A Multi-Word Password Proposal (gridWord) and Exploring Questions about Science in Security Research and Usable Security Evaluation

Kemal Bicakci
TOBB University of Economics and Technology
Ankara, Turkey
bicakci@etu.edu.tr

Paul C. van Oorschot*
School of Computer Science
Carleton University, Ottawa, Canada
paulv@scs.carleton.ca

ABSTRACT

Our agenda is two-fold. First, we introduce and give a technical description of gridWord, a novel knowledge-based authentication mechanism involving elements of both text and graphical passwords. It is intended to address a new research challenge arising from the evolution of Internet access devices, and which may arguably be viewed as motivating a new paradigm: remote access password schemes which accommodate users who alternately login from devices with, and without, full physical keyboards (e.g., users alternating between desktops with easy text input, and mobile devices with tiny or touch-screen virtual keyboards). While the core ideas behind gridWord are well-formed, and may be viewed as a new variation of old (text-based) ideas of building passwords from multiple words, many aspects including recommended parameterization and configuration details, preferred platforms, and primary targets of application remain to be explored in detail. We nonetheless solicit early feedback from the community for several reasons, related to our second agenda item: we use gridWord as a concrete target to focus exploration of a number of questions involving (a) the evaluation of usable security proposals, (b) the often conflicting objectives of various parties involved in the publication of academic research, and (c) the relationship between the design and publication of new security mechanisms and the pursuit of scientific knowledge through experimentation. We believe the second agenda item is important to pursue, given our observation that experts in usability and security have widely varying expectations, and lack consensus on what is important for the evaluation, comparison, and publication of usable security proposals.

*Corresponding author.
Submitted April 11, 2011; revised October 27, 2011.

NSPW'11, September 12–15, 2011, Marin County, California, USA.

Categories and Subject Descriptors

K.6.5 [**Management of Computing and Information Systems**]: Security and Protection—*authentication*; H.5.2 [**Interfaces and Representation**]: User Interfaces—*graphical user interfaces*

General Terms

Security, Human Factors, Design, Experimentation

Keywords

passwords, usable security, evaluation, science

1. INTRODUCTION

Internet authentication continues to be dominated by text passwords, for reasons discussed elsewhere [21]. (For low levels of security, PINs as short as four digits are also common.) Text passwords have a long list of known problems, and new proposals appear regularly—all too frequently, some researchers would say. Indeed we present yet another proposal, albeit packaged in a substantially different way: rather than presenting a new proposal and advertising its merits accompanied by a self-evaluation (naturally positioned as objective, but no doubt containing the inherent biases of its proponents), we present the technical details of a new password scheme and use it as a concrete example from which to pursue a number of questions which in our perception the research community continues to struggle with, related to the introduction and evaluation of such proposals.

We are not only interested in views on how to proceed with completing the design and evaluation of the proposed mechanism, but more generally seek (optimistically) any consensus on what the community should reasonably expect and/or demand, for evaluations accompanying research proposals both in the narrower area of knowledge-based authentication mechanisms, and in the broader field of usability and security. Our own experience suggests surprisingly little consensus on how to properly evaluate usable security proposals, even after substantial growth in the security and usability community, most recognizably as an outgrowth of the annual Symposium on Usable Privacy and Security beginning in 2005. We feel that the lack of resolution of such questions poses roadblocks to the advancement of numerous areas of experimental and empirically-based computer security research involving users. We see this as related to

questions about the lack of science (and scientific experimentation) within security research, as discussed at NSPW 2010 by Maxion et al. [28] and in Longstaff's essay [27].

Regarding our proposed mechanism itself, *gridWord* is a hybrid scheme combining elements of text and graphical passwords, which systematically builds on principles from other recent proposals, and allows choice between text and screen-based (e.g., touch- or stylus-based) input. We believe this latter point makes gridWord of independent interest due to the following issue. The increased popularity of smartphones and mobile devices having less-friendly text input modes than those of desktop computers with full keyboards raises a new and largely unexplored research challenge: the design and deployment of user authentication mechanisms alternative to ordinary text passwords, and which can accommodate a user who accesses the same web site alternately from devices with full physical keyboards and from other devices with less-friendly text input modes. Motivated by diverse (and contradictory) expectations conveyed by referees on previous password mechanism proposals, our approach introducing gridWord herein differs from the usual method of reporting completed work: our evaluation is near its start, and we solicit a priori input on what would be suitable, reasonable, and sufficient evaluation evidence to provide, with the goal of helping consolidate community views.

For example, we are interested in the community's opinion and guidance on what types of evaluation would be acceptable for (a) a single preliminary paper, and (b) a full (possibly multi-paper) exposition and evaluation of gridWord. Here there are numerous conflicting constraints: page limits in conference proceedings and many journals constrain presentation on one hand, even while referees request greater explanation of background, methodology, datasets, implementation and user study details sufficient to independently reproduce results; sound scientific arguments favor systematically pursuing multiple carefully executed experiments controlling single independent variables to confirm or refute specific hypotheses [35], yet there is little visible appetite among computer scientists for the publication of a corresponding set of experiments on a single proposed mechanism. What some argue is systematic exploration supporting accumulation of "reusable scientific knowledge", others lump in with resumé-padding exercises and incremental results. Such divergent views may arise due to lack of reviewer time, lack of subject-area expertise, or differing tastes.

These are difficult questions related to how to carry out scientific research. We suggest they deserve greater discussion within the context of usable security. Our use of gridWord as a concrete example with evaluation in-progress is intended to stimulate discussion, contribute to the appreciation of the many different perspectives, and possibly even lead towards agreement on some matters.

2. GRIDWORD: MOTIVATION AND DESIGN OVERVIEW

Our exposition is facilitated by first giving an outline of gridWord itself, and then considering the motivation with the benefit of context from knowing the design properties.

2.1 Design Overview of gridWord

A gridWord password consists of an ordered set of distinct words chosen from a pre-determined list. (There may be advantages to using words corresponding to "concrete" objects for which visual images are easily formed, e.g., train; studies show their retrieval from memory is far better than abstract words [33].) The login user interface includes a username text field, a set of "combo" boxes (one box for each word) constituting the password, and a 2D grid (see Fig.1). The combo boxes allow the user to either type a word or choose from a (e.g., drop-down) word list. Auto-complete is provided so that (e.g., desktop) users can conveniently enter the password by typing only the first few characters of a word and then hit enter to complete the chosen word. (This does not reduce security as the entire list of possible password components is already available.) Below the combo boxes is a 2D grid composed of numerous cells, with a one-to-one static mapping between words and cells so that a user's password components remain in fixed places from which they can be entered (e.g., click-entered). This design is intended to allow users to leverage spatial memory to find the correct words. On a desktop, users can search and see which word is assigned to a cell by *local exploration*, by pointing the cursor to a cell; on touch-screen smartphones, implementation may allow users to perform this local exploration, for example, by dragging a finger across the grid and lifting it on the correct cell. Each word displayed as a result of local exploration serves as a potential memory cue (or perhaps more as feedback) to corroborate a correct cell location, and vice-versa. Selecting a cell automatically enters its associated word into the corresponding text (combo) box.

User registration involves the user first entering a username (plus any additional required personal information), and then creating a password. To address user choice issues, password creation involves system-suggested passwords (see Fig.2) consisting of n specific cells each with its fixed associated word; as discussed below we consider for illustration $2 \leq n \leq 5$. The user can either accept the suggested password or "shuffle" to get a new suggestion. The idea is to make the selection of a secure (random) password the path of least-resistance [10]. To disallow users from hunting for "hot-spot" (popular) words individually, the system only provides suggestions of full sets of n cells (correspondingly, n component words). Once a user accepts a suggested password, she is directed to a confirmation page (not shown in figure) similar to the login screen of Fig.1. Password creation completes upon successful password confirmation, i.e., upon the user correctly re-entering the password.

