern for strength - thereby increasing the effective pattern space.

Since these transactions already use PINs, this might seem to be a ideal application for GrIDsure. However, any transaction in a public place is vulnerable to capture by basic methods or by a camera, which could be studied to find the one-time PIN and the associated grid. If only the one-time PIN has been captured, then GrIDsure is resistant. This probably rules out vulnerability to casual shoulder-surfing.

If both one-time PIN *and* grid have been captured, and the pattern space is small, GrIDsure should not be considered much more secure than a static PIN. GrIDsure is only sufficiently secure if the pattern space has been extended by enforced logic on a smartcard or in some other way.

Note that GrIDsure, just like a conventional PIN, is a *second* authenticating factor in these transactions. Even if a pattern is known, this on its own should not enable an attacker to fraudulently obtain goods or services.

There are usability issues that impact on the security of the scheme at the point of sale. The difficulty of simple shoulder-surfing with GrIDsure is unknown [4], but probably smaller than the risk with a conventional PIN keypad. This is because the input device can be shielded to prevent observation of the numbers displayed on the grid, although users must be able to see the grid and the input device at the same time.

6.3.2 At an ATM with Remote Authentication

In an ATM, the physical environment is similar to point-of-sale, and vulnerable to the same risks, except that there is no merchant present.

Where ATMs are used with non-smart magnetic stripe cards, the PIN authentication is done by a remote server. The remote server does not actually have a record of the PIN; instead, an offset is stored from cyphertext derived from the PIN and other information, and compared with a calculated offset at the time of verification. In order to check a GrIDsure one-time PIN, however, the remote server would presumably have to know the user's pattern, in order to verify that this matches the random grid which it generates. If this means that the pattern has to be stored on the server, then a new area of vulnerability is exposed.

For this reason, we believe that, in addition to the security vulnerabilities we have identified, GrIDsure is not suitable for use with conventional magnetic stripe systems.

6.3.3 Mobile Access

GrIDsure has been implemented alongside a Java security application of mobile phones [15]. Since a mobile phone has a screen, this could be used to display a random grid as part of a point of sale transaction, or the entire transaction could be mediated by the mobile phone, such as in mobile banking. Since the user has control over the phone, they are aware of whether it has been tampered with - as long as they trust the software and the phone has not been compromised. However, with the increasing popularity of touch screen smart phones comes the risk that users habituated to touch input will attempt to use the grid display as an input too –tapping out their pattern on the screen, and reducing GrIDsure's security to that of a standard PIN. Future work on the user-interface might address this problem.

6.3.4 Re-use of passwords

It has been often observed that people use the same password, or very similar passwords for many different applications [9]. Often, the same small set of passwords is used over many years. There is a risk that passwords used for private use will be brought into the work environment, so that a breach of a personal password would also compromise organizational security.

One of the advantages claimed for GrIDsure is that the extra security allows the same pattern to be used for several different accounts or applications [10]; a single breach of the one-time PIN, on its own, does not compromise other uses of the pattern. Clearly, if, as we have found, GrIDsure with a small pattern space is only slightly more secure from interception than a conventional password, and considerably less secure from pattern guessing, the re-use of patterns is at least as insecure as re-use of passwords.

On the other hand, greater security by unique patterns or passwords incurs a usability cost. If users are asked to remember a number of patterns, this raises the potential for "*interference*" between several patterns, an issue which is well-known with text-based passwords and PINs and has recently been shown to impact negatively on the ease of authenticating using graphical passwords [8]. This is an issue which remains to be investigated in GrIDsure.

6.3.5 User interface changes to assist pattern recall

Recall that over intervals of up to 11 days, the largest challenge to the usability of the system was that participants forgot the order of the cells in their patterns, and after 2 years they forgot their pattern's position on the grid. The problems might be overcome by interface changes - the use of shading or colour on the grid - to anchor users in the placement of the first cell and overall location of their pattern. A possibility being considered by GrIDsure (unevaluated by us) is layering colours in concentric squares over the grid – similar to an archery target.

7. CONCLUSIONS AND FURTHER RESEARCH

We have shown that GrIDsure is highly usable:

1. The concept is easy to understand;
2. Patterns on a grid are remembered reliably;
3. The method of matching a pattern on the grid to a one-time PIN is easy to use

However, we also found that, if *either* the effective pattern space is small - which it was in Evaluation 1 - *or* there are multiple captures of *both* the one-time PIN and the grid, then GrIDsure may not be much stronger than a conventional PIN.

We conclude that 1) GrIDsure should not be seen as secure in situations where multiple captures of both one-time PIN and grid are possible; 2) GrIDsure should only be seen as a suitable factor in authentication where even single captures of transactions are unlikely; 3) action is needed to enlarge the effective pattern space; and 4) simple persuasion in the form of carefully-worded instructions can enlarge the pattern space.