EXPECTED USE OF INTERFACE OPTIONS. The above design is motivated by existing differences between available input modes on different types of end-user devices. While user testing remains, our initial expectation is that the interface options will be used to enter passwords as follows.

- Users with physical keyboards (e.g., desktop users): as primary mode, typing into combo boxes (with auto-completion aiding entry).

- Users with touchscreens (e.g., smartphones, tablets): as primary mode, entry by selecting patterns cells (with local exploration optionally used to improve accuracy or confirm memory of spatial pattern); and as secondary mode, entry by choosing words from word lists ("spinning the combo boxes"). Pulldown lists are likely to be used more if the lists are shorter (with correspondingly more combo boxes), than vice-versa.

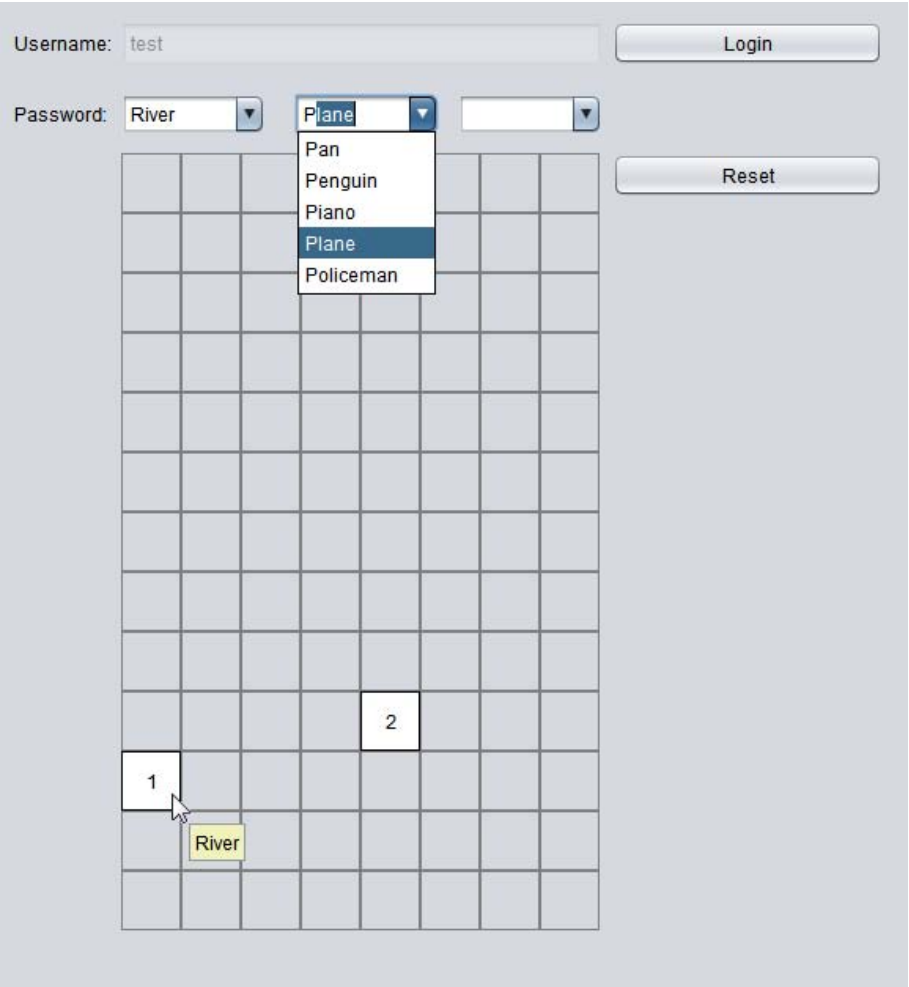

Figure 1: Login interface of gridWord with three words

2.2 Motivation for gridWord

Compared to a full-size physical keyboard common on desktops and laptops, many users find text input on smaller devices (e.g., smart phones and tablet computers which increasingly have soft screen-based keyboards) to be less user-friendly; even miniature physical qwerty keyboards challenge non-experts. The issue is exacerbated for entry of passwords which by historical policy may require mixed-case and special characters, which are harder to locate or take multiple keystrokes on many virtual keyboard layouts.

While upon initial thought there seems a natural match of graphical passwords to mobile devices—e.g., the Android 9-dot login pattern, offering PIN-level security, has become popular for screen-unlocking the local device itself—the issue is more complicated for web site (remote access) passwords, as it is common for a user to alternately access the same password-protected service from her smart phone and desktop machine. As such, ordinary text passwords remain the default mechanism due to historical installed-base issues on the system-side, as well as on the client side (use of physical keyboards on the desktop remains popular, and is often still the original access device from which account registration or password creation is done). As the number of services accessed by mobile devices increases, there will be an increased need to support authentication alternately from these two classes of devices—one with and one without full-size physical keyboards supporting easy input of mixed-case text.

Thus a first motivation for gridWord as a new knowledge-based authentication scheme is that by providing different modes of input producing the same password output to the system, it aims to offer a password system convenient both on devices with full-size physical keyboards and on those for which input of arbitrary text characters is more difficult. Users may choose to enter passwords based on keyboard input or screen-based selection. The text input option allows a password consisting of (e.g., three) ordered words to be entered into three input boxes via keyboard on a desktop machine. Alternately, on touch-screen or stylus-input devices, gridWord supports graphical input through grid cells being selected (optional stretch-and-pinch functionality may enlarge screen portions to explore the grid).

As a second motivation, gridWord potentially offers security and usability advantages over other systems as a standalone scheme in either a desktop (full physical keyboard) or a touch-screen environment. A best-in-class comparison target is PCCP [10].

A third motivation to explore gridWord, and one making it potentially of both academic and practical interest, is that

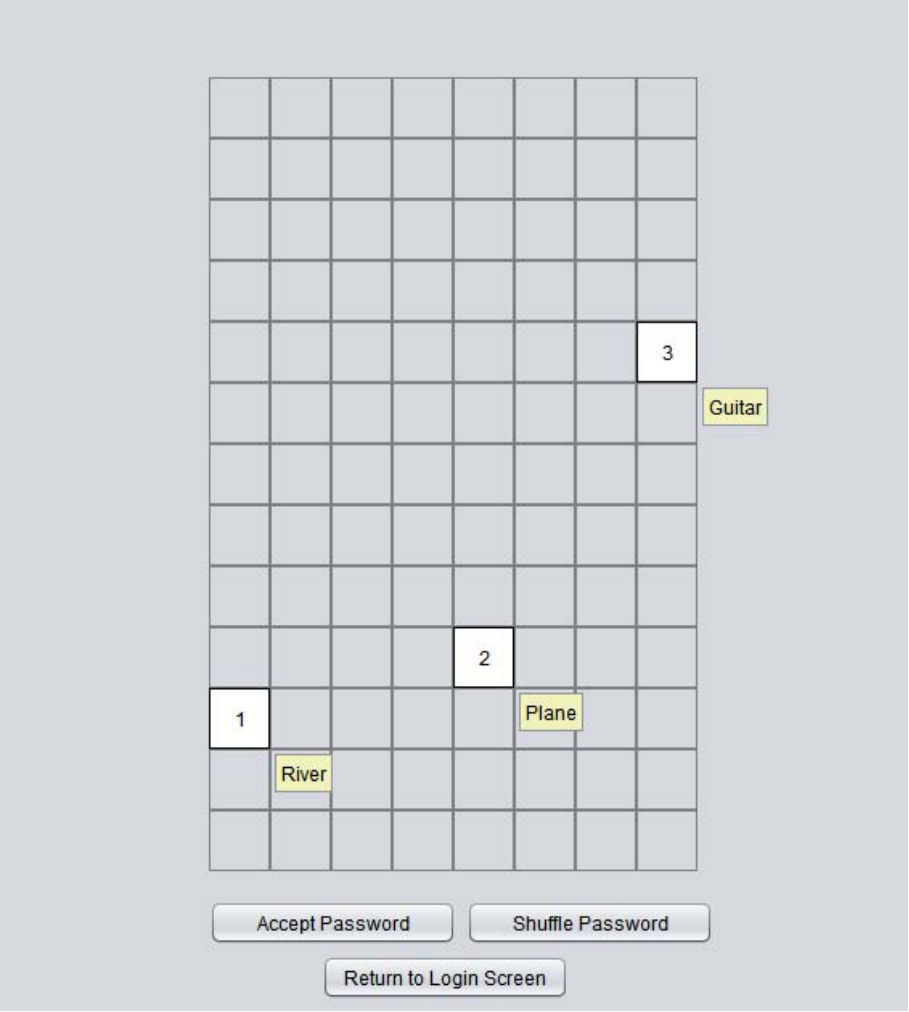

Figure 2: Password creation interface of gridWord with three words.

it may ease a transition to graphical passwords, as now detailed. Among many potential roadblocks to the adoption of graphical passwords are:

(a) inertia—despite frustration and overload, users seem willing to continue typing text passwords;

(b) the lack of a compelling reason for users and sites to move away from text passwords—the research community has failed to provide sufficiently convincing arguments that graphical passwords offer combined security and usability advantages over text passwords; and

(c) the reality that changing password systems requires, in most cases, that users create and remember new passwords after first learning how the new password scheme works, as well as usually requiring changes to system-side authentication infrastructure.

If, for the sake of argument, one assumes that best-in-class graphical password mechanisms do actually offer security and usability advantages—which provides missing motivation per item (b)—then gridWord may help overcome users' resistance to change per items (a) and (c), by softening the transition: the supported modes of password input include the one users are already most familiar with, and transition between two modes requires neither user memorization of a new password, nor the change of back-end systems (as gridWord passwords can be represented as text passwords). Changes to front-end systems may be handled for example by browser extensions or co-operating web sites. Users may opt to continue their existing habit of typing passwords on their desktop systems, and on mobile devices not supporting standard keyboards they can enter passwords either via the substitute keyboard mechanisms (e.g., miniature physical keys, stylus-based keyboards, touch-screen keyboards) or by screen-based selection. While the mode of password entry on the mobile devices would differ from text input, the mental model of the password mechanism would already be familiar, as would the password itself.

Thus if for independent reasons the goal is to deploy a graphical scheme like gridWord, a recommended approach is to first target users to use its graphical mode when logging in from their mobile devices, without trying to change their habit of entering text passwords in desktop environments. This may preclude habit and inertia alone resulting in rejection of the graphical mode.

Advantageous properties unconfirmed. We empha-

size that the possibility (second motivation above) of gridWord as a single-environment scheme being superior to other graphical passwords schemes remains to be explored. This is also true for the first and third motivations—e.g., we presently have no evidence that selecting grid cells is "more usable" on a touch-screen smartphone than typing character-based passwords on a soft keyboard. Indeed, meaningful criteria by which to measure and compare usability and how to weight different factors (time to enter password, user affectation, success rate in terms of accuracy of password recall for passwords of equivalent guessing entropy, etc.) remain unclear, allowing the possibility that selecting grid cells is slower yet still "more usable". Our admittedly optimistic intuition, which remains unconfirmed, is that the use of principles and design aspects learned from previous graphical password schemes [5] will provide advantages that make the image-based gridWord interface on smartphones preferable to text entry thereon, for systems seeking to retain text entry compatibility due to continued use of text passwords on desktop systems. The advantages are expected to arise from the combination of properties as discussed next.

Properties as Related to Prior Work. The design of gridWord includes support for the following features, several of which build on the accumulated scientific knowledge in graphical passwords over the past ten-plus years [5]:

1. persuasion during password creation, including system-suggested full passwords which may be sequentially declined by "shuffling";
2. ability to select words from a list (e.g., dropdown list), rather than typing via keyboard input;
3. local screen exploration with potential component text words displayed on hovering over individual grid cells, providing a combination of discrete memory trigger (cue) and implicit feedback by way of corroborating expected grid location of component words; and
4. a visual component by way of the pattern of word cells on the grid, imperfect memory of which may suffice by refining foggy memory by local screen exploration. (It has been generally claimed that the visual component of graphical password schemes aids memory recall, though this claim seems difficult to verify with respect to individual design features of such schemes.)

Information displayed as part of the local exploration is intended to help the legitimate user with password recall and correctness, but not an attacker. While the system-suggested passwords alone would optimally flatten password distributions, the shuffling feature precludes perfectly equiprobable passwords; nonetheless, the design goal is that user choice issues are significantly ameliorated, due to the suggestion of full passwords. The representation of gridWord passwords as either visual grid patterns or sets of n component words facilitates password back-up and sharing: if desired, users may easily write down the text form of gridWord passwords for storage in a secure location, whereas the inability to easily do this for typical graphical passwords is a known disadvantage. (Though discouraged by security and IT experts, sharing of passwords, e.g., with colleagues to allow temporary access, is in some scenarios an important feature [39].) GridWord may be viewed as combining advantages of text and graphical passwords. For further related work, see §5.

2.3 Parameterization and Password Space

An initial gridWord design for desktop screens might use parameters chosen to facilitate comparison with the cued-recall schemes PCCP and predecessor PassPoints [43], these being among the most studied graphical password schemes to date. GridWord cells at 19x19 pixels would match the commonly studied tolerance region for these, using a total grid size 475x304 pixels would be comparable to the common image sizes (451x331 pixels) they use, and using $25\text{x}16 = 400$ cells would slightly exceed their 391 tolerance squares.

The cardinality of gridWord's theoretical password space is $P(Y, X)$. P denotes permutation, Y the number of cells (words) in the grid, and X the number of constituent words in passwords. $(Y, X) = (400, 5) \approx 2^{43}$ passwords, matching that of the studied systems mentioned above. A gridWord design goal is that the effective password space is close to the full theoretical space, due to the system suggesting complete sets of random component words (albeit allowing shuffling).

Table 1 compares various parameterizations of gridWord to the password spaces of other schemes. As shown, previous studies on PCCP and PassPoints have used 5-click passwords with a theoretical space of 2^{43}. However, evidence suggests that many users today choose passwords with much lower entropy, e.g., the Weir et al. [42] study of 32 million real passwords showed that most had (NIST formula) entropy less than 22 bits; and moreover this crude entropy approximation [31] overestimates password security. Thus 2^{43} is too high a target if aiming for "password equivalent" security to match existing passwords. Similarly, Florencio et al. [18] argue that relatively weak individual passwords of about 20 bits may withstand online attacks when lockout rules are in place. A key point here is that the attack model is important to keep in mind. Thus in practice, depending on the application, a gridWord parameterization with fewer than $X = 5$ words and from grids with fewer than $Y = 400$ cells may be interesting to consider.