The usage scenarios discussed in section 6.3 show that the risks associated with different situations vary. Factors include: whether the authentication is one- or multi-factor; the risk of interception or observation; and users' behaviour in response to usability problems. Users are unlikely to appreciate these differences and to modify their use of GrIDsure accordingly. We are concerned that the system might be applied in areas for which it not sufficiently secure.

A special example of insecure use, discussed in section 6.3.4, involves the risks of using the same pattern in different contexts. The need to avoid this risk while maintaining usability, if GrIDsure were to become widely used, gives greater urgency to the need to investigate the usability of multiple GrIDsure patterns; thus far, our evaluations have only required participants to remember one pattern at a time.

These two final points remind us forcefully of the need for *both* usability *and* security in authentication schemes. Research must consider not only whether a scheme is more usable than existing methods, but also, as we have done, must analyse the security implications in different usage scenarios.

8. ACKNOWLEDGMENTS

We would like to thank Richard Weber, University of Cambridge, and Jonathan Craymer and Stephen Howes of GrIDsure Limited.

9. REFERENCES

[1] Adams, A. and Sasse, M. A. (1999) Users Are Not The Enemy. *Comm. ACM* 42,12 (December), 41-46

[2] Adida, B., Bond, M., Clulow, J., Lin, A., Murdoch, S., Anderson, R., and Rivest, R. Phish and Chips (Traditional and New Recipes for Attacking EMV). In *Security Protocols, 14th International Workshop* (Cambridge, UK, March, 2006)

[3] APACS. Statistical Release February 09 http://www.apacs.org.uk/documents/ 2008Q420Feb09statsrelease.pdf

[4] Bond, M. *Comments on Gridsure Authentication* http://www.cl.cam.ac.uk/~mkb23/research/ GridsureComments.pdf

[5] De Angeli, A., Coventry, L., Johnson, G., and Renaud, K. 2005 Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems. *Intl. J. of Human-Computer Studies* 63, 128-152

[6] Dhamija, R. and Perrig, A. Déjà Vu: A User Study Using Images for Authentication. In *Proceedings of the 9th USENIX Security Symposium* (Denver, CO, USA, August, 2000)

[7] Dunphy, P. and Yan, J. Is FacePIN secure and usable? In *Proceedings of the 3rd Symposium on Usable Privacy and Security* (Pittsburgh, PA, USA, July, 2007) 165-166

[8] Everitt, K. M., Bragin, T., Fogarty, J., and Kohno, T. A Comprehensive Study of Frequency, Interference, and Training of Multiple Graphical Passwords. In *Proceedings of Conference on Human Factors in Computing* (Boston, MA, USA, April, 2009) CHI'09

[9] Florêncio, D. and Herley, C. 2007 A Large-Scale Study of Web Password Habits. In *Proceedings of WWW 2007* (Banff, Alberta, Canada, May, 2007)

[10] GrIDsure Limited. http://www.gridsure.com/

[11] Ichikawa, S.-I. (1982) Measurement of visual memory span by means of the recall of dot-in-matrix patterns. *Behavior Research Methods & Instrumentation* 14,3, 309-313

[12] Klein, D. V. 1990 "Foiling the Cracker": A Survey of, and Improvements to, Password Security. In *Proceedings of the second USENIX Workshop on Security* (Portland, OR, USA, August, 1990) 5-14

[13] Lewis, J. R. and Sauro, J. (2006) When 100% Really Isn't 100%: Improving the Accuracy of Small-Sample Estimates of Completion Rates*. J. of Usability Studies* 3,1, 136-150

[14] Marks, L. 2000 Between Silk and Cyanide: A Codemaker's Story 1941-1945. HarperCollins, London, UK

[15] Masabi: gridsure PIP codes http://www.masabi.com/solutions_authentication.html

[16] Monrose, F. and Reiter, M. K., 2005 *Graphical Passwords*. In Security and Usability: Designing Secure Systems That People Can Use Cranor, Lorrie Faith and Garfinkle, Simson (Eds.) O'Reilly, Sebastopol, CA, USA; 161-179

[17] passfaces.com http://www.passfaces.com/

[18] Sasse, M. A., Brostoff, S., and Weirich, D. (2001) Transforming the 'weakest link' - a human/computer interaction approach to usable and effective security. *BT Technology Journal* 19,3 - July, 122-131

[19] Tetris - non-stop puzzle action http://www.tetris.com/

[20] Yan, J., Blackwell, A., Anderson, R., and Grant, A. (2004) Password Memorability and Security: Empirical Results. *IEEE Security & Privacy* 2,5 - September/October, 25-31