Further exploration is clearly required (e.g., how number and size of cells affects pattern memorability, and ease of physically indicating cells). Pilot studies should identify feasible parameterizations of gridWord for mobile devices, with suitable security-usability characteristics; e.g., while 400 cells may be physically feasible on larger tablets this yields areas too small to easily select on common smartphone screens, where 100–150 cells is a more realistic (physical) upper limit. Table 1 includes 104-cell gridWord, which, e.g., with 3-word passwords gives a password space of about 20 bits, approximating random 6-digit PINs. As noted this may suffice for some applications limited to online attacks, but for others, e.g., if passwords are used to generate crypto keys even key spaces of 43 bits fall far short.

3. USER STUDY CONSIDERATIONS FOR GRIDWORD

From §2.2, gridWord is of interest due to three potential properties:

1. it may be a convenient authentication system for users who wish to access the same web site alternately from desktops and mobile devices;
2. it may offer security and usability advantages over alternatives as a stand alone mechanism; and

Table 1: Bitsize of password space for various schemes and parameters.

	Alphabet size (or number of cells)	Password length	Theoretical password space (base2 log of cardinality)
PINs (if fully random)	10	4 *digits* 6	13.3 19.9
text (lowercase + digits)	36	5 *chars* 8	25.9‡ 41.4‡
text (mixed-case + digits)	62	8	47.6‡
text (keyboard chars)	95	8	52.6‡
PassPoints	391	5 *clickpoints*	43.0‡
PCCP	391	5 *clickpoints*	43.0
gridWord	400 (= 25x16)	5 *cells or words* 4 3 2	43.2 34.6 25.9 17.3
	104 (= 13x8)	5 4 3	33.4 26.7 20.1

‡For user-chosen passwords, theoretical space significantly overestimates effective space (cf. [42])

3. it may ease the transition to graphical passwords (assuming such transition is deemed desirable).

Ideally, a series of suitable experiments would investigate each of these three (possibly several studies for each), to confirm or refute property-specific hypotheses. This is no minor effort, even if it were clear what experiments to run, and the community's enthusiasm for a series of papers pursuing such an agenda is uncertain at best. This raises considerable challenges in responsibly pursuing evaluation. We consider these challenges (specific to gridWord) further in the following subsections.

3.1 Properties to be Explored

Regarding the third potential property above, it seems difficult (if even possible) to design a definitive, convincing experiment to test a hypothesis such as "Using gridWord eases switching from text passwords to graphical passwords". We presently have no planned experiment to test this, but welcome suggestions.

For the first property, we appear to be in slightly better shape. However, "being convenient" is not sufficiently well-defined for scientific measurement. One way forward would be to refine this statement in terms of a specific set of criteria related to "security and usability", drawn from a review of criteria considered in prior literature. The exact criteria to factor in, and how to weight such factors to form a suitable metric, remains an open question. Again, we welcome advice and encourage further discussion on devising meaningful and detailed such criteria for comparing convenience or usability of two or more authentication systems (cf. [5]).

But this moves us closer towards the second item, and to explore it, we consider the following initial plan for a pilot user study: compare gridWord with other systems in terms of measurable criteria such as login times, success rates, etc. This brings the question: *Which other system(s)?* The obvious candidates are standard text passwords, and a best-in-class graphical password scheme. We next consider the pros and cons of each option.

If the chosen target of comparison is text passwords, then we should first answer the following questions about experiments to be conducted:

1. Will the text passwords be system-generated or user-chosen? If user-chosen: (a) How do we control for study participants choosing passwords the same as, or related to, passwords they already use and hence already find easy to recall prior to the experiment; and (b) How do we arrange that (effective) password spaces or guessing entropy of two systems are comparable?

2. How do we design an experiment that accounts for password interference? For example, newly-formed text passwords may have a greater interference effect than gridWord passwords due to the fact that users previously have memorized (only) text passwords.

3. How do we account for learnability effects? (Users have long-standing experience with text passwords, but comparable familiarity with gridWord is hard to arrange.)

4. What gridWord parameters should be used? (cf. §2.3)

All of these are important questions.

On the other hand, if the chosen target of comparison is a graphical scheme like PCCP [10, 12], the easiest path is to choose gridWord parameters that facilitate comparison to earlier published studies (deferring questions on the suitability of those parameters; again, see §2.3). An alternative is to choose new parameterizations of each that cross-calibrate security, and repeat previous experiments on both systems under similar conditions. A drawback of pursuing this path is that it fails to provide convincing evidence that the new system is demonstrably better than text passwords.

3.2 Pilot Study Plan

Considering the numerous issues above, as a strawman for discussion and/or an actual first study we outline here plans for a pilot study. The pilot compares gridWord with PCCP (see §5) which in earlier work was itself compared with PassPoints and argued to have usability and security advantages. Hypotheses to be tested include:

1. Long term login success rates of gridWord will be higher than of a comparable PCCP system.

2. Login times of gridWord will be shorter than those of PCCP on the same devices.

3. The distribution of passwords across users, as determined by cell patterns (gridWord) or clickpoints (PCCP), will be the same or slightly flatter in gridWord than PCCP—that is, similar degrees of "shuffling" are expected with perhaps slightly more in PCCP.

The pilot will follow the methodology of earlier work [10, 12] with necessary changes as given below.[1]

The pilot involves a small lab-based experiment of 10-15 participants, using standalone Java applications for both schemes developed by the same programmer to achieve a comparable look and feel. The PCCP implementation uses the picture set from earlier studies. Parameters in the PCCP and gridWord implementations are calibrated so that they have equivalent password spaces. More specifically, following the discussion in §2.3, for desktop environments we could compare gridWord and PCCP respectively parameterized with three words (on 400 cells) and three click-points (on 391 cells), for about 25.9 bits of password space each. For better suitability to smartphone screen sizes, and to better exploit any potential memorability gain from the visual cell pattern, a preferred parameterization to test for gridWord may be four words (on 104 cells) yielding 26.7 bits. Note that three words over 104 cells yields 20.1 bits.

Participants will take part in two individual sessions scheduled two weeks apart.[2] In the first session, there is a practice session first and then participants are asked to create accounts (by creating and confirming passwords) and login once using both gridWord and PCCP. A within-subject design is used due to the small number of participants. The order of password tasks is balanced between participants. In the second session, participants are asked to login by re-entering their gridWord and PCCP passwords. Each participant is asked to create only one account for each scheme. (Note that this defers the exploration of multiple password interference to future studies; recalling more than one password per system is more challenging, and arguably either more realistic, or an artificially difficult memory task, depending on the environment of use.)

The following measures for usability and security are considered (cf. [5]): login and recall success rates, times for password creation, login and recall, and number of shuffles. To evaluate and compare perceived usability and security, participants are asked to complete a post-task questionnaire. The pilot study focuses on usability. The only security measure involves a simple comparison of number of shuffles, providing a preliminary indication of the relative flatness of password distributions; however, comparable degrees of shuffling does not necessarily imply equivalent levels of security. While not detailing a complete threat model here (which notably, is essential to a full evaluation), we briefly repeat that gridWord appears most appropriate in scenarios and designs not subject to offline attacks, but suitable to face online guessing attacks mitigated by throttling.

[1]Complementing lab-based observation with field studies improves ecological validity. A web-based implementation such as the MVP framework [11] may facilitate this.

[2]Quite low success rates are reported [12] for two-week recall of six 5-click PCCP passwords per user, in an experiment intentionally designed with artificially high cognitive load.

4. DISCUSSION: USABLE SECURITY AND SCIENTIFIC EVALUATION

Evaluation of a new authentication mechanism requires deciding what types of evaluation techniques should be used, what types of experiments and analysis should be done, and what types of user studies are necessary to carry out. There is no consensus yet on the core elements for such evaluation, though many individuals and referees have strong personal ideas. As one example issue, reviewers unfamiliar with the area do not recognize that it is unreasonable to expect authors to "re-run" a user study on short notice, the way that one re-runs a suite of performance tests after modifying software prototypes—it may be more useful for such reviewers to view user studies as being closer to hardware design processes than software re-design. The challenges for this still-young research area are complicated by its interdisciplinary nature with methodological approaches differing vastly between mathematicians, engineers, computer scientists, psychologists, and cognitive researchers. It is thus unsurprising that we lack research community consensus on what is reasonable for evaluating new user authentication mechanisms.

As another issue, one type of criticism we have seen against previous proposals is referee comments along the line of: "The new scheme is not demonstrably better than text passwords (in all aspects)." One of the present authors has the following first reaction to such a response: *Well, if our mechanism was demonstrably better than text passwords, wouldn't we be starting a billion-dollar company instead of trying to publish a paper?* The point here is, the expectation seems inappropriately high, given that text passwords have been by far the dominant means of computer and Internet authentication for 40-plus years, arguably appear destined to continue to be for the foreseeable future, and have a major advantage in any short-term comparison in that many users have had years if not decades of training on them, compared to virtually no training on the new mechanism.

A fair question to ask in response is the following. *Is there no valuable scientific knowledge or research contribution to be had in publishing results about a new practical authentication mechanism unless the overall mechanism is demonstrably superior to text passwords?* That seems too high a bar; but the harder question is, what lower bar is reasonable?

4.1 Specific Questions for Discussion

Regarding scientific evaluation of usable security proposals, many conflicting and open issues arise in considering broad questions on issues such as reasonable expectations of research papers, and types of evaluation to be carried out. Here we isolate and label specific questions to facilitate discussion and feedback at NSPW 2011 and from other readers.

Q1: What level of quality or superiority over alternatives should be shown for new usable security proposals to be considered worthwhile literature contributions?

A narrow version of this question is briefly introduced in the preamble of §4.

Q2: How is usability research incented and rewarded by the peer community, relative to other areas of computer

Table 2: Screen size and resolution of selected smartphones and tablets.

Device Model	Diagonal	Pixels	Pixels/in.
HTC (Dream, Legend)	3.2”	320 x 480	181
RIM BlackBerry Torch 9810	3.2”	480 x 640	253
HTC Touch Diamond2	3.2”	480 x 800	292
Google Nexus One, HTC Droid Incredible	3.7”	480 x 800	252
RIM BlackBerry Torch 9850, 9860	3.7”	480 x 800	253
Samsung Nexus S (SAMOLED, LCD)	4.0”	480 x 800	235
Samsung Galaxy S (I9000)	4.0”	480 x 800	233
HTC Sensation	4.3”	540 x 960	256
Apple iPhone4	3.5”	640 x 960	326
Samsung Galaxy Tab, RIM PlayBook	7.0”	600 x 1024	170
Apple iPad2	9.7”	768 x 1024	132
Samsung Galaxy Tab 10.1	10.1”	800 x 1280	149

and Internet security research? Is this aligned with the historical goals and methods of scientific research?

Similar questions raised in the NSPW 2010 panel [28] and by Longstaff [27], remain far from resolved, and have been considered generically or relative to network security and intrusion detection more than specifically to usable security or authentication proposals.

Q3: How much evaluation suffices, or is reasonable to expect, for papers introducing new usable security mechanisms? Is this comparable to other areas of security?

In a first paper describing a new usable security mechanism, how much emphasis should be given to the mechanism itself and a real-world threat model, vs. a focus on one or more user studies? Should we follow the computer scientist's preferred approach of describing a mechanism, building a prototype, and evaluating performance; or condense these to allow greater emphasis on explaining the methodology and results of one or more scientific experiments [27, 28]?

Many formally refereed venues have page constraints; no single paper (or even series thereof) can answer all questions, in full breadth and depth of exposition. Often the number of imaginable studies is countless; a referee can always ask for more. The mobile world brings additional issues: consider multiplying the typical dimensions of common mobile devices, the number of screen resolutions, and the number of studies possible for desktop machines. Table 2 gives screen size and resolution of selected smartphones and tablets; desktop screens are commonly 72-96 dots per inch (dpi or pixels/inch), with resolutions from SVGA-standard 800x600, to more common XGA-standard 1024x768, with SXGA-standard 1280x1024 or higher preferred by some users. Mobile devices also vary by input means, e.g., mini physical and onscreen keyboards.

Q4: How can we resolve conflicting priorities and views of various stakeholders involved in producing or consuming research? What does each seek in publications?

Expectations of stakeholders vary widely, e.g., authors (undergrad, grad, junior faculty, tenured professors), reviewers, peer academic researchers, industrial research colleagues (start-up, large-cap), user groups, etc., and their incentives and rewards are not always aligned.

Q5: How can we meet different stakeholder discipline goals (e.g., preferred publication venues, requirements)?

This is complicated by the interdisciplinary nature of usable security research; the venues themselves may have vastly different requirements and expectations, some sub-disciplines require real-world deployments but others do not, etc.

Q6: Can we promote research better enabling independent reproduction of results to confirm validity?

Q7: Can we better resolve time-to-publish conflicts?

Researchers often wish to circulate their research as quickly and broadly as possible, for example by posting papers on the web, while professional and corporate organizations may seek to hold publication copyrights in conflict with the distribution goals of authors, or delay publication or dissemination in order to pursue patents or maintain trade secrets.

Q8: How would we best apply scientific method ([35]; cf. [16, 27, 28]) to explore gridWord? What types of detailed experiments are of research interest, with what parameters, how many participants, for what durations?

Possible types of user studies range from field, lab, and web studies to Mechanical Turk studies [1].

Both success and frustration in carrying out and publishing research should be expected in the area of security and usability, as in any other research area. However we believe there is considerable benefit to be had from an open discussion of how better to carry out research in this area such that the results and publications add value to the scientific community. Items to keep in mind include:

1. usable security remains a relatively new academic research area, so the support community is still growing and learning; and
2. interdisciplinary research brings customary challenges including falling between well-defined disciplines and their publication venues. For example, usable security has obvious ties to HCI, but the traditional HCI community values new techniques in user studies more than results specifically related to security. On the other hand, the traditional computer security community is interested in security tools and mechanisms, but has less familiarity and appreciation for user study and data analysis techniques and methods.

4.2 Suggestion: Competition to Evaluate Real-World Usable Security Mechanisms

Three barriers to adopting a scientific approach to information security have been highlighted previously [27]:

1. the time-to-publish effect on academic researchers;
2. expectations and approach biases of peer reviewers (computer scientists typically give more weight to prototype implementation of new "clever" ideas and differentiating them from related tools, than to experimental design and supporting scientific evidence); and
3. implicit expectations of a breakthrough in every paper.

All of these barriers are major and real concerns. The transition to a more scientific approach in security research, if it occurs, is more likely to be through evolution than revolution. In this section, we propose a specific idea towards evolution. It may be considered a thought experiment to drive discussion or comparison (e.g., advantages and disadvantages) with the way we currently referee papers in the area of knowledge-based authentication methods.

First we give an overly simplified summary of the current practice of writing and refereeing papers in this area. Consider a young researcher in her early career. She finally comes up with a new idea after a long time thinking. Not surprisingly, she wants to write a paper about it. She realizes that at least a preliminary usability evaluation is expected for a publication in a reasonably good conference. The details of how it is done are not so relevant for the present discussion, but the user study is eventually conducted despite the researcher not having the experience or expertise required. The paper is submitted. It is not easy to wait for the notification which will be made in 8 to 12 weeks for conference paper review, not to mention 3-12 months or more for a first response on a journal paper (there is also a time-to-referee factor here, for those doing the refereeing).

On the refereeing side, we have already mentioned different perspectives and priorities of referees. The idea itself is the most important factor for some referees (this is perhaps dominant in academic security research). If the idea is novel and seems to make a worthwhile contribution, then the rest of the paper, especially experimental details, is largely unnecessary and even boring to some referees. On the other hand, a minority of referees gives much higher weight to the experiments and the use of scientific methods. Consequently, the destiny of the submission depends heavily on "the luck of the draw"—who the referees are. Regardless of how the decision turns, one thing holds true in the vast majority of cases: the user study is unlikely to be repeated by others (indeed, to date, very few experiments in usable security have been replicated). A commonly held, but less frequently vocalized, view is: *Why waste time for an experiment already done?* (See Feynman's lament [16].) Any form of independent secondary evaluation is relatively rare, as publication efforts related to such efforts are poorly rewarded when competing against results on "new ideas".

We propose the following approach as an interesting alternative world to consider. (This proposal, while independent of NSTIC [32], could become a part of that initiative.) The usable security community announces the organization of a public competition to develop the next generation knowledge-based authentication method(s). An interdisciplinary committee is formed to specify submission requirements and evaluation criteria. Long discussions occur among committe members on issues like: target password space sizes, types and sizes of mobile devices used in evaluation, etc. Reporting usability evaluation results is not mandatory for submission, but open-source implementations must be publicly available to facilitate independent evaluation by others. Similar to the surge of quality research on cryptographic algorithms after NIST announced the original AES competition, and more recently the SHA-3 competition, this competition would ideally attract the attention of many researchers of different specialties. More importantly, the competition could help overcome the barriers to adopting a scientific approach for the following reasons.

1. A multi-year competition timeline could reduce time pressures on researchers. Researchers may find it rewarding to invest time in a high-profile, well-respected long-term activity (like past NIST competitions).
2. While it is hard to eliminate (implicit or subconscious) over-weighting the "cleverness" factor of entries in such competitions, pre-establishing firm evaluation criteria would ideally preclude it from being over-rewarded relative to other evaluation criteria like experimental design and supporting scientific evidence.
3. An expanded competition timeline would ideally provide greater motivation and appetite for publications systematically exploring a broad array of aspects of each of the most promising (short-listed) candidates, and incent unbiased and scientifically rigorous third-party evaluation. (By analogy, compare the depth and extent of analysis of AES to the attention paid to homebrew crypto algorithms announced on mailing lists and non-major conferences.)

4.3 Another Suggestion: Standards

A complementary suggestion emerged from the NSPW 2011 workshop discussion: learn from the maturation path of role-based access control (RBAC) models and mechanisms. A detailed, peer-reviewed academic paper [15] proposed a U.S. government standard for RBAC, as a foundation for commercial product development and evaluation. In the spirit of ISO/IEC security models, the paper defined a core set of RBAC components with a reference model, feature set, and consistent vocabulary (prior to its publication, no single authoritative definition for RBAC existed). Objectives included to unify concepts, models, and ideas from research prototypes and commercial products, including for use in writing government procurement specifications. The effort contributed to an ANSI/INCITS standard in 2004.

The idea is thus that usable security proposals and products, in particular those targeting usable authentication (or more narrowly: password replacement proposals), could similarly benefit from a unified treatment, to facilitate mechanism evaluation, comparison, product selection or procurement. We note two distinct types of standards that would benefit usable security: requirements specifications (e.g., consistent policies or rules across web sites would facilitate tool-generated passwords); and guidelines on how to evaluate or measure usability (e.g., including how to measure user satisfaction). For more on evaluation of knowledge-based authentication mechanisms, see Biddle et al. [5].

5. BACKGROUND AND RELATED WORK

Science and experimentation. Of the rich literature discussing scientific method and progress by experimentation, we have cited only two items herein: the 1964 strong inference paper of Platt [35], and Feynman's 1974 address [16]. Selected works connecting such topics to questions in computer security research are the NSPW 2010 panel report of Maxion et al. [28], and essays of Longstaff [27] and Schell [38].

Graphical passwords. Surveys on graphical passwords are available elsewhere [5, 20]. User choice issues are well-known when users choose their own passwords in typical schemes: the distribution of passwords is then far from equiprobable, increasing the efficacy of guessing attacks. One solution is to flatten the distribution by "persuasive techniques" with the system suggesting or otherwise constraining password choices, allowing users to accept suggested passwords or request alternate suggestions, as per the "shuffle" feature in PCCP graphical passwords [10, 12]. PCCP users choose one clickpoint on each of, e.g., $n = 5$ images displayed sequentially; during password creation the system influences user choice by a viewport mechanism which highlights a randomized portion of each overall picture, temporarily limiting point selection to such portions. Persuasion in gridWord is analogous in spirit, but suggests full passwords rather than constraining individual component elements sequentially.

A graphical password scheme with some similarity to gridWord's use of multi-cell patterns is GrIDsure, wherein passwords involve a 5×5 grid, and users memorize a pattern of some subset of these (e.g., 4 of the 25) in a fixed ordering. At each login, the system displays in each cell a randomized digit, the digits varying across logins. The user enters the corresponding pattern digits by keyboard. GrIDsure has been explored in preliminary user studies [7], and security analysis has reported advantages relative to shoulder surfing [41] as well as weaknesses [6]. The main similarity to gridWord is a visual pattern of cells. Android screen-locking authentication is mentioned in §2.2. Among other examples involving visual memory as part of practical authentication is a scheme wherein users are instructed to recognize pictures prior to entering credentials to banking web sites [37].

Passwords from Words and Phrases. The study of passwords, including usability, is far from new, though many early papers appear little known. Applying methods rooted in cognitive psychology, strongly emphasizing the critical importance of passwords being *both* secure and usable, and warning (already in 1984) of the proliferation of "inexpert users" leading to poor password choices, Barton et al. [3] pursue numerous concrete "user-friendly" methods for password creation and reconstruction leveraging principles of recall and memory aids including episodic memory, private personal experience, environmental cues, and personalized translation rules. They consider, e.g., passwords formed from initial letters of words (acronyms from sentences). A circa 1982 approach related to the latter is *passphrases* [36], with the usability downside that typing long word sequences is time-consuming (frustrates users) and prone to typing errors. Passphrase-based passwords have long been promoted to protect PGP private keys; more recent related studies include Yan et al. [44] and Kuo et al. [26].

Password user studies and analysis of datasets and password mechanisms—which are enjoying renewed interest for both text (e.g., [45]) and graphical passwords [5]—are also not new, though more recent password datasets are often considerably larger [17, 42]. The 1993 paper of Zviran et al. [47] carries out small user studies on a wide spectrum of password approaches, finding that in terms of ability to recall, six approaches explored fell into two distinct groupings: pronounceable, cognitive, and associative passwords each performed much better than passphrases, system-generated passwords, and user-selected passwords. (Here, *cognitive passwords* are those now commonly called passwords based on secret questions, also called password recovery or personal verification questions; and *associative passwords* use words triggered by association with text challenge word cues.)

Building passwords from *pronounceable* words (whether dictionary or nonsense words) is another long-standing idea promoted in various forms to increase memorability and security. Implementations of pronounceable password generators are widely available, many (and a standard [30]) motivated by a PL/1 program designed for Multics by Gasser [19], or Allbery's 1988 `pwgen` program.[3] Pronounceable passwords were popularized by CompuServe in the early 1980's; anecdotal claims remain of users still today recalling their "word-salad" such passwords containing two unrelated system-assigned words separated by a special character. Jobusch et al. [25] discuss password generators, *pass-algorithms, password monitors,* question-answer approaches, human factors issues, and implementation details [24] of the 4.3 BSD UNIX password programs that originated many password rituals.

Passwords, Smartphones, Flexible Data Entry. Text passwords continue to dominate for website access, but pose significant usability issues on smartphones. Cheswick [9] recently proposed user-chosen multi-word passwords specifically for convenient entry on smartphones, e.g., with password creation selecting elements from fixed lists of "smartphone friendly" dictionary words. This would ideally be complemented by phones with spelling correction enabled on password entry and/or word-completion upon typing preliminary characters. (In contrast, spell-correction functionality is typically disabled on password entry, as historical policies and practices discourage use of dictionary words as passwords.) Earlier proposals allowing variability in how passwords are entered include the *pass-sentences* of Spector et al. [40] which focus on semantic meaning allowing variable syntactic representation, the *password-corrective hashing* of Mehler and Skiena [29], and order-independent and error-tolerant passphrases by Bard [2]; see also Brown [8], and Jakobsson's more recent *fastword* multi-word proposal [23].

Further regarding smartphones, and aside from Android screen-locking (which as noted earlier, is for local authentication of owner to device for device unlocking, rather than for remote-access authentication), Dunphy et al. [14] explore the use of recognition-based graphical passwords on mobile devices, and the hybrid object-based ObPwd scheme [4] has been prototyped for Android.[4] User authentication proposals employing special smartphone functionality, e.g., accelerometers for gesture authentication [13], or GPS for location tracking or user profiling [22], are not typically back-end compatible for alternating access from desktop computers. Other smart phone functionality is available for tasks related to security, but unrelated to our immediate work, e.g., involving camera-based functionality or shaking pairs of devices [34] for device association or to prove device possession.

[3] `http://cd.textfiles.com/itools/CISCO/TACACSD.SHA`
[4] `http://www.ccsl.carleton.ca/~mmannan/obpwd/`

6. CONCLUDING REMARKS

Regarding our specific gridWord proposal based on a preliminary prototype design and a plan for pilot testing, we solicit guidance on which design options appear most promising to pursue in detailed user studies, and how best to evaluate and improve the proposal. We hope our work stimulates further research on the general problem of designing password authentication schemes simultaneously convenient for text-based and input-limited devices—given our belief that the speed, familiarity, installed base, and other advantages of text-based passwords make them unlikely to disappear. The difficulty of entering text passwords on mobile devices may change user behavior, e.g., increasing the use of browser password saving/synchronization features, or the use of mobile apps which store their own app passwords. Therefore, a possible alternative to address the usability challenge of password entry is to better secure the access to these devices (e.g., through local device access control passwords or biometrics), with a reduced requirement for manual entry of passwords for the reason mentioned. This would increase the importance of ensuring the security of the authentication scheme used for device access control.

To make significant progress on questions such as those in §4.1—some of which were also raised at NSPW 2010—requires considerably more effort than a slice of a workshop of 35 participants representing only a small subset of subarea experts. Indeed, a workshop dedicated to these issues alone would have a full agenda.

Related to Q2 in §4.1, we see value in the suggestion [27] that one way forward is to foster a sub-community that values and rewards scientific experimentation and evaluation, as separate from technological advances favored by mainstream computer security venues. Also related, we note the following observation by an NSPW 2011 participant: many computer science undergrad and graduate programs, especially in North America, lack appropriate training in scientific methodology and experimentation. In usable security, including authentication mechanisms, we also see (much as in other areas of computer security) a lack of systematic advancement of knowledge, too much re-visiting of past mistakes, and too much "standing on the toes of giants" [28].

While previous work has asked some of the questions considered herein in the general computer security context, our narrower focus on usable authentication makes questions like "Is security not really a scientific discipline at all, but rather a category of enginering technology?" [28] easier to address. While the scientific method may not apply strongly to all areas of computer security, usable security is a sub-area that strongly benefits from, indeed requires, user studies and scientific experimentation. We emphasize a complication that applies more to human-computer interaction than to other scientific disciplines (e.g., physics): experimental results are often very strongly influenced by minute details of user interfaces, instructions to users, experimental design, and prototype implementations. Consequently, it is often very difficult or impossible to generalize results, diminishing the scientific weight of many individual results.

Regarding the difficult question of how to evaluate usable security proposals and mechanisms, whether an evaluation involves a competition run by a third party, a formal standard, or a comparison within an academic paper, the preferred way to systematically compare and evaluate is to use unambiguous criteria. At present, such criteria are missing, even for the highly studied case of authentication.

We encourage further exploration of the many open questions related to how to best evaluate usable security mechanisms, and what role the scientific approach should play in the exposition and evaluation of newly proposed user authentication mechanisms.

Acknowledgments. The second author is Canada Research Chair in Authentication and Software Security, and acknowledges NSERC for funding the chair, a Discovery Grant, and a Discovery Accelerator Supplement. We thank all those who provided comments improving this work, including anonymous referees, Robert Biddle, Sonia Chiasson, Ugur Cil, Alain Forget, M. Mannan, Terri Oda, Anil Somayaji, and all NSPW 2011 participants for guidance, feedback and lively discussion, especially Matt Bishop and Michael Locasto, our shepherd and scribe, respectively.

7. REFERENCES

[1] E. Adar. Why I hate Mechanical Turk research (and workshops). CHI 2011.

[2] G.V. Bard. Spelling-error tolerant, order-independent pass-phrases via the Damerau-Levenshtein string-edit distance metric. Proc. of 5th Australasian Symposium on ACSW Frontiers - volume 68 (ACSW'07), pp.117-124. Australian Computer Society, 2007.

[3] B.F. Barton, M.S. Barton. User-friendly password methods for computer-mediated information systems. Computers & Security v.3(1984):186-195.

[4] R. Biddle, M. Mannan, P.C. van Oorschot, T. Whalen. User Study, Analysis, and Usable Security of Passwords Based on Digital Objects. IEEE Trans. Info. Forensics and Security 6(3):970-979, Sept. 2011.

[5] R. Biddle, S. Chiasson, P.C. van Oorschot. Graphical Passwords: Learning from the First Twelve Years. ACM Computing Surveys 44(4), 2012 (to appear).

[6] M. Bond. Comments on grIDsure authentication. `http://www.cl.cam.ac.uk/~mkb23/research/GridsureComments.pdf`, March 2008.

[7] S. Brostoff, P. Inglesant, and M. A. Sasse. Evaluating the usability and security of a graphical one-time PIN system. BCS-HCI 2010.

[8] C. Brown. *A Meta-Scheme for Authentication using Text Adventures*. Masters thesis, School of Computer Science, Carleton University, Canada, 2010.

[9] B. Cheswick. Rethinking passwords. Invited talk, USENIX LISA 2010 (slides available online). See also summary by Rik Farrow, ;login: (USENIX Magazine) 36(2):68-69, April 2011.

[10] S. Chiasson, A. Forget, R. Biddle, P.C. van Oorschot. Influencing Users Towards Better Passwords: Persuasive Cued Click-Points. BCS-HCI 2008.

[11] S. Chiasson, C. Deschamps, M. Hlywa, G. Chan, E. Stobert, R. Biddle. MVP: A web-based framework for user studies in authentication (poster). SOUPS 2010.

[12] S. Chiasson, E. Stobert, A. Forget, R. Biddle, P.C. van Oorschot. Persuasive cued click-points: Design, implementation, and evaluation of a knowledge-based authentication mechanism. IEEE Trans. Dependable and Secure Computing (to appear, 2011 or later).

[13] M.K. Chong, G. Marsden. Exploring the Use of Discrete Gestures for Authentication. INTERACT 2009, Part II, Springer LNCS 5727, pp.205-213, 2009.

[14] P. Dunphy, A.P. Heiner, N. Asokan. A Closer Look at Recognition-Based Graphical Passwords on Mobile Devices. SOUPS 2010.

[15] D.F. Ferraiolo, R. Sandhu, S. Gavrila, D.R. Kuhn, R.

Chandramouli. Proposed NIST standard for role-based access control. ACM TISSEC 4(3):224-274, 2001.
[16] R. Feynman. Cargo Cult Science Principles of Research. Commencement Address, Caltech, 1974.
[17] D. Florencio, C. Herley. A Large Scale Study of Web Password Habits. WWW 2007.
[18] D. Florencio, C. Herley, B. Coskun. Do strong passwords accomplish anything? USENIX HotSec'07.
[19] M. Gasser. A Random Word Generator for Pronounceable Passwords. Technical Report AD-A017-676, Mitre Corp., Nov.1975.
[20] M.D. Hafiz, A.H. Abdullah, N. Ithnin, and H.K. Mammi. Towards identifying usability and security features of graphical password in knowledge based authentication technique. 2nd Asia Int'l Conf. Modelling & Simulation (2008), 396–403, IEEE.
[21] C. Herley, P.C. van Oorschot. A research agenda acknowledging the persistence of passwords. IEEE Security & Privacy (to appear, Jan/Feb 2012).
[22] M. Jakobsson, E. Shi, P. Golle, R. Chow. Implicit authentication for mobile devices. USENIX HotSec'09.
[23] M. Jakobsson, personal communication, May 2011.
[24] D.L. Jobusch, A.E. Oldehoeft. A survey of password mechanisms: weaknesses and potential improvements (Part 1). Computers & Security v.8(1989):587-604.
[25] D.L. Jobusch, A.E. Oldehoeft. A survey of password mechanisms: weaknesses and potential improvements (Part 2). Computers & Security v.8(1989):675-689.
[26] C. Kuo, S. Romanosky, L.F. Cranor. Human selection of mnemonic phrase-based passwords. SOUPS 2006.
[27] T. Longstaff, D. Balenson, Mark Matties. Barriers to Science in Security. ACSAC 2010 (invited essay).
[28] R. Maxion, T. Longstaff, J. McHugh. Why is there no science in cyber science? NSPW 2010 (panel report).
[29] A. Mehler, S. Skiena. Improving Usability Through Password-Corrective Hashing. SPIRE 2006: String Processing and Info. Retrieval, 13th International Conference. LNCS 4209, pp.193-204, Springer, 2006.
[30] NIST. FIPS 181: Automated Password Generator (APG). Oct.5, 1993.
[31] NIST Special Pub. 800-63-1 (W. Burr, D. Dodson, R. Perlner, W. Polk, S. Gupta, E. Nabbus). Electronic Authentication Guideline. Gaithersburg, April, 2006.
[32] NIST NSTIC. National Strategy for Trusted Identities in Cyberspace: Creating Options for Enhanced Online Security and Privacy. June 25, 2010 draft. `http://www.dhs.gov/xlibrary/assets/ns_tic.pdf`
[33] A. Paivio. Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior*, 1965.
[34] S.N. Patel, J.S. Pierce, F.D. Abowd. A Gesture-based Authentication Scheme for Untrusted Public Terminals. ACM UIST'04.
[35] J.R. Platt. Strong Inference. Science 136(3642): 347-353 (Oct.16, 1964).
[36] S.N. Porter. A password extension for improved human factors. Computers & Security v.1(1982):54-56.
[37] S.E. Schechter, R. Dhamija, A. Ozment, I. Fischer. The Emperor's new security indicators. IEEE Symp. Security and Privacy (Oakland 2007).
[38] R. Schell. Information security: the state of science, pseudoscience, and flying pigs. ACSAC 2001 (invited essay).
[39] S. Singh, A. Cabraal, C. Demosthenous, G. Astbrink, M. Furlong. Password sharing: implications for security design based on social practice. CHI'07.
[40] Y. Spector, J. Ginzberg. Pass-sentence–a new-approach to computer code. Computers & Security v.13(1994):145-160.
[41] R. Weber. The Statistical Security of GrIDsure. Technical report, University of Cambridge, 2006.
[42] M. Weir, S. Aggarwal, M. Collins, H. Stern. Testing Metrics for Password Creation Policies by Attacking Large Sets of Revealed Passwords. ACM CCS 2010.
[43] S. Wiedenbeck, J. Waters , J. Birget, A. Brodskiy, N. Memon. PassPoints: Design and longitudinal evaluation of a graphical password system. Int'l J. Human-Computer Studies 63(1–2), 2005.
[44] J. Yan, A. Blackwell, R. Anderson, A. Grant. Password memorability and security: empirical results. IEEE Security and Privacy magazine 2(5):25-31, 2004.
[45] Y. Zhang, F. Monrose, M.K. Reiter. The Security of Modern Password Expiration: An Algorithmic Framework and Empirical Analysis. ACM CCS 2010.
[46] M.E. Zurko, R.T. Simon. User-centered security. NSPW 1996.
[47] M. Zviran, W.J. Haga. A comparison of password techniques for multilevel authentication mechanisms. *The Computer Journal* 36(3):227-237. Updates report NPS-54-90-014 (June 1990), Naval Postgraduate School, Monterey, California. (Note: the term *multilevel* here relates not to MLS operating systems, but to primary vs. secondary passwords.)

Appendix: Condensed Justification for NSPW

A first motivation is the following research challenge. It is now common that a single user alternately accesses the same web-based accounts from both desktop machines with full physical keyboards, and from mobile devices on which input of mixed-case and special characters (common in passwords) is more difficult and error prone due to tiny physical keyboards or touch-screen virtual keyboards. This creates need for a password mechanism which simultaneously supports convenient access from both types of devices, producing identical passwords from the system viewpoint, and ideally with as little change as possible to current practice.

Herein we present a technical overview of one proposal to address this problem—a hybrid text and graphical password scheme called *gridWord*—but rather than focus on evaluation of the proposal itself, use it as a concrete case study to explore how to best carry out such evaluations, and broader questions involving how the security research community referees (evaluates) work in security and usability. From our own recent research efforts in usable security—a topic historically of interest to NSPW [46]—we see considerable room for improvement in evaluations in this area, and a continuing lack of consensus on many issues, including how to go about evaluation scientifically. This motivates our larger focus on more general questions about usable security research, using gridWord to ground a broader examination.

How should the community actually carry out, improve on, and publish research in security and usability? What is actually expected by peer reviewers? What is rewarded by the community? What is reasonable to expect for publications in this area, in terms of breadth and depth of exposition, experiments, and scientific evaluation [27]? How should we as a community go about the tasks of planning experiments, carrying out, and publishing research in this area? Exploring such questions motivated by this specific context provides an opportunity to progress toward a consensus absent to date. The concreteness may also help advance the NSPW 2010 panel discussion [28].

We believe the focus on the broader questions will prove more profitable and suitable for NSPW than accompanying the outline of our new proposal with results from a first pilot study, or focusing on the design and evaluation of gridWord itself. We nonetheless hope that the larger discussion will inform subsequent detailed experimentation involving gridWord, including design choices, configuration choices, recommended parameterizations, and evaluation approaches. Thus towards the goal of progressing gridWord as a side product, we welcome secondary discussion, guidance, and feedback on gridWord itself.